UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

Andre Seiji Sunahara

Análise da Relação Entre Produtividade e Impacto Científico Via Métodos de Física Estatística

Maringá, fevereiro de 2020.

UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

ANDRE SEIJI SUNAHARA

ANÁLISE DA RELAÇÃO ENTRE PRODUTIVIDADE E IMPACTO CIENTÍFICO VIA MÉTODOS DE FÍSICA ESTATÍSTICA

Dissertação apresentada ao Programa de Pós Graduação em Física da Universidade Estadual de Maringá como requisito parcial para obtenção do título de Mestre em Física.

Orientador: Prof. Dr. Haroldo Valentin Ribeiro

Maringá, fevereiro de 2020.

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

Sunahara, Andre Seiji

S957a Análise da relação entre produtividade e impacto científico via métodos de física estatística / Andre Seiji Sunahara. -- Maringá, PR, 2020.
150 f.: il. color., figs., tabs.

Orientador: Prof. Dr. Haroldo Valentin Ribeiro.
Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Física, Programa de Pós-Graduação em Física, 2020.

1. Sistemas complexos - Método estatístico . 2. Pesquisadores - Impacto científico - Produtividade. 3. Análise logística - Outlier - Produção científica. 4. Física estatística. 5. Análise de dados - Ciência. I. Ribeiro, Haroldo Valentin , orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Física. Programa de Pós-Graduação em Física. III. Título.

CDD 23.ed. 530.13

Rosana de Souza Costa de Oliveira - 9/1366

Agradecimentos

Nesta seção, gostaria de expressar meus agradecimentos a todos que tiveram contribuição direta ou indireta na realização desse trabalho.

Primeiramente, gostaria de agradecer à minha família: Sandra, José e Isadora. Sou grato pelo apoio recebido por meio das mais diversas formas em relação aos estudos. Nos dias atuais, compreendo que é um privilégio ter a possibilidade de estudar sem ter preocupações financeiras.

Agradeço à minha namorada, Thaís, que sempre me apoiou durante o mestrado com muita compreensão e amor. Sei que foi um período difícil para nós dois, mas também acredito que foi uma fase de muito aprendizado.

Agradeço ao professor Haroldo, que esteve presente sempre com entusiasmo no papel de orientador. Aprendi muito em vários aspectos: como redigir de uma maneira mais "científica", como ter disciplina quanto aos estudos e, principalmente, em relação à postura diante da vida.

Agradeço aos meus amigos de pós-graduação. Em especial: Alvaro, Arthur, Denner, Higor, João Victor (JV), Leonardo Cunha (Fera, *the greatest*), Leonardo Mendes, Max e todos da "salinha", que compartilharam comigo muitos momentos. Alguns desses momentos foram bons, outros não tão agradáveis, mas o importante é que sempre estivemos juntos.

Agradeço aos professores Breno, Hatsumi, Renato e Renio pelo apoio de sempre.

Agradeço aos amigos de Lida Kai e Shuyodan, com quem compartilhei momentos de muito aprendizado nesses dois anos.

Agradeço à minha psicóloga Maressa e às amigas, Harumi, Thaís e Yui, que me ajudaram nos momentos mais difíceis.

Agradeço aos amigos Daniel, Fabrício, Lorena, Mrtvi, Thaís e Thiago por estarem presentes em minha vida.

Por fim, agradeço a CAPES e ao CNPq pelo auxílio financeiro que foi essencial para a realização deste trabalho.

Resumo

Nesta dissertação, exploramos aspectos da área de "ciência da ciência" via técnicas e métodos de Sistemas Complexos. Especificamente, analisamos a relação entre produtividade e impacto científico utilizando informações extraídas dos currículos *vitae* de pesquisadores cadastrados na Plataforma Lattes. O impacto científico foi quantificado por meio de indicadores das revistas indexadas na *Web of Science* (fator de impacto) e na *SCOPUS* (indicador SJR), que são bancos de dados de larga escala contendo milhões de artigos científicos. Por meio de medidas de padronização, construímos um plano impacto-produtividade que agrega pesquisadores de diferentes áreas. Esse plano permite identificar os pesquisadores *outliers* nas categorias produtividade e/ou impacto científico. De modo geral, verificamos que pesquisadores que em algum momento da carreira são *outliers* em um quesito tendem a não obter o mesmo desempenho no outro quesito. Investigamos a dinâmica de carreira de pesquisadores de diferentes disciplinas no plano impacto-produtividade. Constatamos que, para algumas áreas, há certa tendência em aumentar a produtividade em detrimento do impacto no decorrer da carreira. Por meio de modelos lineares mistos, avaliamos a hipótese de que a produtividade influencia o impacto científico. Para a maioria das disciplinas, verificamos que a produtividade afeta positivamente o impacto das publicações de pesquisadores não-*outliers*; entretanto, esse efeito varia de área para área. Por fim, constatamos que a flutuação das medidas de impacto como função da produtividade pode ser descrita por uma aproximação exponencial a um valor constante, para a qual a taxa de variação e o limite assintótico dependem da disciplina.

Palavras-chave: Sistemas Complexos. Ciência da Ciência. Análise de Dados. Física Estatística.

Abstract

In this dissertation, we explore different aspects of the discipline of "science of science" via techniques and methods of Complex Systems. Specifically, we analyze the relation between productivity and scientific impact by using data extracted from researchers' curriculum vitae entries on the Lattes Platform. The scientific impact was quantified by using indicators from Web of Science (impact factor) and SCOPUS (SCImago Journal Rank or SJR indicator), which are large scale databases that index millions of scientific papers. By using standard score measures, we propose an impact-productivity plane that aggregates researchers from different areas. Furthermore, the impact-productivity plane allowed us to identify outliers in productivity and/or scientific impact. We verify that researchers, who at some point of their careers are outliers in one category, tend to underperform in the other one. We investigate the career dynamics of researchers from different disciplines in the impact-productivity plane. We find there is a tendency of increasing productivity and diminishing impact in the course of the researchers' careers for some disciplines. By using linear mixed models, we analyze the influence of productivity on scientific impact. We verify that productivity positively affects the impact of publication in most disciplines for non-outlier researchers; however, the intensity of this effect is different among scientific areas. Finally, we find that the fluctuations of the impact measures as a function of productivity are well described by an exponential approach to a constant, for which both the approaching rate and the plateau value are dependent on the area.

Keywords: Complex Systems. Science of Science. Data Science. Statistical Physics.

Sumário

Introdução

De modo geral, o presente trabalho se enquadra na chamada Física de Sistemas Complexos. Por ser uma área relativamente recente e principalmente por seu aspecto intrinsecamente multidisciplinar, é muito difícil definir o espectro de problemas que podem ser abordados pela Física de Sistemas Complexos. Consequentemente, também é difícil estabelecer precisamente o que constitui um Sistema Complexo. Na tentativa de contornar essa dificuldade, vamos considerar dois exemplos de sistemas que podem ser considerados complexos: o sistema imunológico humano e um formigueiro.

O sistema imunológico é responsável pela defesa do corpo humano contra agentes externos e doenças. Na ocorrência de uma infecção bacteriana, por exemplo, as células brancas linfócitos tipo B têm como papel identificar bactérias na corrente sanguínea. Nesse contexto, as bactérias são corpos estranhos ao organismo e, por isso, também são chamadas de antígenos. Em seguida, ocorre a liberação de anticorpos, proteínas que se unem às bactérias, formando complexos antígeno-anticorpo. Esse processo é altamente específico, já que os anticorpos são liberados apenas na presença de determinados estímulos. Eles funcionam como um tipo de "marcador" que auxilia na identificação de tudo que não é pertencente ao corpo. A partir daí, os macrófagos, outro tipo de célula branca, ingerem os complexos antígeno-anticorpo e realizam um processo equivalente à digestão que, nessa escala e situação, é conhecido como fagocitose [1].

Um formigueiro, por sua vez, é caracterizado não só por sua estrutura física, mas também por sua organização social. Em uma colônia de formigas bem desenvolvida, a rainha e os machos férteis têm a responsabilidade de acasalar, a fim de aumentar a população da colônia que, em alguns casos, chega ao número de milhares de indivíduos. Para o restante das formigas, há várias funções a serem exercidas: expansão de túneis subterrâneos; criação e manutenção de uma linha de defesa contra espécies invasoras; busca por comida; construção da câmara para futuras ninhadas – otimizando o uso da radiação solar para que a

temperatura seja ideal para o crescimento das larvas e pupas – etc. Os componentes de um formigueiro desenvolvido possuem a característica de alta especialização. Na realidade, esse é um aspecto particular de colônias mais antigas que, passado o período de formação, conseguem diversificar as funções de seus membros para facilitar a expansão da colônia [2].

Esses dois sistemas, apesar de bastante distintos, compartilham similaridades se analisados como sistemas complexos. Em ambos os casos, os agentes individuais – glóbulos brancos e formigas – atuam sem nenhuma liderança. Porém, o comportamento coletivo resultante é mais complexo do que as ações individuais desses agentes. Em outras palavras, o comportamento global é uma sequência de estímulos químicos que partem de contribuições individuais, formando uma espécie de rede de ações, que acaba por gerar uma resposta complexa. No caso dos glóbulos brancos, cada célula realiza sua própria função. Porém, atuando em conjunto e sem nenhum comando central, o sistema consegue combater os invasores num âmbito geral. O mesmo acontece no formigueiro, pois, apesar de a rainha estar presente, ela não realiza nenhuma ação que expressa comando – as formigas simplesmente "sabem" quais são suas tarefas. Fato é que um cientista, conhecendo apenas das características individuais dos agentes, teria grande dificuldade em prever o comportamento coletivo complexo desses sistemas. Isso é o que chamamos de comportamento *emergente*.

Essa é apenas uma das características dessa classe de sistemas denominados *complexos* em que os dois exemplos anteriores se encaixam. Como já mencionado anteriormente, apesar de não haver consenso em sua definição, isto é, propriedades absolutamente necessárias ou suficientes, existem algumas propriedades presentes na maioria dos sistemas complexos [3, 4, 5]:

- Grande número de agentes interagentes;
- Comportamento emergente;
- Ausência de comando central;
- Dinâmica não-linear e/ou fora do equilíbrio;
- Comportamento auto-organizado;
- Comportamento persistente no tempo;
- Paralelismo nas ações;
- Adaptação, aprendizado e evolução.

Em geral, o estudo de sistemas complexos engloba uma gama de questões cujas respostas são essenciais para o entendimento de seu funcionamento. Com fins ilustrativos, retornemos aos dois exemplos anteriores. No caso do sistema imunológico, por que os linfócitos não

identificam células saudáveis como ameaças? Por que doenças autoimunes ocorrem? Como o sistema "aprende" quais são os agentes nocivos? No caso das colônias de formigas, quais ações individuais desencadeiam o aparecimento de uma estrutura social complexa? Como a colônia se adapta a diferentes condições climáticas? Quais fatores externos e internos são importantes para a descrição do sistema e o estudo de sua evolução temporal? A criação de modelos que, a partir das características do sistema e suas interrelações, conseguem descrever o mecanismo pelo qual ele funciona é uma maneira de responder essas perguntas.

Como exemplos do grande número de sistemas complexos inseridos nas mais diversas áreas do conhecimento, podemos citar alguns trabalhos: o estudo de propriedades de cristais líquidos por meio do uso de técnicas de *machine learning* [6], o estudo do comportamento coletivo em sistemas biológicos [7, 8], a aplicação de leis de escala no estudo do crime [9, 10], o uso de conceitos de entropia e complexidade estatística no estudo de obras de arte [11], a análise de sistemas econômicos [12, 13], a análise linguística empregando métodos estatísticos [14, 15], dentre muitos outros.

Neste trabalho, estudamos um campo de pesquisa inserido na área de sistemas sociais complexos, a *science of science*, ou, abreviadamente, *scisci*. A "ciência da ciência", como o nome sugere, estuda as relações entre pesquisadores, instituições e ideias na ciência [16]. As últimas geralmente são representadas por artigos científicos ou por patentes. Esses agentes interagem formando o grande sistema complexo conhecido como ciência, que possui várias das características descritas anteriormente: grande número de elementos constituintes; capacidade de se auto-organizar mesmo sem uma entidade de controle central; capacidade de adaptação e evolução paralelamente às mudanças no âmbito social-econômico; dentre outras. Esse campo tem sido cada vez mais estudado, principalmente devido ao número crescente de dados disponíveis e pela tendência cada vez mais interdisciplinar da ciência [17]. Como fontes relevantes de dados, podemos citar as bases de dados *Web of Science* (*WoS*), *SCOPUS*, *PubMed* e o *Microsoft Academic Graph* (MAG), que possuem milhares de artigos indexados e, além disso, informações e indicadores a nível de revista científica.

Aqui, utilizamos a base de dados da Plataforma Lattes em conjunto com a *Web of Science* e a *SCOPUS*. Entre os diversos resultados encontrados na literatura de *science of science*, destacamos o crescimento exponencial da produção científica anual dos artigos indexados na base de dados *Web of Science* [18]. Conforme mostra a Figura 1, os dados da Plataforma Lattes também exibem essa impressionante característica. Esse crescimento da produção anual apresenta um tempo característico que faz dobrar seu volume a cada 8.3 anos. O crescimento reflete o desenvolvimento econômico e tecnológico, fenômeno que aconteceu de forma global, facilitando e impulsionando a produção científica de modo geral. A fim de entender como esta dissertação está inserida no campo *science of science*, faremos uma breve revisão de alguns dos principais resultados da área.

Apesar de a produção científica ter crescido exponencialmente no curso do desenvolvi-

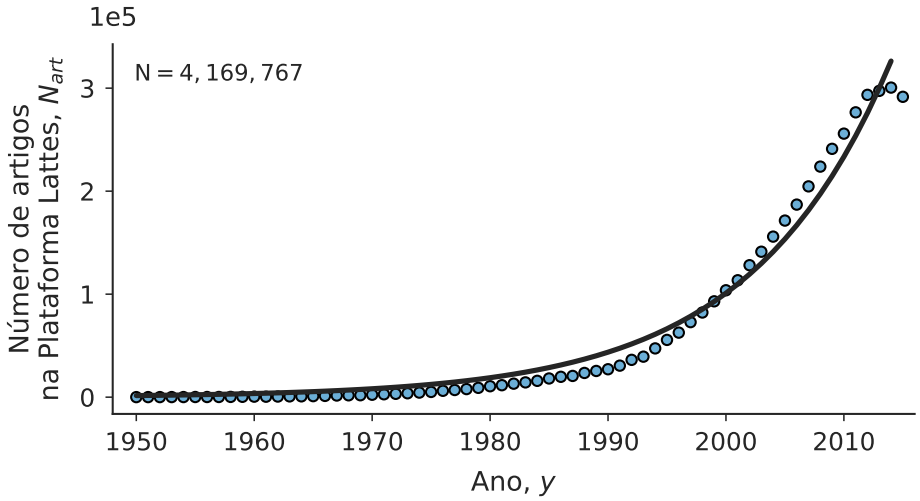

Figura 1: Número de artigos contidos na Plataforma Lattes em função do tempo. Notamos que existe um crescimento aproximadamente exponencial do número de artigos reportados no decorrer do tempo, como mostra a curva contínua ajustada aos dados. O tempo característico para dobrar o volume de produção científica é de aproximadamente 8.3 anos. Somando toda a produção científica reportada na Plataforma Lattes, de 1950 a 2015, há um total de $N = 4\ 169\ 767$ artigos.

mento da ciência, a produção de ideias não tem acompanhado esse ritmo, apresentando um crescimento linear [19]. No trabalho da referência [19], o conceito de "ideia" foi definido como a diversidade léxica em títulos de artigos científicos. Voltando a analisar a produtividade, a crescente no ritmo de publicações faz com que, individualmente, pesquisadores publiquem cada vez mais. Um dos motivos desse comportamento é a mudança no padrão de colaboração: pesquisadores vêm colaborando cada vez mais [20, 21]. A partir da rede de citações, na qual os nós representam pesquisadores e as ligações direcionadas são as citações, estudos mostram que há formação de comunidades de artigos que se citam com maior frequência [22, 23]. Contribuindo com a manutenção dessa estrutura de comunidades, uma análise na área de Biomedicina revela que pesquisadores tendem a escolher tópicos de pesquisa associados ou na mesma linha de sua área atual, isto é, se aventuram pouco em assuntos totalmente novos [24]. Além disso, campos mais antigos e consolidados da ciência possuem o comportamento coletivo de privilegiar o estudo de assuntos bem estabelecidos, em oposição a campos mais novos, que tendem a ser mais disruptivos e inovadores [25]. O padrão de disrupção não está presente apenas entre as áreas, mas também quando comparamos equipes de pesquisa de tamanhos diferentes. Equipes menores tendem a ser mais disruptivas, buscando referências mais antigas, utilizando da interdisciplinaridade e criando conexões entre ideias que não existiam anteriormente; por outro lado, equipes maiores tendem a desenvolver ideias já existentes [26]. Ambos os tipos de equipes são importantes para a ciência, pois promovem o desenvolvimento científico de maneiras diferentes e criam um sistema sustentável e dinâmico. O problema é que há um viés no financiamento desses diferentes "tipos de ciência".

Atualmente, linhas de pesquisa mais consolidadas e menos diruptivas são privilegiadas pelas agências de fomento, pois envolvem menos risco de falha [27, 28, 29]. Nesse sentido, o papel das agências seria de incentivar os diferentes tipos de ciência para possibilitar um desenvolvimento mais diversificado. Resultados mostram que a tolerância por agências de fomento a falhas iniciais possibilita descobertas mais inovadoras [30]. Por outro lado, contratos temporários provocam instabilidade na produtividade, causando o término repentino de carreiras científicas, o que não está associado necessariamente à falta de competência ou perseverança do pesquisador [31]. Além disso, uma pesquisa recente com modelos epidêmicos mostra que existe um viés sistemático na propagação de ideias, pois pesquisadores de universidades de maior prestígio, se comparados aos pesquisadores de universidades de menor prestígio, têm maior chance de disseminar ideias de mesma qualidade, gerando uma vantagem estrutural [32].

A *science of science* também reproduz padrões presentes em outros sistemas sociais. Há desigualdade de gênero, já que mulheres colaboram menos, publicam menos, recebem menos incentivo das agências de fomento e, além disso, quando disputam uma vaga de emprego com homens igualmente qualificados, tendem a não serem escolhidas [33, 34, 35, 36, 37, 38, 39]. O contraste não é apenas na questão do gênero, mas também existe uma assimetria na distribuição de citações. A maioria dos artigos não são citados e uma pequena parcela detém a maior parte das citações, sendo esse comportamento presente em todas as áreas [40]. Além disso, artigos e autores com mais citações tendem a receber ainda mais citações, numa dinâmica denominada de "os ricos ficam mais ricos" (o efeito Matthew) e marcada por uma distribuição cuja cauda segue uma lei de potência [41, 42, 43]. Dessa maneira, o impacto medido em número de citações para autores mais prestigiados é naturalmente maior, principalmente nos primeiros anos após a publicação de um determinado artigo [44]. Portanto, uma medida de impacto não enviesado deveria envolver anos subsequentes à publicação. De modo geral, o impacto é maior logo após a publicação do artigo e a atenção tende a diminuir no decorrer do tempo por causa da obsolescência do conhecimento [45, 46, 47]. Porém, em alguns casos, o impacto não é imediato, ou seja, há artigos que começam a receber citações depois de alguns anos. Esses artigos são usualmente chamados de "belas adormecidas" [48, 49]. Uma investigação também mostra que o trabalho de maior impacto de um pesquisador pode acontecer com igual probabilidade em qualquer período de sua carreira, contanto que ele continue a produzir. De acordo com uma modelagem estocástica, o impacto de um trabalho é uma mistura de "sorte" entre estar estudando um problema com potencial e destreza do pesquisador [50, 51].

Num viés individual, os resultados da literatura de *science of science* ajudam pesquisadores a buscar estratégias para sua carreira. Enquanto isso, globalmente, ajudam a ciência a seguir caminhos mais eficientes por meio dos investimentos e das decisões tomadas por agências de fomento e empresas do setor privado. Nosso trabalho busca contribuir de uma

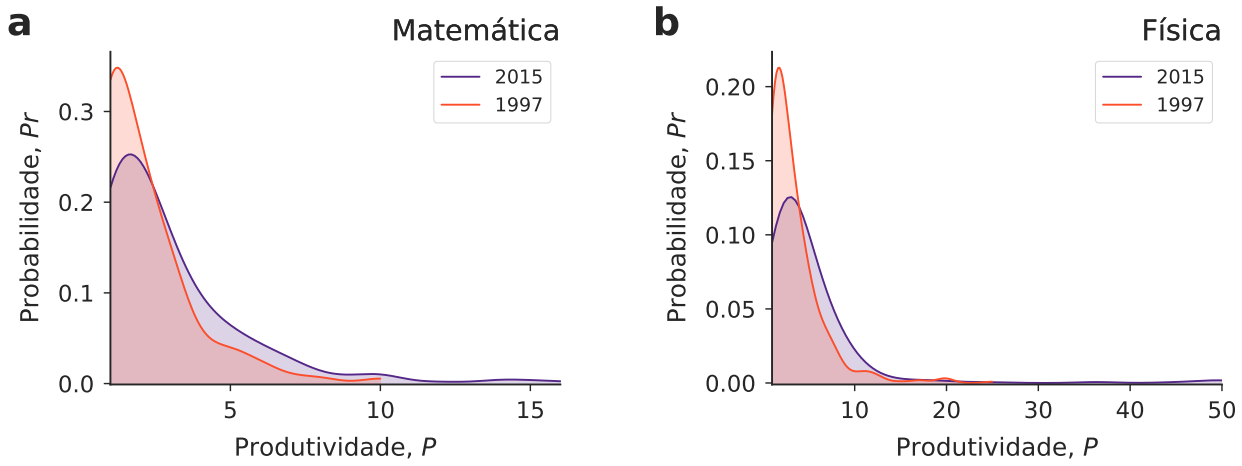

Figura 2: Distribuição de probabilidade da produtividade anual para pesquisadores de duas áreas em 1997 e 2015. Distribuição de probabilidade da produtividade, definida como número de artigos produzidos por ano, de pesquisadores da **(a)** Matemática e **(b)** Física. Há uma diferença intrínseca na produtividade anual das duas áreas, sendo a Física uma área que produz em maior quantidade que a Matemática. Com o decorrer do tempo, também houve uma evolução no número de publicações por ano, já que as curvas se deslocaram para direita e passaram a apresentar caudas mais longas.

maneira diferente, estudando a dinâmica de produtividade e de impacto científico com base em indicadores de impacto dos jornais.

Para ilustrar um pouco melhor alguns aspectos deste trabalho, a Figura 2 mostra a distribuição de probabilidade da produção anual de pesquisadores bolsa produtividade CNPq das áreas de Matemática e Física nos anos 1997 e 2015. Analisando esses resultados, observamos que o volume de produção anual é diferente entre as duas áreas. Portanto, ao comparar as primeiras posições numa hipotética classificação de produtividade de cada área para o ano 2015, um pesquisador que produz treze artigos por ano na Matemática estaria numa posição equivalente a um pesquisador da Física cuja produção é de cinquenta artigos por ano, devido à diferente dinâmica das duas áreas. Também há diferenças entre os anos analisados. A distribuição de 2015 em ambas as áreas é muito mais assimétrica e com cauda mais longa para à direita quando comparada com 1997, o que indica que houve um aumento na produtividade anual de modo geral. Ao realizar a mesma análise com o fator de impacto, observamos um comportamento similar de diferenciação por disciplina e de inflação temporal. Apesar da diferença intrínseca entre a produtividade e impacto de áreas e de períodos diferentes, será que existe um padrão escondido entre essas duas variáveis? Essa é uma das perguntas que nosso trabalho buscou responder.

Outro resultado de nosso trabalho está relacionado à dinâmica da produtividade e "escolha" do jornal no decorrer da carreira, que, aqui, também interpretamos como "potencial impacto" das publicações. Conforme iremos descrever no decorrer desse texto, normalizamos todas as medidas de produtividade e impacto dos jornais para que possamos comparar di-

ferentes áreas em diferentes anos. Com fim de motivar e ilustrar essa dissertação, observe a análise qualitativa da Figura 3. Os símbolos i e p correspondem, respectivamente, à produtividade e ao impacto dos jornais que os pesquisadores publicaram em determinado ano. O símbolo + indica um ano acima da média em tal quesito. Enquanto isso, o símbolo − aponta um ano abaixo da média em determinado quesito. Por fim, o símbolo ++ indica que o pesquisador está muito acima da média naquele quesito, sendo considerado um ano *outlier*. A Figura 3 mostra parte da carreira de dois pesquisadores de influência e grande produtividade da Universidade Estadual de Maringá. O professor Edvani Curti Muniz do Departamento de Química, no período em que temos dados disponíveis, está presente em dois principais tipos de regiões de impacto-produtividade como se pode observar na Figura 3a. Num primeiro momento, apresenta o comportamento de publicar em revistas de alto impacto, alternando entre anos com alta e baixa produtividade. No período subsequente de 2005 em diante, ele começou a produzir em grande quantidade, se tornando um *outlier* em produtividade. Muito provavelmente, esse comportamento é devido ao aumento do número de colaborações que tende a crescer durante a carreira. Por outro lado, a Figura 3b mostra uma parte da carreira da professora Tania Ueda Nakamura do Departamento de Ciências Básicas da Saúde. Os padrões presentes em sua carreira são bem diferentes se comparados aos do professor Edvani, já que, no começo da janela de anos, ela publicou majoritariamente em revistas com baixo impacto e, de 2010 em diante, houve uma migração para revistas de maior impacto. Além disso, ela nunca está presente em regiões *outliers*. De modo mais geral, estudamos qual a dinâmica de carreira nessas faixas de impacto-produtividade analisando o comportamento de diferentes áreas do conhecimento.

Os dois exemplos anteriores ilustram bem os tipos de problema que essa dissertação irá abordar. Para apresentar nossos resultados de forma sistemática, dividimos esse trabalho em quatro partes. Na primeira, detalhamos os métodos estatísticos utilizados em nossas análises. Na segunda, explicamos quais as medidas disponíveis em nossa base de dados para quantificar o impacto de revistas científicas. Na terceira, caracterizamos nossa base de dados e detalhamos filtramos os dados para nossa análise principal. Na quarta e última parte, após a normalização e deflacionamento das medidas de produtividade e impacto dos jornais, procuramos responder as perguntas:

- Qual a distribuição dos anos da carreira de pesquisadores bolsa produtividade do CNPq nos setores impacto-produtividade?
- O que podemos inferir sobre os pesquisadores que são *outliers* em produtividade ou sobre aqueles que são *outliers* em impacto científico?
- Quais são as características de pesquisadores que durante toda carreira estão presentes apenas em setores não-*outliers*?

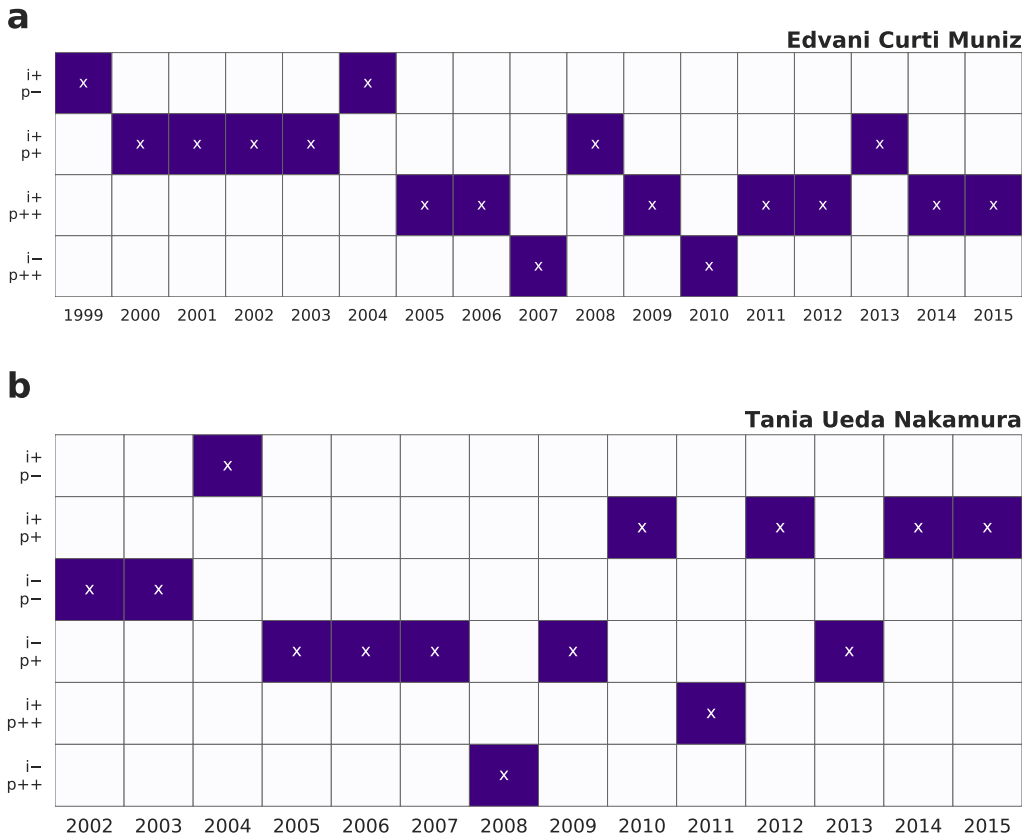

Figura 3: Representação da produtividade do pesquisador e impacto médio das revistas científicas escolhidas para publicação de seus artigos em uma janela de anos. Cada coluna representa um ano do pesquisador. Cada linha corresponde a uma combinação de um padrão de produtividade e um padrão de impacto das revistas em que o pesquisador publica. A sigla i indica o impacto médio das revistas científicas mensurado pelo fator de impacto JCR. A sigla p corresponde à produtividade média mensurada pelo número de artigos produzidos por ano pelo pesquisador. Os sinais $+$ e $-$ correspondem, respectivamente, a publicar acima ou abaixo da média, e o símbolo duplo $++$ indica que o pesquisador é um *outlier*, ou seja, publicou muito acima da média naquele quesito. Com fins ilustrativos, escolhemos dois pesquisadores influentes da Universidade Estadual de Maringá. No painel **(a)** apresentamos alguns anos da carreira do professor Edvani Curti Muniz do Departamento de Química. Na maioria dos anos disponíveis, podemos observar que ele priorizou revistas com impacto elevado. Quanto à produtividade, no final da janela temporal, ele passou a publicar muito mais do que a média se comparado ao início do período apresentado. Já no painel **(b)** está uma janela temporal da carreira da professora Tania Ueda Nakamura do Departamento de Ciências Básicas da Saúde, que tem um comportamento diferente, pois foram raros os anos como *outlier*. Além disso, observamos um padrão de publicação em revistas de maior impacto no final do período disponível.

- Qual é a dinâmica de carreira para pesquisadores de diferentes áreas em relação ao plano impacto-produtividade?
- Existe alguma relação entre produtividade e impacto científico?
- Qual a variabilidade dos indicadores de impacto em cada área?

CAPÍTULO 1

Métodos estatísticos para análise de dados

Neste capítulo, fundamentamos conceitos de diversos tipos de regressão linear que serão utilizados neste trabalho. Em especial, estudamos três modelos: simples, logístico e misto. Além disso, a fim de realizar a estimativa dos parâmetros da regressão mista, apresentamos conceitos sobre amostragem bayesiana. Finalmente, como existem dados *outliers* em nossa base de dados, introduzimos o conceito de estimadores-M que nos auxiliam neste tipo de situação. Caso o leitor tenha familiaridade com esses métodos, recomendamos a leitura a partir do Capítulo 2 e o uso do presente capítulo como referência.

1.1 Regressão linear

Em muitas circunstâncias, é de interesse entender a relação entre duas quantidades. Na Física, existem diversas relações funcionais entre variáveis. Por exemplo, a trajetória unidimensional $y(t)$ de uma partícula em movimento retilíneo uniforme pode ser descrita por uma relação linear com o tempo t,

$$y(t) = y_0 + vt \ , \tag{1.1}$$

em que y_0 é a posição inicial da partícula e v é sua velocidade. Aqui, por relação funcional, entendemos que a variável dependente y está completamente descrita, ponto a ponto, pela variável independente t, gerando uma curva, como mostra a Figura 1.1a. Visualmente, todos os pares ordenados estão sobre a curva que caracteriza a relação funcional.

Em contrapartida, com dados de sistemas reais, existe o que usualmente chamamos de relação estatística. O que distingue a relação estatística de uma relação funcional é que os

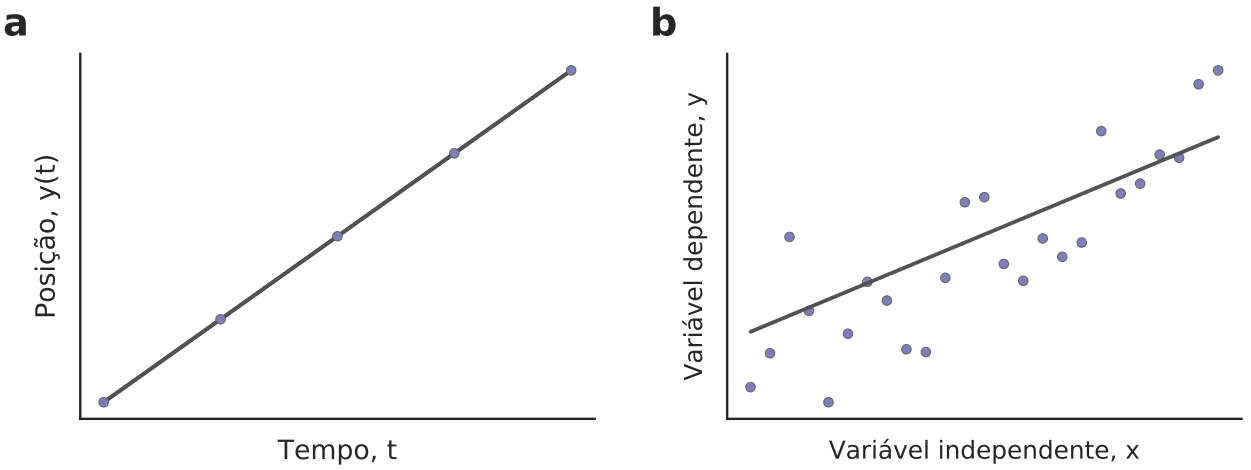

Figura 1.1: Relação funcional e relação estatística. **(a)** Exemplo de relação funcional da posição $y(t)$ em função do tempo t no movimento retilíneo uniforme unidimensional. **(b)** Exemplo de relação estatística entre uma variável dependente $y(x)$ em função de uma variável independente x.

pontos não recaem todos sobre a curva que representa a dependência das variáveis, como mostra a Figura 1.1b. Nesse caso, a "tendência" linear ainda existe, mas a dispersão dos pontos aponta que há outros fatores que influenciam a variável dependente e que não são descritos pela relação estatística.

Neste trabalho, utilizamos o modelo linear misto para entender a relação entre a produtividade de um pesquisador e o impacto científico das revistas em que ele publica. Além disso, utilizamos um modelo logístico para estudar o comportamento de pesquisadores *outliers*. Essas técnicas são uma extensão da regressão linear simples. Por isso, primeiramente, estudaremos este modelo mais elementar. Considerando que os vetores $\boldsymbol{x} = (x_1, ..., x_N)^\top$ e $\boldsymbol{y} = (y_1, ..., y_N)^\top$ representam os dados e N é o tamanho da amostra, queremos um modelo linear do tipo

$$y(x) = \beta_0 + \beta_1 x \ , \tag{1.2}$$

sendo β_0 e β_1 parâmetros do modelo. Porém, a forma da Eq. (1.2) se assemelha à forma da Eq. (1.1), que representa a relação funcional de um movimento uniforme. Para todos os valores da variável independente x, a variável dependente $y(x)$ está definida exatamente. Dessa maneira, devemos procurar alguma forma de incorporar fatores aleatórios à Eq. (1.2). Supondo que existe um espalhamento que ocorre de maneira sistemática, isto é, para cada valor do vetor de variáveis independentes $\boldsymbol{x}$, a dispersão no valor correspondente de $\boldsymbol{y}$ apresenta variância constante, podemos escrever a relação, agora estatística, como [52]

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \tag{1.3}$$

na qual ε_i representa os efeitos aleatórios e o índice i se refere ao i-ésimo ponto do conjunto

de dados. Como veremos mais adiante, é mais conveniente escrever a Eq. (1.3) em notação matricial devido à simplificação em termos algébricos. Para isso, definimos o vetor de variáveis dependentes

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} ,$$

e a matriz de regressores

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} ,$$

que muitas vezes é chamada de matriz *design* [52]. Ela contém uma coluna unitária que representa a hipótese de um intercepto constante. A segunda coluna retrata a variável x, cuja influência em y queremos inferir. Esse modelo envolve apenas uma variável independente, mas poderíamos incorporar outras adicionando novas colunas à matriz *design*. Caso isso aconteça, a regressão então se torna multivariada [53]. Os coeficientes da Eq. (1.3) são representados pelo vetor de parâmetros

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} .$$

No caso de uma regressão linear multivariada, a derivada parcial de y em relação à variável independente x resulta no valor do coeficiente correspondente. Do ponto de vista qualitativo, o significado de um coeficiente corresponde à variação em y devido ao aumento unitário de x, mantendo as demais variáveis constantes. Por causa disso, os coeficientes do vetor $\boldsymbol{\beta}$ também são denominados coeficientes parciais de regressão [53]. Finalmente, podemos escrever em notação matricial

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \beta_0 + x_1\beta_1 \\ \beta_0 + x_2\beta_1 \\ \vdots \\ \beta_0 + x_N\beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix} ,$$

sendo $\boldsymbol{\varepsilon}$ o vetor de termos aleatórios. Intuitivamente, podemos dizer que o valor esperado para $\boldsymbol{Y}$ deve ser

$$\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} ,$$

que são os pontos que compõem a curva da Figura 1.1b. Como consequência da nossa

hipótese, a variância de $\boldsymbol{Y}$ e o valor esperado de $\boldsymbol{\varepsilon}$ devem ser

$$Var(\boldsymbol{Y}) = \mathbf{0}$$
$$\text{e} \quad \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$$

em que $\mathbf{0}$ é o vetor nulo de dimensão N.

Seguindo a construção do modelo, é necessário especificar a natureza da distribuição do termo aleatório. Uma hipótese razoável seria supor a normalidade. A fundamentação dessa conjectura vem do fato de que, para explicar o comportamento de y, existem múltiplos fatores a serem considerados e não apenas um como admitimos. Por exemplo, podemos modelar a influência da situação financeira de uma pessoa em sua saúde mental. No entanto, existem muitos outros fatores que poderiam influenciar seu bem-estar (satisfação com seu emprego, atividade social, prática de atividades físicas, tempo em família etc.). De acordo com o modelo, não temos controle sobre essas variáveis adicionais. Elas flutuam aleatoriamente e, supondo sua independência, no limite em que o número de fatores se torna muito grande, o erro se torna normal pelo Teorema Central do Limite [52]. Dessa maneira, vamos assumir que o erro $\boldsymbol{\varepsilon}$ é distribuído normalmente com variância σ^2, isto é,

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) \ ,$$

sendo $\boldsymbol{I}$ a matriz identidade de dimensão N. Consequentemente, o vetor de variáveis dependentes $\boldsymbol{Y}$ é distribuído

$$\boldsymbol{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \ ,$$

o que significa que cada um dos y_i é uma variável aleatória com sua média e variância correspondentes.

Neste caso, podemos dizer que esse modelo supõe a existência de uma relação estatística linear entre as duas variáveis x e y. O intercepto da reta é representado pelo coeficiente β_0 e sua inclinação pelo coeficiente β_1. Ainda, a estrutura do erro é bem definida: uma distribuição normal com média nula e variância σ^2 para cada valor x_i, como mostra a Figura 1.2. Uma das possibilidades para obter uma estimativa dos parâmetros do modelo (σ^2, β_0 e β_1) é empregar o método de máxima verossimilhança [54]. Para esse fim, precisamos investigar a densidade de probabilidade de cada y_i, dada por

$$f(y_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \right\} \ , \tag{1.4}$$

em que os parâmetros β_0, β_1 e σ^2 aparecem após o ponto e vírgula para indicar que não foram determinados.

A Eq. (1.4) representa a distribuição de probabilidade de que a amostra (cada par or-

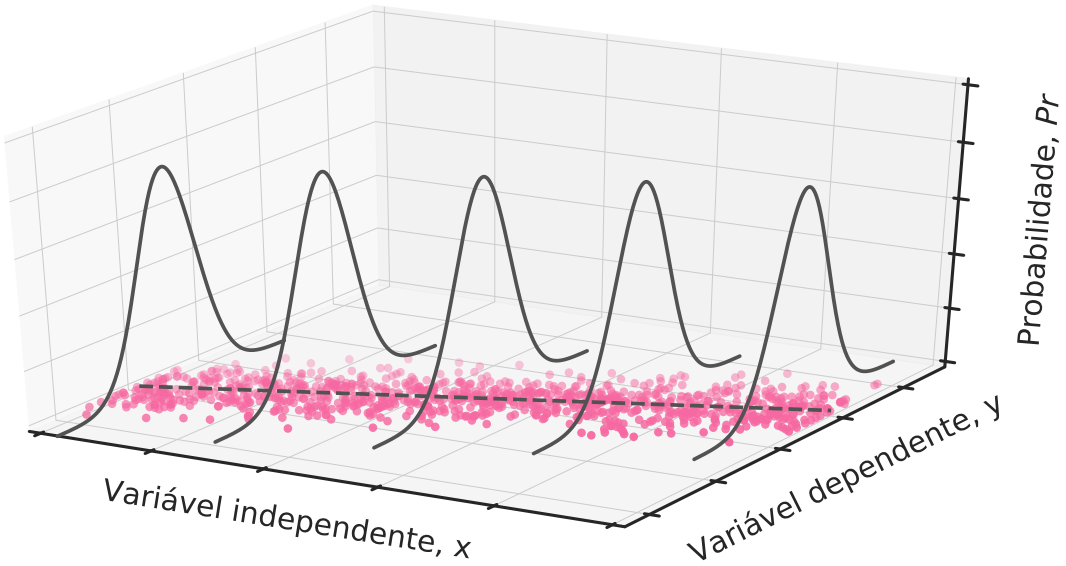

Figura 1.2: Exemplo de regressão linear simples. A reta pontilhada ilustra um modelo linear e as curvas representam a suposição do erro distribuído normalmente.

denado x_i e y_i) siga o modelo proposto com parâmetros arbitrários. Sendo assim, podemos escrever a verossimilhança do modelo como sendo a distribuição conjunta de todos os y_i. Se as variáveis forem independentes, a densidade conjunta se torna o produto das densidades individuais

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}|^2\right] , \tag{1.5}$$

sendo $\mathcal{L}(\boldsymbol{\beta}, \sigma^2)$ a verossimilhança que depende dos parâmetros $\boldsymbol{\beta}$ e σ^2. Para facilitar os cálculos, podemos maximizar o logaritmo da verossimilhança. Essa transformação preserva a localização do máximo da função e, além disso, transforma o produto da Eq. (1.5) em somatória. Dessa forma, estimamos os valores mais prováveis de $\boldsymbol{\beta}$ e σ^2 de acordo com nosso modelo, isto é, de

$$\frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}} = \frac{\partial |\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}|^2}{\partial \boldsymbol{\beta}} = \boldsymbol{0} ,$$

encontramos

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} .$$

Similarmente, temos

$$\hat{\sigma}^2 = \frac{|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}|^2}{N - 2} ,$$

em que os símbolos dotados de "chapéu" indicam que esse resultado é uma estimativa do parâmetro. No caso de $\hat{\sigma}^2$, o denominador foi modificado para remover o viés que advém do tamanho da amostra [52].

Esses estimadores são os mesmos obtidos via método dos mínimos quadrados [52]. Porém,

a vantagem do método da máxima verossimilhança é que podemos obter o intervalo de confiança dessas estimativas. Foge do escopo deste trabalho a explanação sobre esse assunto, mas esses intervalos podem ser determinados, por exemplo, como mostram as referências [53, 52].

Em resumo, as hipóteses do modelo de regressão linear simples são [55]:

- Especificação correta do modelo;
- Normalidade do erro;
- Homoscedasticidade, isto é, a variância constante;
- Independência das variáveis.

1.2 Regressão logística

A regressão logística surge no estudo de variáveis binárias (ou dicotômicas), sendo o modelo padrão nesse contexto [56]. O valor da variável dependente y_i representa a realização ($y_i = 1$) ou não realização ($y_i = 0$) de um evento. A Figura 1.3a mostra um *scatter plot* de uma variável dependente binária y em função de uma variável independente contínua x. Um desses eventos pode ser interpretado como um ensaio de Bernoulli do tipo [57]

$$P(y) = \pi(x)^y [1 - \pi(x)]^{1-y} \ , \tag{1.6}$$

em que $\pi(x)$ é a probabilidade de sucesso e $[1 - \pi(x)]$ a probabilidade de fracasso do experimento. Aqui, vamos modelar a probabilidade $\pi(x)$ como dependente de uma variável arbitrária x. Quantitativamente, isso significa que para cada valor de x pode existir uma proporção diferente de "sucessos" e "fracassos". Assim, a probabilidade $\pi(x)$ corresponde a essa proporção como função de x. Podemos parametrizar o valor de y por meio da função sigmoide [57]

$$S(y) = \frac{\exp y}{1 + \exp y} \ .$$

Desse modo, a variável parametrizada se torna

$$\hat{\pi}(x) = S(\beta_0 + \beta_1 x) = \frac{\exp\left[\beta_0 + \beta_1 x\right]}{1 + \exp\left[\beta_0 + \beta_1 x\right]} \ .$$

Se invertermos a equação para obter uma relação linear, temos

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 x \ , \tag{1.7}$$

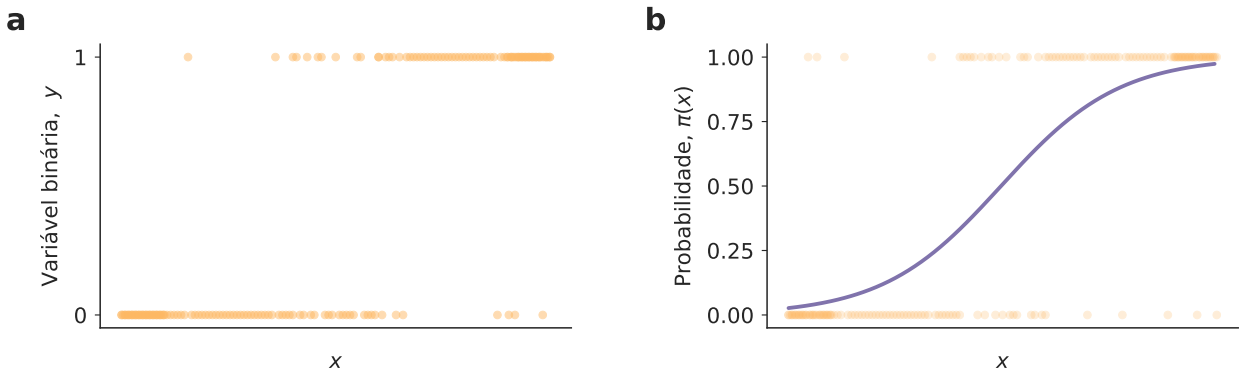

Figura 1.3: Exemplo de regressão logística. (a) *Scatter plot* de uma variável binária y em relação a uma variável contínua x. **(b)** Exemplo de ajuste de uma curva logística.

em que definimos a função *logit* (logística) [57]. Dessa forma, a Eq. (1.7) representa uma parametrização para escrever a probabilidade $\pi(x)$ no intervalo $[0, 1]$ como uma função linear de $x \in (-\infty, \infty)$. Em outras palavras, a parametrização possibilita tratarmos esse modelo como uma regressão linear do tipo $\beta_0 + \beta_1 x$. A Figura 1.3b mostra a curva logística ajustada para um conjunto de dados binários relacionados com uma variável contínua arbitrária x. Nesse caso, podemos interpretar qualitativamente a relação linear como o logaritmo da chance. A chance é definida como a razão entre as probabilidades dos dois eventos, isto é,

$$\text{chance} = \frac{\pi}{1-\pi} \ .$$

Por exemplo, se um evento ocorre três vezes em quatro circunstâncias, a chance de ele acontecer é de três para um (ou 3:1). Dessa maneira, o coeficiente β_1 controla o efeito de x sobre a chance e β_0 está associado ao valor da chance para $x = 0$.

Podemos também definir o modelo de maneira similar ao caso da regressão simples. Neste caso, porém, é preciso escrever a relação para cada x_i, uma vez que $\pi(x)$ depende de x, ou seja,

$$\sum_{i=1}^{k} y_i = \mathbb{E}[y|x] + \varepsilon_i = k\pi(x) + \varepsilon_i \ , \tag{1.8}$$

sendo k o número total de eventos para determinado valor de x e ε_i o erro correspondente para determinado x. Sabendo que existe uma probabilidade $\pi(x)$ de sucesso para cada um dos k ensaios de Bernoulli, podemos caracterizar o erro ε_i como uma distribuição binomial [58] com valor esperado

$$\begin{aligned} \mathbb{E}[\varepsilon_i] &= \mathbb{E}\left[\sum_{i=1}^{k} y_i\right] - \mathbb{E}[k\pi(x)] \\ &= k\pi(x) - k\pi(x) \\ &= 0 \ , \end{aligned} \tag{1.9}$$

e variância

$$\begin{aligned} Var[\varepsilon_i] &= Var\left[\sum_{i=1}^{k} y_i\right] - Var[k\pi(x)] \\ &= k\pi(x)[1-\pi(x)] - 0 \\ &= k\pi(x)[1-\pi(x)] \ . \end{aligned} \tag{1.10}$$

Esse modelo faz parte de uma classe mais abrangente de regressões denominada regressão linear generalizada [56], que, resumidamente, é realizada por algum tipo de parametrização, também chamada de *link* nesse contexto. Além disso, uma distribuição de erro da família exponencial pode ser especificada de acordo com as características de cada problema. No caso da regressão linear simples, o *link* é a identidade (não há parametrização) e a distribuição do erro é normal.

Para estimar os parâmetros, recorremos novamente à maximização da verossimilhança. Supomos um conjunto de dados do tipo (x_i, y_i) com $i = 1, 2, \ldots, N$, em que y_i é a i-ésima amostra da variável binária e x_i é a i-ésima amostra da variável contínua. Utilizamos a Eq. (1.6) com propósito de estimar as contribuições de cada par (x_i, y_i) para a verossimilhança, ou seja,

$$\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \ , \tag{1.11}$$

lembrando que cada y_i admite o valor 1 ou 0. Assumindo a independência das observações, a verossimilhança pode ser obtida pelo produto de todos os termos individuais e escrita como

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \ . \tag{1.12}$$

Em termos do logaritmo da verossimilhança, que possibilita trabalhar com uma somatória, temos

$$\log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \{y_i \log \pi(x_i) + (1-y_i)\log[1-\pi(x_i)]\} \ . \tag{1.13}$$

Diferenciamos a Eq. (1.13) em relação aos parâmetros de interesse β_0 e β_1 e igualamos as expressões a zero. Dessa maneira, podemos encontrar os valores que maximizam a verossimilhança do problema. As expressões resultantes são

$$\begin{aligned} &\sum_{i=1}^{N}[y_i - \pi(x_i)] = 0 \quad \text{e} \\ &\sum_{i=1}^{N} x_i[y_i - \pi(x_i)] = 0 \ . \end{aligned} \tag{1.14}$$

No caso da regressão linear simples, encontramos expressões analíticas para os parâmetros.

No caso da regressão logística, isso não é possível, posto que as expressões na Eq. (1.14) não são lineares nos parâmetros. Para superar essa dificuldade, é comum a utilização de métodos numéricos na estimação de β_0 e β_1. Em nossos resultados, utilizamos o pacote *statsmodels* [59] do *Python* que implementa essas rotinas numéricas.

1.3 Regressão linear mista

Para entender a importância de aperfeiçoarmos e generalizarmos o modelo linear simples discutido na Seção 1.1, apresentamos uma análise ilustrativa de dois sistemas: o primeiro baseado em dados reais (Caso 1) e o segundo referente a uma situação hipotética (Caso 2).

1.3.1 Caso 1: Evolução da equidade de gênero

Mesmo que as variáveis estudadas se relacionem por meio de um comportamento linear, o modelo de regressão linear simples ainda apresenta limitações decorrentes da hipótese de independência estatística de variáveis aleatórias identicamente distribuídas. Para ilustrar esse fato, vamos analisar um conjunto de dados disponibilizado pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE). Esses dados contêm informações sobre a evolução da diferença salarial de mulheres e homens de 43 países com o passar dos anos [60]. Definimos d como uma medida de desigualdade de gênero, representando a diferença percentual da mediana dos ganhos entre homens e mulheres (adotando como base salários do gênero masculino). A abrangência da base de dados é de 1970 até 2018, mas alguns países possuem mais pontos que outros. Além disso, o dado é dividido em duas categorias: trabalhadores empregados ou profissionais liberais. No total, há $N = 1055$ pares ordenados.

A Figura 1.4a mostra um *scatter plot* dos pontos para todos os países. Ao examinar esse gráfico, é razoável afirmar que existe uma tendência em direção ao decréscimo da desigualdade de gênero com o passar do tempo. Uma análise interessante seria entender com que passo essa desigualdade está diminuindo. Utilizando um modelo simples, podemos supor que a relação é expressa por

$$\boldsymbol{d} = \beta_0 \boldsymbol{I} + \beta_1 \boldsymbol{y} + \boldsymbol{\varepsilon} \ , \tag{1.15}$$

em que d é a diferença percentual do salário entre homens e mulheres, y o ano, ε o erro residual e β_0 e β_1 são parâmetros do modelo. A curva preta da Figura 1.4a representa o resultado da regressão pelo método da máxima verossimilhança, com estimativas $\hat{\beta}_0 = 609.23$ e $\hat{\beta}_1 = -0.29$. O valor negativo do coeficiente angular β_1 indica a diminuição da desigualdade com o tempo.

Porém, será que essa estimativa realmente retrata o comportamento médio global? A resposta é negativa. Um indicativo para respondermos melhor essa pergunta está na Figura 1.4b, na qual efetuamos as regressões lineares para algumas nações separadamente.

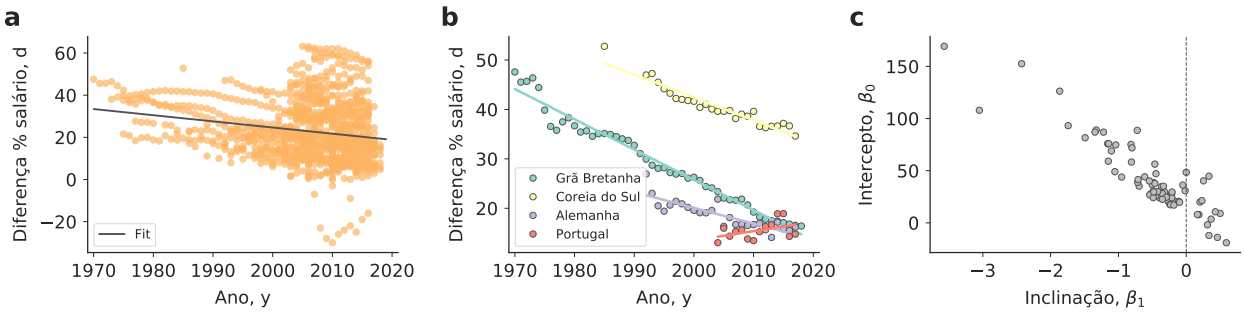

Figura 1.4: Desigualdade de gênero nos países da OCDE. **(a)** *Scatter plot* e ajuste linear da diferença percentual da mediana dos salários entre os gêneros (tendo como base o gênero masculino) em função do tempo. **(b)** *Scatter plot* e ajuste linear para dados de países específicos (Grã Bretanha, Coreia do Sul, Alemanha e Portugal). **(c)** Parâmetros do modelo linear estimados para todos os países individualmente.

Percebemos que as inclinações são diferentes, podendo inclusive apresentar valores positivos. A Figura 1.4c mostra os resultados das regressões para os estimadores de todos os países. Percebemos que os valores do intercepto (β_0) e da inclinação (β_1) variam bastante. Além disso, há outro problema: como existem países que possuem mais pontos, como a Grã Bretanha na Figura 1.4b, o comportamento global na regressão total será muito mais influenciado por eles. No outro extremo, o fato de existir pouca informação para alguns países pode ocultar seu comportamento verdadeiro. Em termos estatísticos, esses fatos violam algumas hipóteses do modelo linear. Primeiramente, os pontos não são independentes, já que existe uma maior correlação para dados do mesmo país. Além disso, os dados não são identicamente distribuídos porque cada país apresenta sua própria distribuição, com sua própria média e variância. Foi com o intuito de resolver essas dificuldades que o modelo linear misto surgiu na literatura [61].

1.3.2 Caso 2: Relação do tempo de estudo com desempenho

Em uma regressão linear simples, os parâmetros de interesse, β_0 e β_1, da Eq. (1.3) correspondem a efeitos médios na população. Eles também são denominados *efeitos fixos* exatamente por refletirem somente informações globais que independem da estrutura hierárquica do dado. Porém, como vimos no caso anterior, quando temos uma estrutura hierárquica, as observações dentro de um grupo (país) podem ser mais similares do que entre os grupos (países), indicando que há violação da hipótese de independência das variáveis. Em outras palavras, podemos dizer que elas obedecem a suposição de independência apenas dentro do grupo. Uma abordagem ingênua para tentar entender o comportamento global de um sistema é realizar o procedimento do caso anterior: (i) uma regressão para todas as observações, sem distinção dos grupos – o que desrespeita a suposição de independência (Figura 1.4a); (ii) várias regressões, uma para cada grupo – o que não aproveita a informação da relação

entre grupos e não gera um resultado no âmbito global (Figura 1.4c). A solução oferecida pelos modelos lineares mistos pode ser interpretada como o meio termo entre essas duas perspectivas. Afirmamos que cada um dos grupos possui seus próprios parâmetros que, nesse contexto, se tornam variáveis aleatórias. Consideramos que eles vêm de distribuições de média μ_0 e μ_1 que, respectivamente, representam o comportamento global dos parâmetros β_0 e β_1. Dessa forma, os parâmetros que possuem ambos comportamentos, local e global, formam o que chamamos de *efeitos aleatórios* [62].

Com o propósito de conhecermos os modelos que compõem essa classe de formulações, imagine que estamos interessados em investigar como o tempo individual de estudo influencia o desempenho mensurado pela nota média final de estudantes de todo país. Inicialmente, é razoável supor que estudantes de diferentes escolas terão comportamentos distintos, assim como os países tiveram no exemplo anterior.

Intercepto aleatório

Um primeiro modelo seria imaginar que o impacto do tempo de estudo x sobre o desempenho y é constante, isto é, a taxa de aprendizado independe do local onde o estudante se encontra. Todos os estudantes evoluem da mesma maneira, dependendo apenas do tempo de estudo, ou seja, o sistema é meritocrático. Porém, a relação pode ter um intercepto (que representa a menor nota dentro de uma escola) diferente para cada local, como ilustra a Figura 1.5a. A causa disso pode ser associada a distintas características socioeconômicas tanto da escola quanto dos alunos. Podemos imaginar que estudantes com mais condições financeiras frequentam colégios melhor estruturados. Estes, por sua vez, conseguem assistir os alunos de forma mais eficiente. Por isso, a nota mais baixa de cada escola pode variar e não é necessariamente nula. Esse conjunto de interceptos (notas mínimas) representa o efeito aleatório do modelo. Para um grande número de escolas, modelamos esse efeito como sendo normalmente distribuído com média μ_0 e variância σ_0, isto é,

$$\beta_0 \sim \mathcal{N}(\mu_0, \sigma_0) \ .$$

Podemos escrever a equação como

$$y_i = \mu_0 + b_{0j} + \beta_1 x_i + \varepsilon_i \ ,$$

em que o índice i corresponde à i-ésima observação, o índice j corresponde à j-ésima escola e b_{0j} representa a variação no intercepto da escola j. Em notação mais resumida, temos

$$\begin{aligned} y_i &= \beta_{0j} + \beta_1 x_i + \varepsilon_i \\ \beta_{0j} &= \mu_0 + b_{0j} \ . \end{aligned}$$

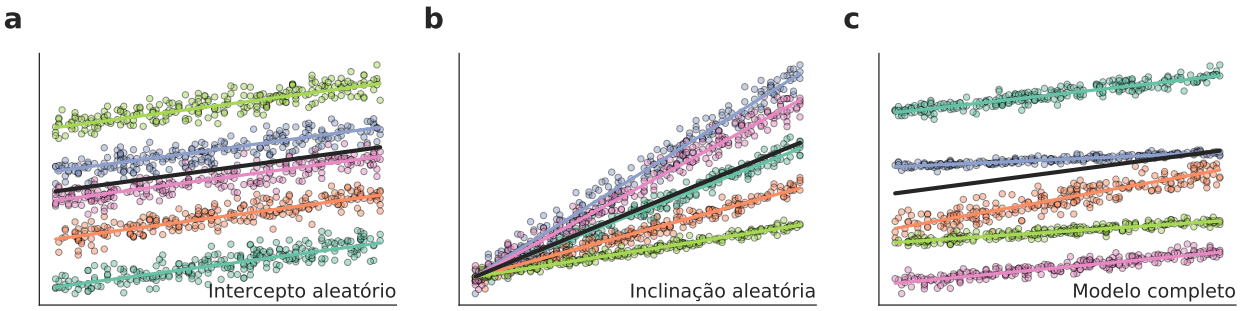

Figura 1.5: Ilustração dos modelos lineares mistos para o aprendizado. **(a)** Intercepto aleatório. **(b)** Inclinação aleatória. **(c)** Modelo completo com inclinação e intercepto aleatórios.

Com esse modelo, podemos entender como a distribuição de notas mínimas varia entre as escolas (linhas coloridas da Figura 1.5a) e, além disso, inferir o comportamento médio nacional de maneira mais precisa (linha preta da Figura 1.5a). Se utilizássemos o modelo linear simples, a estimativa de ambos os parâmetros não corresponderia ao valor verdadeiro. A diferença entre as notas mínimas de diferentes escolas provocaria uma alavancagem da reta estimada para longe do comportamento real. Dessa forma, podemos dizer que o modelo de interceptos aleatórios considera a estrutura hierárquica dos dados para estimar corretamente e concomitantemente os comportamentos locais e global. Mesmo que exista um desequilíbrio em relação ao número de dados disponíveis para cada grupo, o modelo não é afetado.

Inclinação aleatória

Outra hipótese de modelagem seria afirmar que a nota mínima é a mesma para todos os alunos que não se dedicam fora da sala de aula. Porém, podemos supor que a taxa de aprendizado varia entre os alunos de diferentes escolas devido às variadas metodologias empregadas, diferentes condições de infraestrutura e apoio familiar. O reflexo disso se manifesta na presença de diferentes inclinações para cada escola como mostra a Figura 1.5b. Nesse caso, as inclinações estariam distribuídas normalmente

$$\beta_1 \sim \mathcal{N}(\mu_1, \sigma_1) \ ,$$

com respectivas equações

$$y_i = \beta_0 + \beta_{1j} x_i + \varepsilon_i \ ,$$

e

$$\beta_{1j} = \mu_1 + b_{1j} \ ,$$

em que b_{1j} representa a variação na inclinação da escola j. A Figura 1.5b mostra os comportamentos locais (linhas coloridas) e global (linha preta) descritos por esse modelo.

Inclinação e intercepto aleatórios

O caso mais realístico é considerar que diferentes escolas têm notas mínimas distintas e alunos possuem taxas de aprendizados diferentes. Dessa maneira, podemos incorporar a aleatoriedade em ambos os parâmetros como mostra a Figura 1.5c. Nesse caso, o modelo completo pode ser escrito como

$$y_i = \beta_{0j} + \beta_{1j} x_i + \varepsilon_i \ .$$

É importante possuir um conhecimento razoável sobre o sistema estudado para escolher o modelo mais adequado de acordo com a interpretação dos parâmetros β_0 e β_1.

A aplicabilidade do modelo linear misto não se restringe apenas a diferentes grupos espaciais, mas se estende também a repetidas observações do mesmo sistema, o que é conhecido como "análise longitudinal". Por exemplo, em um laboratório, muitas vezes é necessário repetir o mesmo experimento para que se obtenha uma estimativa confiável dos parâmetros de interesse. Nesses experimentos, é comum que os resultados sejam diferentes, pois muitos fatores aleatórios podem estar envolvidos (por exemplo, temperatura, umidade do ar e concentração do pesquisador). Assim, poderíamos incorporar esses fatores aleatórios por meio do modelo linear misto e estimar mais precisamente os parâmetros.

1.3.3 Descrição matemática do modelo linear misto

Com o intuito de descrever matematicamente o modelo linear misto, utilizaremos a notação usada no manual do pacote *lme4* da linguagem R [63]. Podemos definir o modelo como sendo a distribuição condicional de Y, a variável dependente aleatória, dado que $\mathcal{B} = \boldsymbol{b}$, sendo $\mathcal{B}$ o vetor de efeitos aleatórios, isto é,

$$(Y|\mathcal{B} = \mathbf{b}) \sim \mathcal{N}(\boldsymbol{\mu}_{Y|\mathcal{B}=\mathbf{b}}, \sigma^2 \mathbf{W}^{-1}) \ ,$$

sendo

$$\boldsymbol{\mu}_{Y|\mathcal{B}=\mathbf{b}} = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{o} + \boldsymbol{\varepsilon} \ ,$$

em que $\boldsymbol{\mu}_{Y|\mathcal{U}=\mathbf{u}}$ é o vetor de preditores lineares, $\mathbf{Z}\mathbf{b}$ é a contribuição dos efeitos aleatórios, $\boldsymbol{Z}$ é a matriz modelo para o vetor de efeitos aleatórios $\boldsymbol{b}$, $\mathbf{o}$ é o *offset* fixado se houver informações previamente conhecidas e $\boldsymbol{W}$ é o vetor diagonal de pesos preestabelecidos para a variância, posto que é possível modelar a estrutura de variância do modelo.

A distribuição dos efeitos aleatórios $\mathcal{B}$ é suposta ser multivariada e normalmente distribuída com a matriz de covariância positiva e semi-definida $\boldsymbol{\Sigma}$, isto é,

$$\mathcal{B} \sim \mathcal{N}(0, \boldsymbol{\Sigma}) \ .$$

O nosso objetivo é encontrar as melhores estimativas para os parâmetros $\boldsymbol{\beta}$, σ^2 e os elementos da matriz de covariância $\boldsymbol{\Sigma}$. A fim de permitir a singularidade de $\boldsymbol{\Sigma}$, podemos definir $\boldsymbol{\Sigma}$ em termos de um fator relativo de covariância $\boldsymbol{\Lambda_\theta}$, cujos parâmetros $\boldsymbol{\theta}$ representam os elementos da matriz covariância dos efeitos aleatórios a menos de uma escala da variância populacional σ^2 [64], ou seja,

$$\boldsymbol{\Sigma_\theta} = \sigma^2 \boldsymbol{\Lambda_\theta} \boldsymbol{\Lambda_\theta}^\top \ .$$

Supondo que $\mathcal{U}$ seja uma variável esférica dos efeitos aleatórios, tal que

$$\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) \ ,$$

podemos realizar a transformação $\mathcal{B} \to \mathcal{U}$ como

$$\mathcal{B} = \boldsymbol{\Lambda_\theta} \mathcal{U} \ .$$

A distribuição de $\mathcal{B}$ é definida como uma função de $\mathcal{U}$. Isso é necessário, pois, no caso em que $\boldsymbol{\Lambda_\theta}$ é singular e definimos $\mathcal{U}$ em função de $\mathcal{B}$, a distribuição esférica não poderia ser calculada. Nessas condições, o modelo fica definido por

$$\begin{aligned} (Y|\mathcal{U} = \mathbf{u}) &\sim N(\boldsymbol{\mu}_{Y|\mathcal{U}=\mathbf{u}}, \sigma^2 \mathbf{W}^{-1}) \\ \boldsymbol{\mu}_{Y|\mathcal{U}=\mathbf{u}} &= \boldsymbol{X\beta} + \mathbf{Z\Lambda_\theta u} + \mathbf{o} + \boldsymbol{\varepsilon} \end{aligned} \ ,$$

em que $\boldsymbol{\mu}_{Y|\mathcal{U}=\mathbf{u}}$ é a média condicional da variável aleatória esférica $\mathcal{U}$ dado os valores observados do vetor de variáveis dependentes.

1.3.4 Estrutura dos efeitos aleatórios

Para entender melhor a estrutura do modelo, vamos analisar a forma dos componentes que diferem da regressão linear simples. Tomemos como exemplo o sistema descrito anteriormente em que tentamos prever a nota final. Seja y_i a nota de um aluno de acordo com o tempo de estudo x_i numa amostra de N observações. Nesse caso, o modelo linear simples é escrito como

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon} \ .$$

Incorporando efeitos aleatórios, nosso modelo teria a seguinte estrutura

$$\underbrace{\boldsymbol{Y}}_{N\times 1} = \underbrace{\overbrace{\boldsymbol{X}}^{N\times p}\ \overbrace{\boldsymbol{\beta}}^{p\times 1}}_{N\times 1} + \underbrace{\overbrace{\boldsymbol{Z}}^{N\times q}\ \overbrace{\boldsymbol{\Lambda_\theta}}^{q\times q}\ \overbrace{\boldsymbol{u}}^{q\times 1}}_{N\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{N\times 1} \ , \tag{1.16}$$

em que N é o número de observações (dados), p o número de parâmetros, $q = lp'$ o número de l grupos (escolas) vezes o número de parâmetros p' modelados como efeitos aleatórios.

Para analisar a forma de $\boldsymbol{Z}$ e $\boldsymbol{\Lambda_\theta}$, suponha que tenhamos um conjunto de dados estruturados tal como mostra a Figura 1.6, e queremos modelar a taxa de aprendizado como um efeito aleatório. O vetor modelo de efeitos aleatórios $\boldsymbol{Z}$ poderia ser descrito como

$$\boldsymbol{Z} = \begin{pmatrix} \underbrace{x_1}_{\text{Escola 1}} & \underbrace{0}_{\text{Escola 2}} & \underbrace{0}_{\text{Escola 3}} & \underbrace{0}_{\text{Escola 4}} \\ x_2 & 0 & 0 & 0 \\ x_3 & 0 & 0 & 0 \\ 0 & x_4 & 0 & 0 \\ 0 & x_5 & 0 & 0 \\ 0 & 0 & x_6 & 0 \\ 0 & 0 & x_7 & 0 \\ 0 & 0 & 0 & x_8 \\ 0 & 0 & 0 & x_9 \\ 0 & 0 & 0 & x_{10} \end{pmatrix}$$

Se, além disso, quisermos incorporar o intercepto (nota mínima) como efeito aleatório, resultando no modelo completo, o vetor modelo $\boldsymbol{Z}$ se torna

$$\boldsymbol{Z} = \begin{pmatrix} x_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ x_2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ x_3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & x_4 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & x_5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & x_6 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & x_7 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_8 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_9 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_{10} & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \\ \underbrace{\qquad}_{\text{Escola 1}} \quad \underbrace{\qquad}_{\text{Escola 2}} \quad \underbrace{\qquad}_{\text{Escola 3}} \quad \underbrace{\qquad}_{\text{Escola 4}} \tag{1.17}$$

Para o modelo completo descrito pela matriz da Eq. (1.17) (intercepto e inclinação alea-

$$\boldsymbol{x} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\} \qquad N = 10$$
$$\boldsymbol{Y} = \{y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10}\} \qquad l = 4$$

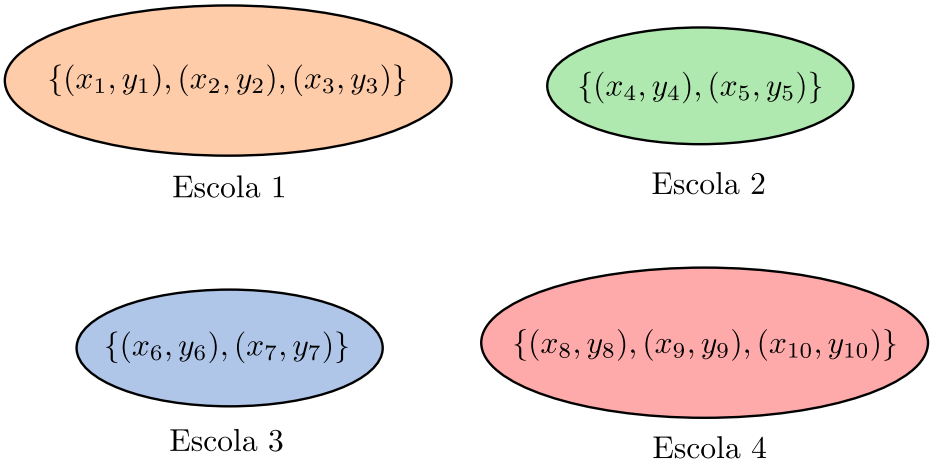

Figura 1.6: Disposição hipotética dos dados das escolas. Nesse caso, o conjunto de dados consiste de $N = 10$ observações que estão distribuídas entre $l = 4$ grupos (escolas).

tórios), teremos uma matriz de covariância relativa com a seguinte estrutura

$$\boldsymbol{\Lambda_\theta} = \begin{pmatrix} a & . & . & . & . & . & . & . \\ c & b & . & . & . & . & . & . \\ . & . & a & . & . & . & . & . \\ . & . & c & b & . & . & . & . \\ . & . & . & . & a & . & . & . \\ . & . & . & . & c & b & . & . \\ . & . & . & . & . & . & a & . \\ . & . & . & . & . & . & c & b \end{pmatrix},$$
$$\underbrace{\quad}_{\text{Escola 1}} \quad \underbrace{\quad}_{\text{Escola 2}} \quad \underbrace{\quad}_{\text{Escola 3}} \quad \underbrace{\quad}_{\text{Escola 4}}$$

em que os elementos da diagonal representam os parâmetros de variância e os elementos restantes são parâmetros de covariância. Os parâmetros $\boldsymbol{\theta}$ deste modelo são, portanto,

$$\boldsymbol{\theta} = (a, b, c) \ .$$

De maneira geral, o número de parâmetros m do vetor $\boldsymbol{\theta}$ é dado por

$$m = \binom{p+1}{2} = \frac{(p+1)!}{2!(p-1)!} \ .$$

Finalmente, especificados os parâmetros do modelo, nosso objetivo se torna encontrar as melhores estimativas de $\boldsymbol{\beta}$, σ^2 e $\boldsymbol{\theta}$. O método usado pelo pacote *lme4* da linguagem *R* é o da maximização do perfil de verossimilhança (assim como utilizado para regressão simples e logística). Essa abordagem é a vertente frequentista em que as estimativas dos parâmetros são valores pontuais com a incerteza dada por um intervalo de confiança e o p-valor. Porém, conforme veremos, a metodologia que utilizamos é a bayesiana, em que resultado não é composto por valores pontuais, mas sim de uma distribuição de probabilidade para cada parâmetro. Esse método está se popularizando nos últimos anos e oferece uma outra perspectiva sobre o assunto. Na próxima seção, entenderemos quais as principais diferenças entre as duas abordagens.

1.4 Abordagens frequentista e bayesiana

Quando estamos estudando um sistema, é de interesse quantificar suas características para podermos prever e entender seu comportamento. Supondo um sistema hipotético, podemos representar suas características por meio de um vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^\top$, sendo n o número total de parâmetros a serem estimados. Designamos por *amostra* as informações que possuímos sobre o sistema e que compõem nosso conjunto de dados. Assim, por meio de sua utilização, podemos aplicar modelos estatísticos em que incorporamos nossas hipóteses sobre o comportamento desse sistema. Em sua totalidade, esse processo é denominado *inferência estatística*. Existem duas abordagens principais para realizar a inferência: a *frequentista* e a *probabilística* (*bayesiana*) [65]. A seguir, apresentaremos as principais características de cada uma dessas abordagens.

1.4.1 Abordagem frequentista

A abordagem frequentista considera que a amostragem é o resultado de um processo probabilístico, no qual os parâmetros estudados são considerados fixos. Para entender o que isso significa, suponha que queremos quantificar a probabilidade θ de obtermos a face cara para cima no lançamento de uma moeda não-enviesada. A partir de um número N de lançamentos que constituem uma amostra, a estimativa de θ pode ser escrita como

$$\hat{\theta} = \frac{N_c}{N} \ , \tag{1.18}$$

em que $\hat{\theta}$ é a estimativa da probabilidade de obter cara e N_c é o número de lançamentos em que obtemos cara na amostra. Como o cálculo do estimador é realizado diretamente pela razão entre a frequência de caras e a frequência total de lançamentos, é comum denominar essa abordagem de *frequentista*. Além disso, podemos considerar que calculamos uma estimativa pontual, já que obtemos um valor específico para o parâmetro θ. Esse não será o

caso para a perspectiva bayesiana.

Podemos imaginar que a probabilidade de se obter cara a partir de uma moeda justa é de 50%, ou seja, o resultado cara é tão provável quanto coroa. Porém, a Figura 1.7 revela que esse não é o caso quando realizamos o "experimento". Em nossos experimentos numéricos, consideramos mil lançamentos de uma moeda justa para estimar a probabilidade $\hat{\theta}$ a partir da Eq. (1.18), repetindo o procedimento um total de cem realizações. A Figura 1.7 mostra que o valor da probabilidade varia ao redor da fração que esperaríamos obter no decorrer das realizações, mas não é exatamente igual a 1/2.

De acordo com a perspectiva frequentista, apesar dessa aparente discrepância, o valor do parâmetro θ é definido e vale 1/2. O que acontece é que as amostras vêm de uma distribuição populacional fixa de θ, mas, por serem de tamanho finito, possuem intrinsecamente um aspecto ruidoso. Isso acaba por impossibilitar a igualdade da estimativa do parâmetro com seu valor teórico. Como mencionado anteriormente, a amostragem pode ser considerada como resultado de um processo probabilístico, no qual o parâmetro possui um valor fixo. No limite de infinitas realizações do experimento, esperamos que a média das estimativas de $\hat{\theta}$ corresponda ao valor verdadeiro de θ, isto é,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \hat{\theta} = \theta = 0.50 \ ,$$

sendo N o total de realizações do experimento. Na prática, ao calcularmos essa média com centenas de realizações do experimento, já encontramos valores bem próximos de 0.50.

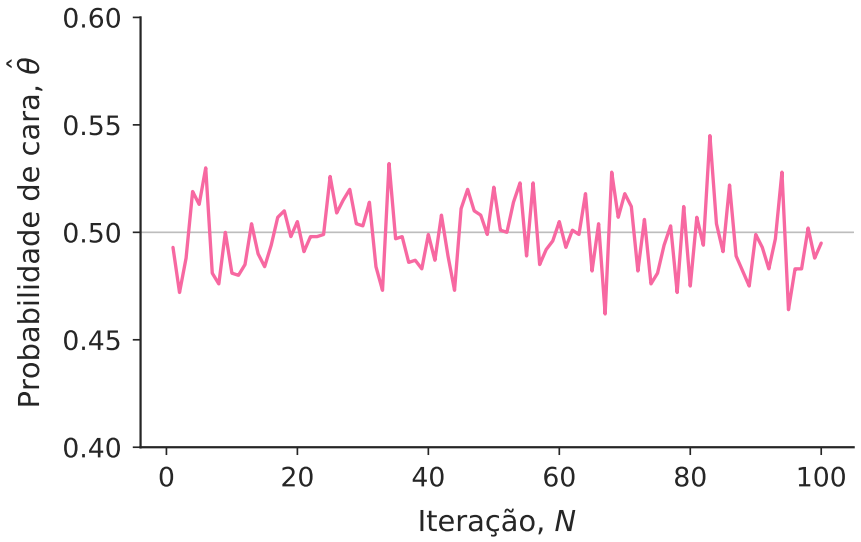

Figura 1.7: Estimativa da probabilidade de se obter cara no lançamento de uma moeda. Calculamos o valor do estimador $\hat{\theta}$ dessa probabilidade para várias realizações do processo de $N = 1000$ lançamentos de uma moeda justa.

Uma abordagem muito comum na perspectiva frequentista da inferência estatística é

considerar a distribuição de probabilidade que representa a *verossimilhança* [52, 53, 56, 58] (assim como consideramos para a regressão linear e para regressão logística). A verossimilhança indica qual a probabilidade de um modelo ser o correto de acordo com as evidências, caracterizadas pelo conjunto de dados. Para obter uma estimativa dos parâmetros, maximizamos o perfil de verossimilhança. De fato, ao usar a Eq. (1.18) para o caso de uma moeda não-enviesada estamos maximizando esse perfil. Para perceber essa correspondência, notamos que a verossimilhança para N lançamentos é escrita como

$$\mathcal{L}(\theta) = p(\theta) = \binom{N}{N_c}\theta^{N_c}(1-\theta)^{N-N_c} \ . \tag{1.19}$$

Maximizando a função com relação a θ, isto é,

$$\frac{\mathrm{d}\mathcal{L}(\theta)}{\mathrm{d}\theta} = \binom{N}{N_c}\frac{\mathrm{d}[\theta^{N_c}(1-\theta)^{N-N_c}]}{\mathrm{d}\theta} = 0 \ , \tag{1.20}$$

obtemos

$$\hat{\theta} = \frac{N_c}{N} \ . \tag{1.21}$$

Portanto, a estimativa da Eq. (1.18) é exatamente a mesma obtida pela maximização do perfil de verossimilhança.

Contudo, raramente temos conhecimento do valor verdadeiro do parâmetro de antemão. Por isso, precisamos de alguma maneira para calcular a credibilidade de nossa estimativa. Para isso, é comum realizarmos um processo de re-amostragem, que possibilita o cálculo da estatística correspondente ao grau de confiabilidade do resultado. Chamamos essa medida de p-valor. Essa quantidade pode ser definida como a probabilidade de que uma estimativa seja tão extrema quanto ou maior do que a que realmente obtemos. Se essa probabilidade é maior do que um nível de significância α pré-estabelecido pelo experimentador, rejeitamos a estimativa. Caso contrário, dizemos que ela é estatisticamente significante. No caso da moeda, não há necessidade de realizar uma re-amostragem para avaliar se ela é enviesada ou não, pois podemos calcular as probabilidades diretamente. Por exemplo, ao realizar um experimento com dez lançamentos, no qual obtemos sete caras, podemos calcular o valor da probabilidade de obtermos um valor tão extremo como esse, isto é,

$$\begin{aligned} p\text{-valor} &= P(N_c \geq 7) \\ &= \frac{1}{2^{10}}\left[\binom{10}{7} + \binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right] \\ &= 0.17 \ . \end{aligned}$$

Se estabelecermos um nível de significância de 5% (uma escolha comum na literatura), não podemos descartar a hipótese de que essa moeda é justa, uma vez que obteríamos um número

de caras tão extremo quanto sete em 17% dos casos, mesmo que a moeda fosse não-enviesada.

Nesta seção, tentamos explicitar duas características pertencentes à visão frequentista da inferência estatística: os parâmetros são considerados como constantes universais e as amostras têm um caráter aleatório, proveniente das distribuições populacionais dos parâmetros verdadeiros. Essas características serão contrastantes com o ponto de vista bayesiano, conforme veremos adiante.

1.4.2 Abordagem bayesiana

Do ponto de vista bayesiano ou probabilístico, os parâmetros investigados não são constantes universais. Em vez disso, consideramos que os parâmetros são representados por distribuições de probabilidade. Nesse sentido, as probabilidades são indicativos de nossa certeza ou incerteza acerca dos parâmetros. Um valor próximo da unidade indica maior grau de certeza em determinada hipótese. Por outro lado, um valor distante da unidade sugere um certo grau de incerteza. Podemos atribuir duas causas a essa natureza. Primeiro, podemos considerar que os parâmetros são probabilísticos por natureza: não sabemos ao certo seu verdadeiro valor. Numa segunda alternativa, podemos imaginar que os parâmetros são fixos, mas, devido ao nosso conhecimento imperfeito da situação, eles acabam por ter uma interpretação probabilística. Do ponto de vista prático, as duas visões acabam por ser matematicamente indistinguíveis. Em contrapartida com a visão frequentista, o conjunto de dados pode ser entendido como fixo na perspectiva probabilística.

Para entender como a visão probabilística se adequa a um problema, podemos analisar uma corrida presidencial em determinado país, como sugere Lambert [65]. Anteriormente ao período das eleições, são realizadas pesquisas de intenção de voto que constituem nossas amostras e, a partir delas, tentamos prever o resultado das eleições. Dentro dessa perspectiva, o conjunto de dados é fixo, pois é muito improvável que se repitam as mesmas circunstâncias associadas a uma amostra. Uma nação é um agente dinâmico, cujas características (economia, política, cultura etc.) mudam com o tempo. Antes do resultado da eleição, nossas previsões são probabilísticas por natureza, pois não podemos determinar o vencedor com certeza. Do ponto de vista frequentista, a amostra consistiria de elementos aleatórios de uma distribuição populacional e a nossa previsão do resultado seria um candidato único dentre todos eles. Porém, é difícil imaginar que existe uma distribuição populacional de vários processos eleitorais que ocorrem em determinado ano, pois existe apenas uma único desses processos. Esse, portanto, é um exemplo em que a visão probabilística se adequaria melhor ao sistema em análise.

As probabilidades dos parâmetros que investigamos são vistas como expressões de uma "crença" subjetiva. Essas "crenças" podem ser atualizadas conforme novas informações são obtidas. Para isso usamos o Teorema de Bayes, que dá nome a essa linha de pensamento e

pode ser expresso como

$$P(\boldsymbol{\theta}|D) = \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(D)} \ , \tag{1.22}$$

em que:

- $P(D|\boldsymbol{\theta})$ é a *verossimilhança*: a distribuição de probabilidade da nossa amostra D supondo que o modelo para os parâmetros $\boldsymbol{\theta}$ é o correto;
- $P(\boldsymbol{\theta})$ é a *priori*: a distribuição de probabilidade dos parâmetros do modelo antes de sermos expostos aos dados;
- $P(D)$ é a probabilidade de obtermos uma determinada amostra sob qualquer hipótese;
- $P(\boldsymbol{\theta}|D)$ é a *posteriori*: a distribuição de probabilidade dos parâmetros $\boldsymbol{\theta}$ do modelo depois de incorporarmos as informações contidas no conjunto de dados D.

Podemos dizer que a Eq. (1.22) representa a interpretação diacrônica do teorema de Bayes [66], pois a probabilidade muda conforme obtemos mais informações sobre o sistema. Estritamente falando, o significado da Eq. (1.22) corresponde à atualização da "crença" sobre a distribuição de probabilidade dos parâmetros $\boldsymbol{\theta}$ (a *priori*) a partir de evidências experimentais sobre o modelo (a verossimilhança). Dessa maneira, multiplicando essas duas quantidades obtemos uma distribuição de probabilidade atualizada (a *posteriori*), a menos de um fator de normalização.

Podemos derivar o Teorema de Bayes considerando dois eventos A e B, com probabilidades $P(A)$ e $P(B)$, escrevendo suas probabilidades condicionais como

$$\begin{aligned} P(A|B) &= \frac{P(A,B)}{P(B)} \\ P(B|A) &= \frac{P(B,A)}{P(A)} \end{aligned} \ , \tag{1.23}$$

em que $P(A,B)$ é a probabilidade conjunta de ambos os eventos ocorrerem. Utilizando o fato de que $P(A,B) = P(B,A)$ nas expressões da Eq. (1.23), obtemos o Teorema de Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \ . \tag{1.24}$$

De modo geral, consideramos que os elementos das expressões (1.22) e (1.24) podem ser interpretados como probabilidades ou distribuições de probabilidade.

É interessante notar que na perspectiva frequentista consideramos apenas a verossimilhança para estimar os parâmetros de interesse. O significado da verossimilhança é simplesmente considerar de antemão que o modelo escolhido é o verdadeiro e utilizar os dados para aplicá-lo. A estimativa dos parâmetros é realizada maximizando a verossimilhança e a

incerteza é dada pelo p-valor calculado posteriormente. Do ponto de vista bayesiano, não consideramos o modelo correto desde o princípio. O que fazemos é inverter a lógica, isto é, estimar a distribuição de probabilidade dos parâmetros do modelo dadas as informações disponíveis, que corresponde à distribuição a *posteriori*. A verossimilhança é utilizada apenas como evidência e ajuda a determinar a forma da distribuição a *posteriori*. No caso probabilístico, a incerteza que estimamos pelo p-valor na vertente frequentista é atribuída na distribuição a *priori* e, dessa maneira, não precisamos especificar um nível de significância. Toda a informação sobre os parâmetros e sua incerteza está contida na distribuição a *posteriori*.

A perspectiva bayesiana de inferência estatística é alvo de críticas devido a um certo grau de subjetividade na escolha da distribuição a *priori*. No entanto, é importante ressaltar que o ponto de vista frequentista também apresenta um grau de subjetividade. Por exemplo, na escolha do nível de significância [67]: o que torna um nível de significância de 5% diferente de 5.05%? De fato, em qualquer metodologia, sempre existe algum grau de subjetividade. Porém, é importante ressaltar que no olhar probabilístico as incertezas são explicitamente estabelecidas na forma da distribuição a *priori*. Isso pode ser considerado como um ponto positivo, pois essa informação está abertamente disponível para o leitor. Além disso, é importante mencionar que quanto mais dados estão disponíveis para a análise, menor é o impacto da escolha da distribuição a *priori* no resultado final.

Utilizando o Teorema de Bayes

A fim de compreender a dinâmica subjacente ao uso do Teorema de Bayes, vamos propor um problema simples que ajudará a explorar as características de cada termo da expressão da Eq. (1.22). Suponha que estamos novamente interessados em estimar a probabilidade de obter cara no lançamento de uma moeda, sendo essa descrita por uma variável θ. Inicialmente, sem qualquer informação sobre a moeda, podemos considerar que ela é não-enviesada, ou seja, a cada lançamento, a probabilidade é igual para ambas as faces. Essa informação corresponde ao que chamamos de distribuição a *priori*, o termo $P(\theta)$ da Eq. (1.22). Podemos modelar essa informação usando a distribuição de probabilidade beta

$$f(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \ ,$$

com parâmetros $\alpha = \beta = 2$, o que corresponderia obter uma cara e uma coroa em dois lançamentos e garantiria um certo grau de incerteza, visto que ainda não temos nenhuma informação sobre a moeda. Sendo assim, escrevemos

$$P(\theta) = f(\theta; \alpha = 2, \beta = 2) = \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} \theta(1-\theta) \ . \tag{1.25}$$

Agora, vamos analisar o comportamento do termo $P(D)$ da Eq. (1.22). Esse termo não depende do parâmetro θ e, portanto, pode ser interpretado como um fator de normalização. Ele é necessário para permitir que a *posteriori* tenha uma interpretação probabilística. Se o considerarmos como um fator de normalização, podemos calculá-lo usando

$$P(D) = \int P(D|\theta)P(\theta)d\theta \ . \tag{1.26}$$

Por enquanto, nada afirmamos sobre o experimento em si, que inicia quando lançamos a moeda e contabilizamos o número de caras. Portanto, suponha que lançamos uma moeda quatro vezes e obtemos cara em três oportunidades. Essa informação possibilita estimarmos a verossimilhança, que é a informação necessária para atualizar nossas "crenças" iniciais sobre o valor de θ, Eq. (1.25). Para esse resultado de lançamentos, podemos escrever a verossimilhança como

$$P(D|\theta) = f(\theta; \alpha = 4, \beta = 2) = \frac{\Gamma(6)}{\Gamma(4)\Gamma(2)}\theta^3(1-\theta) \ .$$

Além disso, integrando a expressão (1.26) no intervalo $[0, 1]$, obtemos

$$P(D) = 20 \ .$$

Nesse caso, conseguimos escrever a *posteriori* a partir da Eq. (1.22) como

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{\theta^4(1-\theta)^2}{20} \ .$$

A Figura 1.8a mostra as três distribuições de probabilidade (*priori*, verossimilhança e *posteriori*) para o caso descrito anteriormente. Podemos notar que o aspecto da distribuição *priori* denota um alto grau de incerteza associado à largura da distribuição. Se analisarmos a verossimilhança, notamos que essa distribuição é mais concentrada se comparada com a *priori*, uma vez que realizamos um total de quatro lançamentos no experimento (o dobro do suposto para definir a *priori*). Finalmente, a *posteriori* é uma mistura das duas distribuições e indica a atualização das nossas "crenças" à luz das novas evidências. De modo geral, como realizamos somente quatro lançamentos, a *posteriori* é uma distribuição pouco localizada. A densidade de probabilidade para o valor de $\theta = 1/2$ é relativamente alta, não indicando que a moeda possa ser enviesada. Os efeitos de localização da distribuição ficam mais evidentes quando consideramos mais lançamentos, como mostra a Figura 1.8b. Nesse cenário, observamos que a distribuição de probabilidade da verossimilhança se torna muito mais estreita após a ocorrência de 21 caras em 28 lançamentos. Observamos que a *posteriori* fica muito próxima da verossimilhança, ilustrando o que já mencionamos anteriormente: quanto mais dados estão disponíveis, menor é o impacto da escolha da distribuição a *priori*

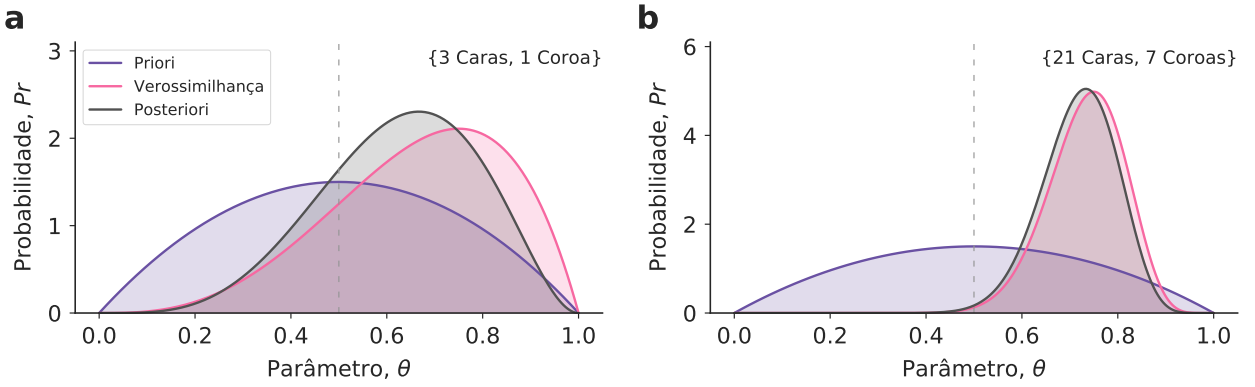

Figura 1.8: Um exemplo de inferência bayesiana: o lançamento de uma moeda. Distribuições de probabilidade da *priori*, verossimilhança e *posteriori* para **(a)** um experimento de 4 lançamentos (3 caras) e **(b)** um experimento com 28 lançamentos (21 caras). Para ambos os experimentos, consideramos como *priori* a distribuição beta com $\alpha = \beta = 2$. Essa distribuição representa nossa esperança de que a moeda seja justa, isto é, para cada dois lançamentos, obtém-se uma cara e uma coroa.

no resultado da inferência. Além disso, a probabilidade para $\theta = 0.50$ é muito pequena, indicando que a moeda analisada não deve ser justa.

Como os parâmetros são descritos por distribuições de probabilidade, precisamos de algum procedimento para indicar valores pontuais para os parâmetros na inferência bayesiana. Assim, podemos comparar resultados anteriores com abordagens frequentistas e, muito além disso, utilizar nossas estimativas para aprimorar ou prever comportamentos do sistema em estudo. As estatísticas mais utilizadas [65] são, nessa ordem de importância, (i) a média no caso contínuo

$$\mathbb{E}[\boldsymbol{\theta}|D] = \int_{\text{todo } \boldsymbol{\theta}} \boldsymbol{\theta} P(\boldsymbol{\theta}|D) d\boldsymbol{\theta} \ , \tag{1.27}$$

e no caso discreto

$$\mathbb{E}[\boldsymbol{\theta}|D] = \sum_i \boldsymbol{\theta}_i p(\boldsymbol{\theta}_i|D) \ , \tag{1.28}$$

(ii) a mediana definida como o valor que divide a distribuição de probabilidade na metade em relação à sua massa e (iii) o estimador de máxima *posteriori*

$$\text{MAP}[\boldsymbol{\theta}|D] = argmax_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|D) \ ,$$

definido como o valor da variável $\boldsymbol{\theta}$ em que a *posteriori* apresenta a maior densidade. É importante ressaltar que estimativas pontuais não são suficientes para representar a *posteriori* de maneira resumida. Estatísticas de dispersão da distribuição de probabilidade, como a

variância,

$$\begin{aligned} Var[\boldsymbol{\theta}|D] &= \mathbb{E}[\boldsymbol{\theta}^2|D] - (\mathbb{E}[\boldsymbol{\theta}|D])^2 \\ &= \int \boldsymbol{\theta}^2 P(\boldsymbol{\theta}|D) d\boldsymbol{\theta} - \left[\int \boldsymbol{\theta} P(\boldsymbol{\theta}|D) d\boldsymbol{\theta}\right]^2 , \end{aligned} \tag{1.29}$$

também são costumeiramente apresentadas. Ademais, momentos superiores também podem vir a ser relevantes e, inevitavelmente, teremos que calcular essas integrais. Em nossos resultados empíricos, escolhemos utilizar a média como medida representativa das estimativas dos parâmetros devido à sua interpretação mais intuitiva.

Estimando a *posteriori* em problemas reais

A relativa simplicidade na utilização do Teorema de Bayes no exemplo anterior é ilusória. Ao investigar sistemas com mais parâmetros ou com distribuições mais complicadas, o cálculo analítico da *posteriori* é muito trabalhoso e, muitas vezes, impraticável. A primeira dificuldade está na determinação do termo de normalização, isto é,

$$P(D) = \int P(D|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \ .$$

A solução analítica dessa integral é possível somente para distribuições mais simples. Além disso, soluções numéricas são computacionalmente viáveis apenas para contextos que envolvam poucas dimensões. Para todos os outros casos, não conseguimos explicitamente calcular ou estimar o valor de $P(D)$. Como se não bastasse, mesmo que conseguíssemos avaliar a integral, precisaríamos calcular também as integrais do primeiro e segundo momentos, definidos pelas Eqs. (1.27) e (1.29), o que acaba por tornar essa abordagem ainda mais inviável.

Na prática, não estamos interessados em estimar o denominador do Teorema de Bayes, pois ele é apenas uma constante de normalização. Como a forma da *posteriori* é proporcional ao numerador, é comum escrever

$$\begin{aligned} P(\boldsymbol{\theta}|D) &= \frac{P(D|\boldsymbol{\theta}) P(\boldsymbol{\theta})}{P(D)} \\ &\propto P(D|\boldsymbol{\theta}) P(\boldsymbol{\theta}) , \end{aligned} \tag{1.30}$$

de modo que precisamos apenas de algum procedimento para estimá-lo. A Figura 1.9 mostra três *posterioris*, duas delas não-normalizadas e uma normalizada. Apesar de possuírem diferentes amplitudes, a forma das distribuições se mantém. A área abaixo da curva pode não corresponder à unidade, mas a diferença de altura entre dois pontos é proporcional entre as diferentes curvas. Dessa maneira, qualquer das três "distribuições" representa bem a *posteriori*. Assim, uma alternativa à solução analítica é replicar a distribuição de probabilidade verdadeira por meio de métodos de amostragem.

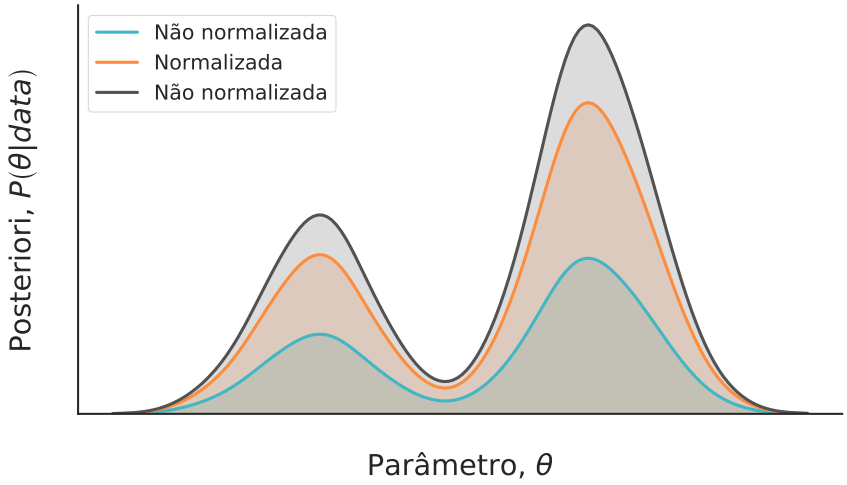

Figura 1.9: Três distribuições representando uma *posteriori* arbitrária. As distribuições representadas pelas curvas verde e cinza não são normalizadas. Enquanto isso, a distribuição em laranja retrata uma distribuição de probabilidade.

A título de exemplo, considere que queremos avaliar qual a distribuição de probabilidade $p(x_i)$ de uma variável X_j que descreve o resultado do lançamento de um dado de seis faces, sendo x_i o número do i-ésimo lado ($i = 1, 2, \ldots, 6$). Podemos lançar o dado um número N de vezes e estimar a probabilidade para cada um dos lados usando

$$\hat{p}(x_i) = \frac{N_i}{N} \ , \tag{1.31}$$

em que $\hat{p}(x_i)$ é a estimativa da probabilidade para o lado x_i, N_i é a quantidade de vezes que o dado caiu com o i-ésimo lado voltado para cima e N é o número total de lançamentos. A Figura 1.10 mostra a estimativa da distribuição de probabilidade amostrada para $N = 1000$ lançamentos. A linha tracejada indica o valor verdadeiro da distribuição (1/6). Observamos que a função é bem representada para uma amostra com um número relativamente baixo de lançamentos. Um número maior de lançamentos faria com que a estimativa definida pela Eq. (1.31) se aproximasse mais ainda de seu valor teórico. Denominamos esse tipo de amostragem de amostragem independente, pois um lançamento não afeta o resultado de nenhum outro lançamento. Usando a estimativa da distribuição de probabilidade, podemos calcular as estatísticas pertinentes, como a média

$$\mathbb{E}(x) = \sum_{i=1}^{6} x_i p(x_i) \approx \frac{1}{N} \sum_{j=1}^{N} X_j \ ,$$

em que X_j é o valor da face do dado no j-ésimo lançamento e N é o tamanho da amostra. Também podemos generalizar essa aproximação para o caso contínuo. Por exemplo, imaginando que uma variável $X_j \sim p(x)$ possa ser amostrada independentemente, podemos

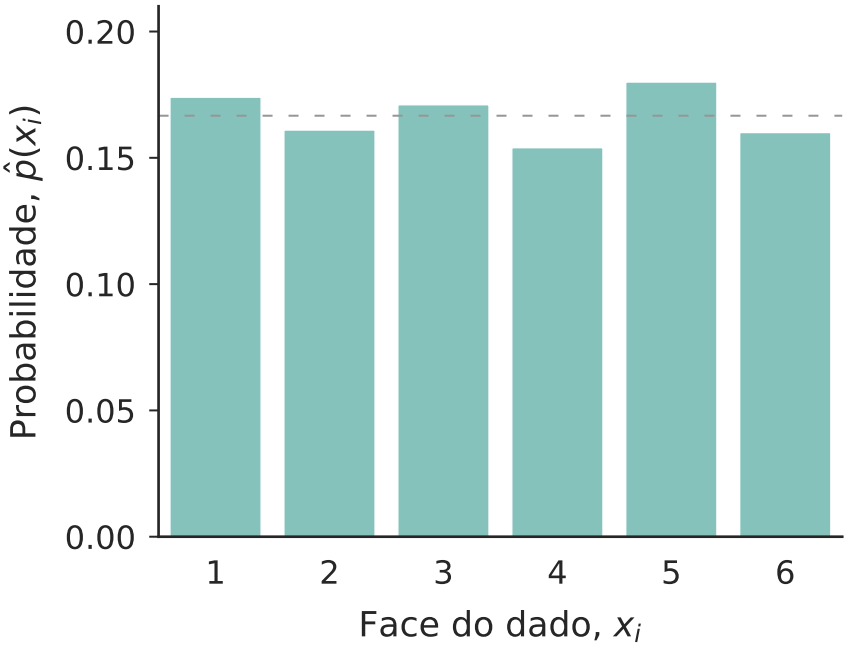

Figura 1.10: Amostragem independente de um dado de seis lados. Estimativa da probabilidade $\hat{p}(x_i)$ de ocorrência da face x_i calculada via amostragem independente de um dado lançado $N = 1000$ vezes. A linha tracejada indica o valor verdadeiro $p(x_i) = 1/6$.

avaliar sua média via

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xp(x)dx \approx \frac{1}{N}\sum_{j=1}^{N} X_j \ . \tag{1.32}$$

De modo geral, podemos aproximar o valor médio para qualquer função $f(\boldsymbol{x})$ em um número arbitrário de dimensões, isto é,

$$\mathbb{E}[f(\boldsymbol{X})] = \int_{-\infty}^{\infty} f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \approx \frac{1}{N}\sum_{j=1}^{N} f(\boldsymbol{X_j}) \ . \tag{1.33}$$

A Eq. (1.33) permite calcularmos a estimativa de qualquer momento a partir da distribuição de probabilidade amostrada. Por exemplo, a variância pode ser aproximada por

$$\begin{aligned} Var(\boldsymbol{X}) &= \mathbb{E}(\boldsymbol{X}^2) - [\mathbb{E}(\boldsymbol{X})]^2 \\ &\approx \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{X}_j^2 - \left[\frac{1}{N}\sum_{j=1}^{N} \boldsymbol{X}_j\right]^2 \ . \end{aligned} \tag{1.34}$$

Aparentemente, podemos realizar amostragens independentes da *posteriori* e, com certa precisão, estimá-la. De fato, é possível estimar a distribuição de probabilidade de um dado, assim como é possível amostrar uma distribuição de probabilidade uniforme. Entretanto, em todas essas situações, sabemos exatamente a forma da distribuição. Quando a distribuição é

desconhecida, não existe uma maneira simples de realizar amostragens independentes. Poderíamos utilizar uma distribuição conhecida, por exemplo, a distribuição uniforme, mapeada na distribuição desejada $P(\theta|D)$ dado uma condição de rejeição que adequasse a amostra à $P(\theta|D)$. Porém, esse processo apresenta um alto custo computacional que cresce exponencialmente com o número de dimensões do sistema [65]. A resposta mais atual para o problema de amostragem está em métodos que não fazem uso de informações sobre a estrutura global da *posteriori* (que usualmente é muito complexa) e focam em passos locais: técnicas denominadas de amostragem dependente, que apresentaremos a seguir.

1.5 Métodos de amostragem MCMC

Os métodos de amostragem estocásticos *Markov Chain Monte Carlo* (MCMC) [68] realizam amostragem a partir da construção de cadeias de Markov cuja distribuição de equilíbrio é a distribuição que desejamos amostrar, em nosso caso, a *posteriori*. Nesse contexto, as cadeias de Markov são caminhantes aleatórios que exploram o espaço de parâmetros condicionados por algum algoritmo que procura por regiões que contribuem mais para distribuição alvo. Além disso, são um caso particular de processos estocásticos que não possuem memória. Em outros termos, a distribuição de probabilidade condicional do estado seguinte da cadeia depende somente do estado atual. O MCMC difere da amostragem independente pelo fato de que sucessivos passos possuem um certo grau de correlação. O benefício dessa abordagem é que não precisamos saber, de antemão, a forma da *posteriori*. De maneira mais formal, partimos de uma distribuição arbitrária $\pi(\theta)$ e realizamos transições markovianas de acordo com um operador de transição $\mathcal{T}(\theta'|\theta)$. O operador de transição representa a distribuição de probabilidade condicional que delineia a probabilidade da cadeia transitar de θ para θ'. Assim, a integral que representa a distribuição da cadeia após o primeiro passo é

$$\pi(\theta') = \int d\theta \mathcal{T}(\theta'|\theta)\pi(\theta) \ ,$$

em que a integral cobre todos os valores de θ. Se considerarmos um conjunto de n passos sucessivos, podemos estimar a distribuição alvo $P(\boldsymbol{\theta}|D)$ via

$$\begin{aligned} \pi(\theta'') &= \int d\theta' \mathcal{T}(\theta''|\theta') \int d\theta \mathcal{T}(\theta'|\theta)\pi(\theta) \\ &\vdots \\ P(\theta|D) &\approx \int d\theta'^{[n]} \mathcal{T}(\theta'^{[n]}|\theta'^{[n-1]}) \ ... \int d\theta \mathcal{T}(\theta'|\theta)\pi(\theta) \ . \end{aligned}$$

No limite de um grande número de transições, a cadeia entra em equilíbrio e converge para a distribuição alvo. Assim, a *posteriori* é obtida por meio do registro dos passos da cadeia. Os

diversos algoritmos MCMC diferem pela forma com que as transições são realizadas. Aqui, apresentaremos alguns amostradores de maneira mais qualitativa. Neste trabalho, utilizamos as implementações disponíveis no pacote *pymc3* [69] da linguagem *Python*.

1.5.1 Amostrador de Metropolis

Um dos primeiros métodos MCMC foi proposto por Metropolis em 1953 [70] e, por conta disso, recebeu o nome de amostrador de Metropolis. Para ilustrar esse procedimento, consideramos uma situação hipotética em que queremos explorar uma região montanhosa, como proposto por Lambert [65]. Podemos comparar esse problema com a estimação de uma distribuição de probabilidade em duas dimensões. As regiões mais altas corresponderiam às áreas de maior probabilidade, em que, do ponto de vista probabilístico, gostaríamos de passar mais tempo. Nesse sentido, utilizamos o algoritmo de Metropolis para completar essa tarefa. Nessa ilustração, podemos imaginá-lo como um processo intermediário entre uma caminhada aleatória e uma escalada de alpinista. Na caminhada aleatória, andamos em qualquer direção com igual probabilidade e, portanto, não conseguiríamos identificar as regiões mais altas com facilidade. Do ponto de vista do amostrador, o resultado corresponderia a uma distribuição uniforme no espaço dos parâmetros. Na escalada, por sua vez, podemos impor que o único tipo de passo permitido é aquele em que subimos a montanha. Nesse caso, ascenderíamos em direção ao pico (moda da distribuição) e ali permaneceríamos por tempo indefinido. Esse não é o comportamento desejado, pois não teríamos informações sobre a distribuição como um todo. O algoritmo de Metropolis resolve essa dificuldade por meio de uma probabilidade de aceitação do passo definida por

$$\alpha = \min\left[1, \frac{P(\theta_{i+1}|D)}{P(\theta_i|D)}\right] \ ,$$

em que $P(\theta_i|D)$ é o valor da distribuição a *posteriori* no ponto θ_i do espaço de parâmetros e $P(\theta_{i+1}|D)$ tem a mesma interpretação para θ_{i+1}. Sorteamos um número aleatório uniformemente distribuído entre 0 e 1 – $u \sim \mathcal{U}(0,1)$ – e o comparamos com a probabilidade α: se $u < \alpha$, realizamos o passo; caso contrário, permanecemos na mesma posição. Quando a probabilidade no ponto seguinte é maior, isto é, $\alpha = 1$, movemos deterministicamente para aquela direção. Assim, o amostrador tende sempre a visitar regiões de maior densidade. Porém, quando a probabilidade é menor, movemos de forma probabilística com probabilidade α. Dessa maneira, conseguimos explorar, de acordo com os respectivos pesos, tanto pontos mais altos quanto mais baixos da montanha (distribuição de probabilidade alvo).

No caso unidimensional, partindo de um ponto arbitrário do espaço de parâmetros, é comum propormos o passo seguinte θ_{i+1} sorteando um valor de uma distribuição normal centrada em θ_i, $\mathcal{N}(\theta_i, \sigma)$. Podemos denominar esta distribuição de *distribuição de proposição*.

Nesse caso, ela é simétrica, uma vez que

$$\mathcal{N}(\theta_{i+1} - \theta_i, \sigma) = \mathcal{N}(\theta_i - \theta_{i+1}, \sigma) \ .$$

Essa simetria garante que a cadeia de Markov obedeça o princípio de balanço detalhado [65]. Em outras palavras, no equilíbrio, cada transição ocorre com igual probabilidade, qualquer seja sua direção. Então, quando $N \to \infty$, a distribuição estacionária obtida corresponde à distribuição da *posteriori*. Para outros casos em que as transições não são reversíveis, é necessário inserir um termo na distribuição de proposição para corrigir a assimetria. O Algoritmo 1 mostra a implementação desse procedimento de amostragem.

A dispersão σ da distribuição de proposição representa a amplitude do passo. Se o passo for muito grande, conseguimos facilmente encontrar regiões de maior densidade – também chamadas de *conjunto típico* – no entanto, é difícil explorá-las. Se o passo for pequeno, muitas iterações são necessárias para encontrar o conjunto típico, porém, quando encontrado, é bem explorado. Existe um tamanho ideal de σ que faz com que exploremos o espaço de parâmetros de maneira mais efetiva. Por isso, é comum executar previamente o algoritmo em uma fase de aquecimento, variando esse parâmetro a fim de determinar seu valor ótimo. Em uma dimensão, a taxa de aceitação ótima é cerca de 44% [65].

Algoritmo 1 Amostrador de Metropolis.

1: Inicialização da variável: $\theta_0 \leftarrow \theta$
2: **for** $i = 1, 2, \ldots$ **do**
3: Propor o passo: $\theta_{prop} \sim q(\theta_i|\theta_{i-1})$
4: Definir da probabilidade de aceitação: $\alpha = \min\left[1, \dfrac{P(\theta_{prop}|D)}{P(\theta_{i-1}|D)}\right]$
5: $u \sim \mathcal{U}(0, 1)$
6: **if** $u < \alpha$ **then**
7: Aceitar a proposição: $\theta_i \leftarrow \theta_{prop}$
8: **else**
9: Rejeitar a proposição: $\theta_i \leftarrow \theta_{i-1}$

1.5.2 Amostrador de Gibbs

É comum utilizar o método de Gibbs para estimar distribuições multivariadas com parâmetros $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$ [71]. Porém, é necessário conhecer as distribuições condicionais univariadas dos parâmetros para utilizar o algoritmo. Considerando um modelo com três parâmetros, isso significa que conseguimos determinar exatamente as expressões $p(\theta_1|\theta_2, \theta_3, D)$, $p(\theta_2|\theta_1, \theta_3, D)$ e $p(\theta_3|\theta_1, \theta_2, D)$. Além disso, devemos ser capazes de gerar amostras independentes a partir delas. Partindo de um ponto arbitrário do espaço de parâmetros, realizamos amostragens independentes para cada parâmetro e atualizamos nossa estimativa da distribuição a *posteriori*. Dessa maneira, a atualização da *posteriori* ocorre em uma dimensão por

vez no espaço de parâmetros. O Algoritmo 2 mostra o procedimento mínimo para realizar a amostragem de Gibbs.

A grande desvantagem do amostrador de Gibbs é que não conseguimos amostrar as distribuições de probabilidade condicionais na maioria dos casos. Por essa razão, é comum utilizar o método apenas para as variáveis que podem ser amostradas via esse método. Assim, aproveitamos o conhecimento da estrutura matemática do problema. Para os parâmetros restantes, utilizamos outros tipos de amostradores, como o algoritmo de Metropolis. Além disso, se a geometria da distribuição a *posteriori* não é favorável, o algoritmo falha em explorar regiões de alta correlação [65]. Um *software* que foi amplamente utilizado na inferência bayesiana com amostrador de Gibbs é o BUGS (*Bayesian inference Using Gibbs Sampling*) [72]. Atualmente, existem técnicas mais eficientes para estimação da *posteriori*, tal como o amostrador de Monte Carlo Hamiltoniano que será apresentado na próxima seção.

Algoritmo 2 Amostrador de Gibbs para três variáveis.

1: Inicialização aleatória da variável: $\boldsymbol{\theta}^0 \leftarrow (\theta_1^0, \theta_2^0, \theta_3^0)$
2: **for** $i = 1, 2, \ldots$ **do**
3: Amostragem independente: $\theta_1^i \sim P(\theta_1 | \theta_2^{i-1}, \theta_3^{i-1}, D)$
4: Amostragem independente: $\theta_2^i \sim P(\theta_2 | \theta_1^i, \theta_3^{i-1}, D)$
5: Amostragem independente: $\theta_3^i \sim P(\theta_3 | \theta_1^i, \theta_2^i, D)$

1.5.3 Amostrador de Monte Carlo Hamiltoniano (HMC)

A necessidade de um amostrador mais eficiente surge quando trabalhamos com distribuições multidimensionais da *posteriori*. A Figura 1.11a ilustra uma determinada região da *posteriori* em coordenadas esféricas (hipoteticamente em 1, 2 e 3 dimensões) correspondente à faixa modal da distribuição, também denominada conjunto típico, que estamos interessados em explorar. Para uma dimensão, a região de interesse possui o mesmo "volume" que seu entorno. No caso bidimensional, observamos que a vizinhança apresenta "volume" maior que a região de interesse. A Figura 1.11a também mostra que a diferença entre os volumes é ainda maior no caso tridimensional. De fato, esse comportamento escala rapidamente com o número crescente de dimensões [73]. Quanto maior a dimensionalidade da distribuição, maior é a discrepância entre o hiper-volume do conjunto típico e seus arredores. Como estamos interessados em determinar integrais do tipo

$$\mathbb{E}(q) = \int dq \pi(q) q \ , \tag{1.35}$$

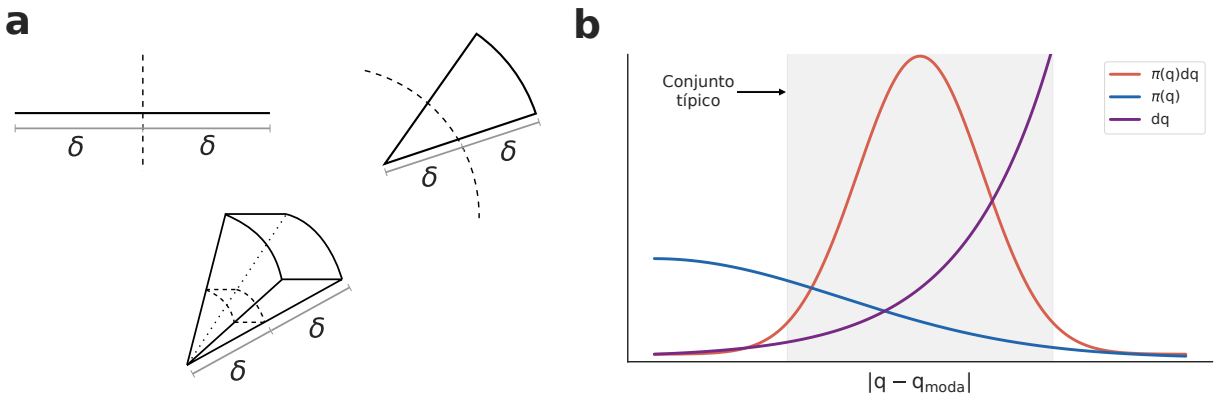

Figura 1.11: Geometria de distribuições multidimensionais. **(a)** O aumento da dimensão faz com que o volume externo à região de interesse seja drasticamente maior do que o da própria região. **(b)** Equilíbrio entre as grandezas volume dq e densidade $\pi(q)$ em relação ao integrando da Eq. (1.35).

é importante entender a relação entre o elemento de volume dq e a densidade $\pi(q)$ da distribuição alvo[1]. A Figura 1.11b mostra o equilíbrio entre a densidade e volume. Regiões próximas à moda possuem grande densidade, mas pequeno volume. Regiões distantes da moda apresentam baixa densidade, porém grande volume. O conjunto típico (região sombreada) representa a faixa em que existe equilíbrio entre as duas quantidades. Para casos multidimensionais, as regiões que apresentam simultaneamente volume e densidade elevados se tornam escassas. A consequência disso é o fenômeno conhecido como "concentração de medida", em que o conjunto típico se torna progressivamente mais singular [73]. Dessa maneira, acabamos por desperdiçar processamento computacional avaliando regiões que não contribuem para estimação das integrais, como a da Eq. (1.35).

O algoritmo de Metropolis não considera a geometria da distribuição. Por isso, sua eficiência cai drasticamente com o aumento de dimensões. Como o volume externo ao conjunto típico é maior, as cadeias tendem a migrar para essas regiões. Nessa faixa, a densidade dos pontos é tão pequena que praticamente todas as proposições são rejeitadas e as cadeias não conseguem avançar. O método de Monte Carlo Hamiltoniano (HMC) é uma alternativa ao problema que advém da dimensionalidade. Essa técnica leva em conta a geometria da distribuição da *posteriori* para realizar os passos.

O amostrador HMC realiza uma analogia física para propor os passos da cadeia de Markov [74, 75]. Nesse contexto, os parâmetros da distribuição *posteriori* são interpretados como a posição de um sistema físico de mecânica clássica. O objetivo é encontrar uma maneira de visitar diversas regiões da *posteriori* de modo determinístico. Dessa maneira, uma possível abordagem é realizar o cálculo da trajetória de um sistema da Física Clássica de acordo com suas equações de Hamilton. As equações de Hamilton são utilizadas porque possibilitam ex-

[1]Mudamos a notação do parâmetro da *posteriori* de θ para q, pois essa é a representação usualmente empregada para descrição do amostrador hamiltoniano.

plorar o espaço da *posteriori* segundo sua geometria. Para isso, consideramos a distribuição de probabilidade conjunta, dada por

$$\begin{aligned}\pi(\boldsymbol{q},\boldsymbol{p}) &= \pi(\boldsymbol{p}|\boldsymbol{q})\pi(\boldsymbol{q}) \\ &= \pi(\boldsymbol{p}|\boldsymbol{q})[p(D|\boldsymbol{q})p(\boldsymbol{q})] \ ,\end{aligned} \tag{1.36}$$

em que $\boldsymbol{q}$ são as coordenadas de posição (os parâmetros da *posteriori*), $\boldsymbol{p}$ é um conjunto de variáveis auxiliares que representam as coordenadas de momento de mesma dimensão que a posição, $\pi(\boldsymbol{p}|\boldsymbol{q})$ é a distribuição do momento condicionado à posição e $\pi(\boldsymbol{q})$ é a *posteriori*. O hamiltoniano é definido como

$$\mathcal{H}(\boldsymbol{q},\boldsymbol{p}) = -\log\pi(\boldsymbol{q},\boldsymbol{p}) \ , \tag{1.37}$$

de tal maneira que podemos interpretar o sistema como um *ensemble* canônico, isto é, a probabilidade pode ser escrita como

$$\begin{aligned}P &\propto e^{-\mathcal{H}(\boldsymbol{q},\boldsymbol{p})} \\ P &\propto \pi(\boldsymbol{q},\boldsymbol{p}) \ .\end{aligned} \tag{1.38}$$

Os dois termos do hamiltoniano correspondem à energia cinética e potencial do sistema, ou seja,

$$\mathcal{H}(\boldsymbol{q},\boldsymbol{p}) = K(\boldsymbol{p},\boldsymbol{q}) + V(\boldsymbol{q}) \ , \tag{1.39}$$

em que $K(\boldsymbol{p},\boldsymbol{q}) = -\log\pi(\boldsymbol{p}|\boldsymbol{q})$ é a energia cinética e $V(\boldsymbol{q}) = -\log\pi(\boldsymbol{q})$ a energia potencial. A energia potencial corresponde ao próprio logaritmo da *posteriori*. A energia cinética não possui restrições e podemos escolhê-la de maneira conveniente [73]. Ao considerarmos um sistema sem atrito, a energia é constante e todas as trajetórias estão confinadas num nível energético específico, isto é,

$$\mathcal{H}^{-1}(E) = \{\boldsymbol{q},\boldsymbol{p}|\mathcal{H}(\boldsymbol{q},\boldsymbol{p}) = E\} \ , \tag{1.40}$$

que resulta em hiper-superfícies $(2D-1)$-dimensionais no espaço de fase. Podemos decompor a distribuição canônica em termos microcanônicos por meio de

$$\pi(\boldsymbol{q},\boldsymbol{p}) = \pi(q_E|E)\pi(E), \tag{1.41}$$

sendo $\pi(q_E|E)$ a distribuição microcanônica e $\pi(E)$ a distribuição marginal das energias. De modo geral, o HMC consiste de um ciclo de duas etapas: a primeira delas é o cálculo determinístico das trajetórias; a segunda é a exploração estocástica dos níveis de energia por meio da variação do momento. A Figura 1.12a ilustra as duas etapas no espaço de fase. As elipses concêntricas representam energias totais $\mathcal{H}$ e as respectivas coordenadas permitidas.

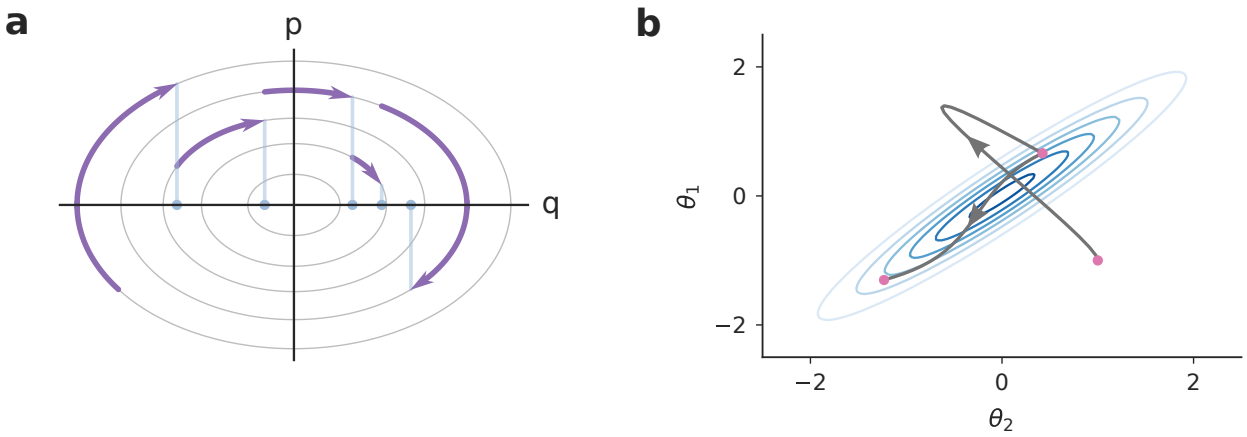

Figura 1.12: Amostrador de Monte Carlo Hamiltoniano. **(a)** Espaço de fases e transições energéticas do algoritmo HMC. **(b)** Trajetória da partícula no espaço de parâmetros de acordo com o algoritmo HMC.

A primeira fase é representada pelas curvas roxas e indicam a trajetória da partícula no espaço de fase. Após o término do movimento, armazenamos as coordenadas da partícula para construir a distribuição $\pi(\boldsymbol{q})$. A parte estocástica do método se refere ao sorteio de outro momento $\boldsymbol{p}$ que provoca a mudança do nível energético (elipse no caso da figura). Por isso, a escolha da energia cinética deve contribuir para que possamos explorar os níveis de energia de maneira eficiente. Desejamos que a distribuição de transições energéticas seja similar à distribuição marginal de energias uma vez que a distribuição marginal é representativa de todas as energias relevantes ao sistema. Nesse caso, as amostragens seriam realizadas de maneira idealmente independente [73]. Existem duas escolhas comuns para a energia cinética: Gaussiana-Euclidiana e Gaussiana-Riemanniana [73].

Uma energia cinética Gaussiana-Euclidiana utiliza a métrica euclidiana para gerar energias do tipo

$$K(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{M}^{-1}\boldsymbol{p} + \log|\boldsymbol{M}| + \textit{constante} \; . \tag{1.42}$$

A matriz $\boldsymbol{M}$ é conhecida como matriz de massa que representa transformações lineares realizadas na *posteriori* [76]. Os termos de variância têm o efeito de esticar ou comprimir a *posteriori* para que todos os parâmetros apresentem a mesma escala. Já os elementos de covariância rotacionam a *posteriori* para que os parâmetros possam ser considerados independentes. Se a matriz de massa apresenta forma semelhante à matriz de covariância da *posteriori*, conseguimos realizar amostragens independentes. O problema é que não possuímos essa informação de antemão.

A energia cinética Gaussiana-Riemanniana utiliza a métrica de Riemann para construção da matriz de massa e é definida por

$$K(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{2}\boldsymbol{p}^\top [\boldsymbol{\Sigma}(\boldsymbol{q})]^{-1}\boldsymbol{p} + \log|\boldsymbol{\Sigma}(\boldsymbol{q})| + \text{constante} \; , \tag{1.43}$$

em que $\boldsymbol{M} = \boldsymbol{\Sigma}(\boldsymbol{q})$ é uma matriz de massa dependente da posição $\boldsymbol{q}$. Nesse caso, podemos dizer que a matriz de massa e a métrica em si dependem da posição no espaço [73]. O efeito prático disso é a maior eficiência em regiões de alta curvatura espacial.

Para ilustrar o funcionamento do amostrador HMC, optamos por utilizar a energia gaussiana mais simples dada por

$$K(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{p} + \text{constante} \ , \tag{1.44}$$

que corresponde à escolha de uma distribuição gaussiana para momento do tipo $\boldsymbol{p} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Com essa escolha, voltamos nossa atenção para as equações de Hamilton

$$\begin{aligned} \frac{d\boldsymbol{q}}{dt} &= \frac{\partial \mathcal{H}(\boldsymbol{p}, \boldsymbol{q})}{\partial \boldsymbol{p}} = \frac{\partial K(\boldsymbol{p})}{\partial \boldsymbol{p}} + \frac{\partial V(\boldsymbol{q})}{\partial \boldsymbol{p}} \\ \frac{d\boldsymbol{p}}{dt} &= -\frac{\partial \mathcal{H}(\boldsymbol{p}, \boldsymbol{q})}{\partial \boldsymbol{q}} = -\frac{\partial K(\boldsymbol{p})}{\partial \boldsymbol{q}} + -\frac{\partial V(\boldsymbol{q})}{\partial \boldsymbol{q}} \ , \end{aligned} \tag{1.45}$$

que se reduzem a

$$\begin{aligned} \frac{d\boldsymbol{q}}{dt} &= \boldsymbol{p} \\ \frac{d\boldsymbol{p}}{dt} &= -\frac{\partial V(\boldsymbol{q})}{\partial \boldsymbol{q}} \ . \end{aligned} \tag{1.46}$$

Desse modo, a implementação do algoritmo consiste basicamente de três passos:

- Amostrar o momento $\boldsymbol{p} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$;
- Simular $\boldsymbol{q}(t)$ e $\boldsymbol{p}(t)$ usando as equações de Hamilton por T unidades de tempo;
- Armazenar os valores finais de $\boldsymbol{q}(T)$.

Para realizar o segundo passo, é necessário solucionar as equações diferenciais, o que não é realizável de modo analítico. Por isso, devemos recorrer a métodos numéricos que discretizam as equações de Hamilton. Vamos considerar o tamanho de cada passo discreto como ϵ e o número total de passos como L (a duração da trajetória). Um algoritmo muito utilizado para solução do conjunto de Eqs. (1.46) é chamado *leapfrog* [77] e pode ser descrito como

$$\begin{aligned} \boldsymbol{p}_{t+\epsilon/2} &= \boldsymbol{p}_t + (\epsilon/2)\nabla_{\boldsymbol{q}} V(\boldsymbol{q}_t) \\ \boldsymbol{q}_{t+\epsilon} &= \boldsymbol{q}_t + \epsilon \boldsymbol{p}_{t+\epsilon/2} \\ \boldsymbol{p}_{t+\epsilon} &= \boldsymbol{p}_{t+\epsilon/2} + (\epsilon/2)\nabla_{\boldsymbol{q}} V(\boldsymbol{q}_{t+\epsilon}) \ , \end{aligned} \tag{1.47}$$

sendo o índice das variáveis referente ao número de iterações do procedimento numérico e

∇_q o diferencial em relação à posição, isto é,

$$\nabla_{\boldsymbol{q}} V(\boldsymbol{q}_t) \to \frac{\partial V(q_{t,i})}{\partial q_i} \ , \tag{1.48}$$

em que o índice i corresponde a i-ésima coordenada. O integrador *leapfrog* utiliza as próprias coordenadas para atualizar a posição no espaço de fase. Isso garante que as atualizações preservem o volume e será importante para que a cadeia respeite o princípio de balanço detalhado [78]. Quando terminamos o cálculo da trajetória, realizamos o passo com probabilidade

$$\alpha = \min\left(1, \frac{\exp V(\tilde{\boldsymbol{q}}) - \frac{1}{2}\tilde{\boldsymbol{p}} \cdot \tilde{\boldsymbol{p}}}{\exp V(\boldsymbol{q}_0) - \frac{1}{2}\boldsymbol{p}_0 \cdot \boldsymbol{p}_0}\right) \tag{1.49}$$

em que $\tilde{\boldsymbol{p}}$ é o momento no último passo, $\tilde{\boldsymbol{q}}$ a posição no último passo, $\boldsymbol{p}_0$ o momento sorteado no início e $\boldsymbol{q}_0$ a posição inicial. A razão de probabilidades da Eq. (1.49) representa a perda da energia durante a trajetória de 0 a $T = \epsilon L$. Como o integrador é uma aproximação discreta, é natural que exista uma diferença entre a trajetória estimada e a verdadeira [73]. Se conseguíssemos calcular a dinâmica analiticamente, a razão seria sempre $\alpha = 1$ e as proposições seriam sempre aceitas. Comparativamente, o algoritmo de Metropolis utiliza apenas a energia potencial no cálculo de α, aceitando deterministicamente proposições para as quais essa energia diminui, e com probabilidade α quando ela aumenta. O algoritmo 3 mostra o código mínimo para implementação do HMC. É importante notar que, caso a proposição seja aceita ao final da trajetória, as coordenadas de momento armazenadas têm o sinal trocado, para que exista reversibilidade temporal e o balanço detalhado valha [73].

Como as proposições são aleatórias no caso dos amostradores de Metropolis e Gibbs, o caminhante pode ter dificuldade para propor passos que levam a faixas distantes do estado atual dependendo da geometria do sistema. Caso isso aconteça, o amostrador demora muito tempo para se deslocar entre regiões devido à alta autocorrelação entre os passos [76]. É importante notar que todos os métodos de amostragem apresentados aqui possuem certo grau de autocorrelação entre seus passos. Porém, dentre todos, o método HMC é o que possui menor grau de autocorrelação. O caminhante HMC consegue propor passos em praticamente qualquer ponto do espaço da *posteriori*. A Figura 1.12b ilustra o comportamento de uma trajetória calculada após duas iterações do Algoritmo 3 para uma distribuição bidimensional. Podemos observar que os passos são realizados para regiões distintas no espaço da *posteriori*.

O algoritmo HMC é eficiente em explorar o espaço da *posteriori*, mas possui desvantagens por ser um método paramétrico. De fato, a técnica é muito sensível à escolha dos parâmetros ϵ (o tamanho do passo) e L (o número total de passos) [78]. Se a trajetória for curta, as amostras sucessivas serão tão próximas que o comportamento será equivalente a de um caminhante aleatório. Se L for muito grande, as trajetórias podem entrar num ciclo e acabam

Algoritmo 3 Amostrador de Monte Carlo Hamiltoniano

Inicialização das variáveis: $\boldsymbol{q}_0$, ϵ, L
for $i = 1, 2, \ldots$ **do**
    Amostrar o momento: $\boldsymbol{p}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
    Definir: $\tilde{\boldsymbol{q}} \leftarrow \boldsymbol{q}_{i-1}, \tilde{\boldsymbol{p}} \leftarrow \boldsymbol{p}_0$
    for $j = 1$ to L **do**
        Definir: $\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}} \leftarrow \text{Leapfrog}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}}, \epsilon)$
    Definir a probabilidade de aceitação: $\alpha = \min\left(1, \dfrac{\exp V(\tilde{\boldsymbol{q}}) - \frac{1}{2}\tilde{\boldsymbol{p}} \cdot \tilde{\boldsymbol{p}}}{\exp V(\boldsymbol{q}_{i-1}) - \frac{1}{2}\boldsymbol{p}_0 \cdot \boldsymbol{p}_0}\right)$
    $u \sim \mathcal{U}(0, 1)$
    if $u < \alpha$ **then**
        Aceitar a proposição: $\boldsymbol{q}_i \leftarrow \tilde{\boldsymbol{q}}, \boldsymbol{p}_i \leftarrow -\tilde{\boldsymbol{p}}$
    else
        Rejeitar a proposição: $\boldsymbol{q}_i \leftarrow \boldsymbol{q}_{i-1}, \boldsymbol{p}_i \leftarrow \boldsymbol{p}_{i-1}$
function Leapfrog$(\boldsymbol{q}, \boldsymbol{p}, \epsilon)$
Definir: $\tilde{\boldsymbol{p}} \leftarrow \boldsymbol{p} + (\epsilon/2)\nabla_{\boldsymbol{q}} V(\boldsymbol{q})$
Definir: $\tilde{\boldsymbol{q}} \leftarrow \boldsymbol{q} + \epsilon\tilde{\boldsymbol{p}}$
Definir: $\tilde{\boldsymbol{p}} \leftarrow \tilde{\boldsymbol{p}} + (\epsilon/2)\nabla_{\boldsymbol{q}} V(\tilde{\boldsymbol{q}})$
return $\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}}$

por explorar as mesmas regiões sem necessidade. Para contornar essa situação, foi criado um algoritmo chamado *NUTS* (*No U-Turn Sampler*) [78], que evita que as trajetórias deem meia volta.

Nesse procedimento, a trajetória é construída acumulando passos em direções temporais opostas e aleatoriamente escolhidas. Em outras palavras, primeiramente, o movimento é calculado realizando dois passos para frente e, em seguida, a próxima direção é definida a partir do sorteio de um número aleatório $u \sim \mathcal{U}(\{-1, 1\})$. Na próxima etapa, realizamos o dobro de passos na direção sorteada e, assim, sucessivamente. O número de dobras é chamado de *comprimento da árvore*. Esse algoritmo permite a criação de trajetórias que não são tão longas e nem tão curtas [76]. Sendo $\boldsymbol{z}_-(t)$ e $\boldsymbol{z}_+(t)$ as fronteiras da trajetória no tempo t e $\boldsymbol{q}_\pm(t)$ e $\boldsymbol{p}_\pm(t)$ as respectivas posições e momentos, um critério de parada é satisfeito caso o caminhante execute meia volta, isto é,

$$\begin{aligned} &\boldsymbol{p}_+(t)^\top \cdot [\boldsymbol{q}_+(t) - \boldsymbol{q}_-(t)] < 0 \\ \text{e} \quad &\boldsymbol{p}_-(t)^\top \cdot [\boldsymbol{q}_-(t) - \boldsymbol{q}_+(t)] < 0 \end{aligned}, \tag{1.50}$$

em que os momentos em ambas as pontas estão alinhados de maneira oposta à linha que une as posições [73]. Outro critério de parada é a divergência do erro de aproximação, isto é, valores da energia total $\mathcal{H} \to \infty$. Para o algoritmo NUTS, a amostra da i-ésima iteração é escolhida aleatoriamente como qualquer posição por onde o caminhante passou.

Se o valor de ϵ for muito grande, conseguimos explorar melhor a distribuição. Porém, o crescente erro de aproximação pode levar a maiores variações na energia total de tal maneira que $\mathcal{H} \to \infty$. Denominamos essa situação como uma *transição divergente*, que acarreta baixas taxas de aceitação [76]. Por outro lado, se ϵ é pequeno, o processamento computacional será desperdiçado dando pequenos passos. Portanto, é necessário encontrar o tamanho do passo ideal. Para isso, é comum escolher o valor de ϵ adaptativamente. O processo é realizado seguindo uma adaptação do algoritmo de Nesterov [79], que passaremos a descrever a seguir. Definindo a estatística G_t como

$$G_t = \delta - \alpha_t \ , \tag{1.51}$$

em que δ é a probabilidade de aceitação desejada e α_t é a probabilidade de aceitação no tempo t. O valor esperado dessa quantidade é

$$\mathbb{E}_t[G_t|\epsilon] = g(\epsilon) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[G_t|\epsilon] \ . \tag{1.52}$$

No caso em que $g(\epsilon)$ é uma função não decrescente, se realizarmos as atualizações seguindo

$$\epsilon_{t+1} \leftarrow \epsilon_t - \nu_t G_t \ , \tag{1.53}$$

sendo ν_t o tamanho do passo dessas iterações, conseguimos fazer com que $g(\epsilon) \to 0$. O que é equivalente a alcançar taxa de aceitação desejada $\delta \approx \alpha_t$. Essas condições são satisfeitas se $\sum_t \nu_t \to \infty$ e $\sum_t \nu_t^2 < \infty$ [79]. Especificamente, a escolha $\nu_t = t^{-\kappa}$ com $\kappa \in (0.5, 1]$ vai ao encontro dessas condições [79]. Na prática, como mencionado anteriormente, o procedimento mais utilizado é uma adaptação do acima descrito. Esse procedimento recebe o nome de algoritmo de dupla média (*dual-averaging algorithm*) e suas atualizações são realizadas de acordo com a equação [79]

$$\epsilon \leftarrow \mu - \frac{\sqrt{t}}{\gamma} \frac{1}{t + t_0} \sum_{i=0}^{t} G_i \ , \tag{1.54}$$

em que μ é o valor escolhido para o qual as estimativas de ϵ_t são reduzidas, γ é um parâmetro que define a intensidade de concentração para μ, t_0 estabiliza as iterações iniciais e a somatória faz parte da média sobre os valores de G_t. Dessa maneira, podemos estimar uma quantidade média $\overline{\epsilon}_t = \epsilon_t$ e realizar atualizações via

$$\overline{\epsilon}_{t+1} \leftarrow \nu_t \epsilon_{t+1} + \overline{\epsilon}_t - \nu_t \overline{\epsilon}_t \ , \tag{1.55}$$

com $\nu_t = t^{-\kappa}$. Como no algoritmo NUTS não existe uma probabilidade de aceitação, utilizamos a taxa de aceitação média do algoritmo HMC para última iteração dobrada [78]. Para decidir o tamanho do passo, realizamos várias iterações de aquecimento para estimar o tama-

nho do passo ótimo para amostragem. Dessa maneira, utilizando o NUTS e o *dual-averaging algorithm*, conseguimos estimar o tamanho do passo e o comprimento da trajetória ideais para o amostrador hamiltoniano. As Figuras 1.13a-c mostram a performance dos amostradores nas 100 primeiras iterações para uma distribuição bidimensional do tipo

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{\mu} = \begin{bmatrix}0\\ 0\end{bmatrix}, \boldsymbol{\sigma} = \begin{bmatrix}1 & 0.95\\ 0.95 & 1\end{bmatrix}\right) ,$$

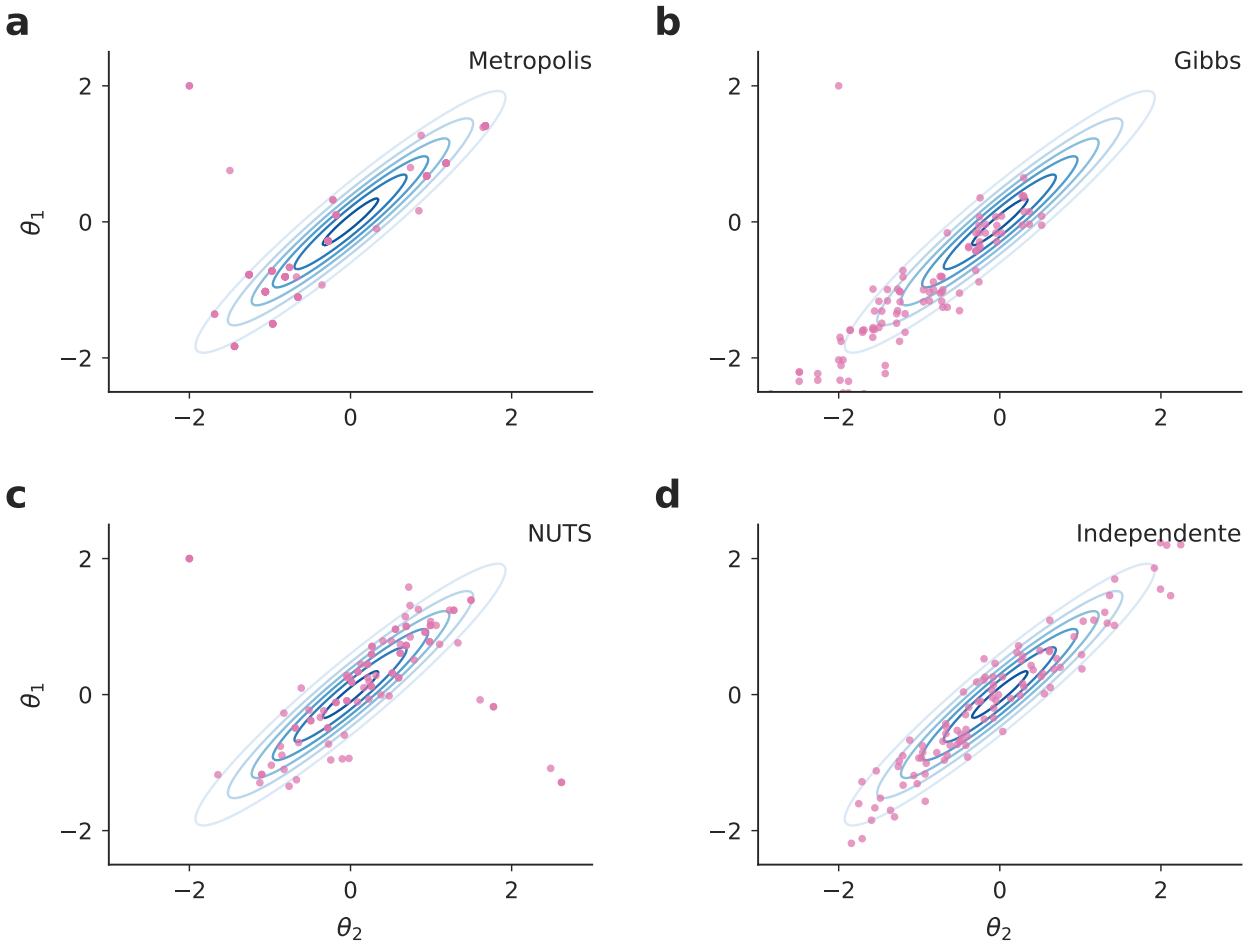

Figura 1.13: Tipos de amostradores e sua performance. **(a)** Amostragem de Metropolis; **(a)** Amostragem de Gibbs; **(a)** Amostragem NUTS; **(a)** Amostragem independente.

em que há correlação entre as variáveis θ_1 e θ_2. Todos os amostradores começam do mesmo ponto inicial $(-2, 2)$. A Figura 1.13d mostra como seria idealmente a amostragem independente. O amostrador de Metropolis visivelmente rejeita mais proposições, visto que a quantidade de pontos na Figura 1.13a é muito menor do que para os outros casos. Isso possui reflexo na eficiência computacional do algoritmo, que leva mais tempo para estimar a *posteriori*. O amostrador de Gibbs resolve esse problema por meio da amostragem das distribuições condicionais. Porém, ele não consegue lidar bem com a geometria da distri-

buição com poucas iterações, e é enviesado na direção inferior da distribuição. O melhor amostrador é o HMC com algoritmo NUTS, que traça trajetórias realizando uma analogia física e evitando trajetórias redundantes. Dentre todos os amostradores da Figura 1.13, ele é o mais próximo da amostragem independente. Para um número grande de iterações, todos os amostradores tendem a aproximar bem a *posteriori* quando a geometria é simples, como a distribuição da Figura 1.13. No entanto, quando temos *posterioris* mais complicadas, o NUTS exibe uma melhor performance[2].

Avaliando a convergência da cadeia de Markov

Na prática, nunca sabemos ao certo o formato da *posteriori* (se soubéssemos, não estaríamos estudando o sistema em questão), então precisamos buscar alguma forma de quantificar a convergência das cadeias de Markov. De fato, não existe um método infalível para comprovar a convergência. Uma prática muito comum é executar diversas cadeias de Markov e calcular duas quantidades. Uma delas é a variância dentro da cadeia que, para um sistema unidimensional com parâmetro θ, pode ser definida como

$$W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_{ij} - \overline{\theta}_j)^2 \ , \tag{1.56}$$

sendo m o número total de cadeias e n o total de iterações para cada cadeia. Aqui, o índice j se refere às cadeias, o índice i se refere à i-ésima amostra e $\overline{\theta}_j$ é a média do parâmetro para cadeia j. A outra quantidade é variância entre cadeias, que pode ser escrita como

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\overline{\theta}_j - \overline{\theta})^2 \ ,$$

sendo $\overline{\theta}$ a média do parâmetro para todas as cadeias. Quando W e B apresentam valores próximos, temos indícios de que as cadeias estão bem "misturadas". Assim, podemos concluir que elas provavelmente chegaram ao estado de equilíbrio, sendo a *posteriori* bem estimada. Essa ideia é resumida pela estimativa da variância da *posteriori* proposta por Gelman e Rubin [80]

$$\begin{aligned} var(\hat{\theta}|D) &= \frac{n-1}{n} W + \frac{1}{n} B \\ &= W + \frac{1}{n}(B - W) \ , \end{aligned} \tag{1.57}$$

que superestima a variância, mas é não-enviesada quando as cadeias atingem o estado estacionário. Se $B \to W$ ou se $n \to \infty$, a variância da *posteriori* corresponde exatamente à

[2]Uma visualização interessante para comparação das técnicas de amostragem de diferentes distribuições é o aplicativo criado por Chi Feng (https://chi-feng.github.io/mcmc-demo/app.html).

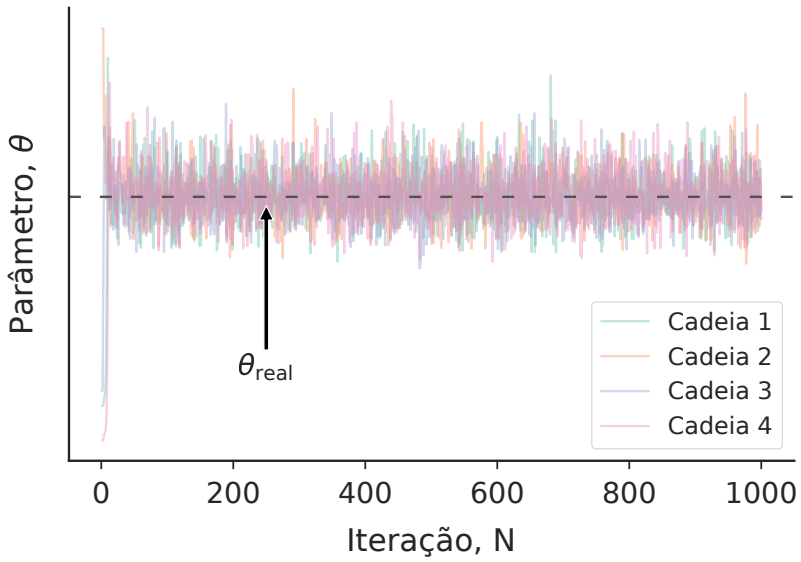

Figura 1.14: Mistura das cadeias de Markov. Um *trace plot* de um processo ilustrativo de amostragem usando quatro cadeias de Markov.

variância interna $var(\hat{\theta}|D) \to W$, indicando a mistura das cadeias. Gelman e Rubin também sugerem o cálculo da quantidade

$$\hat{R} = \sqrt{\frac{W + \frac{1}{n}(B - W)}{W}} \ , \tag{1.58}$$

que representa a razão entre a estimativa da variância e a variância desejada W. No início da amostragem, como $B \gg W$ temos que o valor $\hat{R} \gg 1$. Porém, no decorrer do processo de amostragem, a tendência é que $B \to W$. Nesse caso, a estatística $\hat{R} \to 1$, que é o valor desejado. É de praxe interpretar um valor de $\hat{R} \approx 1.1$ como suficiente para considerar as cadeias misturadas [65]. A Figura 1.14 mostra o *trace plot* (série temporal da estimativa do parâmetro amostrado) que indica o comportamento ideal de mistura das cadeias em relação à estimativa de um parâmetro θ. Nesse caso, todo o espaço da *posteriori* parece ter sido explorado.

Como estamos usando cadeias de Markov, existe uma autocorrelação entre passos sucessivos. Para estimar a quantidade de passos efetivos, podemos calcular o tamanho efetivo da amostra por [81]

$$n_{ef} = \frac{mT}{1 + 2\sum_{\tau=1}^{\infty} \rho_\tau} \ , \tag{1.59}$$

em que m é o número de cadeias, T o número de passos de cada cadeia e ρ_τ a autocorrelação com *lag* τ. Usualmente, não sabemos o valor de ρ_τ e, por isso, utilizamos uma estimativa amostral $\hat{\rho}_\tau$. Dessa maneira, n_{ef} corresponde à quantidade de passos que foram realizados

de modo independente para o amostrador baseado em cadeias de Markov.

1.6 Modelos hierárquicos bayesianos

Neste trabalho, utilizamos modelos lineares mistos com amostragem bayesiana para avaliar a influência da produtividade sobre o impacto. Com intuito de realizar a amostragem, precisamos escrever a distribuição *posteriori* do modelo hierárquico descrito pela Eq. (1.16), isto é,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\boldsymbol{u} + \boldsymbol{\varepsilon} \ , \tag{1.60}$$

em que $\boldsymbol{Y}$ é a variável dependente, $\boldsymbol{X}$ a variável independente, $\boldsymbol{\beta}$ os coeficientes do modelo (intercepto e inclinação), $\boldsymbol{Z}$ a matriz modelo de efeitos aleatórios, $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ a matriz de covariância, $\boldsymbol{u}$ uma variável aleatória esférica e $\boldsymbol{\varepsilon}$ o erro do modelo.

Considerando um sistema de dois níveis, como aquele representado na Figura 1.6, podemos supor que o conjunto de dados possui uma estrutura hierárquica genérica com l grupos. Existem $\{n_1, \ldots, n_l\}$ observações respectivas aos l grupos. Dessa forma, o número total de dados deve ser $N = \sum_{j=1}^{l} n_j$. A verossimilhança da i-ésima amostra e j-ésimo grupo pode ser expressa como

$$\boldsymbol{Y}_{ij}|\boldsymbol{\theta}_j \sim P(y_{ij}|\boldsymbol{\theta}_j) \ . \tag{1.61}$$

Supondo que as observações de cada grupo sejam independentes entre si, podemos escrever a verosimilhança do j-ésimo grupo como o produto das verossimilhanças das amostras [81], ou seja,

$$P(\boldsymbol{y}_j|\boldsymbol{\theta}_j) = \prod_{i=1}^{n_j} P(y_{ij}|\boldsymbol{\theta}_j) \ . \tag{1.62}$$

Na abordagem bayesiana, os "parâmetros dos parâmetros" são usualmente chamados de hiper-parâmetros, enquanto as distribuições a *priori* correspondentes são denominadas distribuições a *hiper-priori* [65]. Em particular, no modelo hierárquico, os parâmetros $\boldsymbol{\beta}$ advêm da distribuição a *hiper-priori* $P(\boldsymbol{\phi})$ dos hiper-parâmetros $\boldsymbol{\phi}$ e, por isso, a distribuição a *priori* depende dessa distribuição subjacente. A distribuição a *hiper-priori* permite que grupos com pouca quantidade de dados "emprestem força estatística" dos grupos com maior disponibilidade de dados [82]. Considerando que os parâmetros específicos de cada j-ésimo grupo são independentes, podemos escrever a distribuição a *priori* como [81]

$$P(\boldsymbol{\theta}|\boldsymbol{\phi}) = \prod_{j=1}^{l} P(\boldsymbol{\theta}_j|\boldsymbol{\phi}) \ . \tag{1.63}$$

A partir da escolha das distribuições a *priori* e *hiper-priori* adequadas, a distribuição a

posteriori completa para um modelo hierárquico de dois níveis pode ser descrita como [81]

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\phi}|D) &\propto P(D|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{\phi})P(\boldsymbol{\phi}) \\ &\propto p(\boldsymbol{\phi}) \prod_{j=1}^{l} P(\boldsymbol{\theta}_j|\boldsymbol{\phi})P(\boldsymbol{y}_j|\boldsymbol{\theta}_j) \ , \end{aligned} \tag{1.64}$$

para a qual aplicamos os métodos de amostragem da seção anterior a fim de estimá-la.

1.7 Estimadores-M

Quando estamos analisando um conjunto de dados, é de interesse calcular medidas que resumam suas características para que possamos analisar o sistema numa escala global. A média de uma amostra de uma variável unidimensional y,

$$\mu = \frac{1}{N-1} \sum_{i=1}^{N} y_i \ , \tag{1.65}$$

em que N é o tamanho da amostra e y_i é a i-ésima amostra, é, talvez, a estatística de localização mais utilizada para esse propósito. Para representar o grau de dispersão da amostra, também chamado de escala, é comum estimarmos sua variância

$$\sigma^2 = \frac{1}{N-1} \sum_{i} (y_i - \mu)^2 \ . \tag{1.66}$$

Essas duas medidas têm desempenho ótimo quando utilizadas para um conjunto de dados cuja distribuição é normal. No entanto, sua sensibilidade é alta e a presença de *outliers*[3] pode fazer com que elas não representem bem a amostra. Por exemplo, na presença de um único *outlier* divergente $y_k \to \infty$, essas duas estatísticas também divergem.

Quando trabalhamos com uma amostra em que há *outliers*, como é o caso deste trabalho, surge a necessidade de utilizar estatísticas robustas a esse tipo de comportamento. Uma medida de localização que poderíamos empregar é a mediana, definida como o valor que divide a amostra em duas metades. Essa quantidade é robusta à presença de *outliers*. Além disso, para estimar a escala, podemos empregar o desvio da mediana (MAD), definido como a mediana dos desvios absolutos da mediana da amostra, isto é,

$$\text{MAD} = k \ \text{mediana}(|y_i - \text{mediana}(y)|) \ , \tag{1.67}$$

em que $k = 1.4826$ é uma constante que torna o MAD um estimador consistente com o desvio

[3] *Outliers* são pontos que apresentam valores muito discrepantes quando comparados ao restante da amostra.

padrão. Conforme veremos, nosso conjunto de dados é constituído principalmente por duas variáveis: a produtividade anual e o impacto médio das revistas em que um pesquisador publica no ano respectivo. A natureza dessas variáveis é, respectivamente, discreta e quasi-discreta. Por causa disso, não é interessante trabalhar com a mediana, uma vez que não estamos interessados em saber o ponto médio da amostra, mas sim o seu comportamento médio de acordo com o ano e a área do pesquisador. Assim, precisamos de estatísticas que intuitivamente se pareçam com a média e a variância, mas sejam robustas à presença de *outliers*.

Para descrever um conjunto de medidas com essas propriedades, os chamados estimadores-M, considere que $f(y;\mu,\sigma)$ é a distribuição de probabilidade que gera uma amostra da variável aleatória y. Os parâmetros de localização μ e de escala σ são inicialmente desconhecidos. Para determiná-los, consideramos a verossimilhança da amostra dada por

$$\mathcal{L}(\mu,\sigma) = \sum_i \sigma^{-1} f\left(\frac{y_i-\mu}{\sigma}\right) \ , \tag{1.68}$$

em que escrevemos a distribuição de probabilidade normalizada e centrada na origem, com a somatória abrangendo todo conjunto de dados. Como a transformação logarítmica preserva as características da verossimilhança, podemos considerar o logaritmo negativo da verossimilhança, isto é,

$$\rho = -\log\mathcal{L}(\mu,\sigma) = \sum_i \left[\log\sigma - \log f\left(\frac{y_i-\mu}{\sigma}\right)\right] \ . \tag{1.69}$$

Assim, podemos tornar este um problema de maximização da verossimilhança[4]. De fato, a origem do nome estimadores-M está diretamente associado ao "M" no termo maximização da verossimilhança. Qualquer variável definida pela maximização da expressão

$$\sum_i \psi(y_i;\theta) = 0 \ , \tag{1.70}$$

em que $\psi(y_i;\theta)$ é a derivada de ρ em relação ao parâmetro θ, pode ser considerada como um estimador-M [83]. Especificamente, se considerarmos a Eq. (1.70) para os casos da localização μ e escala σ definidos pela Eq. (1.69), temos

$$\begin{aligned} &\sum_i \psi\left(\frac{y_i-\mu}{\sigma}\right) = 0 \\ &\sum_i \left[\left(\frac{y_i-\mu}{\sigma}\right)\psi\left(\frac{y_i-\mu}{\sigma}\right) - 1\right] = 0 \ . \end{aligned} \tag{1.71}$$

[4]De fato, ao tomar o negativo da verossimilhança, estamos tornando esse um problema de minimização. Decidimos adotar essa linha de raciocínio para manter a notação adotada por Huber [83].

Para resolver as Eqs. (1.71), é importante escolhermos a função ψ adequada [84]. Nesse contexto, a função ψ representa a função geradora da amostra e diferentes situações podem requerer diferentes escolhas para ψ. Por exemplo, a função

$$\psi(y) = \begin{cases} y & \text{se } |y| < c \\ 0 & \text{caso contrário} \end{cases}, \tag{1.72}$$

representa a média cortada, na qual designamos peso nulo aos *outliers* definidos como valores maiores que uma constante arbitrária c. Dessa maneira, a constante c pode ser considerada como o ponto de corte. Outra possibilidade é considerar

$$\psi(y) = \begin{cases} -c & \text{se } y < -c \\ y & \text{se } |y| < c \\ c & \text{se } y > c \end{cases}, \tag{1.73}$$

em que atribuimos a todos os *outliers* o peso constante definido pelo valor de corte c. Essa é a função proposta por Huber [85]. Podemos integrar ψ para obter a distribuição de probabilidade geradora a menos de uma constante. Em ambos os casos anteriores, constatamos que a parte central da distribuição corresponde a uma distribuição gaussiana. No caso da proposta de Huber, as caudas são distribuições exponenciais duplas, isto é,

$$\rho_H(y) = \begin{cases} y^2 & \text{se } |y| < c \\ c(2|y| - c) & \text{caso contrário} \end{cases}, \tag{1.74}$$

em que ρ_H é o logaritmo da verossimilhança para definição de Huber. Outras escolhas para a função ψ incluem a função de duplo peso de Tukey

$$\psi(y) = y\left[1 - \left(\frac{y}{R}\right)^2\right]_+^2, \tag{1.75}$$

em que R é uma constante e o símbolo + corresponde à parte positiva da função, e a função de Hampel

$$\psi(y) = sign(x) \begin{cases} |y| & \text{se } 0 < |y| < a \\ a & \text{se } a < |y| < b \\ a(c - |y|)/(c - b) & \text{se } b < |y| < c \\ 0 & \text{se } c < |y| \end{cases}, \tag{1.76}$$

em que a, b e c são constantes. A Figura 1.15 ilustra as quatro possibilidades descritas anteriormente. Em nosso trabalho, escolhemos a função de Huber para estimar os parâmetros de localização e escala. Nesse contexto, precisamos estimar conjuntamente ambos os parâ-

metros. Dessa forma, é preciso realizar uma pequena correção na equação de maximização do parâmetro de escala, a fim de que sua estimativa não seja enviesada para distribuição normal [84, 86]. Essa modificação corresponde a

$$\begin{aligned} \sum_i \left[\left(\frac{y_i - \mu}{\sigma} \right) \psi \left(\frac{y_i - \mu}{\sigma} \right) \right] &= (n-1)a(c) \\ \sum_i \psi \left(\frac{y_i - \mu}{\sigma} \right) &= 0 \end{aligned} \quad , \tag{1.77}$$

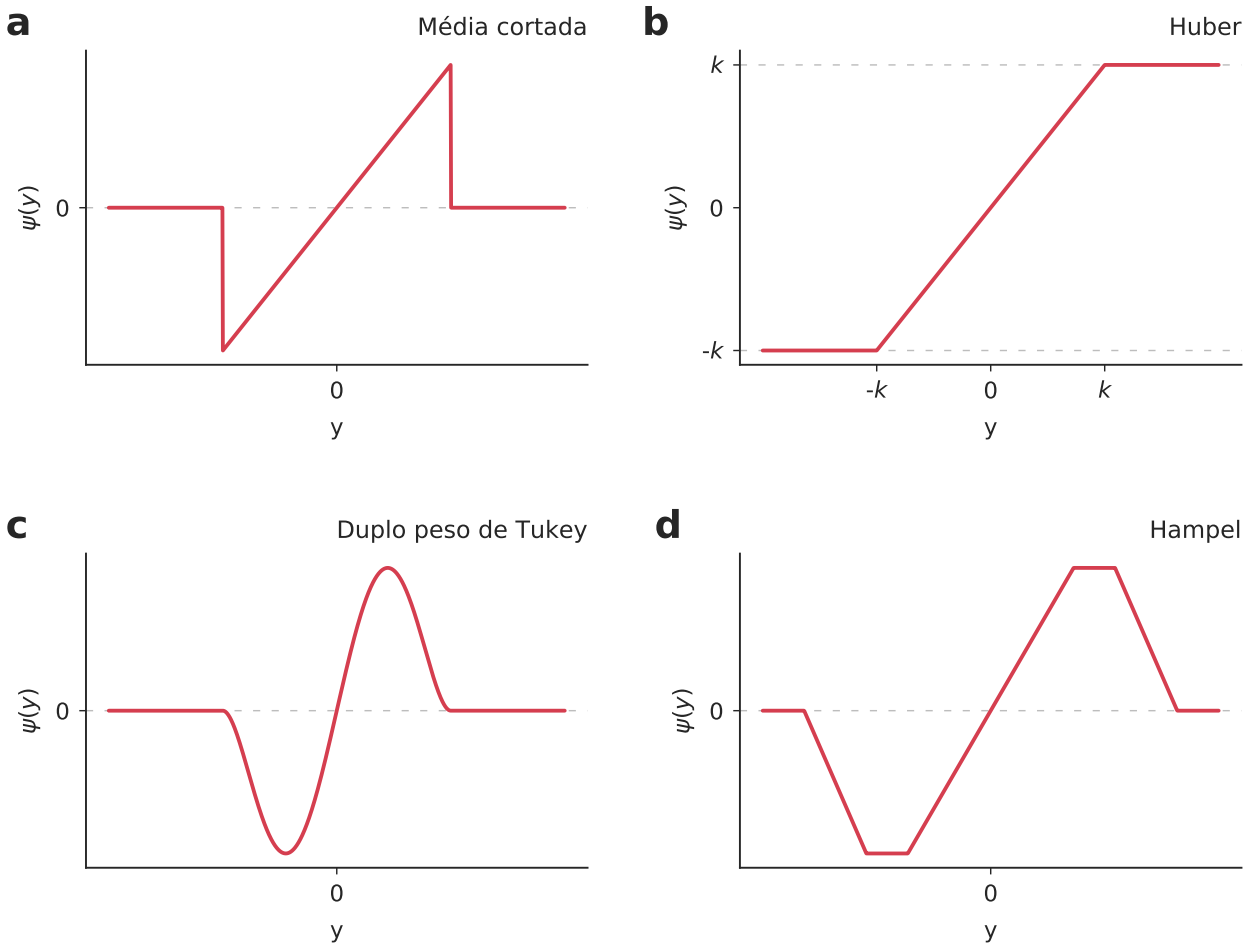

Figura 1.15: Diferentes funções $\psi(y)$ para determinação do estimador-M. (a) Média cortada. **(b)** Função de Huber. **(c)** Função duplo peso de Tukey. **(d)** Função de Hampel.

em que $a(c)$ é uma constante escolhida para que a estimativa não seja enviesada. Computacionalmente, para avaliar as raízes das Eqs. (1.77), utilizamos o método numérico de Newton [87]. Partindo de um valor inicial próximo do valor verdadeiro do parâmetro, o método consiste em aproximar a função como a reta tangente a ela. Em seguida, calculamos a raiz para essa aproximação e repetimos o processo até que uma variação mínima entre sucessivas iterações seja satisfeita. Para uma função arbitrária $h(x)$, a reta tangente estimada

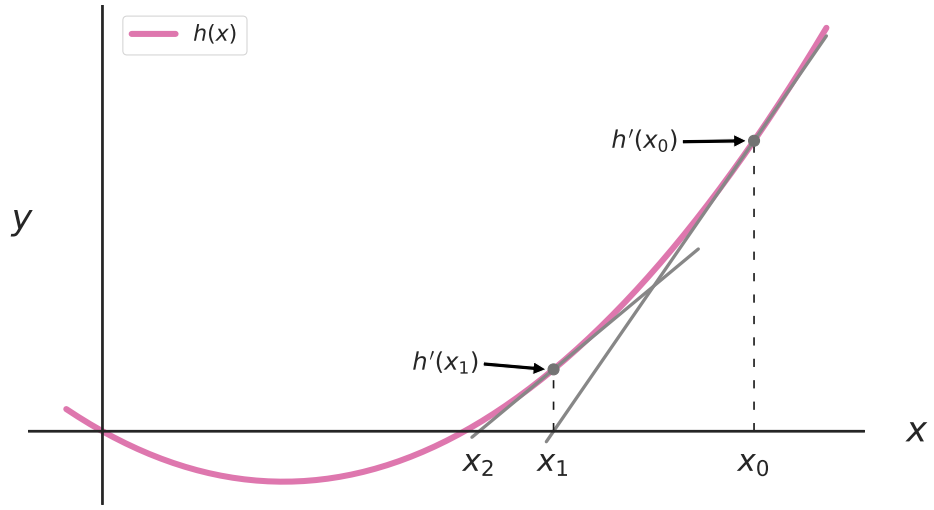

Figura 1.16: Ilustração do método de Newton para uma função arbitrária $h(x)$.

a partir do valor inicial x_n é descrita por

$$y = h'(x_n)(x - x_n) + h(x_n) \ . \tag{1.78}$$

O valor do intercepto no eixo x, isto é, a raiz estimada a partir dessa aproximação é obtida igualando a Eq. (1.78) a zero, o que nos leva a

$$x_{n+1} = x_n - \frac{h'(x_n)}{h(x_n)} \ . \tag{1.79}$$

A Figura 1.16 mostra o processo para duas iterações do método de Newton para uma função arbitrária $h(x)$. Em nossos resultados, consideramos as estimativas iniciais como sendo a mediana e o MAD para estimar os parâmetros de localização μ e de escala σ na proposta de Huber. Nesse caso, as iterações são realizadas simultaneamente pelas equações

$$\begin{aligned} [\sigma_{n+1}]^2 &= \frac{1}{(n-1)a(c)} \sum_i \psi^2(y_{n,i})[\sigma_n]^2 \\ \mu_{n+1} &= \mu_n + \frac{\sum_i \psi(y_{n,i})\sigma_n}{\psi'(y_{n,i})} \ , \end{aligned} \tag{1.80}$$

em que ψ é dado pela Eq.(1.73) e $a(c)$ é otimizada considerando a eficiência assimptótica de μ e o limite inferior da função de influência [86]. Para nossos resultados empíricos, utilizamos o pacote *statsmodels* [59] do *Python* para realizar o cálculo dessas medidas. Neste pacote, a constante $|c|$ é igual a 1.5.

CAPÍTULO 2

Descrição dos dados da Plataforma Lattes, fator de impacto e indicador SJR

Neste capítulo, apresentaremos as bases de dados utilizadas em nossa investigações. Da plataforma Lattes, obtivemos os currículos de todos os doutores cadastrados na base de dados, contendo seus respectivos históricos e metadados de publicações. Além disso, obtivemos dados referentes ao impacto das revistas científicas a partir da *Web of Science* (fator de impacto) e *SCOPUS* (indicador SJR).

2.1 Fator de impacto das revistas científicas

O conhecimento, que nos primórdios era transmitido primariamente por meio da oralidade, passou, em grande parte, a ser registrado de maneira escrita. Com o passar do tempo, o número crescente de documentos escritos trouxe naturalmente consigo uma necessidade de se criar mecanismos para organizá-los de maneira sistemática. Foi dessa maneira que ocorreu o advento dos índices bibliográficos. Existem vários tipos de índices, por exemplo, existem índices temáticos, em que as obras são divididas e listadas conforme o teor do seu conteúdo; os índices de citações, em que, para cada obra, são listados todos os documentos que a referenciam etc. Os primeiros índices, em sua maioria, diziam respeito a escritos religiosos e possuíam o intuito de impulsionar o progresso da religião. O *Index Librorum Prohibitorum* (índice de livros proibidos pela Inquisição) pode ser considerado como um exemplo do caso religioso. Além disso, é interessante mencionar que, na tradição judaica, a religião tem uma interpretação legal. Dessa maneira, os índices de citações eram úteis no sentido de mostrar o desenvolvimento dessas leis e serviam como meio de legitimar a religião de modo verti-

cal [88]. Um exemplo de outro contexto é o *índice de citações de Shepard*. Ele foi criado em 1973 para organizar os casos da suprema corte de Illinois. Mas, logo, o índice passou a abranger processos de todos os estados norte-americanos. Para cada caso, existia uma lista de documentos – processos que o referenciaram, decisões da corte que o afetaram e outras referências que poderiam ser de valia para o advogado na avaliação do caso [89, 90]. No mundo jurídico, o registro de referências é importante uma vez que as decisões do direito são baseadas em precedentes. Com inspiração no *índice de Shepard*, Eugene Garfield (1925-2017) propôs a criação de um índice de citações para o meio científico em 1955 [89]. As motivações de Garfield para sua proposição incluem [89, 91]:

(i) Acompanhar o desenvolvimento do conhecimento de maneira mais efetiva. Como milhares de artigos já eram publicados anualmente naquela época, se tornou humanamente impossível acompanhar as novidades na ciência;

(ii) Facilitar a comunicação entre pesquisadores que trabalham com a mesma linha de estudo;

(iii) Entender a rede de conexão entre revistas científicas;

(iv) Possibilitar o resgate das origens de uma invenção ou ideia;

(v) Em relação às revistas, manter um registro referente ao número total de documentos, número de cada tipo de documento (artigos, revisões, comentários etc.) e número total de citações;

(vi) Em relação aos documentos, quantificar o número de citações, que pode ser usado como uma medida para mensurar a significância do trabalho em questão. Além disso, podemos considerar que existe uma relação semântica entre o material que cita e aquele que é citado.

Desse modo, o conhecimento poderia ser organizado e teríamos uma visão mais global de seu avanço. Em 1960, com esse intuito, foi criado o *Institute for Scientific Information* (ISI) e, logo em seguida, o índice de citações *Science Citation Index* (SCI) em 1964 [92].

No âmbito de um artigo científico, o número de citações é uma medida relativa de seu impacto e pode ser associado à significância do trabalho de seu autor. Por outro lado, quando vamos analisar uma revista científica, nem sempre o número de citações reflete seu impacto. Por exemplo, considere duas revistas A e B que publicam, respectivamente, cerca de 400 e 200 000 artigos ao ano. Suponha que o número médio de citações de um artigo da revista A é de 100 citações por ano numa janela de 4 anos após os artigos serem publicados. Enquanto isso, um artigo da revista B possui em média 12 citações por ano na mesma janela temporal. Nesse sentido, podemos dizer que, apesar de a revista A produzir em menor quantidade, seu impacto é maior se considerarmos o volume proporcional de citações. Assim, percebemos

que o número bruto de citações de uma revista não reflete efetivamente seu impacto. Em 1972, usando os artigos indexados no SCI de 1969, Garfield propôs uma medida de influência para o contexto de revistas científicas denominada fator de impacto e definida por

$$(\text{Fator de impacto})_{i,y} = \text{IF}_{i,y} = \frac{(\text{citações})_{i,y-1} + (\text{citações})_{i,y-2}}{(\text{publicações})_{i,y-1} + (\text{publicações})_{i,y-2}} \ , \qquad (2.1)$$

ou seja, o $\text{IF}_{i,y}$ da revista i no ano y é a soma do número de citações no ano y referentes a artigos publicados nos dois anos anteriores, $y-1$ e $y-2$, dividido pela soma dos itens citáveis (artigos, notas e revisões) do jornal nesse mesmo período [93].

Inicialmente, essa medida foi criada com o objetivo de auxiliar bibliotecas na escolha das revistas mais apropriadas para atualizar seus acervos, visto que o grande número de revistas impossibilitava uma análise objetiva na ausência de um índice de qualidade. Além disso, o fator de impacto auxiliava pesquisadores na escolha da revista mais adequada para publicação de seus estudos [94]. Atualmente, o fator de impacto é computado por meio dos artigos indexados na *Web of Science* (*WoS*) da *Clarivate Analytics*, uma vez que, em 1997, houve a incorporação do SCI a *Web of Science Core Collection*. O conjunto de dados da *WoS* contém mais de 1.7 bilhões de citações de cerca de 159 milhões de documentos [95]. A divisão responsável pelo cálculo do fator de impacto é a *Journal Citation Reports* (JCR) que faz parte da *WoS*.

2.2 Indicador SJR de revistas científicas

No período em que surgiu o fator de impacto, a indexação de todos os documentos e revistas era feita manualmente, diferentemente de hoje em que as bases de dados são armazenadas virtualmente. Numa tentativa de otimizar o procedimento, Garfield chegou a sugerir o uso de cartões perfurados para facilitar a computação das citações [89]. Ainda assim, o processo era muito trabalhoso e demandava bastante tempo. Parcialmente devido a essa dificuldade, não era possível considerar medidas alternativas mais complexas do ponto de vista matemático, pois acarretaria a adição de mais fatores a um procedimento que já era demasiadamente laborioso. Assim, a escolha e popularização do fator de impacto são em grande parte devido à sua simplicidade (uma razão entre dois números). Atualmente, essa simplicidade tem suscitado muitas críticas, uma vez que o fator de impacto não leva em conta muitas considerações importantes acerca da natureza das distribuições das citações e dos jornais. Mais adiante, comentaremos sobre as principais ressalvas. Por isso, além do fator de impacto, consideramos uma medida alternativa para determinar a influência de um jornal chamada de *SCImago Journal Rank* (SJR), a qual passaremos a descrever.

O indicador SJR foi criado em 2004 juntamente com a base de dados *SCOPUS* da editora Elsevier. Os documentos indexados nessa base, distintos da *WoS*, são utilizados como fonte

para o cálculo dessa medida. O índice SJR é inspirado no algoritmo *PageRank* [96], utilizado para determinar a relevância dos elementos de um conjunto de *websites* presentes na internet. Medidas desse tipo também são denominadas "baseadas em autovetores" [97]. O *PageRank* considera as páginas como uma rede complexa, na qual os nós são os *sites* e as ligações são arestas direcionadas e pesadas representando o número de citações de um *site* a outro. De forma simplificada, o *PageRank* pode ser descrito pela equação

$$\mathrm{PR}(u) = c \sum_{v \in B_u} \frac{\mathrm{PR}(v)}{N_v} \ , \tag{2.2}$$

em que $\mathrm{PR}(u)$ é o *PageRank* do *site* u, $\mathrm{PR}(v)$ é o *PageRank* do *site* v, B_u é o conjunto de todos os *sites* que referenciam u, N_v é o número de citações que partem de v e c é uma constante de normalização. Estritamente falando, o *PageRank* de um *site* é a probabilidade de encontrar o site ao navegar aleatoriamente pela rede de citações. Portanto, a soma de todos os *PageRanks* de um conjunto de páginas deve ser unitária e o maior valor corresponde ao *website* mais importante dentro do conjunto. O método é iterativo e, assim, devemos atribuir um valor inicial para cada nó da rede. O valor inicial que usualmente é escolhido considera que todos os sites têm a mesma importância. Analisando o lado direito da Eq. (2.2), observamos que o *PageRank* da página u depende do *PageRank* de todas as páginas v que o referenciam, isto é, leva em consideração a importância dos *sites* que o citam. Esse número está dividido pelo número de ligações saindo do site v e, assim, indica a importância proporcional de cada referência. Dessa maneira, o valor final do *PageRank* depende de dois fatores: o número de citações recebidas e a importância de cada uma dessas citações.

A Figura 2.1 mostra uma rede de *sites* hipotética contendo quatro *sites*: 1, 2, 3 e 4, que são representados por vértices da rede. As ligações são direcionadas do *site* que referencia para o site que é referenciado. A Tabela 2.1 mostra o cálculo do *PageRank* para cada iteração do método numérico. Na iteração inicial, supomos que todas as páginas têm a mesma importância e atribuímos as probabilidades 1/4 para o *PageRank* de cada página. Conforme o algoritmo vai evoluindo, observamos que o *site* 4 tem um aumento no valor de seu *PageRank*, uma vez que ele é citado por todos os outros *sites* e acaba com o maior valor ($\mathrm{PR}(4) = 0.390$). Os outros *sites* recebem apenas uma citação cada e, se o *PageRank* dependesse exclusivamente do número de citações, eles teriam a mesma importância. Porém, não é isso que acontece, como a página 2 é referenciada pela página 4 (o *site* mais importante da rede), o seu *PageRank* é mais alto ($\mathrm{PR}(2) = 0.344$), pois a referência recebida possui um peso maior do que, por exemplo, uma referencia do *site* 3. O *PageRank* de 2 é comparável ao de 4, porque podemos imaginar que, se o *site* 4 usou o segundo como referência, parte da popularidade do *site* 4 pode ser atribuída ao *site* 2, tornando a importância do *site* 2 similar a do site 4. Levando em consideração essa lógica, podemos verificar que o *ranking* das páginas em ordem decrescente de importância é: 4, 2, 1 e 3.

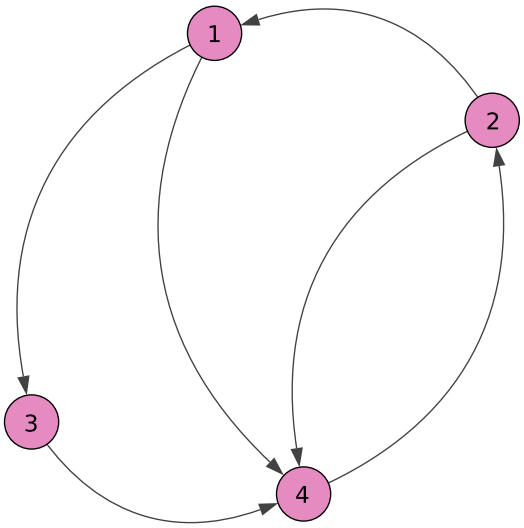

Figura 2.1: Rede de sites hipotética. Os nós representam os *sites* e as arestas direcionadas representam as referências de uma página à outra.

	0	1	2	3	4	5
PR(1)	0.25	0.125	0.125	0.25	0.156	0.188
PR(2)	0.25	0.25	0.5	0.313	0.375	0.344
PR(3)	0.25	0.125	0.063	0.062	0.125	0.078
PR(4)	0.25	0.5	0.312	0.375	0.344	0.390
Soma	1.0	1.0	1.0	1.0	1.0	1.0

Tabela 2.1: **Simulação do algoritmo *PageRank* na rede hipotética de *sites* da Figura 2.1.** As colunas correspondem às diferentes iterações e as linhas representam cada página. Como o magnitude da medida é equivalente à probabilidade de o *site* ser escolhido, a soma de todos os *PageRanks* é igual à unidade, conforme indica a última linha da tabela.

Comparado ao fator de impacto, notamos que o algoritmo *PageRank* é muito mais sofisticado e leva em consideração a estrutura da rede de citações. Podemos considerar que esse procedimento é uma consequência do desenvolvimento tecnológico que facilitou o armazenamento dos dados e a realização de cálculos de estatísticas mais complexas e em larga escala.

Conforme já mencionado, o indicador SJR é baseado no algoritmo *PageRank*, sendo aplicado à rede de jornais científicos ativos num ano y da base *SCOPUS* [98]. As somas das citações entre jornais de artigos dos últimos três anos são arestas pesadas e direcionadas das revistas cujos artigos citam para as revistas cujos artigos recebem as citações. Assim como o *PageRank*, o cálculo do SJR é iterativo. Porém, nesse caso, é realizado em duas etapas. Na primeira, calculamos uma medida de prestígio da revista que é dependente do seu tamanho, o prestígio SJR. Na segunda etapa, normalizamos o prestígio SJR para obter o indicador SJR.

A expressão para calcular o prestígio SJR de um jornal consiste de três partes: (a) o valor mínimo de prestígio que o jornal possui apenas por estar incluso na base de dados; (b) o valor de prestígio associado à fração relativa de artigos que o jornal i contribuiu à totalidade de artigos da base de dados; (c) uma quantidade semelhante ao *PageRank*, para qual estimamos o prestígio de uma revista relativo ao número de citações recebidas, importância e semelhança (em relação à área do conhecimento) dos jornais que o citam. Inicialmente, o prestígio de todos os jornais é definido como sendo $1/N$, sendo N o número total de jornais. Esses valores mudam a cada iteração de acordo com as características da rede. Matematicamente, podemos escrever o prestígio SJR como

$$\text{Prestígio SJR}_{i,y} = \text{PSJR}_{i,y} = \overbrace{\frac{1-d-e}{N}}^{(a)} + \overbrace{e\frac{N_{a,i}}{\sum_{j=1}^{N} N_{a,j}}}^{(b)} + \overbrace{\frac{d}{\text{PSJRD}}\sum_{j=1}^{N} \text{Coef}_{ji}\ \text{PSJR}_j}^{(c)} \ , \quad (2.3)$$

em que $\text{PSJR}_{i,y}$ é o prestígio SJR do jornal i no ano y, $d = 0.9$ e $e = 0.0999$ são duas constantes, N o número total de revistas na base de dados, $N_{a,i}$ o número de artigos citáveis (artigos, revisões, avaliações curtas e artigos de conferência) do jornal i numa janela de três anos anteriores a y. A quantidade PSJRD é o prestígio total distribuído em cada iteração e é definido pela equação

$$\text{PSJRD} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\cos_{ji}\ C_{ji}}{\sum_{h=1}^{N} \cos_{jh}\ C_{jh}} \text{PSJR}_j \ , \quad (2.4)$$

sendo C_{ij} as citações do jornal j para o jornal i na janela de três anos e $\cos_{ij}$ o "cosseno" dos perfis de cocitação dos jornais i e j (não incluindo eles mesmos) definido por

$$\cos_{ij} = \frac{\sum_{h=1,h\neq i,h\neq j}^{N} \text{Cocit}_{ih}\ \text{Cocit}_{hj}}{\sqrt{\sum_{h=1,h\neq i,h\neq j}^{N} (\text{Cocit}_{ih})^2}\sqrt{\sum_{h=1,h\neq i,h\neq j}^{N} (\text{Cocit}_{jh})^2}} \ , \quad (2.5)$$

em que Cocit_{ih} é a cocitação entre os jornais i e h. A cocitação é a frequência com que os documentos da revista i e da revista h são simultaneamente citados por outros artigos. O termo $\cos_{ij}$ é usado para averiguar a proximidade entre duas revistas com respeito a suas cocitações dentro da janela de três anos. Assim, duas revistas fazem parte de uma mesma área do conhecimento se elas possuem cocitações das mesmas revistas. Quanto maior esse número, maior será a similaridade entre elas traduzida matematicamente pela generalização do cosseno.

Os coeficientes Coef_{ji} na Eq. (2.3) são calculados a cada iteração por

$$\text{Coef}_{ji} = \frac{\cos_{ji}\ C_{ji}}{\sum_{h=1}^{N} \cos_{jh}\ C_{jh}} \ , \quad (2.6)$$

introduzindo a ideia de similaridade e a importância do número de citações de j para i no prestígio SJR. Os valores dos coeficientes são limitados a um máximo de 0.5 ou $0.1C_{ji}$ [98].

A segunda parte do cálculo, envolve a normalização do prestígio $\mathrm{PSJR}_{i,y}$, já que ele depende do tamanho dos jornais (em número de artigos e citações). Do jeito que foi construída a ideia de prestígio, jornais terão maiores valores de prestígio pelo fato de terem publicado mais artigos. Dessa maneira, para podermos realizar uma comparação justa entre jornais de tamanhos diferentes, dividimos o prestígio pela razão entre os documentos citáveis do jornal e o total de documentos na base *SCOPUS*, isto é,

$$\text{Indicador SJR}_{i,y} = \mathrm{SJR}_{i,y} = \frac{\mathrm{PSJR}_{i,y}}{N_{a,i}/\sum_{j=1}^{N} N_{a,j}} \ . \tag{2.7}$$

Dessa maneira, o indicador SJR é equivalente ao prestígio proporcional do jornal considerando a fração de documentos que ele inclui na base de dados. Supondo um indicador SJR de um jornal igual à unidade, podemos dizer que esse jornal possui um prestígio por documento igual à média do prestígio dos artigos da base de dados. Enquanto isso, um indicador SJR de magnitude 2 indica que o prestígio por documento é duas vezes a média.

De modo geral, o indicador SJR é uma medida muito mais complexa e completa do que o fator de impacto, pois leva em consideração a estrutura da rede de citações e a similaridade entre os jornais.

2.3 Abrangência das bases de dados

Ambos os indicadores de jornais, fator de impacto e SJR, dependem de uma base de dados para serem computados. O fator de impacto de cada revista é calculado, como já mencionado, a partir dos documentos indexados na *Web of Science*, sendo pioneira nesse sentido. Além do fator de impacto, a JCR também informa o fator de impacto de cinco anos, o fator de impacto pesado, o *eigenfactor* e o *eigenfactor* normalizado, sendo estes dois últimos inspirados no *PageRank*. Porém, em contraste com o fator de impacto tradicional, esses últimos indicadores estão disponíveis para um número reduzido de revistas. Em 2004, a editora Elsevier inaugurou a base de dados *SCOPUS* e, por meio dela, informa anualmente as seguintes estatísticas: o indicador SJR, o índice h e o fator de impacto para janelas de dois e três anos. No mesmo ano, o Google lançou sua própria base de dados, o *Google Scholar*, baseado nos mecanismos de busca do Google, que acessa todas as páginas e documentos presentes na internet que têm uma estrutura parecida com a de documentos acadêmicos, indexando suas informações [99]. O *Google Scholar*, no âmbito da análise do impacto dos jornais, fornece o índice h dos últimos 5 anos de cada jornal. Mais recentemente, em 2015, a Microsoft criou o *Microsoft Academic Graph* (MAG) que também indexa milhares de artigos usando seu mecanismo de busca *Bing* [100]. O MAG emprega sua medida baseada em

autovetor denominada *saliência* e sua versão normalizada [97] para avaliar os jornais. Todas as bases são consideradas *open access*, com exceção da *WoS* [101, 97].

Em relação à abrangência das bases de dados, estudos constatam que, entre as bases *WoS*, *SCOPUS* e *Google Scholar*, o *Google Scholar* é o mais abrangente em relação à quantidade de documentos presentes [99, 102]. Porém, como não existe uma *API* para acessar os documentos, essas investigações são baseadas em amostras específicas, não sendo possível saber ao certo o quão maior é a cobertura do *Google Scholar* como um todo. Há evidências de que o *Google Scholar* engloba mais citações na área de humanidades, enquanto a *WoS* e *SCOPUS* são deficientes nesse sentido [99, 103]. Além disso, esses estudos também comprovam que há uma maior cobertura de documentos em outras línguas no *Google Scholar*, sendo chinês a segunda maior língua, após o inglês [99]. Porém, um fator negativo do *Google Scholar* é a duplicação de documentos e a inclusão de *sites* da internet, teses e dissertações [102]. Dessa maneira, para termos uma estimativa mais confiável da influência dos jornais, escolhemos as bases de dados *WoS* e *SCOPUS* para realizar nossa análise.

Acessamos os *sites* do *SCImago Journal Ranking*[1] e da JCR[2] e realizamos o *download* das informações sobre o impacto dos respectivos jornais de cada base. Para o *SCImago*, a janela temporal abrangeu os anos de 1999 a 2017. Já para o JCR, o período é um pouco maior, de 1997 a 2017. A Figura 2.2 mostra a cobertura das duas bases em número de jornais únicos. A *WoS* possui 13 984 revistas únicas indexadas nesse período, sendo 1298 revistas exclusivas dessa base. A *SCOPUS*, por outro lado, engloba um total de 31 644 revistas únicas, sendo 18 958 revistas exclusivas.

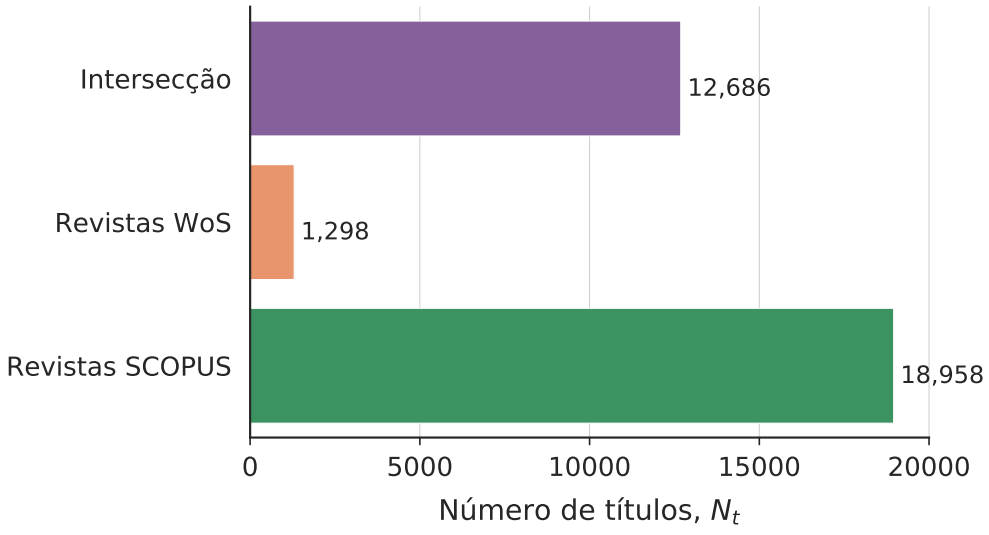

Figura 2.2: Abrangência das base de dados *WoS* e *SCOPUS*. Em roxo, os jornais que estão presentes nas duas base de dados; em laranja, as revistas exclusivas da *WoS*; em verde, as revistas exclusivas da *SCOPUS*.

[1] https://www.scimagojr.com
[2] https://jcr.clarivate.com

2.4 Críticas ao uso do fator de impacto

Há muita discordância acerca do uso do fator de impacto. Parte dessas críticas decorre de sua natureza simplista. A seguir mencionamos algumas das principais críticas [104, 105, 106, 107]:

(a) De maneira ideal, para que o fator de impacto representasse bem o impacto de um jornal em número de citações, esperaríamos que a distribuição das citações dos artigos desse jornal seguisse uma distribuição normal. No entanto, não é isso que ocorre. As distribuições são assimétricas, representando a natureza do sistema, isto é, muitos artigos têm poucas citações e poucos artigos têm muitas citações. Dessa maneira, os artigos com poucas ou sem citações acabam recebendo o crédito por estarem em um jornal de alto impacto;

(b) O denominador da Eq. (2.1) contém apenas itens "citáveis", cujo significado é definido pela própria *Clarivate Analytics* e não é explicitamente divulgado;

(c) As autocitações estão inclusas no numerador da Eq. (2.1);

(d) Artigos de revisão são naturalmente mais citados, já que são usados como referência em uma quantidade maior de trabalhos. No cálculo do fator de impacto, isso não é levado em consideração e suas citações têm o mesmo peso de um artigo comum. Isso faz com que algumas revistas incentivem a confecção de artigos desse gênero com intuito de aumentar seu fator de impacto e em detrimento do avanço da ciência. Além disso, com o mesmo intuito, as revistas tendem a estimular artigos que a citem e também a produção de artigos menores que não entram no cálculo do denominador do fator de impacto, mas cujas citações entram no numerador. Revistas que adotam esse tipo de estratégia são denominadas jogadoras do "jogo do fator de impacto";

(e) Áreas mais dinâmicas e inovadoras, como bioquímica, tendem a possuir uma taxa de obsolescência mais rápida. Sendo assim, os artigos mais citados são aqueles de períodos mais próximos. Por isso, as suas citações são capturadas pela janela bianual do fator de impacto. Por outro lado, áreas mais conservadores, como a matemática, têm uma fração menor das citações referentes a artigos recentes. Consequentemente, o fator de impacto não captura essa dinâmica, pois a janela temporal acaba sendo muito curta;

(f) Existe um viés temporal em relação ao fator de impacto. Como ele depende do número de citações e essas têm aumentado com o passar do tempo, o fator de impacto de um ano não é comparável com de um ano distinto;

(g) Muitas agências de fomento utilizam o fator de impacto das publicações dos pesquisadores analisados para avaliar a pesquisa em si. Essa prática pode ser problemática uma vez que o fator de impacto pode não refletir o impacto do trabalho.

Conhecendo esses problemas distintos associados ao uso do fator de impacto, procuramos ferramentas estatísticas e abordagens alternativas que nos auxiliassem a interpretar essa medida da maneira mais correta. Para o ponto (a), é importante entender o quê a medida consegue ou não expressar [105]. Mesmo que as distribuições sejam assimétricas, existe uma relação implícita com o prestígio da revista. Por exemplo, o alto fator de impacto da revista *Nature*, sem dúvidas, reflete seu prestígio. Estudos muito recentes, que utilizaram *preprints* do ArXiv como medida de impacto desacoplado do impacto da revista em que posteriormente seriam publicados, comprovam que a publicação de artigos em jornais de prestígio é muito correlacionada com um maior número de citações [108, 109]. A causa do maior impacto pode ser atribuída a diversos fatores como processo de *peer-review* mais estrito, seleção de trabalhos com maior potencial, maior visibilidade, maior circulação e maior qualidade dos trabalhos. Se o número de citações de um artigo científico é considerado nossa medida de impacto, então não podemos ignorar as métricas de jornais, uma vez que essas medidas estão correlacionadas ao número de citações. Além disso, um estudo baseado em dados de universidades italianas mostra que o fator de impacto pode ser preferível para publicações mais "jovens" em detrimento do número de citações, principalmente em áreas cujo ciclo de vida das citações é mais longo como a Matemática [110]. Outra investigação que emprega simulações numéricas sugere que o uso do fator de impacto para estimar a qualidade de um artigo não é necessariamente errado [111]. Na verdade, isso depende de quão precisamente o número de citações e o processo de *peer-review*, de fato, refletem a qualidade de um artigo. Aqui, vamos interpretar essa medida como o "potencial impacto" de um trabalho medido pelo prestígio do jornal que o pesquisador escolheu e conseguiu publicar o artigo. No decorrer do texto, o "potencial impacto" será, então, interpretado como "impacto científico".

É importante deixar claro que não incentivamos o uso exclusivo de indicadores de impacto das revistas para avaliação de pesquisadores por instituições de fomento à pesquisa. No entanto, acreditamos que sua utilização é importante desde que seja munida de argumentos e com o auxílio de outras métricas pertinentes. Os indicadores de impacto são alvo de muita controvérsia, existem muitas pessoas e instituições que apoiam seu uso, mas também muitas outras que são contra. No final das contas, tratam-se de números e não podemos imaginar que a qualidade do trabalho seja completamente representada pelo jornal no qual a publicação foi realizada ou pelo número de citações do trabalho. Porém, podemos usar a informação contida na métrica para entender o comportamento dos pesquisadores.

Para contornar os itens (b)-(c), procuramos utilizar uma medida alternativa na nossa análise. O indicador SJR soluciona parcialmente aqueles problemas, pois atribui pesos diferentes para cada citação e inclui todos os documentos citáveis da base de dados. Analisando a Figura 2.3, observamos que para as duas medidas existem mais jornais com impacto baixo e poucos jornais com alto impacto. Além disso, o *scatter plot* entre as $N = 12\,686$ revistas presentes em ambas as bases revela que existe uma correlação linear entre as duas medidas.

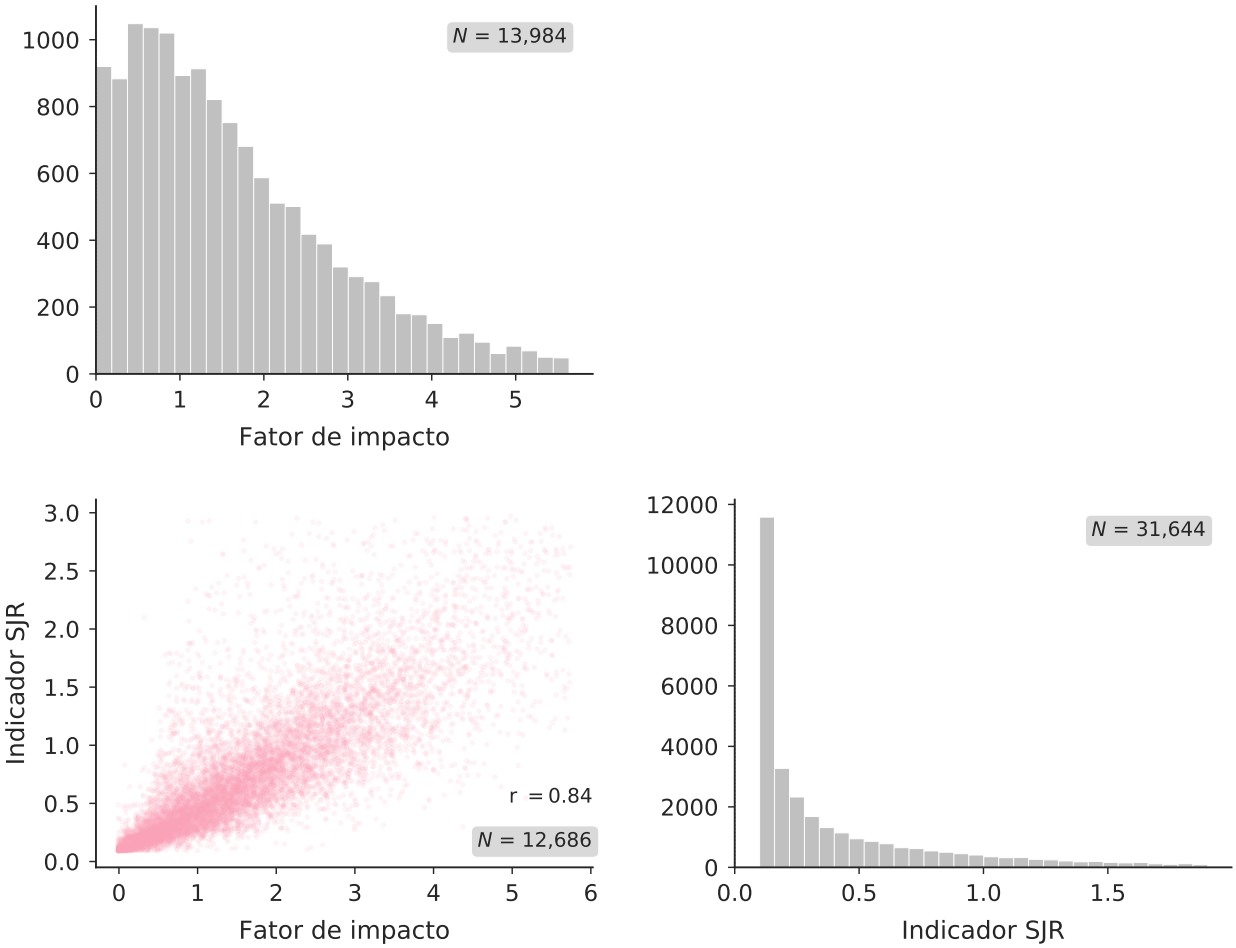

Figura 2.3: Fator de impacto e indicador SJR. Histogramas das distribuições do fator de impacto e indicador SJR (limitadas ao percentil 95) e *scatter plot* entre as $N = 12\,686$ revistas presentes nas duas bases de dados que resultam num coeficiente de Pearson $r = 0.84$.

Essa correlação pode ser quantificada pelo coeficiente de Pearson $r = 0.84$, calculado pela Eq. (B.1) (Apêndice B). Esse fato foi constatado anteriormente em uma pesquisa envolvendo os 20 jornais de maior fator de impacto e indicador SJR no ano 2006, mostrando que 13 deles têm a mesma posição em ambos os *rankings* [101]. As discrepâncias ocorrem devido à diferença nas revistas indexadas em cada base de dados (a *SCOPUS* possui um volume maior) e também devido à natureza das medidas. No entanto, constatamos que existe grande correlação entre as duas medidas.

Como o fator de impacto é uma medida mais difundida e a janela de anos em que ele está disponível é maior por dois anos, escolheremos esse indicador para as análises reportadas no texto principal. As análises correspondentes ao indicador SJR são apresentadas na seção complementar (Apêndice C) e são importantes, pois, como veremos adiante, dão informação para áreas que não estão presentes para o fator de impacto.

Finalmente, para contornar os pontos (e)-(f), utilizaremos o *z-score*, descrito pela Eq. (A.1),

em cada ano e para cada área, a fim de remover a tendência temporal de crescimento dos indicadores e, além disso, para poder comparar áreas distintas.

2.5 Plataforma Lattes

Em 1999, a plataforma Lattes[3] foi criada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) com intuito de estabelecer um modelo de currículo científico virtual que seria utilizado, entre outros fins, para auxiliar na gestão e planejamento estratégico realizado por órgãos de fomento da ciência brasileira [112]. O nome é uma homenagem ao pesquisador Césare Giulio Lattes (1924-2005), proeminente físico brasileiro, que teve contribuições para física de partículas (descobrimento da partícula méson pi) e foi um dos idealizadores do Centro Brasileiro de Pesquisas Físicas (CBPF) no Rio de Janeiro [112]. No meio acadêmico brasileiro, o currículo Lattes se tornou não apenas uma referência, como também obrigatório para qualquer pesquisador interessado em pleitear financiamento das agências de fomento de ciência, uma vez que ele é utilizado na análise de mérito e competência do pesquisador. Assim sendo, apesar de ser preenchido manualmente, podemos considerar que o currículo Lattes representa bem a carreira de pesquisadores, quando frequentemente atualizado. É importante destacar que algumas pesquisas já utilizaram dados dessa plataforma [113, 114]. Aqui, propomos sua utilização a fim de mapear a carreira de pesquisadores de uma variedade de áreas. Em geral, existe uma dificuldade em encontrar bases de dados que contêm informações precisas sobre a carreira dos pesquisadores [115].

Em novembro de 2019, a plataforma continha 6 523 167 currículos, incluindo estudantes, graduados, técnicos, mestres e doutores. Por meio do filtro de busca presente no *site*, focamos nossa análise nos doutores que possuem bolsa produtividade do CNPq. A bolsa produtividade é oferecida para pesquisadores que se destacam em suas respectivas áreas segundo critérios normativos do CNPq (produção científica, formação de recursos humanos, inovação, participação em projetos de pesquisa etc.) [116]. Sendo assim, esses pesquisadores podem ser considerados como a elite de pesquisadores em atividade no Brasil. Eles estão divididos em duas categorias: pesquisador nível 1 e pesquisador nível 2, sendo o primeiro mais elevado que o segundo nessa hierarquia. Além disso, pesquisadores nível 1 são subdivididos em 1A, 1B, 1C, 1D, em ordem decrescente de importância.

Obtivemos de forma automatizada os currículos de 14 487 pesquisadores bolsa produtividade em formato XML, disponibilizados pela plataforma em maio de 2017. Em cada arquivo, vários dados do pesquisador estão presentes: nome, nacionalidade, data da última atualização, grande área de atuação, área de atuação, data da obtenção dos títulos, artigos publicados etc, como mostra a Figura 2.4. Por meio do pacote *ElementTree* [117] da linguagem *Python*, selecionamos as informações relevantes para nossa análise. Em relação

[3] www.lattes.cnpq.br

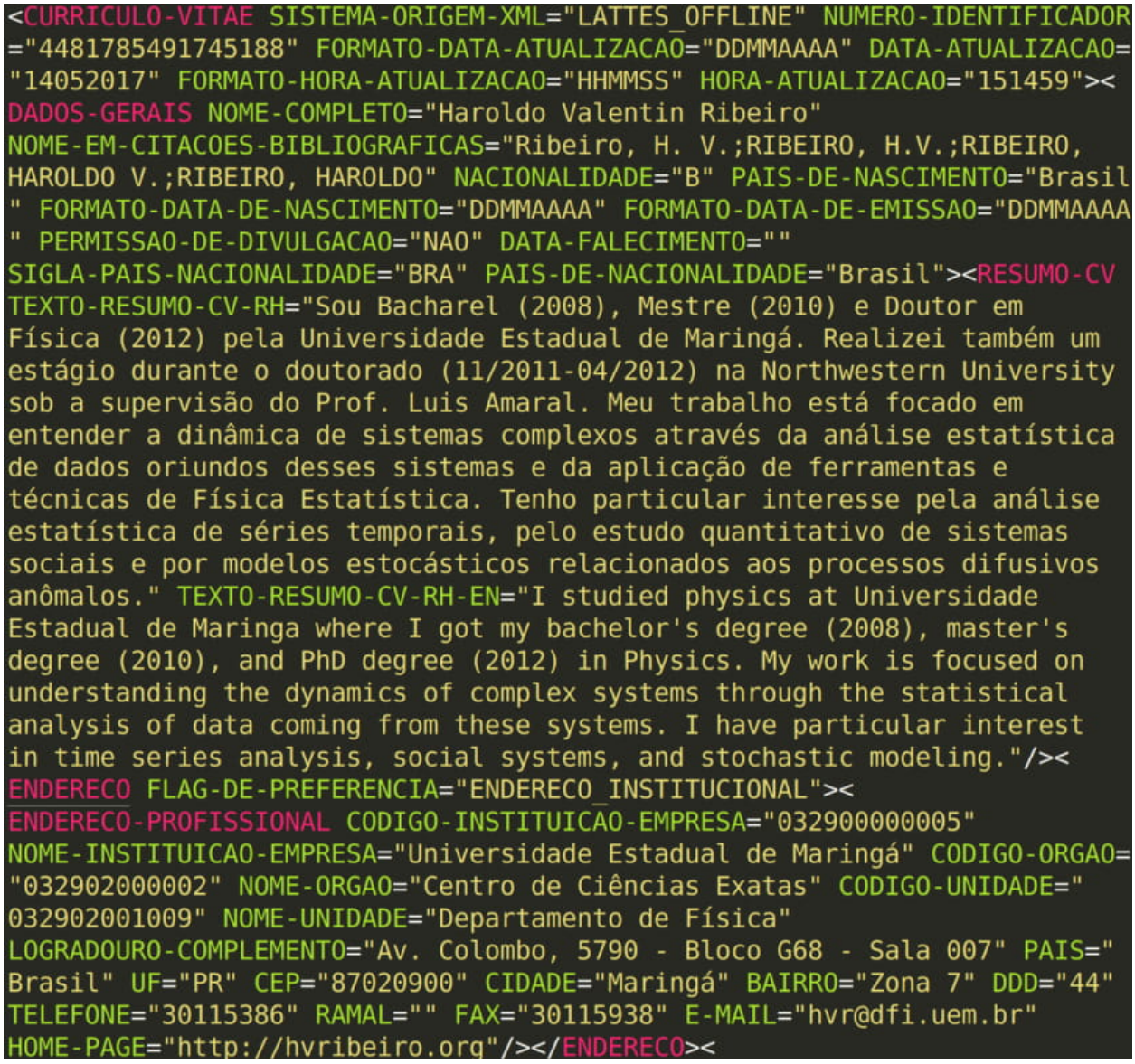

Figura 2.4: Exemplo do arquivo XML obtido do currículo Lattes de um pesquisador. O arquivo está organizado em uma estrutura hierárquica de árvore. Em rosa, podemos visualizar as *tags* descrevendo os elementos "raiz" que são categorias mais abrangentes. Em verde, estão as *tags* representando os elementos "filho" que são as subcategorias que contêm os valores correspondentes. Utilizando o pacote *ElementTree*, conseguimos acessar as *tags* e extrair as informações de interesse.

aos artigos científicos extraídos dos currículos, eliminando materiais de divulgação científica, obtivemos um montante de 1 121 652 documentos no período de 1954 a 2017.

De posse desses dados, aplicamos os seguintes filtros no conjunto de artigos:

- **Pesquisadores que possuem área de atuação no currículo.** Procuramos entender o comportamento de cada área do conhecimento. Por isso, pesquisadores sem essa informação no currículo foram descartados;

- **Pesquisadores que possuem data de obtenção do título de doutorado.** Con-

sideramos a definição de carreira de cada pesquisador como sendo os anos a partir da data de obtenção do título de doutor. Por isso, pesquisadores sem essa informação no currículo foram removidos de nossa análise;

- **Pesquisadores com data de última atualização do currículo igual ou superior a 2016.** Observando a Figura 2.5, percebemos que a fração de currículos atualizados decresce bruscamente com o tempo. Esse comportamento é esperado uma vez que os currículos foram obtidos no começo de 2017. Dessa maneira, para englobar uma quantidade maior de currículos atualizados, estabelecemos o ano de 2016 como limiar;

- **Artigos publicados entre 1997 e 2015.** Como a base de dados JCR abrange apenas o fator de impacto de jornais de 1997 em diante, artigos anteriores a essa data foram descartados. Além disso, escolhemos o ano de 2015 como limite superior, pois é possível que aqueles que atualizaram o currículo em 2016 – data limiar para última atualização do currículo – tenham produzido nesse ano, mas os dados dos artigos não tenham sido adicionados à Plataforma Lattes;

- **Artigos com data posterior à data de obtenção do título de doutorado.** Por fim, consideramos apenas os artigos produzidos dentro da janela temporal que definimos como carreira do pesquisador, isto é, nos anos seguintes à obtenção do título de doutor.

Após a aplicação dos filtros, o número bruto de $N_a = 1\ 121\ 652$ artigos reduz para 862 958, um decréscimo de 23%, como mostra a Figura 2.6a. No entanto, o número de pesquisadores tem um decréscimo relativamente pequeno de 2.3%, diminuindo de $N_r = 14\ 487$ para 14 146 pesquisadores, como mostra a Figura 2.6b, refletindo a obrigatoriedade da atualização do currículo para obtenção da bolsa produtividade do CNPq. Podemos constatar também que a fração de artigos e pesquisadores é maior para a categoria 2, mas isso não necessariamente significa que a produtividade desses pesquisadores seja maior. Examinando a Figura 2.6c, verificamos que a produtividade anual média do grupo 1A é superior aos demais grupos. Pesquisadores dessa categoria publicam em torno de sete artigos por ano, enquanto pesquisadores do nível 2 publicam cerca de quatro artigos anualmente. A Figura 2.6d mostra quantos anos da janela de 1997-2015 estão disponíveis em média para cada grupo. O número é maior para pesquisadores da categoria 1A, indicando que suas carreiras são mais longas. Analisando o ano médio de obtenção do título de doutor, observamos que a categoria 1A, em média, obteve o título em 1987. Por outro lado, pesquisadores nível 2 adquiriram o título em 2001 na média. Essa diferença de estágios da carreira será importante, pois mais adiante analisaremos a dinâmica de publicação em diferentes estágios da carreira.

Para totalidade dos 862 958 artigos, utilizamos seus metadados (ISSN, *DOI* e nome da revista) para pesquisar nas bases de dados da *Web of Science* e *SCImago*. Assim, procuramos

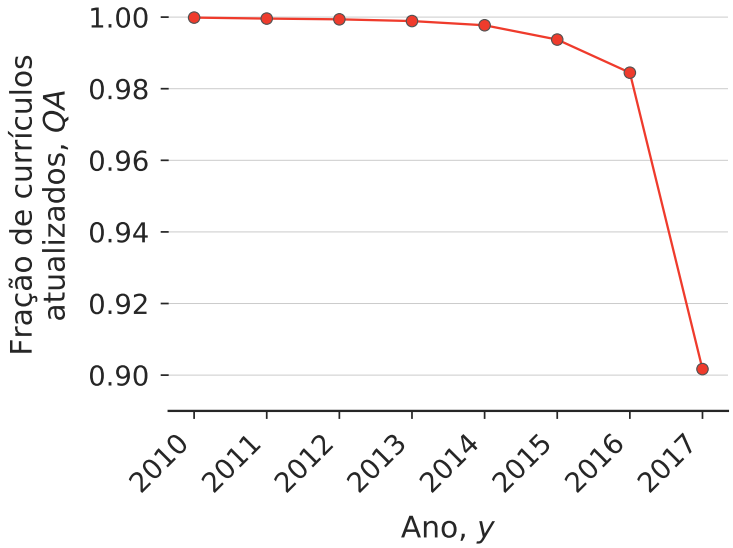

Figura 2.5: Fração de pesquisadores com currículo Lattes atualizado para cada ano entre 2010 a 2017. Observamos que existe uma queda na fração de currículos atualizados do ano de 2016 para 2017, que ocorre devido à data de *download* do dado (maio de 2017).

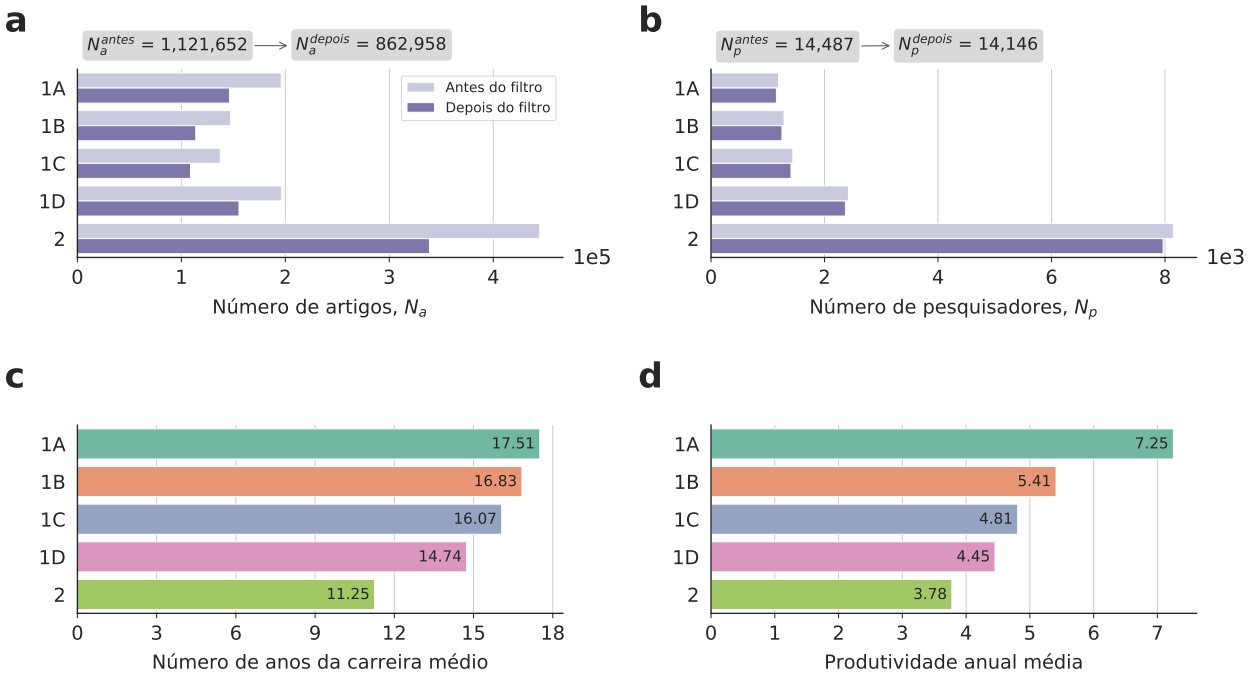

Figura 2.6: Informações para diferentes níveis de pesquisadores bolsa produtividade do CNPq. **(a)** Número de artigos antes e após a aplicação dos filtros para as diferentes categorias. **(b)** Número de pesquisadores antes e após a aplicação dos filtros para as diferentes categorias. **(c)** Produtividade anual média para as diferentes categorias. **(d)** Número médio de anos disponíveis na janela 1997-2015 para as diferentes categorias.

atribuir as medidas de impacto das respectivas revistas a cada um dos jornais. O ISSN gerou correspondência direta com as medidas de impacto, pois é característica exclusiva dos jornais. O *DOI*, sigla para *digital object identifier*, é característico de cada artigo e permitiu a localização do ISSN das revistas. O nome da revista, por sua vez, foi utilizado para realizar a correspondência direta com os nomes descritos nas bases. Para a medida do fator de impacto, obtemos um total de 501 296 artigos e 12 370 pesquisadores, ao passo que para o indicador SJR foram 563 586 publicações e 13 512 pesquisadores. Grande parte dos artigos possuía os metadados, mas não houve correspondência nas bases dos indicadores de impacto. Em outras palavras, as revistas não estavam indexadas naquele ano em específico. Além do mais, uma pequena parcela dos artigos (cerca de 70 mil) não apresentou metadados de ISSN e *DOI*, mas possuíam o nome da revista, com o qual tentamos realizar a correspondência direta com as bases. A Figura 2.7 mostra a quantidade de artigos e pesquisadores com fator de impacto para cada área. As diferentes cores representam as grandes áreas do CNPq. A Figura C.1 mostra as mesmas informações para os artigos com indicador SJR disponível. Notamos que as grande áreas de humanidades (Ciências Humanas, Ciências Sociais Aplicadas e Linguística, Letras e Artes), em geral, têm volume menor de artigos e pesquisadores financiados pelo CNPq se comparados com as demais áreas.

A nossa análise busca investigar aspectos da produtividade e do impacto das revistas em que os artigos foram publicados. Porém, como veremos mais adiante, essas duas medidas sofrem com inflacionamento ao decorrer do tempo. De modo geral, a produtividade cresceu devido ao avanço tecnológico que facilitou a pesquisa científica e viabilizou colaborações. Já o impacto das revistas, de acordo com as medidas que temos, depende do número de citações que os artigos recebem. Como houve um inflacionamento das citações, os indicadores também cresceram com o decorrer do tempo. Dessa maneira, precisamos considerar um mecanismo para deflacionar essas medidas e é com esse intuito que aplicamos um outro filtro nos dados, cujo significado se tornará claro no próximo capítulo. Selecionamos apenas áreas para as quais todos os anos em que temos dados do fator de impacto e indicador SJR tenham ao menos 50 pesquisadores com pelo menos uma publicação. A Figura 2.8 mostra as quantidades de artigos e pesquisadores no caso do fator de impacto para as 14 áreas selecionadas. O total de artigos é de $N_a = 315\ 771$ e de pesquisadores $N_p = 6\ 028$. Para o indicador SJR, como mostra a Figura C.2, a quantidade total de artigos é de $N_a = 444\ 819$ e o número de pesquisadores é de $N_p = 8\ 465$. Além disso, a quantidade de áreas é 25 para o SJR, sendo maior que a quantidade obtida para o fator de impacto. Provavelmente, isso acontece devido à diversidade superior de revistas disponíveis na base *SCOPUS*.

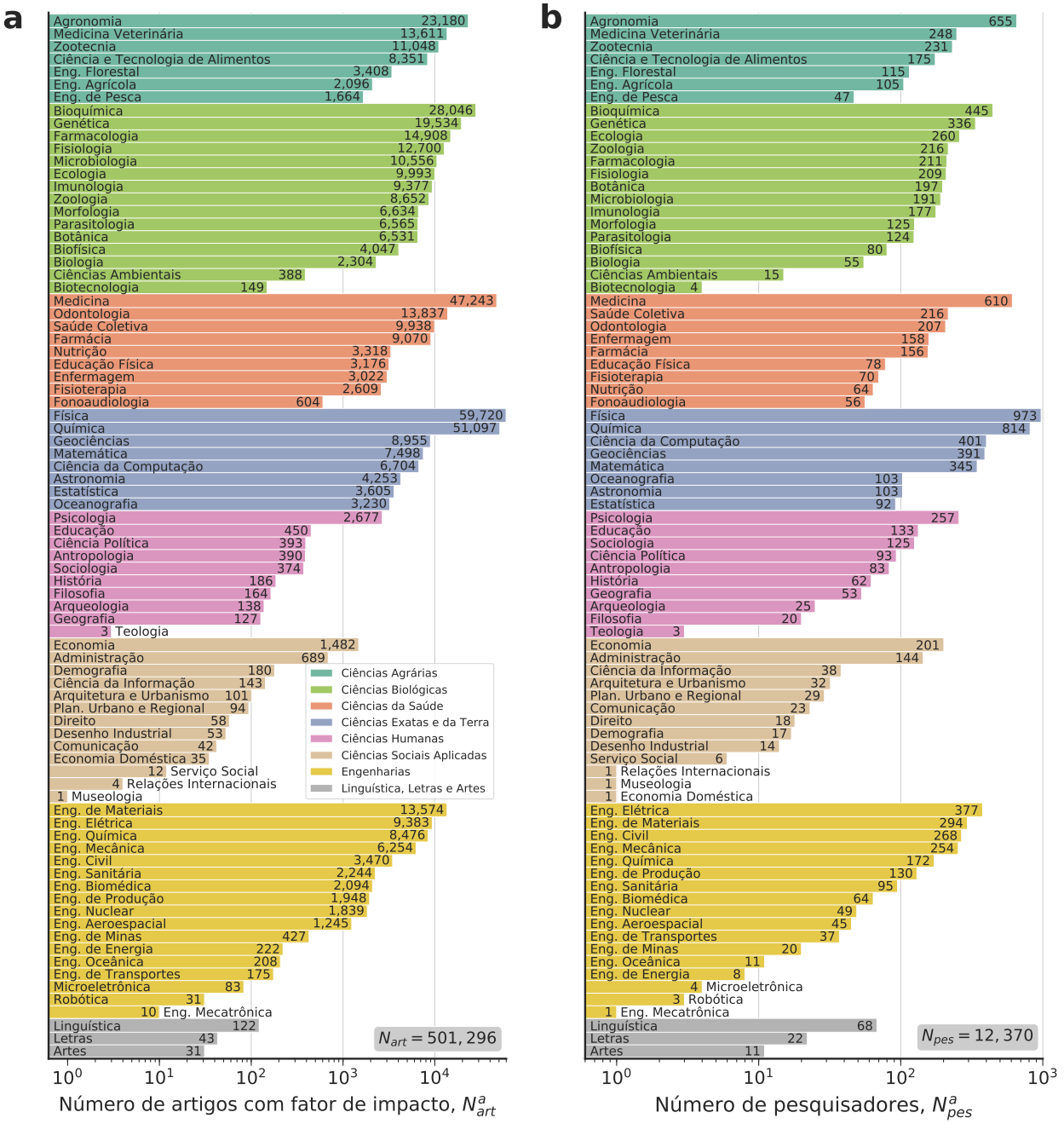

Figura 2.7: Caracterização das áreas em relação aos artigos com fator de impacto. **(a)** Número de artigos com fator de impacto em cada área, sendo o total de artigos $N_a = 501\,296$. **(b)** Número de pesquisadores em cada área, sendo o total de pesquisadores $N_p = 12\,370$. As cores representam as diferentes grandes áreas conforme indicado pela legenda.

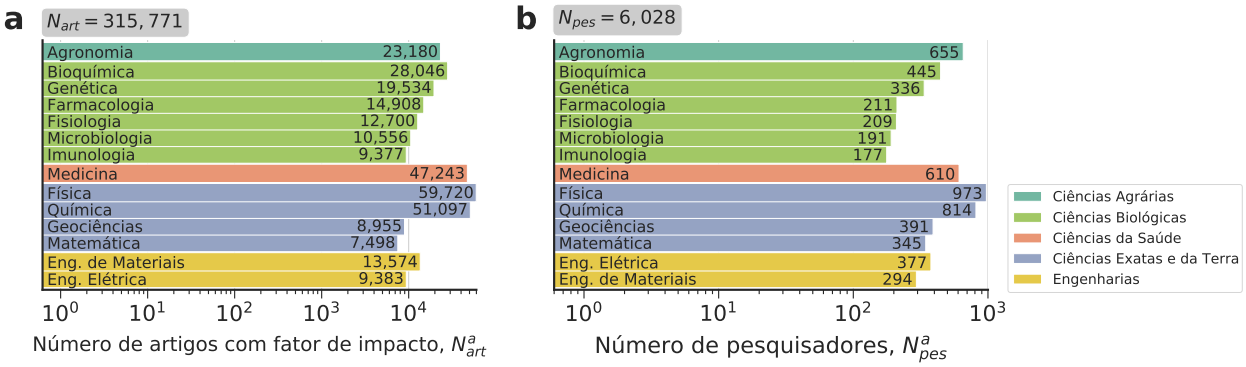

Figura 2.8: Caracterização das 14 áreas selecionadas após o filtro em relação aos artigos com fator de impacto. **(a)** Número de artigos com fator de impacto em cada área, sendo o total de artigos $N_a = 315\ 771$. **(b)** Número de pesquisadores em cada área, sendo o total de pesquisadores $N_p = 6\ 028$. As cores representam as diferentes grandes áreas conforme indicado pela legenda.

CAPÍTULO 3

A relação entre produtividade e impacto

Neste capítulo, investigamos aspectos sobre a produtividade e o impacto científico dos pesquisadores com bolsa produtividade em pesquisa do CNPq. Começamos constatando a inflação temporal da produtividade e dos indicadores de impacto. Em seguida, utilizamos medidas de padronização para agrupar todos os pesquisadores num plano impacto-produtividade. A mesma ideia é aplicada com medidas deflacionadas que geram planos impacto-produtividade específicos para cada disciplina. A medida de padronização utilizada para todas as áreas permite a definição do conceito de *outlier*. Desse modo, foi possível analisar a dinâmica de *outliers* em produtividade e impacto. Também, exploramos a dinâmica dos pesquisadores não-*outliers* e como os anos de suas carreiras estão distribuídos nos setores do plano impacto-produtividade. Essa análise motivou o estudo da dinâmica de carreira de diferentes áreas no plano impacto-produtividade. Além disso, investigamos a influência da produtividade no impacto científico dos pesquisadores por meio de um modelo linear misto. Por fim, avaliamos a variabilidade dos indicadores de impacto de diversas áreas por meio de um modelo exponencial, considerando o desvio padrão do impacto anual médio individual.

3.1 Definição das variáveis agregadas

Como estamos interessados em investigar relações entre a produtividade e o impacto de pesquisadores cadastrados na Plataforma Lattes, é necessário que a medida de impacto seja agregada, por exemplo, dentro de uma janela anual, para que a produtividade correspondente seja definida. Com o objetivo de caracterizar o ano y de cada pesquisador, agregamos os indicadores de impacto por intermédio de sua média. Dessa maneira, o indicador de impacto

médio $I_i(y)$ do pesquisador i no ano y é dado por

$$I_i(y) = \frac{1}{P_i(y)} \sum_{j=1}^{P_i(y)} \tilde{I}_{i,j}(y) \ , \tag{3.1}$$

sendo $P_i(y)$ a produtividade (número de artigos publicados no ano y) e $\tilde{I}_{i,j}(y)$ o indicador de impacto do j-ésimo artigo dentre os $P_i(y)$ artigos.

3.2 Inflação da produtividade e do impacto científico

Há evidências empíricas relacionadas ao aumento do volume da produção científica com o decorrer do tempo [18]. Esse aumento está associado ao desenvolvimento tecnológico que, entre outros efeitos, modernizou a maneira como a ciência é desempenhada. De fato, a realização de colaborações científicas foi facilitada e impulsionou a produtividade de modo geral [17]. Como consequência, indicadores como o fator de impacto e o indicador SJR apresentaram inflação temporal, uma vez que são sensíveis ao crescimento do número de citações decorrente do aumento da produtividade [118]. As Figuras 3.1b e 3.1e mostram, respectivamente, a tendência média de crescimento do fator de impacto (I) e da produtividade (P) para as disciplinas de Genética e Física. Dado que estamos interessados em investigar a relação entre produtividade e impacto científico, a inflação não permite uma comparação direta entre diferentes períodos. Além disso, é razoável supor que essas variáveis tenham comportamentos específicos para diferentes disciplinas, impossibilitando sua comparação. Dessa maneira, é necessário uso de medidas de padronização com objetivo de remover esses dois efeitos. Aqui, propomos a utilização do *z-score* que efetua a padronização por intermédio de medidas de localização e escala. Nesse caso, essas duas estatísticas devem ser específicas de cada ano y e de cada área a com intuito de eliminar ambos os vieses.

No entanto, existe um problema que precisa ser levado em conta. Uma análise dos *boxplots* das Figuras 3.1a e 3.1d revela que há *outliers* para ambas as variáveis[1]. Como podemos observar, a amplitude dos *outliers* é excessivamente superior aos valores das variáveis no intervalo interquartil. Para a produtividade, a amplitude chega a ser quase cem vezes maior no caso da Física. A presença de *outliers* pode prejudicar completamente a estimativa da média uma vez que sua estimativa diverge se apenas um elemento da amostra divergir [84]. Nessas condições, a média não representa o "comportamento médio" da amostra. Podemos visualizar o efeito da presença de *outliers* por meio de picos na estimativa média, porém, o efeito fica mais evidente para o segundo momento. As Figuras 3.1c e 3.1f mostram que o desvio padrão é muito mais sensível à presença de *outliers*. Existem picos acentuados justamente nos anos com grande quantidade de *outliers*. Para contornar essa situação, é

[1] Os *outliers* são ilustrados pelos pontos que extrapolam os fios do *boxplot*.

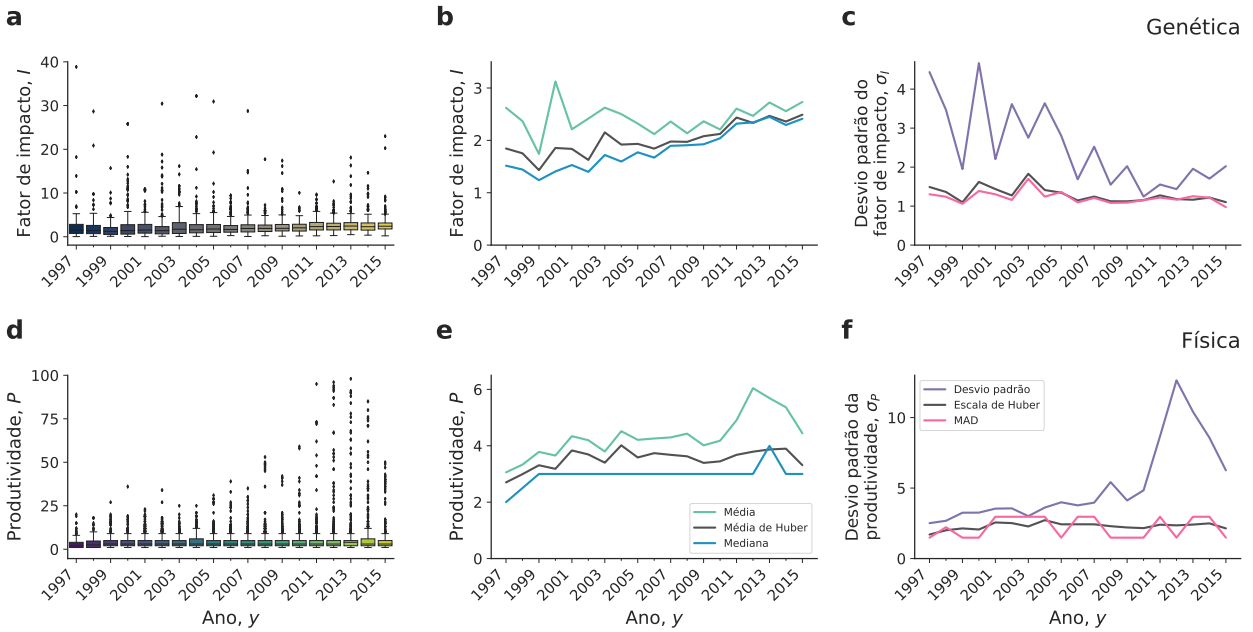

Figura 3.1: Detecção de *outliers* e medidas robustas. **(a)** *Boxplot*, medidas de **(b)** localização e **(c)** escala do fator de impacto para pesquisadores da Genética em função do tempo. **(d)** *Boxplot*, medidas de **(e)** localização e **(f)** escala da produtividade para pesquisadores da Física em função do tempo.

necessário utilizar medidas de localização e escala robustas a *outliers*.

Uma primeira tentativa é o uso da mediana para localização e do desvio absoluto da mediana (MAD) para escala [84, 86]. Como característica específica, essas duas estatísticas necessariamente devem resultar em valores que compõem a amostra. No caso da produtividade, isso se apresenta como uma desvantagem visto que o dado é discreto. A Figura 3.1e indica que existe uma inflação representada pela média. Porém, a mediana não consegue capturar esse comportamento. A solução que propomos é a utilização dos estimadores-M [83]. Mais especificamente, propomos a utilização das medidas de localização e escala de Huber que foram apresentadas na Seção 1.7. Desse modo, as estimativas são robustas e, ao mesmo tempo, possuem interpretação similar a da média e do desvio padrão. Visualmente, o estimador de localização de Huber captura o padrão de crescimento e se apresenta entre a média e a mediana, como mostram as Figuras 3.1b e 3.1e. Por outro lado, o estimador de escala de Huber se assemelha ao MAD, todavia não exibe um aspecto discreto, como mostram as Figuras 3.1c e 3.1f.

Com auxílio das medidas de localização de Huber, podemos visualizar o padrão de inflação referente ao fator de impacto (Figura 3.2a) e à produtividade (Figura 3.2b). Nessas figuras, as linhas acinzentadas indicam o comportamento individual de cada uma das disciplinas. Em azul, temos o comportamento da disciplina de Física. Como mencionado anteriormente, as áreas apresentam taxas de crescimento específicas representadas por diferentes "inclinações". Por sua vez, a linha escura ilustra a média global de crescimento no período analisado. Esses

resultados mostram que existe uma tendência de crescimento do fator de impacto com o tempo. O indicador SJR apresenta um comportamento similar, apesar de existirem algumas áreas com tendência decrescente (Figura C.3a). O padrão de crescimento da produtividade se mantém para todas as áreas quando utilizamos o indicador SJR (Figura C.3b).

Uma possibilidade para estimar a taxa de crescimento dos indicadores de impacto e produtividade para cada uma das áreas é usar modelos de regressão linear simples definidos como

$$\begin{aligned} I_a(y) &= \alpha_0^a + \alpha_1^a y \\ P_a(y) &= \beta_0^a + \beta_1^a y \end{aligned}, \tag{3.2}$$

sendo $I_a(y)$ o impacto médio da área a no ano y calculado pela medida de localização de Huber, α_0^a o intercepto do modelo para o impacto e α_1^a a taxa de crescimento do impacto para área a. De modo similar, $P_a(y)$ representa a produtividade média da área a no ano y calculada pela medida de localização de Huber, β_0^a é o intercepto do modelo para a produtividade e β_1^a é a taxa de crescimento da produtividade para área a.

A Figura 3.1c mostra os valores das taxas de crescimento do fator de impacto (α_1^a) multiplicados por 10 anos, isto é, as taxas de crescimento por década para cada área. Dentre todas as áreas, o maior crescimento é da Química com cerca de ≈ 0.90 unidades de impacto por década. Em contrapartida, os menores ritmos de crescimento são da Agronomia (≈ 0.29 unidades por década) e Matemática (≈ 0.33 unidades por década). Observamos que todas as taxas de crescimento são positivas. Os resultados para o indicador SJR são mostrados no gráfico de barra da Figura C.3c. Nesse caso, a maior taxa de crescimento é da Medicina ($\approx$ 0.38 unidades por década); por outro lado, a menor taxa é da Engenharia dos Materiais (≈ -0.08 unidades por década).

A produtividade média apresentou um crescimento para todas as áreas conforme ilustram as Figuras 3.2d e C.3d. Referente ao conjunto de dados do fator de impacto, a Medicina apresentou maior ritmo de crescimento com uma inflação de ≈ 3.30 artigos por década. Para o indicador SJR, a Odontologia foi o destaque com ≈ 4.81 artigos por década. Em ambos os casos, a diferença na inflação da produtividade entre o primeiro e segundo colocados é grande: cerca de um artigo a mais por década. Quando analisamos as menores taxas de crescimento no conjunto de dados do fator de impacto, observamos que a produtividade de pesquisadores da Engenharia Elétrica cresceu apenas ≈ 0.30 artigos por década. Enquanto isso, para os dados do SJR, a menor taxa de crescimento foi para a Física com ≈ 0.26 artigos por década.

Em termos de produtividade, é interessante notar que as disciplinas das áreas de saúde e as biológicas apresentaram as maiores inflações tanto para o conjunto de dados do fator de impacto quanto para os dados do indicador SJR. Enquanto isso, as disciplinas de exatas e ciências naturais apresentaram as menores taxas de crescimento. Esse comportamento se

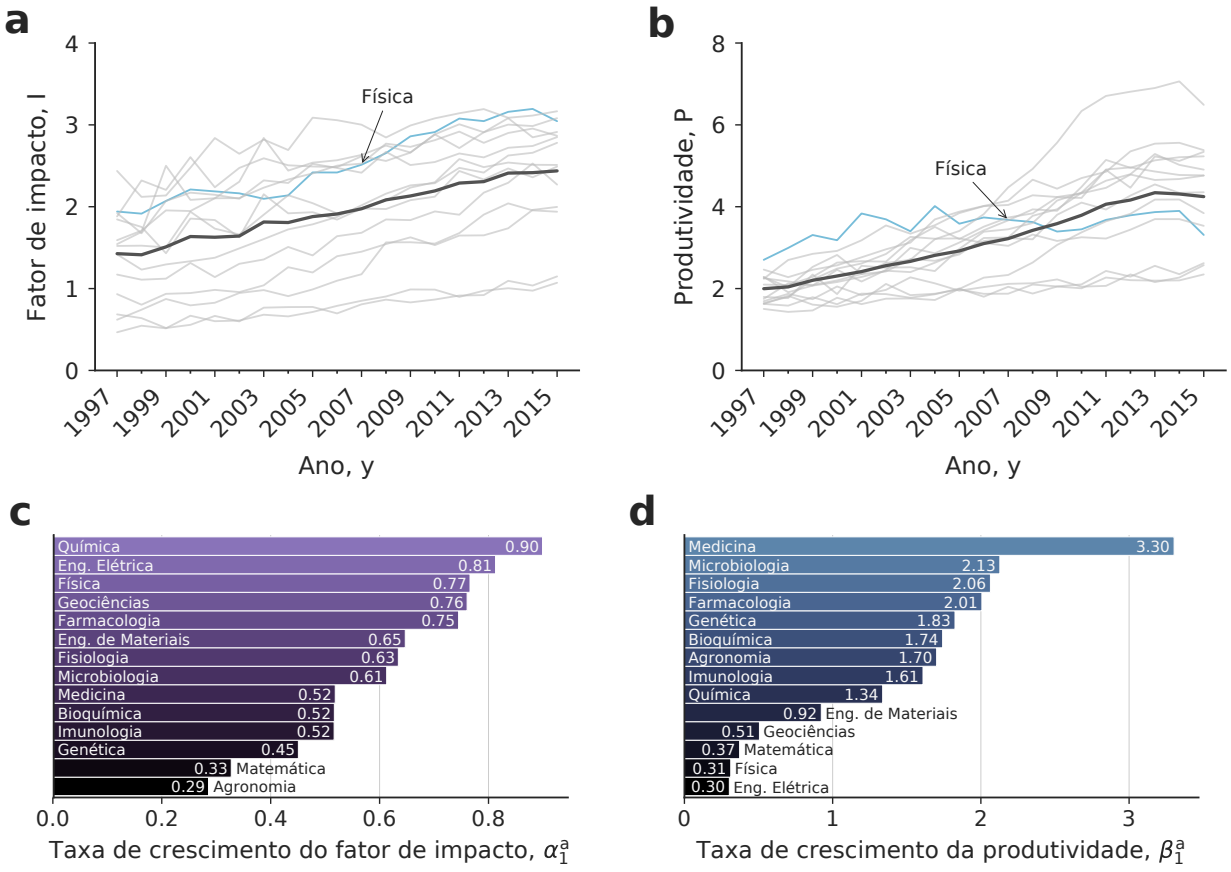

Figura 3.2: Inflação do fator de impacto e da produtividade entre as diferentes áreas do conhecimento. **(a)** Padrão de crescimento do fator de impacto com o decorrer do tempo. A linha escura representa o comportamento global de todas as disciplinas. As linhas claras representam o comportamento individual de cada área. A linha azul representa o comportamento da Física. **(b)** Padrão de crescimento da produtividade com o decorrer do tempo. **(c)** Taxas de crescimento por década do fator de impacto para diferentes áreas. **(d)** Taxas de crescimento por década da produtividade por década referente ao conjunto de dados do fator de impacto para diferentes áreas.

repete para o padrão de inflação do indicador SJR. O fator de impacto, por outro lado, tem como destaque as áreas de ciências naturais e engenharias.

Em um primeiro momento, com propósito de comparar e agregar diferentes áreas na mesma análise, propomos a utilização de um *z-score* robusto, como definido no Apêndice A. A padronização de uma variável $x_i(y)$, que corresponde ao impacto ou à produtividade de um pesquisador i no ano y, é realizada por meio de

$$(\text{z-score robusto})_i^{a,y} = \frac{x_i(y) - \mu_{a,y}}{\sigma_{a,y}} \ , \tag{3.3}$$

em que o índice a corresponde à área, o índice y corresponde ao ano, e $\mu_{a,y}$ e $\sigma_{a,y}$ são, respectivamente, a média e desvio padrão de Huber (para o indicador de impacto ou produtividade) da área a no ano y.

Em um segundo momento, com intuito de quantificar o efeito da produtividade no impacto, propomos a utilização de uma medida deflacionada cuja interpretação é a mesma das variáveis originais. Considerando a variável genérica $x_i(y)$, a medida deflacionada em relação ao ano de 2015 é dada por

$$x_i^{(\text{def})}(y) = x_i(y)\frac{\mu_{a,2015}}{\mu_{a,y}} \ , \tag{3.4}$$

em que $\mu_{a,y}$ é a medida de localização de Huber da área a no ano y. Dessa maneira, a medida deflacionada representa a variável $x_i(y)$ em termos do impacto ou produtividade do ano de 2015 considerando um crescimento linear.

3.3 Plano impacto-produtividade de todas as áreas

A partir da padronização das variáveis de produtividade e impacto por meio da Eq. (3.3), podemos construir um plano impacto-produtividade agrupando os anos das carreiras de pesquisadores de todas as disciplinas, conforme ilustra a Figura 3.3 para o fator de impacto e a Figura C.4 para o indicador SJR. Nesse plano, cada ponto representa a performance do pesquisador em determinado ano em relação aos quesitos produtividade e impacto científico. De modo específico, o valor nulo indica um desempenho médio dentro de sua área naquele ano. Por outro lado, um valor positivo indica que o pesquisador obteve um desempenho acima da média em unidades de desvio padrão. A mesma interpretação é válida para valores negativos que, dessa vez, correspondem a um desempenho baixo se comparado com pesquisadores da mesma área.

Além disso, é possível definir um limite de *z-score* em que os pesquisadores podem ser considerados *outliers*. Com esse objetivo, é muito comum a adoção do valor 3.5 na literatura [119]. Esse valor de *z-score* corresponde à interpretação de que os *outliers* são os 0.05% valores mais extremos de uma distribuição normal. Denominamos os *outliers* produtividade como hiperprolíficos. Enquanto isso, podemos dizer que os *outliers* impacto produzem predominantemente em revistas de altíssimo impacto.

Para entender melhor a dinâmica no plano impacto-produtividade, podemos analisar a distribuição dos pontos em cada uma das seções. Considerando $ZI_i(y)$ o *z-score* do impacto do i-ésimo pesquisador no ano y e $ZP_i(y)$ o *z-score* da sua respectiva produtividade, definimos os setores e suas respectivas representações da seguinte maneira:

- Impacto abaixo da média (i$-$): $ZI_i(y) < 0$;
- Impacto acima da média (i+): $0 \leq ZI_i(y) < 3.5$;
- *Outlier* em impacto (i++): $ZI_i(y) \geq 3.5$;
- Produtividade abaixo da média (p$-$): $ZP_i(y) < 0$;

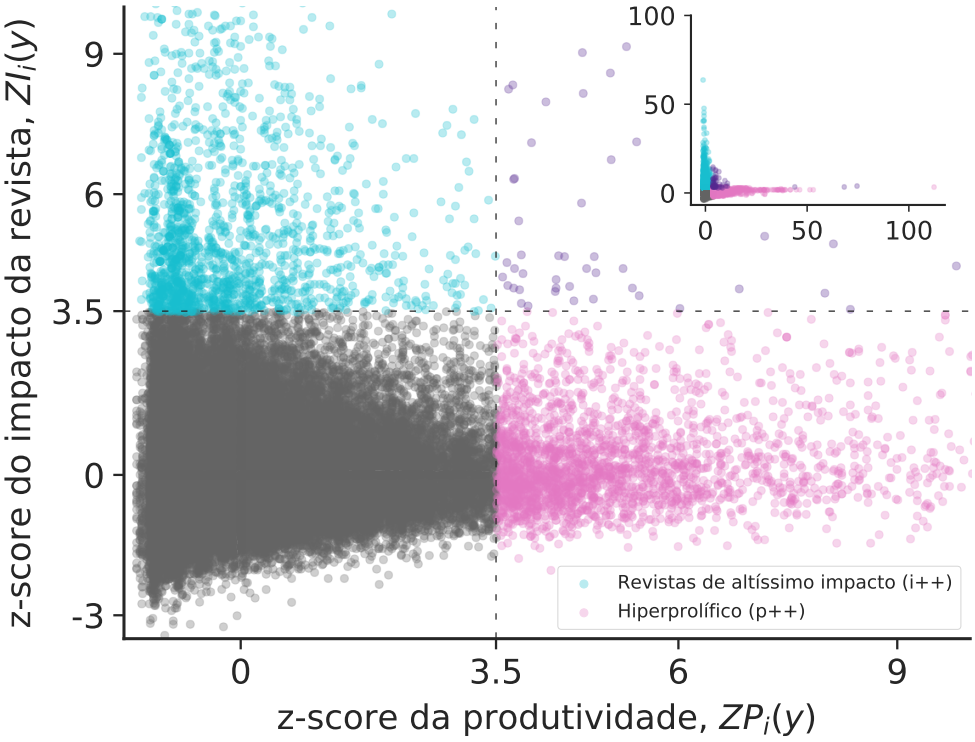

Figura 3.3: Plano impacto-produtividade agregando todas as áreas do conhecimento. A partir da interpretação do *z-score*, dividimos o plano impacto-produtividade em setores. Valores acima de 3.5 indicam que o pesquisador foi *outlier* naquele quesito. Os pontos rosas indicam anos de pesquisadores hiperprolíficos. Enquanto isso, os pontos azuis indicam que o pesquisador publicou apenas em revistas de grande fator de impacto para sua área naquele ano. Valores positivos (negativos) indicam que o pesquisador esteve acima (abaixo) da média naquele ano, em sua área e em determinado quesito. O gráfico mostra valores de *z-score* até 10. O *inset* mostra o plano inteiro.

- Produtividade acima da média (p+): $0 \leq ZP_i(y) < 3.5$;
- *Outlier* em produtividade ou hiperprolífico (p++): $ZP_i(y) \geq 3.5$.

A Figura 3.4 ilustra essas regiões do plano impacto-produtividade. Não nos preocupamos com *outliers* negativos, pois naturalmente existe uma limitação inferior para produtividade (um artigo por ano) e para o impacto (zero).

O gráfico de barras apresentado na Figura 3.5 mostra a distribuição dos pontos no plano impacto-produtividade em relação a cada um dos setores para o banco de dados do fator de impacto. A Figura C.5 mostra as mesmas informações para o indicador SJR[2]. Uma vantagem do uso da medida de localização de Huber (em comparação com a mediana) é que a estatística não necessariamente está presente na amostra. Portanto, a escolha da seção que contém o zero não afeta de maneira efetiva os resultados da análise. Para os dois indicadores

[2]Para que o texto não fique repetitivo, referiremos à base de dados do fator de impacto e indicador SJR como "para o fator de impacto" e "para o indicador SJR" em alguns momentos.

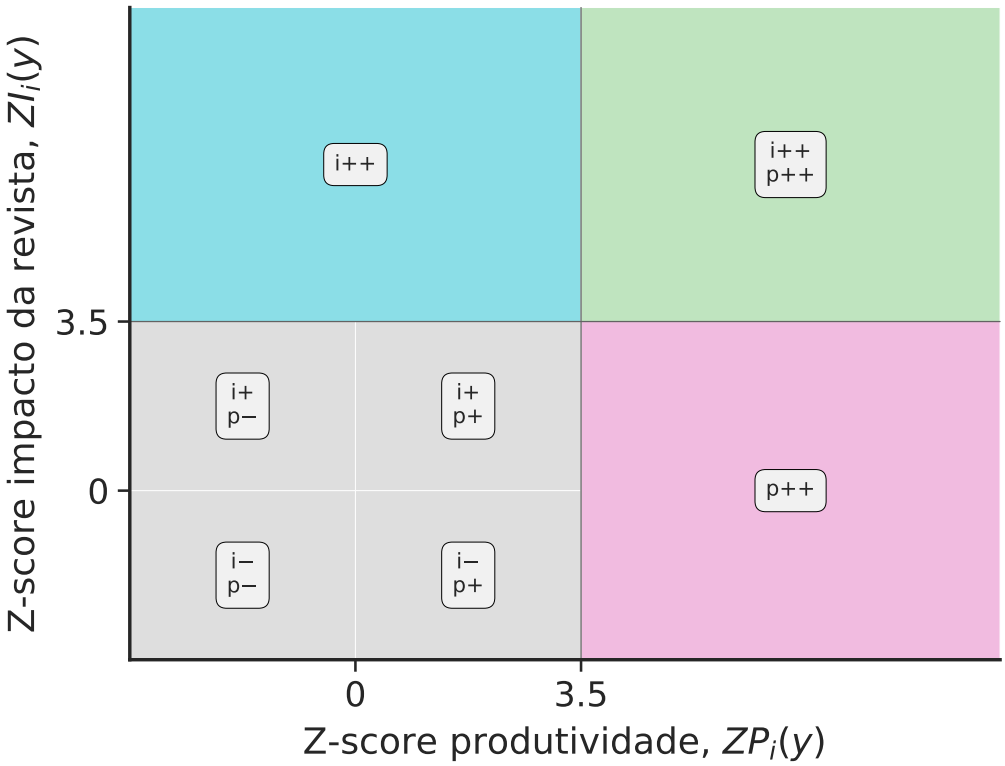

Figura 3.4: Regiões do plano impacto-produtividade.

de impacto, notamos que os setores não-*outliers* apresentam muito mais pontos. Além disso, existe uma tendência de que os setores mais povoados sejam aqueles em que os pesquisadores estão abaixo da média (−) para produtividade e impacto (i−, p−). De fato, essa discrepância está bem representada pela diferença entre o setor (i−, p−) que tem 24 762 pontos e o setor (i+, p+) que possui apenas 13 429 pontos para o fator de impacto. Para o indicador SJR, essa diferença também está presente, sendo que a região (i−, p−) possui 33 031 pontos e a região (i+, p+) possui 17 859 pontos.

Os setores *outliers* são naturalmente pouco povoados por conta de sua definição. Visualmente, notamos que existe uma nítida separação entre os dois tipos de *outlier* no plano impacto-produtividade. Essa separação evidencia que existem poucos pesquisadores que são *outliers* em ambos os quesitos em algum ano de suas carreiras. De fato, existem apenas 58 pontos na região (i++, p++) para o fator de impacto. Similarmente, 56 pontos estão presentes nesse setor para o indicador SJR. De modo qualitativo, isso significa que é muito difícil produzir em grande quantidade publicando predominantemente em revistas de alto impacto. Na seção 3.5, vamos investigar como esses padrões emergem para diferentes áreas do conhecimento.

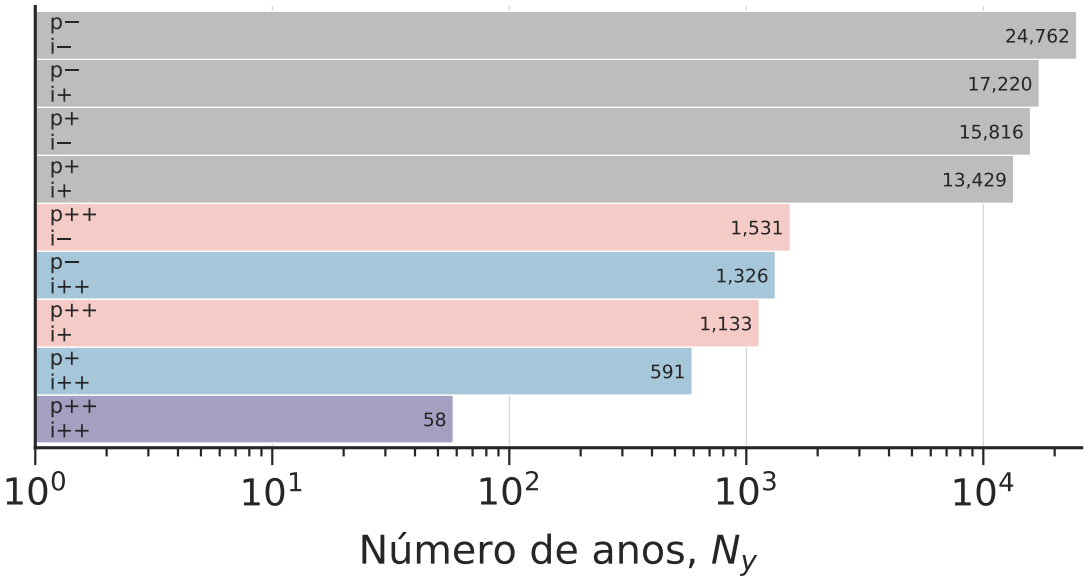

Figura 3.5: Quantidade de anos da carreira dos pesquisadores em cada setor do plano impacto-produtividade.

3.4 Plano impacto-produtividade para cada área

O plano impacto-produtividade pode ser construído em termos da produtividade e impacto deflacionados por meio da Eq. (3.4). Porém, não é possível agregar todas as disciplinas num plano único, pois o limiar que define um *outlier* é distinto entre as diferentes áreas. A Figura 3.6 mostra o plano das áreas de Física, Matemática, Medicina e Química para o fator de impacto. A Figura C.6 mostra os mesmos planos em termos do indicador SJR. O valor médio foi calculado pela estatística de localização de Huber aplicada em todas as medidas deflacionadas de determinada área. Os *outliers* foram definidos usando como referência os valores de produtividade e impacto para os quais o *z-score* é igual a 3.5 e mapeados em unidades deflacionadas referentes ao ano de 2015. Percebemos que os valores de produtividade limiar de *outlier* são sistematicamente maiores para o indicador SJR. Isso acontece devido à maior abrangência da base *SCOPUS*. O que acaba por permitir a contabilidade de mais artigos a nível de pesquisador, fazendo com que esse limiar seja maior.

A dinâmica de produção científica entre as áreas é diferente de tal maneira que, se compararmos o limiar *outlier* de Medicina com de Matemática, percebemos que o valor do primeiro é três vezes o do segundo. De outra maneira, um *outlier* da Matemática produziria como um pesquisador "comum" da Medicina. As Figuras 3.7a e 3.7b mostram, respectivamente, os limiares *outliers* para o fator de impacto (O^a_{imp}) e produtividade (O^a_{prod}) para cada disciplina a. Em outras palavras, estão representados os valores acima dos quais o pesquisador pode ser considerado um *outlier* em impacto ou produtividade. As Figuras C.7a e C.7b apresentam a mesma informação para o banco de dados do indicador SJR. Além disso, as Figuras 3.7c e

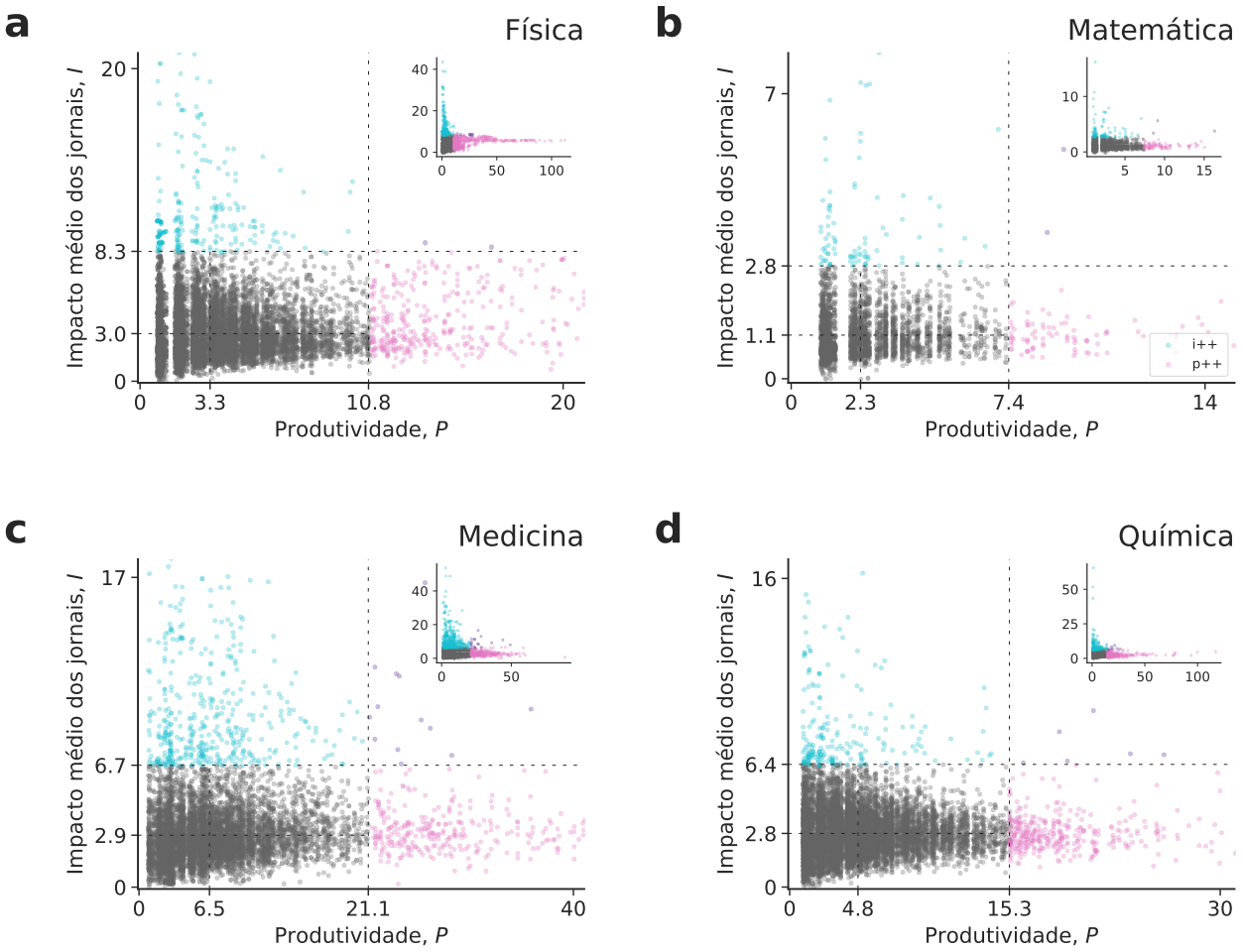

Figura 3.6: Plano impacto-produtividade deflacionado para diferentes áreas do conhecimento. Os painéis mostram o plano impacto-produtividade deflacionado para as disciplinas de **(a)** Física, **(b)** Matemática, **(c)** Medicina e **(d)** Química. Os valores médios de cada área foram calculados por meio da medida de localização de Huber e os limiares *outliers* são os valores deflacionados correspondentes ao valor do *z-score* igual a 3.5.

3.7d informam o comportamento médio do fator de impacto (μ^a_{imp}) e produtividade (μ^a_{prod}) de cada área a. Enquanto isso, as Figuras C.7c e C.7d o fazem para o indicador SJR. É possível observar que disciplinas com baixos valores de produtividade média possuem baixos valores médios dos indicadores de impacto. Isso é uma consequência da dependência da produtividade da área (quando é maior, acarreta num número maior de citações) na definição dos indicadores.

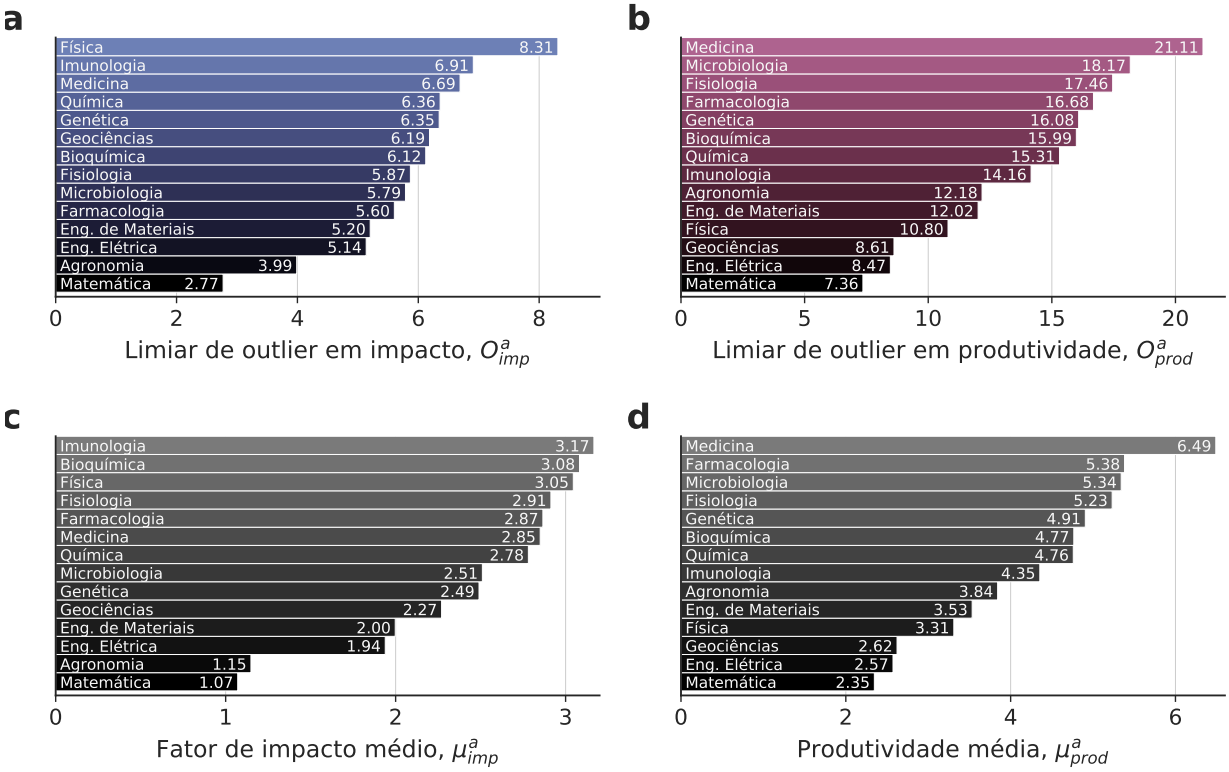

Figura 3.7: Caracterização das disciplinas de acordo com as medidas deflacionadas. Como as áreas apresentam comportamentos diferentes, os limiares de *outlier* para impacto e produtividade diferem. Os painéis mostram os limiares de *outlier* para **(a)** o fator de impacto e **(b)** a produtividade, além dos valores médios de Huber para **(c)** o fator de impacto e **(d)** a produtividade para cada área a.

3.5 Análise dos pesquisadores *outliers*

A investigação do plano impacto-produtividade mostrou evidências de que existe uma segmentação entre dois tipos de pesquisadores *outliers* num âmbito global. Os pesquisadores hiperprolíficos não conseguem produzir predominantemente com altíssimo impacto e pesquisadores que publicam predominantemente em revistas de altíssimo impacto não conseguem ser hiperprolíficos. No entanto, o plano impacto-produtividade não nos fornece informações referentes ao comportamento dos pesquisadores do ponto de vista individual. Com esse intuito, construímos um diagrama de Venn com todos os pesquisadores *outliers*, como mostra a Figura 3.8a para o fator de impacto e a Figura C.8a para o indicador SJR. Na construção desse diagrama, não utilizamos informações sobre o setor (i++, p++), uma vez que praticamente não existem pontos nessa região. Considerando apenas as seções (i++) e (p++), existem um total de 1774 pesquisadores com ao menos um ano *outlier* para o fator de impacto e 2576 pesquisadores para o indicador SJR. Desses pesquisadores, apenas 9.3% estão presentes em ambas as categorias como *outlier* para o fator de impacto e 6.7% para o indicador SJR. Dessa maneira, há indícios fortes de que o comportamento persiste também

do ponto de vista individual.

As Figuras 3.8b e C.8b mostram distribuições de probabilidade do *z-score* da produtividade agora considerando os pontos da região (i++, p++). Quando analisamos os *outliers* de impacto (i++) que extrapolam o limiar para *outlier* de produtividade, observamos que a produtividade respectiva em termos do *z-score* não é muito elevada. Isso indica que, mesmo quando o pesquisador *outlier* impacto consegue ser hiperprolífico, é pouco provável que sua produtividade seja altíssima, não ultrapassando valores de 10 desvios padrões. O inverso também é verdade: quando o pesquisador é hiperprolífico, é pouco provável que seu impacto seja muito elevado.

Também podemos caracterizar alguns padrões microscópios do comportamento individual dos pesquisadores. Dado um conjunto de probabilidades $\{P_i\}$ (com $i = 1, 2, \ldots, n$) que indicam a fração de anos dos pesquisadores em cada uma das seções *outliers*, podemos calcular a entropia normalizada de Shannon H [120] associada a essas probabilidades via

$$H = -\frac{1}{\log n} \sum_{i=1}^{n} P_i \log P_i \ . \tag{3.5}$$

O valor da entropia normalizada está contido no intervalo $[0, 1]$. O extremo inferior ($H = 0$) indica que as probabilidades são totalmente desiguais, isto é, existe um certo $P_i = 1$, enquanto todos os outros $P_j = 0$ para todo $i \neq j$. O extremo superior ($H = 1$) aponta que as probabilidades são uniformemente distribuídas entre as categorias, ou seja, $P_i = 1/n$ sendo $n = 2$ o número total de categorias. Em nossa análise, a entropia normalizada de Shannon para cada pesquisador representa a concentração ($H \approx 0$) ou não ($H \approx 1$) dos anos da carreira de um pesquisador em uma categoria *outlier*. Dessa maneira, conseguimos inferir se existe um comportamento temporal persistente para pesquisadores que foram *outliers* em algum momento da carreira. A Figura 3.8c mostra a distribuição de probabilidade da entropia normalizada de Shannon para pesquisadores *outliers* da base de dados do fator de impacto. A Figura C.8c mostra uma distribuição similar para o indicador SJR. Observamos que existe uma maior densidade de probabilidade ao redor de valores pequenos de H, indicando que a maioria dos pesquisadores tende a ser *outlier* em apenas um quesito ao longo de suas carreiras. Portanto, seu comportamento é persistente com o tempo. Ainda, realizamos um teste de permutação para verificar se a existência de pesquisadores que possuem anos *outliers* nos dois quesitos acontece ao acaso. Para isso, atribuímos os anos dos pesquisadores *outliers* aleatoriamente para cada um dos setores. Neste contexto, o p-valor corresponde à probabilidade de obter um número de pesquisadores presentes em ambos os setores *outliers* menor ou igual ao que realmente ocorre. Para 1000 repetições, o p-valor é virtualmente nulo, o que indica que a dicotomia é significativa e não pode ser explicada pelo acaso. De outra maneira, a presença de pesquisadores com anos em ambos os setores no decorrer da carreira ocorre ao acaso.

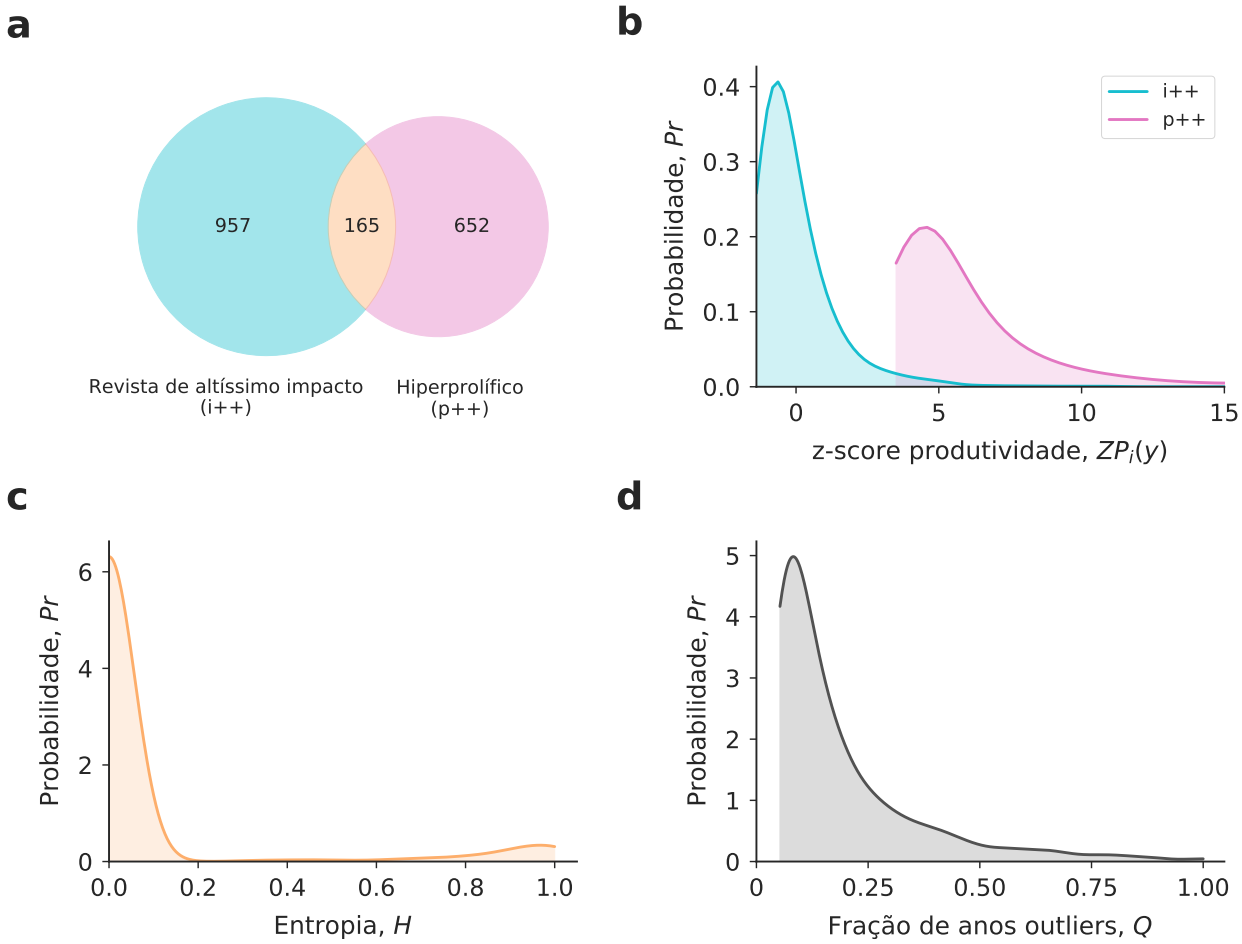

Figura 3.8: Análise dos setores *outliers*. **(a)** Diagrama de Venn indicando quantos pesquisadores estão presentes em cada setor. A intersecção do diagrama significa que, no decorrer de sua carreira, o pesquisador esteve presente no setor *outlier* produtividade e no setor *outlier* impacto ao menos um ano. A maioria dos pesquisadores *outliers* está presente exclusivamente em um setor. **(b)** Distribuição de probabilidade do *z-score* produtividade para as duas categorias *outliers*. O gráfico mostra que a intersecção das distribuições de probabilidade é muito pequena. Além disso, notamos que a maioria dos anos *outliers* impacto apresenta produtividade baixa. Os *outliers* de ambos os quesitos não possuem produtividade com valor tão acentuado, ultrapassando o limiar com apenas algumas unidades de desvio padrão. **(c)** Distribuição de probabilidade da entropia normalizada dos pesquisadores que, no decorrer da carreira, foram *outliers* em ambos os quesitos. Como a maioria da densidade se concentra em valores pequenos de entropia, os pesquisadores apresentam um comportamento persistente ao longo das carreiras. **(d)** Distribuição de probabilidade da fração de anos *outliers* por pesquisador. Os anos *outliers* geralmente representam uma pequena parcela da carreira dos pesquisadores.

Até agora, descobrimos que existem evidências acerca da dicotomia entre *outliers* produtividade e impacto. Mesmo os pesquisadores que conseguem se destacar nos dois quesitos não o fazem de maneira exacerbada. Além disso, por meio da entropia de Shannon, verificamos que o comportamento do pesquisador *outlier* é persistente durante sua carreira. A seguir,

vamos investigar aspectos mais quantitativos como a fração de anos *outliers* (produtividade ou impacto) na carreira dos pesquisadores. A Figura 3.8d mostra a distribuição de probabilidade da fração de anos *outliers* por pesquisador para o fator de impacto. A Figura C.8d mostra a mesma quantidade para o indicador SJR. Existem picos para valores pequenos da fração, revelando que o comportamento padrão é de que anos *outliers* são raros no decorrer da carreira. Apesar disso, existem casos em que o pesquisador consegue consistentemente manter uma carreira "fora da curva", mas esse é um comportamento bastante raro.

Na continuação da investigação de aspectos quantitativos dos *outliers*, realizamos uma regressão logística (conforme descrito na Seção 1.2) para avaliar a influência da quantidade de anos *outliers* produtividade na probabilidade de o pesquisador ser *outlier* impacto. Para isso, consideramos uma variável binária y_i que assume o valor $y_i = 1$ se o i-ésimo pesquisador foi *outlier* impacto no decorrer da carreira e o valor $y_i = 0$ caso contrário. Além disso, utilizamos a variável discreta $N_{p,i}$ que indica a quantidade de anos *outliers* produtividade do i-ésimo pesquisador. O modelo logístico é dado por

$$\log\left\{\frac{P(\text{outlier impacto})}{P(\text{não-outlier impacto})}\right\} = \beta_0 + \beta_1 N_p \tag{3.6}$$

e os parâmetros do modelo β_0 e β_1 foram estimados a partir da maximização da verossimilhança.

Estritamente falando, a constante β_0 é o valor do logaritmo da chance quando $N_p \approx 0$. Dessa maneira, a exponencial da constante e^{β_0} é igual à chance de o pesquisador ser *outlier* impacto quando não conseguiu produzir como *outlier* produtividade. Porém, como estamos considerando a parcela de pesquisadores que, de fato, são *outliers*, naturalmente a fração correspondente a $N_p \approx 0$ é igual à unidade uma vez que todos os pesquisadores devem ser *outliers* impacto. Portanto, a estimativa do coeficiente β_0 é influenciada negativamente pelas diferentes frações apenas quando $N_p > 0$. Diferentemente da interpretação usual da constante β_0, a interpretação que adotamos é de que quanto maior o valor de β_0, mais fortes são as evidências de que existe uma chance maior de ser *outlier* impacto quando $N_p \approx 0$. O coeficiente logístico β_1, por sua vez, indica a rapidez com que a curva logística decresce e satura. Se $\beta_1 \ll -1$, a existência de anos como *outlier* produtividade tem o caráter de impossibilitar a existência de anos como *outlier* impacto.

Para o conjunto total de dados, os parâmetros estimados são $\beta_0 = 1.35$ e $\beta_1 = -0.61$ para o fator de impacto e $\beta_0 = 1.65$ e $\beta_1 = -1.20$ para o indicador SJR. As Figuras 3.9a e C.9a ilustram a curva logística para os respectivos parâmetros estimados. Para base de dados do fator de impacto, o modelo aponta que existe probabilidade de os pesquisadores produzirem com alto impacto até $N_p \approx 5$ anos como *outlier* produtividade. A chance para $N_p \approx 0$ é de 385:100. A situação é ainda mais pronunciada para o indicador SJR. Considerando valores de $N_p \approx 1$ ano, o modelo logístico indica que a probabilidade de ser *outlier* de impacto tende

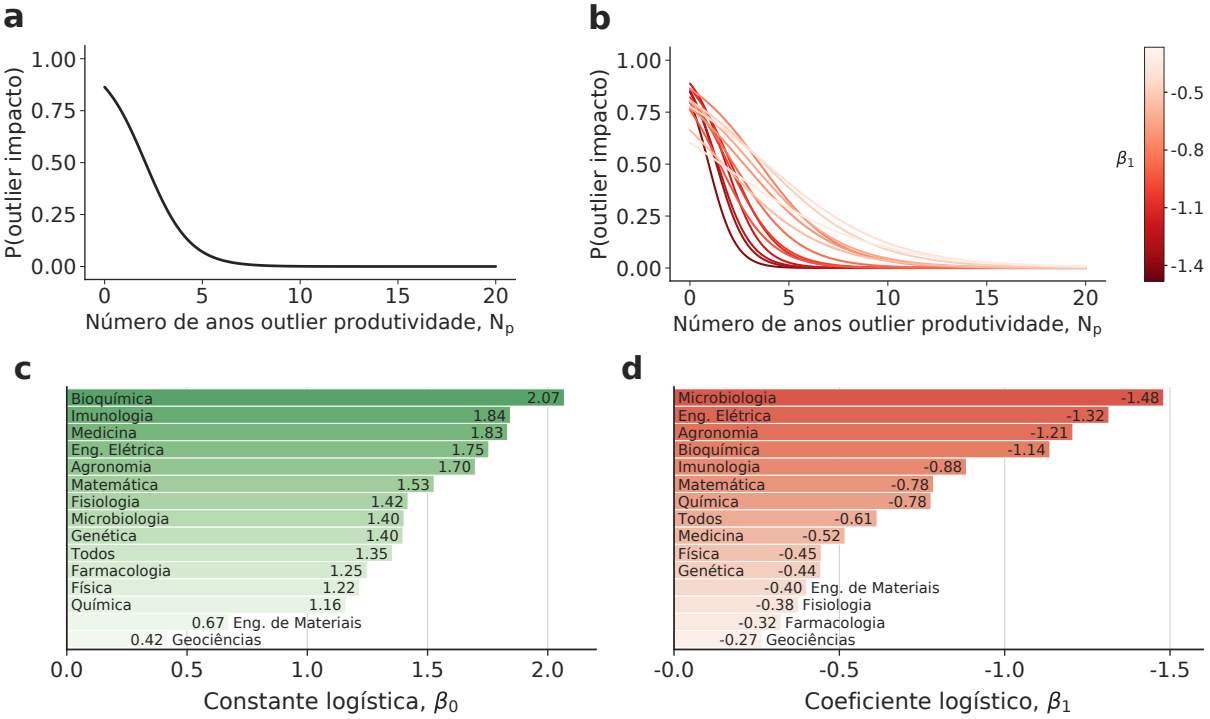

Figura 3.9: Regressão logística para análise de *outliers*. **(a)** Regressão logística global agregando todas as áreas do conhecimento. **(b)** Regressão logística por área do conhecimento. A escala de cor vermelha representa diferentes valores de β_1. **(c)** Valores da constante β_0 para as diferentes áreas. **(d)** Valores do coeficiente β_1 para as diferentes áreas.

a zero. Ou seja, existe uma separação completa do ponto de vista global: ser hiperprolífico impede a produção em altíssimo impacto. A chance de ser *outlier* impacto para $N_p \approx 0$ é de 520:100 para a base de dados do indicador SJR.

Em seguida, empregamos o modelo logístico para estimar os parâmetros β_0 e β_1 para cada uma das áreas. As Figuras 3.9b e C.9b mostram as curvas ajustadas para cada área. Em primeiro lugar, notamos que o valor da chance para $N_p \approx 0$ é diferente para cada área. Por exemplo, a chance quando $N_p \approx 0$ da Bioquímica é de aproximadamente 792:100 para os dados referentes ao fator de impacto e 25:1 para o indicador SJR. No outro extremo, o valor dessa chance para Geociências é menor e aproximadamente igual a 15:10 para o fator de impacto e 18:10 para o indicador SJR. Dessa maneira, existem evidências de que a chance de ser *outlier* impacto é maior quando $N_p \approx 0$ para algumas áreas em comparação com as demais disciplinas. Os gráficos de barra das Figuras 3.9c e C.9c mostram os valores de todas as constantes logísticas β_0.

O coeficiente β_1 também é diferente entre as áreas. As Figuras 3.9b e C.9b mostram que existem diferentes taxas de saturação das curvas logísticas representadas pela diferentes escalas de vermelho. Algumas áreas possuem um efeito mais brando de saturação conforme o aumento do número de anos como *outlier* produtividade. Esse é o caso de pesquisadores

da disciplina Geociências com $\beta_1 = -0.41$ para o fator de impacto e da Parasitologia com $\beta_1 = -0.30$ para o indicador SJR. Por outro lado, existem áreas que saturam rapidamente com o aumento do número de anos como *outlier* produtividade. A Microbiologia apresenta um valor de $\beta_1 = -1.48$ para o fator de impacto e a Bioquímica é o destaque com $\beta_1 = -3.44$ para o indicador SJR. Assim, notamos que existem áreas em que é possível ser *outlier* impacto (em algum momento da carreira) mesmo produzindo como *outlier* produtividade. Da mesma maneira, existem áreas em que a produção como *outlier* produtividade praticamente impede a performance como *outlier* impacto. As Figuras 3.9d e C.9d mostram os valores estimados de todos os coeficientes β_1.

3.6 Análise dos pesquisadores não-*outliers*

Nesta seção, investigamos o comportamento dos pesquisadores que não estão presentes nas regiões com *z-score* maior do que 3.5. Para isso, selecionamos apenas os pesquisadores que estiveram em regiões não-*outliers* durante toda sua carreira (presente em nossa base de dados). Nesse sentido, podemos considerá-los como pesquisadores com comportamento mais próximo do comportamento "médio" da área. Além disso, definimos um limiar de pelo menos cinco anos para carreira a fim de considerar as frações de anos em cada seção. Para o fator de impacto, existe um número total de 4275 pesquisadores não-*outliers*. Para o indicador SJR, esse número é de 5924 pesquisadores.

Em termos da fração em nível global, existe uma tendência de maior ocupação da seção (i−, p−), como mostram as Figuras 3.5 e C.5. Podemos também averiguar se o comportamento individual em relação à fração de anos em cada seção segue a mesma tendência. As Figuras 3.10a e C.10a mostram o comportamento individual dos pesquisadores em termos das distribuições de probabilidade das frações em cada seção. O *inset* mostra as frações médias para cada seção não-*outlier*. A seção com distribuição mais uniforme é a seção (i−, p−), isto é, existem pesquisadores com frações de anos elevadas e frações baixas nessa seção, apesar de existir uma leve assimetria que favorece as pequenas frações. O pico dessa distribuição acontece em ≈ 0.40 que corresponde ao valor médio da variável para o fator de impacto. Para o indicador SJR, o valor médio é de ≈ 0.39. Para as outras seções, as distribuições apresentam sempre uma assimetria em direção à esquerda. Em especial, a distribuição mais assimétrica é a do setor (i+, p+). Do ponto de vista qualitativo, os pesquisadores apresentam frações majoritariamente baixas referentes a esses setores, o que é consequência da maior ocupação do setor (i−, p−).

Um pesquisador não-*outlier* pode transitar entre as quatro seções de impacto e produtividade durante sua carreira. De modo similar ao que fizemos na Seção 3.5 para os pesquisadores *outliers*, podemos usar a entropia normalizada de Shannon, Eq. (3.5), para estimar se existe uma tendência na permanência em determinada seção ($H \approx 0$) ou se as quatro seções são

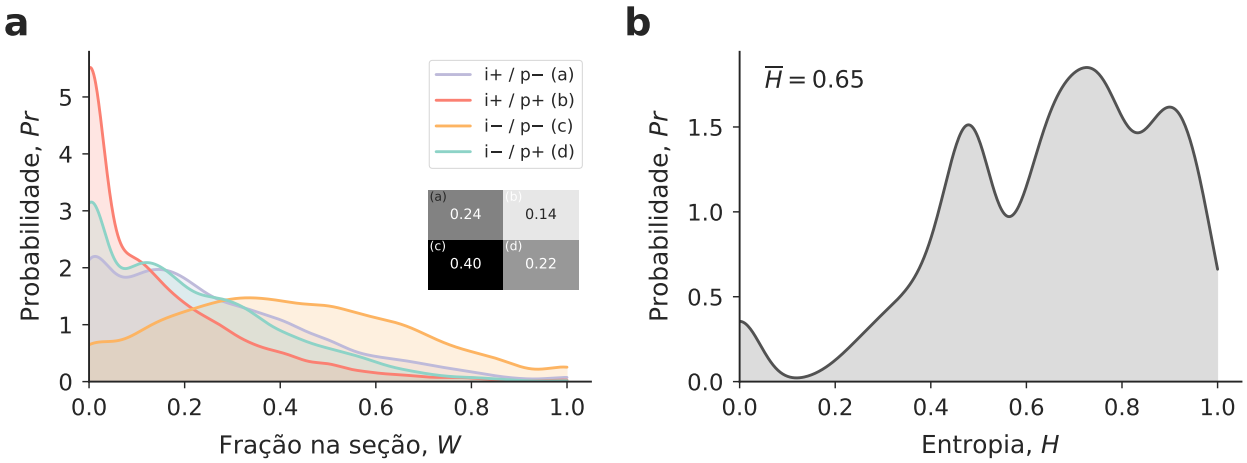

Figura 3.10: Análise dos setores não-*outliers*. **(a)** Distribuição de probabilidade dos setores não-*outliers*. A letra i na legenda representa o impacto dos jornais em um ano e o símbolo p a produtividade. O símbolo $+$ indica que o pesquisador publicou acima da média naquele quesito. Por outro lado, o símbolo $-$ indica que o pesquisador publicou abaixo da média em determinada categoria. O *inset* mostra a fração dos anos em cada seção correspondente. **(b)** Distribuição de probabilidade da entropia normalizada da distribuição dos anos dos pesquisadores nas seções não-*outliers*. Os valores elevados indicam que os anos da carreira não estão concentrados em apenas uma seção, isto é, pesquisadores tendem a transitar entre as seções no decorrer de suas carreiras.

igualmente visitadas no decorrer da carreira ($H \approx 1$). As Figuras 3.10b e C.10b mostram as distribuições de probabilidade da entropia H. Os valores mais elevados (> 0.50) dominam a distribuição, indicando que o comportamento mais comum é o de transitar entre as seções. Apesar disso, existem pesquisadores que permanecem em apenas algumas seções durante toda sua carreira, mas esses não representam o comportamento mais usual nas carreiras científicas. A entropia média é de 0.65 para o fator de impacto e de 0.66 para o indicador SJR.

3.7 Análise da carreira

Na seção anterior, verificamos por meio da entropia normalizada de Shannon que pesquisadores da Plataforma Lattes transitam entre diferentes setores não-*outliers* no plano impacto-produtividade. Podemos estender essa ideia e realizar uma análise mais detalhada em relação aos padrões temporais das frações em cada setor. Dessa maneira, podemos compreender melhor a dinâmica da carreira dos pesquisadores no plano impacto-produtividade. Como o intervalo temporal é limitado pela disponibilidade da bases de dados, não conseguimos extrair informações sobre carreiras completas dos pesquisadores de maneira individualizada. No entanto, a existência de pesquisadores em diferentes estágios da carreira advém naturalmente da estratificação do sistema de bolsas do CNPq. Isso possibilita uma análise

de carreiras completas sob a perspectiva das disciplinas. Para isso, consideramos o ano de obtenção do título de doutor como primeiro ano da carreira. Em seguida, dividimos os períodos subsequentes em janelas quadrienais. Se o pesquisador não produziu em algum dos anos, descartamos tal período. As frações médias de cada setor j são calculadas por

$$(\text{fração do setor } j)_{a,r} = \sum_{i=1}^{(n_{pes})_{a,r}} \frac{(\text{número de anos em } j)_i}{4(n_{pes})_{a,r}} \ , \tag{3.7}$$

em que o índice i se refere ao i-ésimo pesquisador, o índice a se refere à área, o índice r se refere ao intervalo de quatro anos[3] e $(n_{pes})_{a,r}$ é o número de pesquisadores da área a no intervalo r.

Além disso, calculamos o coeficiente de Gini (G) para as frações das seções não-*outliers*. Em seguida, estimamos o valor médio para todos os pesquisadores. Para isso, consideramos apenas os períodos de quatro anos em que os pesquisadores estiveram exclusivamente em setores não-*outliers*. O coeficiente de Gini é uma medida de dispersão estatística que, originalmente, foi criada para quantificar a desigualdade da distribuição de renda de um país [121] e é definida como [122]

$$G = \frac{\sum_{j=1}^{n}(2j - n - 1)x_j}{n\sum_{j=1}^{n} x_j} \ , \tag{3.8}$$

sendo x_j o valor da fração do j-ésimo setor não-*outlier* e $n = 4$ uma vez que existem quatro setores. Em nosso contexto, o coeficiente G mede a desigualdade entre as frações dos setores. Se o valor do coeficiente é $G = 0$, as frações estão uniformemente distribuídas, ou seja, a fração em cada setor é $1/4$. Por outro lado, valores elevados do coeficiente apontam que uma das frações é dominante e a distribuição não é uniforme. Dessa maneira, conseguimos também estimar como a desigualdade das frações não-*outliers* evolui com o decorrer da carreira.

As Figuras 3.11 e 3.12 mostram mapas de calor para todas as disciplinas. As colunas contêm informações sobre as frações médias em cada setor do plano impacto-produtividade e sobre o coeficiente de Gini médio para cada período da carreira. As linhas representam os períodos quadrienais da carreira. O número de pesquisadores utilizado no cálculo das frações está especificado entre parênteses ao final de cada linha. Os limites das escalas de cor são específicas para os três grupos: (i) setores não-*outliers*, (ii) setores *outliers* e (iii) coeficiente de Gini, para que as variações sejam mais perceptíveis. Os valores das frações representam o comportamento global dos pesquisadores de cada disciplina.

[3]Por exemplo, se $r = 1$, então consideramos o período de 1 a 4 anos; se $r = 2$, consideramos o período de 5 a 8 anos e assim por diante.

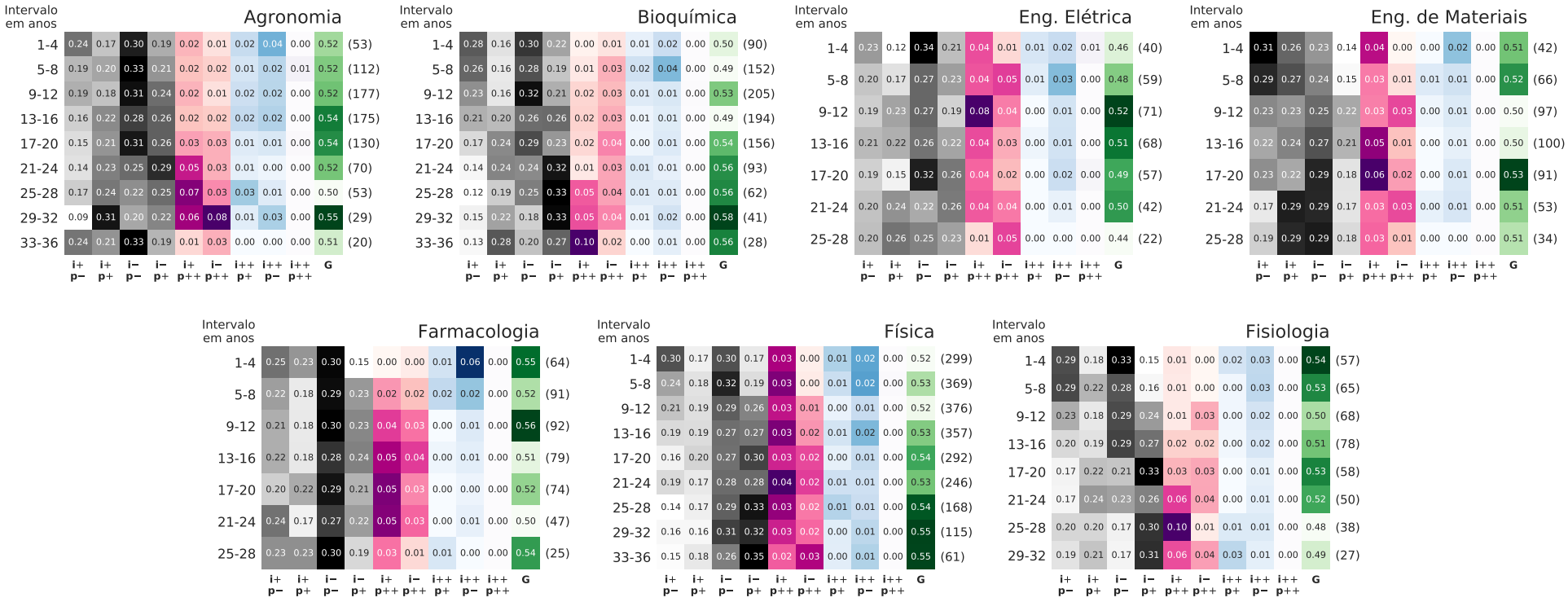

Figura 3.11: Análise das frações dos setores impacto-produtividade ao longo das carreiras de pesquisadores de diferentes áreas (Parte 1). Dividimos a carreira dos pesquisadores em janelas de quatro anos, contando como primeiro ano a data de obtenção do título de doutor do pesquisador e, assim, calculamos a fração média de anos em cada seção para cada janela de tempo. As linhas representam os períodos da carreira do pesquisador em determinada área. As nove primeiras colunas representam frações médias em cada uma das seções e a última coluna é o coeficiente de Gini dos setores não-*outliers*. O número de pesquisadores em cada janela temporal é indicado entre parênteses ao final das linhas. As janelas temporais varrem um intervalo de tempo que é superior ao disponível em nossa base de dados (19 anos, 1997-2015) porque existe uma variedade de pesquisadores em épocas diferentes de suas carreiras.

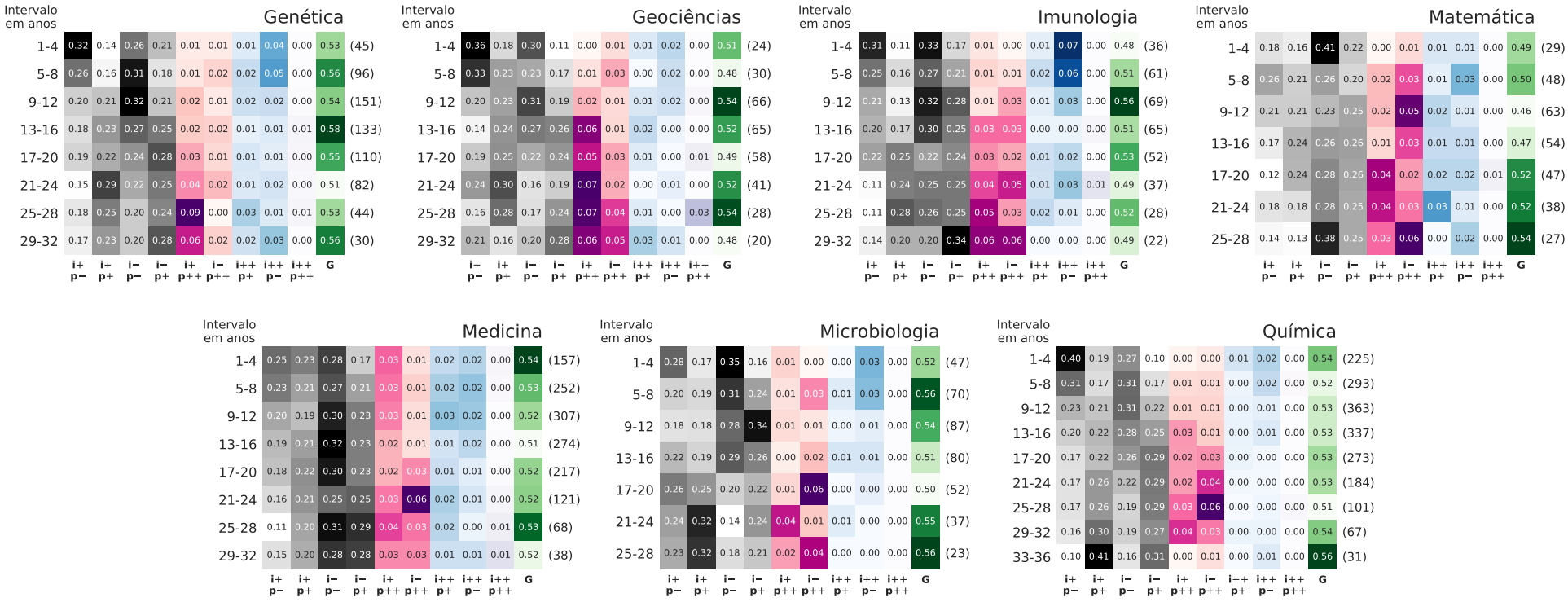

Figura 3.12: Análise das frações dos setores impacto-produtividade ao longo das carreiras de pesquisadores de diferentes áreas (Parte 2). Dividimos a carreira dos pesquisadores em janelas de quatro anos, contando como primeiro ano a data de obtenção do título de doutor do pesquisador e, assim, calculamos a fração média de anos em cada seção para cada janela de tempo. As linhas representam os períodos da carreira do pesquisador em determinada área. As nove primeiras colunas representam frações médias em cada uma das seções e a última coluna é o coeficiente de Gini dos setores não-*outliers*. O número de pesquisadores em cada janela temporal é indicado entre parênteses ao final das linhas. As janelas temporais varrem um intervalo de tempo que é superior ao disponível em nossa base de dados (19 anos, 1997-2015) porque existe uma variedade de pesquisadores em épocas diferentes de suas carreiras.

Começamos analisando as frações de setores não-*outliers*. Primeiro, nos concentramos no grupo composto pelas disciplinas de Agronomia, Bioquímica, Fisiologia, Genética, Geociências, Imunologia e Química. No começo da carreira, os pesquisadores desse grupo apresentam uma tendência de povoarem setores de baixa produtividade (p−). Em estágios mais avançados da carreira, existe uma migração para setores de alta produtividade (p+). Esse comportamento pode ser justificado pelo maior número de colaborações e maturidade adquiridos pelos pesquisadores em etapas posteriores da carreira. Num segundo momento, analisamos o grupo formado por pesquisadores das áreas de Física e Medicina. Os pesquisadores dessas duas disciplinas mantém frações médias mais ou menos constantes para os setores (i+, p+) e (i−, p−). Além disso, é visível que existe uma transição da seção (i+, p−) para (i−, p+). De modo similar ao grupo anterior, existe um aumento da produtividade; entretanto, esse aumento ocorre em detrimento do impacto das revistas em que majoritariamente publicam. Podemos analisar também os pesquisadores da disciplina de Matemática. Os matemáticos não mostram um comportamento monotonicamente crescente ou decrescente. O início da carreira é marcado por alta fração média na região de baixa produtividade e baixo impacto (i−, p−), enquanto a região de alta produtividade e impacto (i+, p+) é subpovoada. Em fases intermediárias da carreira, ocorre uma inversão da situação: existe um pico ao redor dos 20 anos da carreira para a fração na seção (i+, p+) e um vale para a seção (i−, p−). Porém, curiosamente, a tendência inicial retorna em períodos posteriores. As demais áreas não apresentam um comportamento tão regular, sendo descritas por padrões mais individualizados.

Em seguida, vamos analisar as frações de setores *outliers*. Em primeiro lugar, percebemos que a fração no setor (i++, p++) é quase sempre nula para todas as disciplinas. Esse resultado indica que a dificuldade de os pesquisadores se destacarem em ambos os quesitos persiste ao longo da carreira. Além disso, quando agregamos os setores hiperprolíficos (p++) observamos que existe uma tendência crescente no decorrer da carreira para todas as disciplinas, com exceção da Engenharia Elétrica. Por outro lado, quando agregamos os setores de altíssimo impacto (i++), observamos que existe uma tendência de decrescimento das frações no decorrer da carreira, com exceção da Matemática que apresenta frações mais ou menos constantes.

Como descrito anteriormente, o coeficiente de Gini representa a desigualdade das frações nas seções não-*outliers*. Assim, podemos associar valores pequenos do coeficiente (uniformidade das frações) a uma certa "indecisão" em relação à estratégia de publicação adotada pelos pesquisadores. No caso extremo em que $G = 0$, essa estratégia de publicação pode ser considerada aleatória. De maneira oposta, valores elevados (desigualdade nas frações) indicam que os pesquisadores podem escolher uma estratégia específica. Nesse contexto, essa estratégia de publicação representa a "escolha" referente à quantidade de artigos produzidos e ao impacto das revistas em que o pesquisador publica. Notamos que existe uma tendência

crescente no coeficiente de Gini médio para pesquisadores das áreas de Bioquímica e Física. Isso indica que os pesquisadores não-*outliers* dessas disciplinas tendem a adotar uma estratégia de publicação mais específica com a progressão da carreira. Alternativamente, as áreas de Medicina e Química apresentam coeficiente de Gini aproximadamente constantes. Dessa maneira, pesquisadores dessas áreas mantém o mesmo nível de determinação em relação à estratégia durante toda carreira. As demais disciplinas não apresentam um comportamento tão regular quanto as demais e devem ser analisadas individualmente.

A base de dados do fator de impacto engloba 14 disciplinas, como mostram as Figuras 3.11 e 3.12. Realizamos a mesma análise para o indicador SJR que compreende um total de 25 áreas, conforme mostram as Figuras C.11, C.12 e C.13. De modo geral, observamos que os resultados são similares para as disciplinas presentes em ambas bases de dados. Entretanto, os resultados para o indicador SJR mostram as particularidades das demais áreas.

3.8 Influência da produtividade no impacto científico

Uma inspeção visual do plano impacto-produtividade da Figura 3.3 aponta que pode existir uma relação entre as duas variáveis. De acordo com a disposição dos pontos, maiores valores de produtividade aparentam estar associados a menores valores de impacto. Para testar essa hipótese, decidimos empregar um modelo linear. Porém, é necessário fazer algumas considerações para determinar o conjunto de dados que se adequa melhor a esse modelo. Especificamente:

- Os *outliers* são valores extremos e, por isso, podem ser considerados como medidas influentes em um modelo regressor [53]. Eles podem distorcer a estimativa dos parâmetros e alavancar a reta da regressão linear em uma direção que não corresponde ao verdadeiro comportamento médio [52]. Por essa razão, aplicamos o modelo apenas para o conjunto de pontos dos setores não-*outliers*[4];

- O conjunto de dados não é balanceado com relação ao tamanho das carreiras. Dessa maneira, indivíduos com carreiras mais longas terão maior influência na regressão, enviesando as estimativas dos parâmetros e novamente ocultando o comportamento médio verdadeiro. Na realidade, podemos dizer que esse fato viola o princípio de homogeneidade do modelo linear simples;

- Os pares ordenados de impacto e produtividade de um determinado pesquisador estão correlacionados localmente: o conjunto de dados de um mesmo pesquisador exibe uma tendência mais similar se comparado com de outro pesquisador. Dessa maneira, o princípio de independência estatística do modelo linear simples também é violado.

[4]Uma análise dos pesquisadores *outliers* foi realizada na Seção 3.5.

Conseguimos resolver o problema dos *outliers* selecionando a porção correta dos dados para realização da regressão. No entanto, a violação dos princípios do modelo linear simples não possui uma solução simples. Com intuito de levar em consideração a estrutura hierárquica a nível de pesquisador em nosso favor e, além disso, lidar com os problemas descritos acima, propomos a utilização de um modelo linear misto do tipo

$$\boldsymbol{I} = \boldsymbol{P}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\boldsymbol{u} + \boldsymbol{\varepsilon} \ , \tag{3.9}$$

em que $\boldsymbol{I}$ é o impacto, $\boldsymbol{P}$ a produtividade, $\boldsymbol{\beta}$ os coeficientes do modelo (intercepto e inclinação), $\boldsymbol{Z}$ a matriz modelo de efeitos aleatórios, $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ a matriz de covariância, $\boldsymbol{u}$ uma variável aleatória esférica e $\boldsymbol{\varepsilon}$ o erro do modelo.

Com o modelo linear misto, levamos em conta a estrutura hierárquica que, neste caso, é representada pelo agrupamento dos dados por pesquisador. De modo explícito, assumimos que existem parâmetros médios populacionais: μ_0 para o intercepto e μ_1 para a inclinação. Além disso, afirmamos que esses mesmos parâmetros são variáveis aleatórias que diferem entre os grupos. Dessa maneira, os dados podem ser considerados aproximadamente independentes no nível dos pesquisadores. Assim, supomos que o intercepto β_0 é distribuído normalmente como

$$\beta_0 \sim \mathcal{N}(\mu_0, \sigma_0) \ ,$$

com média μ_0 e variância σ_0. Por sua vez, a inclinação β_1 é distribuída normalmente como

$$\beta_1 \sim \mathcal{N}(\mu_1, \sigma_1) \ ,$$

com média μ_1 e variância σ_1. Assim como no modelo linear simples, o erro continua sendo normal com média nula e variância σ, isto é,

$$\varepsilon \sim \mathcal{N}(0, \sigma) \ .$$

Esse tratamento também possibilita resolvermos o problema do desbalanceamento do dado, visto que pesquisadores com quantidades de dados diferentes têm a mesma importância no modelo. Porém, precisamos que haja pontos suficientes por pesquisador para que seu intercepto e inclinação sejam estimados com confiança. Por isso, selecionamos apenas pesquisadores com no mínimo cinco anos nos setores não-*outliers* do plano impacto-produtividade.

De modo que a abordagem hierárquica fique mais clara, podemos reescrever o modelo como

$$I_i = \beta_{0j} + \beta_{1j} P_i + \varepsilon_i \ , \tag{3.10}$$

em que o índice i representa o i-ésimo dado da amostra e j representa o j-ésimo pesquisador.

Os coeficientes podem ser interpretados como

$$\begin{aligned} \beta_{0j} &= \mu_0 + b_{0j} \\ \beta_{1j} &= \mu_1 + b_{1j} \end{aligned} ,$$

em que b_{0j} e b_{1j} representam as variações dos coeficientes β_0 e β_1 para o j-ésimo pesquisador e μ_0 e μ_1 são as médias dos mesmos coeficientes β_0 e β_1.

Decidimos utilizar o método bayesiano para amostrar os parâmetros do modelo linear misto. A abordagem bayesiana possibilita a obtenção de uma distribuição de probabilidade para cada parâmetro. No caso da maximização da verossimilhança da vertente frequentista, obtemos sempre uma estimativa pontual. Como pesquisadores de uma mesma área podem apresentar diferentes comportamentos, a modelagem bayesiana aparenta ser mais adequada, visto que podemos analisar não apenas a tendência global, mas também peculiaridades em nível individual por meio das distribuições de probabilidade a *posteriori* dos parâmetros. Dessa forma, estimamos os parâmetros ($\boldsymbol{\theta} = \{\sigma, \beta_{0j}, \beta_{1j}\}$) e hiper-parâmetros ($\boldsymbol{\phi} = \{\mu_0, \mu_1, \sigma_0, \sigma_1\}$) do modelo por intermédio do módulo *pymc3* [69] com o algoritmo HMC-NUTS para amostragem da *posteriori* que é dada pela Eq. (1.64), isto é,

$$P(\boldsymbol{\theta}, \boldsymbol{\phi}|D) \propto P(D|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{\phi})P(\boldsymbol{\phi}) \ , \tag{3.11}$$

em que $P(\boldsymbol{\theta}, \boldsymbol{\phi}|D)$ é a distribuição a *posteriori*, $P(D|\boldsymbol{\theta})$ é a verossimilhança, $P(\boldsymbol{\theta}|\boldsymbol{\phi})$ é a distribuição a *priori* e $P(\boldsymbol{\phi})$ é a distribuição a *hiper-priori*. A Figura 3.13 mostra a escolha das distribuições a *priori* (em roxo) e a *hiper-priori* (em verde) para os parâmetros e hiper-parâmetros do modelo. Optamos por distribuições de probabilidade não-informativas, para que a escolha não influencie na estimação da *posteriori*. Primeiramente, para a variância do erro, definimos a distribuição a *priori* como

$$\sigma \sim \mathcal{U}(0, 100) \ .$$

Para os hiper-parâmetros que correspondem à média dos coeficientes, estipulamos a *hiper-priori* como sendo uma distribuição normal com alta variabilidade, isto é,

$$\begin{aligned} \mu_0 &\sim \mathcal{N}(0, 10^5) \\ \mu_1 &\sim \mathcal{N}(0, 10^5) \end{aligned} .$$

Por outro lado, como o hiper-parâmetro de variância deve ser estritamente positivo, não podemos utilizar uma distribuição normal, pois ela é uma distribuição simétrica que envolve valores negativos. Por isso, decidimos utilizar a distribuição gama inversa [123] com

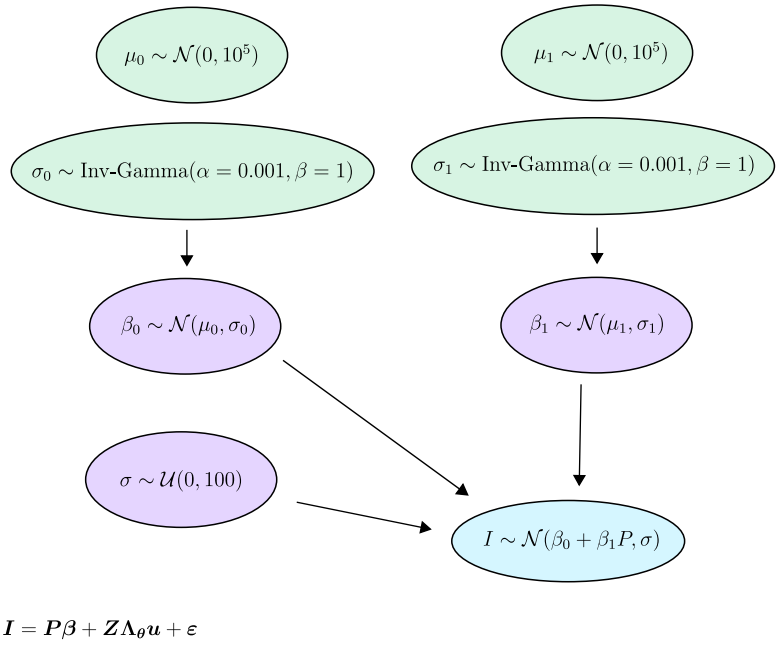

Figura 3.13: Escolha das distribuições a *priori* (em roxo) para os parâmetros e escolha das distribuições a *hiper-priori* (em verde) para os hiper-parâmetros.

parâmetros $\alpha = 0.001$ e $\beta = 1$, ou seja,

$$\begin{aligned} \sigma_0 &\sim \text{Inv-Gamma}(\alpha = 0.001, \beta = 1) \\ \sigma_1 &\sim \text{Inv-Gamma}(\alpha = 0.001, \beta = 1) \end{aligned} .$$

Para uma análise preliminar, utilizamos as medidas padronizadas pelo *z-score* para avaliar o efeito global da produtividade sobre o impacto científico por meio do modelo descrito pela Eq. (3.9). Assim, conseguimos comparar a influência da produtividade para diferentes disciplinas numa mesma escala. A Figura 3.14 mostra as distribuições de probabilidade marginais das inclinações (que denominamos "efeito da produtividade no impacto") para todas as disciplinas. As Figuras C.14 e C.15 mostram as distribuições de probabilidade marginais para o indicador SJR. A distribuição marginal é obtida mediante integração da distribuição a *posteriori* em relação aos demais parâmetros. Para o fator de impacto, observamos que existe uma tendência majoritária de aumento do impacto científico com o aumento da produtividade em oposição ao que supomos inicialmente. Para a Farmacologia e Imunologia, no entanto, existe uma correlação negativa. A Matemática apresenta o maior efeito positivo da produtividade no impacto para o modelo padronizado do fator de impacto.

A situação é relativamente diferente para o indicador SJR: há mais áreas que apresentam associações negativas entre impacto científico e produtividade. Esse é o caso para as áreas de Engenharia Elétrica, Engenharia de Materiais, Engenharia Mecânica, Farmacologia, Geoci-

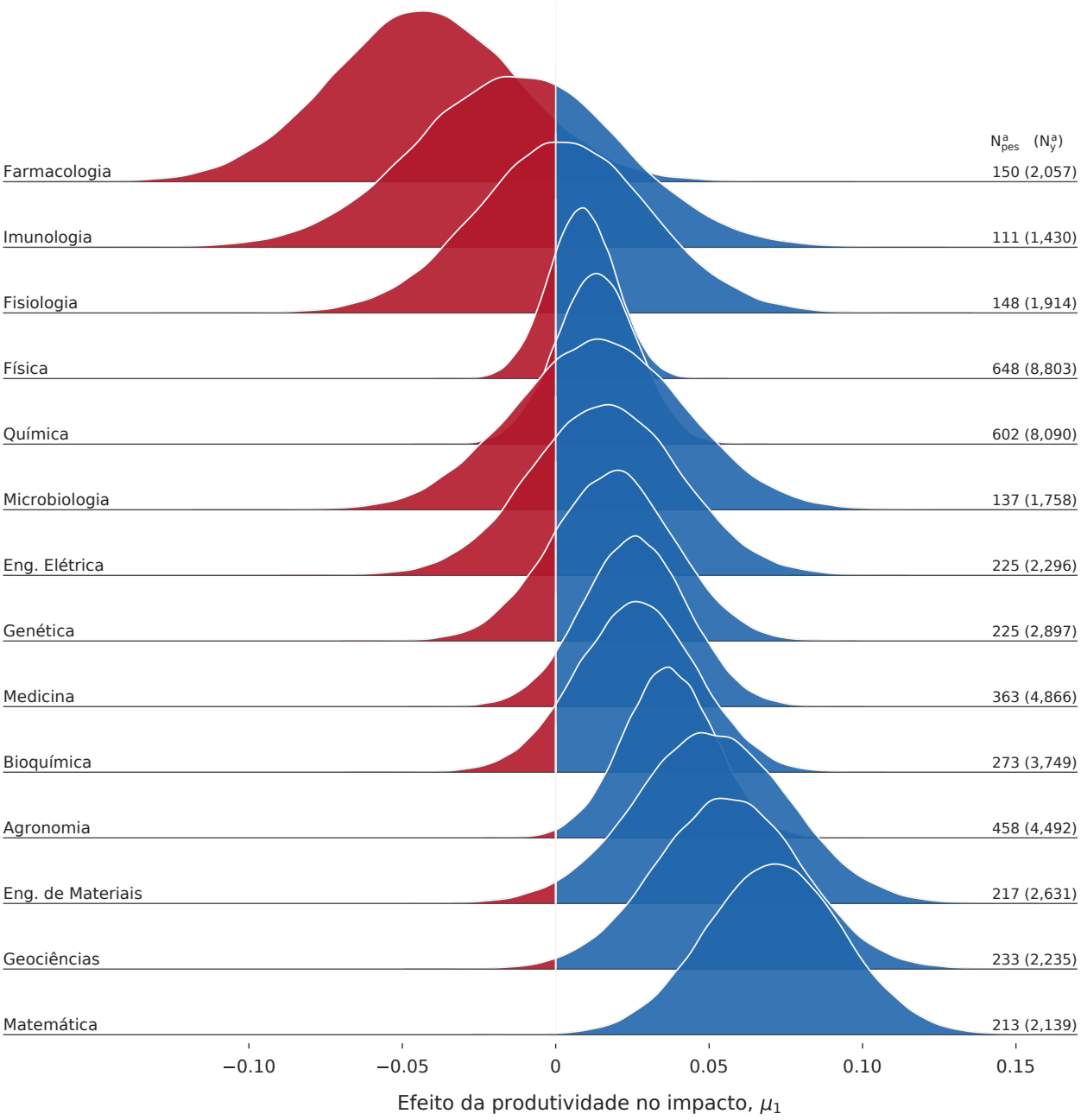

Figura 3.14: Distribuições de probabilidade do parâmetro β_1 do modelo linear misto padronizado para cada disciplina. As distribuições de probabilidade marginais são obtidas integrando a distribuição a *posteriori* em relação aos demais parâmetros. À direita de cada distribuição, estão especificados o número de pesquisadores (N^a_{pes}) e, entre parênteses, o número de pontos (N^a_y) utilizados na realização da regressão para cada área a.

ências e Zoologia. Notamos que ocorre uma inversão no sinal dos parâmetros estimados para as disciplinas de Engenharia Elétrica, Engenharia de Materiais, Geociências e Imunologia se compararmos os resultados dos dois indicadores. Conforme veremos mais adiante, algumas dessas inversões ocorrem novamente para regressão deflacionada. Na parte à direita de cada distribuição, estão especificados o número de pesquisadores (grupos, N^a_{pes}) e o número total

de anos (pontos, N_y^a) entre parênteses. Observamos que as quantidades de pesquisadores e de anos tornam as distribuições mais ou menos concentradas. Isso ocorre porque, com aumento da quantidade de dados, temos mais evidências da hipótese e, consequentemente, a incerteza é menor. Por exemplo, como mais dados estão disponíveis para as áreas de Física e Química, suas distribuições são mais localizadas. Além disso, os resultados para disciplinas presentes em ambas as bases de dados podem variar, pois os indicadores não são perfeitamente correlacionados, como descrito na Figura 2.3. Também é necessário lembrar que os bancos de dados englobam diferentes revistas em diferentes períodos. Desse modo, esses três fatores – disponibilidade de dados, definição dos indicadores e diferença nas bases de dados – poderiam explicar a inversão de sinal para as áreas mencionadas anteriormente. Por fim, a disciplina de Saúde Coletiva apresenta a maior influência positiva da produtividade para o indicador SJR e, em seguida, aparece a Matemática.

Num segundo momento, aplicamos novamente o modelo descrito pela Eq. (3.9) para realizar uma regressão com as medidas deflacionadas. Dessa maneira, conseguimos estimar o efeito da produtividade sobre o impacto em termos reais para cada disciplina. Antes de expor os resultados, analisaremos aspectos qualitativos do modelo por meio da Figura 3.15 para compreender melhor a modelagem linear mista. O algoritmo HMC amostra os parâmetros de acordo com a geometria da distribuição a *posteriori* e encontra, para cada um dos pesquisadores, um valor de intercepto β_0 e de inclinação β_1. A Figura 3.15a mostra o conjunto de inclinações para pesquisadores da disciplina de Física. Existem $N_+ = 394$ pesquisadores que apresentam inclinação positiva e $N_- = 254$ pesquisadores que apresentam comportamento negativo. A Figura 3.15b exemplifica o melhor ajuste de um pesquisador com correlação positiva, enquanto a Figura 3.15c o faz para um pesquisador com tendência negativa. Assim, o modelo linear misto é capaz de incorporar ambos os comportamentos (positivo e negativo), mesmo que a tendência global, representada pelo valor médio μ_1, seja positiva no caso da Física.

A Figura 3.16a mostra o efeito real da produtividade no impacto médio para todas as disciplinas de acordo com o modelo linear misto. As barras de erro representam a região de maior densidade da distribuição a *posteriori* correspondente aos percentis 5 e 95. Observamos que apenas a Farmacologia apresenta um decréscimo do impacto com o aumento da produtividade ($\mu_1 \approx -0.11$ unidades de fator de impacto por artigo publicado[5]). Todas as demais disciplinas apresentam uma influência positiva da produtividade sobre o impacto. A área com maior aumento médio do impacto com a produtividade é Geociências com $\mu_1 \approx 0.042$, seguido pela Matemática com $\mu_1 \approx 0.026$ unidades de fator de impacto por artigo publicado. De modo geral, a análise para o indicador SJR concorda com os resultados obtidos para o fator de impacto. Entretanto, ocorrem novamente algumas inversões de sinal na relação entre

[5]Daqui em diante, quando escrevermos $\mu_n \approx$ valor, esse corresponde ao valor médio do parâmetro β_n (sendo $n = 0, 1$) calculado pela Eq. (1.32).

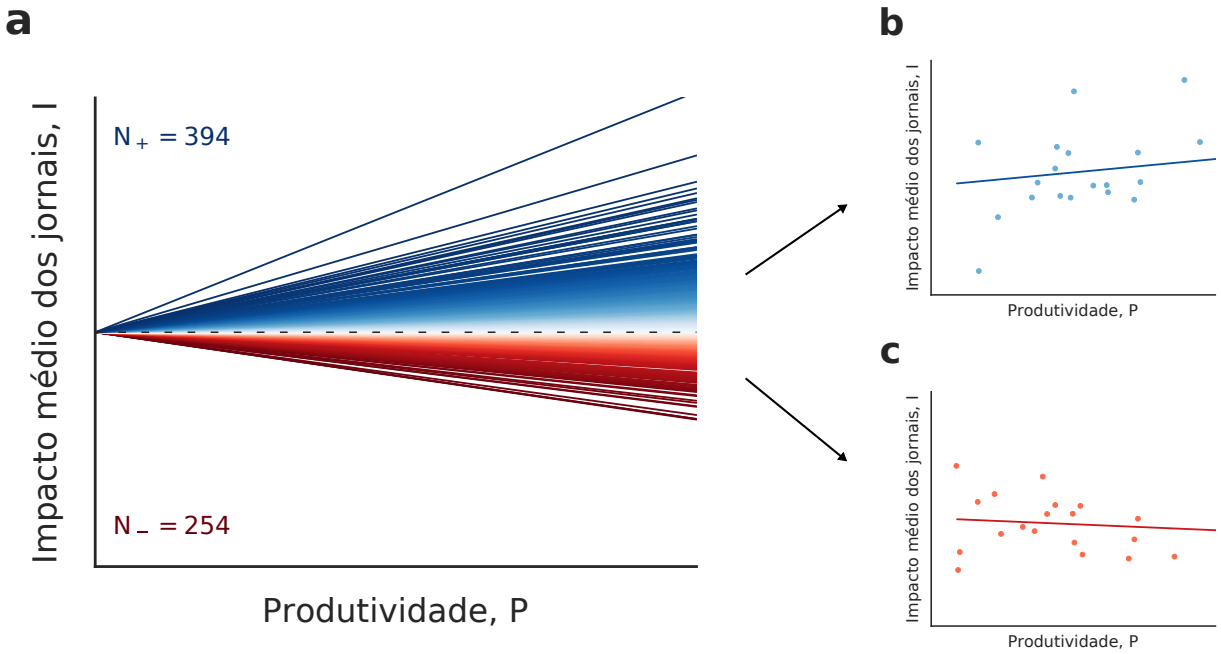

Figura 3.15: Exemplificando a regressão linear mista por pesquisador. **(a)** Estimativa das inclinações obtidas por meio do modelo linear misto para $N = 648$ pesquisadores da Física. Para esse caso, $N_+ = 394$ pesquisadores apresentam uma inclinação positiva, enquanto que $N_- = 254$ pesquisadores apresentam uma inclinação negativa. Ilustração da inclinação e intercepto ajustados para **(b)** um pesquisador com comportamento positivo e **(c)** um pesquisador com comportamento negativo. Nesses últimos dois painéis, cada ponto representa a produtividade e o fator de impacto médio num ano da carreira dos pesquisadores.

produtividade e impacto científico quando comparamos resultados de diferentes regressões para certas disciplinas. É o que acontece para as áreas de Engenharia Elétrica, Engenharia dos Materiais, Geociências, Imunologia e Zoologia. Nesses casos, verificamos também um grande espalhamento na distribuição a *posteriori* dos valores de μ_1 (indicado pelas grandes barras de erro das Figuras 3.16a e C.16a). Assim, não podemos ser tão confiantes sobre como é a relação entre produtividade impacto para esses casos. Apenas com a inclusão de mais dados é que poderíamos, a princípio, esclarecer melhor essa diferença. Isto é, poderíamos esclarecer melhor se a divergência entre os sinais ocorre somente pela escassez de dados ou se acontece também pela natureza intrinsecamente distinta dos indicadores e bases de dados.em que escrevemos a distribuição de probabilidade normalizada e centrada na origem, com a somatória abrangendo todo conjunto de dados. Como a transformação logarítmica preserva as características da verossimilhança, podemos considerar o logaritmo da verossimilhança, isto é,

Quando propomos o modelo linear misto, supomos que o parâmetro β_1 estava distribuído normalmente com

$$\beta_1 \sim \mathcal{N}(\mu_1, \sigma_1) \ .$$

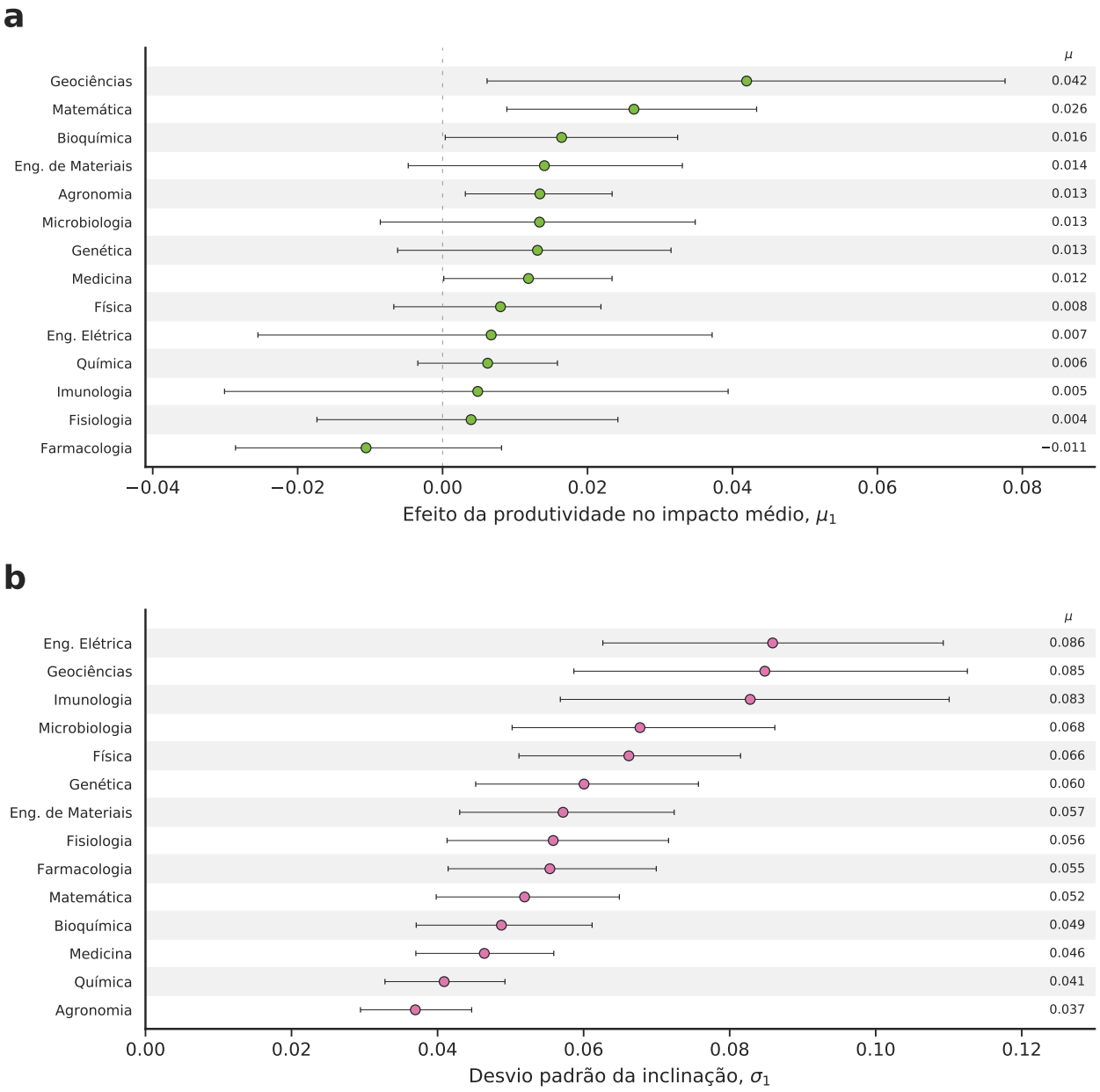

Figura 3.16: Estimativa bayesiana dos parâmetros de localização e de escala da distribuição das inclinações β_1. Os painéis apresentam valores médios para **(a)** o valor médio μ_1 e **(b)** o valor do desvio padrão σ_1 do efeito da produtividade no impacto. As barras de erro ilustram a região de maior densidade da distribuição a *posteriori*.

Nesse sentido, o desvio padrão σ_1 representa a variabilidade do conjunto de inclinações de pesquisadores de uma determinada área, isto é, o quão diferentes as inclinações são entre os pesquisadores. A Figura 3.16b mostra a estimativa dos valores de σ_1 para todas as disciplinas. Observamos que a área com maior variabilidade das inclinações entre os pesquisadores é a área de Engenharia Elétrica com $\sigma_1 \approx 0.086$ unidades de fator de impacto por artigo publicado. Por outro lado, pesquisadores da Agronomia demonstram a menor variabilidade entre suas inclinações com $\sigma_1 \approx 0.037$ unidades de fator de impacto por artigo publicado.

A Figura C.16b exibe os desvios padrões das disciplinas presentes no banco de dados do indicador SJR. A área com maior variabilidade é a Parasitologia com $\sigma_1 \approx 0.058$ unidades de indicador SJR por artigo publicado. A área com a menor variabilidade é, novamente, a Agronomia com $\sigma_1 \approx 0.017$ unidades de indicador SJR por artigo publicado. Notamos também que as disciplinas com resultados conflitantes entre o fator de impacto e o indicador SJR, e entre as regressões padronizada e deflacionada são as que apresentam maiores valores de σ_1, sugerindo novamente que a relação entre produtividade e impacto nessas áreas não está bem definida em nossos dados.

3.9 Variabilidade dos indicadores de impacto

Nas análises realizadas anteriormente, percebemos que as disciplinas possuem comportamentos característicos. Por exemplo, os limiares *outliers* impacto e produtividade são diferentes para cada área, como mostra a Figura 3.7. Portanto, pesquisadores que produzem em grande quantidade em uma certa disciplina não necessariamente são considerados *outliers* nos padrões de outra. Uma maneira de investigar mais aspectos quantitativos dessa diferenciação por área consiste em analisar o desvio padrão do impacto médio anual como função da produtividade. Podemos definir o desvio padrão do impacto médio anual $\sigma_I^i(y)$ como

$$\sigma_I^i(y) = \sqrt{\frac{1}{P_i(y)-1}\sum_{j=1}^{P_i(y)}\left[\tilde{I}_{i,j}(y) - I_i(y)\right]^2}\ , \tag{3.12}$$

em que $I_i(y)$ é o indicador de impacto médio do pesquisador i no ano y, $P_i(y)$ é a produtividade respectiva e $\tilde{I}_{i,j}(y)$ o indicador de impacto do j-ésimo artigo dentre os $P_i(y)$ artigos publicados no ano y.

Para compreender a variabilidade de valores do indicador de impacto para determinada disciplina, modelamos a relação entre os valores médios do desvio padrão do impacto σ_I e a produtividade deflacionada P por intermédio de um modelo exponencial definido como

$$\sigma_I = Ae^{-\frac{P}{n_s}} + k\ , \tag{3.13}$$

em que A é uma constante, n_s é o coeficiente de saturação exponencial e k é a constante relacionada ao limiar de saturação exponencial. A Figura 3.17 mostra um *scatter plot* do desvio padrão do impacto médio, definido pela Eq. (3.12), *versus* a produtividade (P) para a disciplina de Física.

A Figura 3.18 mostra o comportamento médio do desvio padrão do impacto anual em relação à produtividade para todas as disciplinas. As Figuras C.17 e C.18 mostram as relações para o indicador SJR. Para as curvas alaranjadas, realizamos uma média por janela do conjunto de pontos mostrado na Figura 3.17, selecionando os pontos para cada disciplina.

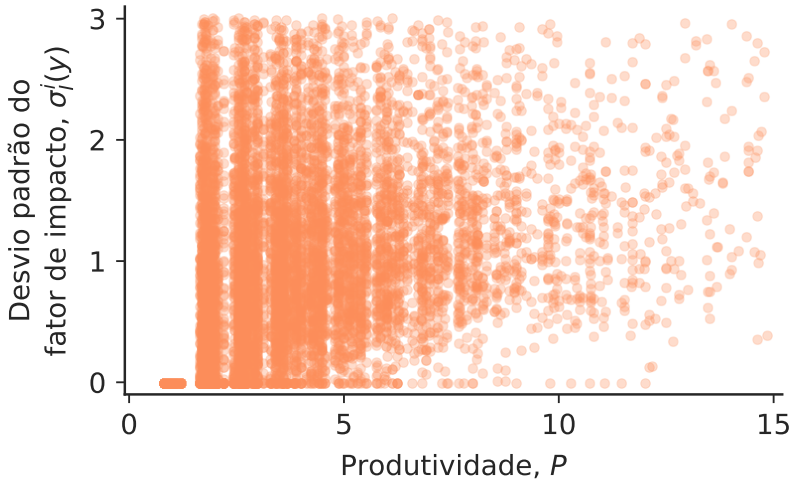

Figura 3.17: *Scatter plot* do desvio padrão do fator de impacto médio e produtividade deflacionada para a disciplina de Física. O painel contém apenas os pontos da região em que a relação exponencial foi ajustada para área de Física.

Para isso, dividimos o conjunto de dados em dez partes de acordo com o intervalo percentil e calculamos a média de cada intervalo. As curvas escuras mostram os ajustes exponenciais dos dados para o modelo descrito pela Eq. (3.13). Os valores dos parâmetros estão especificados para cada disciplina. Primeiramente, percebemos que existe uma diferença clara entre os limites de produtividade para diferentes áreas. Além disso, os valores nos quais as curvas exponenciais saturam e as taxas de variação do impacto com a produtividade também variam entre as disciplinas.

A Figura 3.19 mostra o valor dos parâmetros do modelo exponencial para todas as disciplinas. A Figura C.19 mostra os valores dos parâmetros para o indicador SJR. O coeficiente exponencial n_s indica a taxa de variação das curvas com a produtividade. Dessa maneira, valores maiores de n_s indicam uma saturação mais branda com a produtividade, enquanto valores menores de n_s apontam uma saturação mais rápida com a produtividade. Para o fator de impacto, a saturação mais suave é da Medicina ($n_s \approx 3.51$). Por outro lado, a saturação do desvio padrão do impacto anual é mais rápida para Física ($n_s \approx 0.74$). Podemos interpretar o valor de n_s como uma medida de "variedade média de valores de fator de impacto". Por exemplo, para Medicina é necessário que se produza cerca de nove artigos anualmente para que toda variedade média de valores de fator de impacto seja varrida. Já para Física, produzindo cerca de quatro artigos por ano, é possível cobrir toda variedade média de fatores de impacto. Para o indicador SJR, a disciplina com saturação mais branda é a Saúde Coletiva ($n_s \approx 4.68$), enquanto a Física ($n_s \approx 0.69$) novamente aparece como área de saturação mais rápida com a produtividade.

A constante de saturação exponencial k determina em que valor do desvio padrão a função exponencial tende a saturar. Nesse sentido, a constante k é uma medida que quantifica o valor médio da variabilidade máxima em unidades do indicador de impacto. Assim, disciplinas

com maior valor de k apresentam maior variação máxima do fator de impacto das revistas. Para o fator de impacto, a área com maior variabilidade média é a Medicina ($k \approx 2.70$ unidades de fator de impacto), enquanto isso, a disciplina com menor variabilidade média é a Matemática ($k \approx 0.60$ unidades de fator de impacto). A Matemática é uma disciplina cujas revistas apresentam valores menores para o fator de impacto devido principalmente às menores taxas de produtividade e de citação da área. Assim, o resultado que obtivemos reforça essa noção. Para o indicador SJR, a área com maior variabilidade média é a Ecologia ($k \approx 1.31$ unidades de indicador SJR) e a Engenharia de Materiais ($k \approx 0.43$ unidade de indicador SJR) apresenta o menor valor.

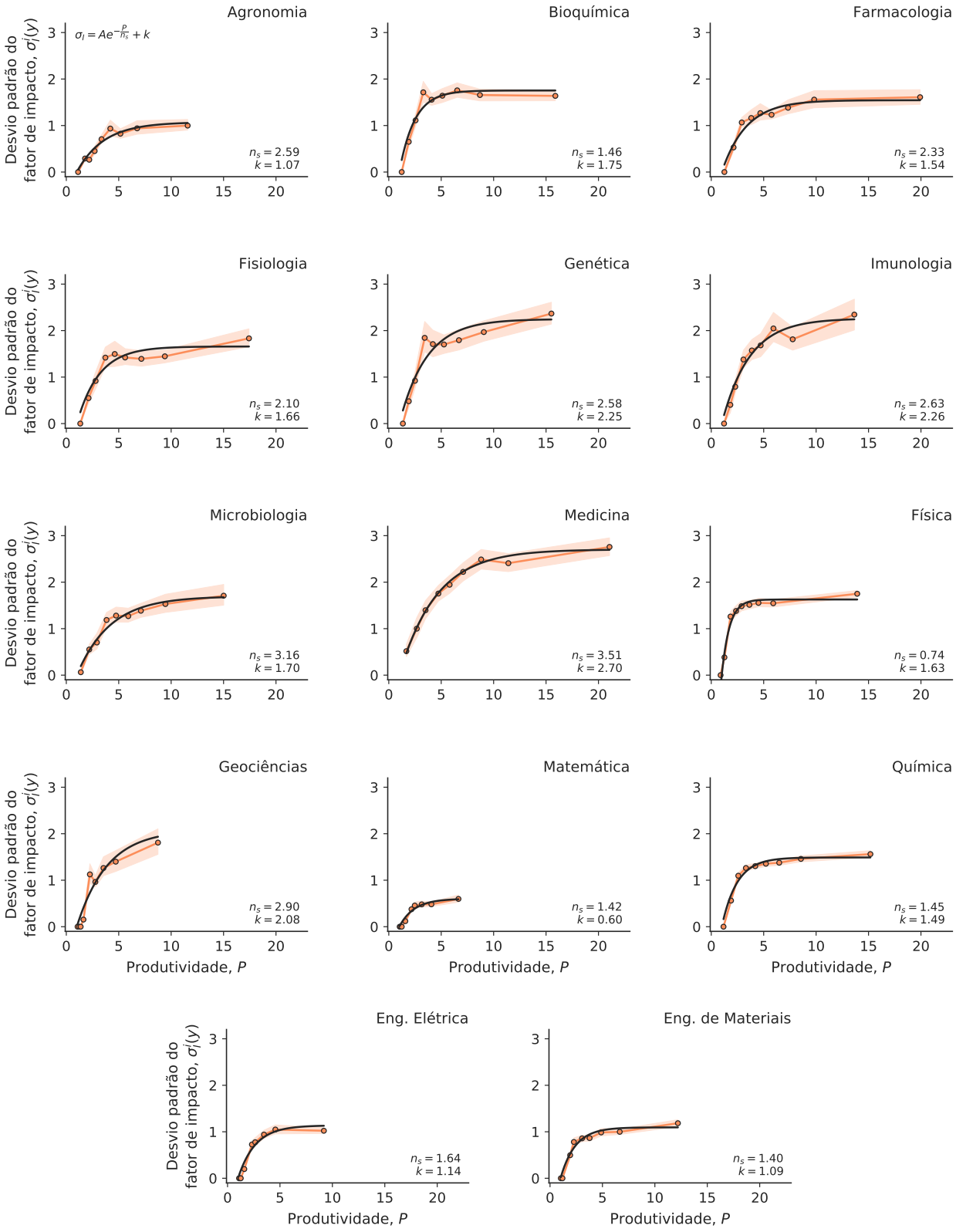

Figura 3.18: Variabilidade de valores de fator de impacto. Realizamos um ajuste exponencial para entender a relação média entre o desvio padrão anual $\sigma_I^i(y)$ do fator de impacto e a produtividade P do pesquisador. O coeficiente de saturação n_s indica o quão rápido o desvio padrão satura com o aumento da produtividade. A constante de saturação k indica qual é o valor de saturação para grandes valores de produtividade.

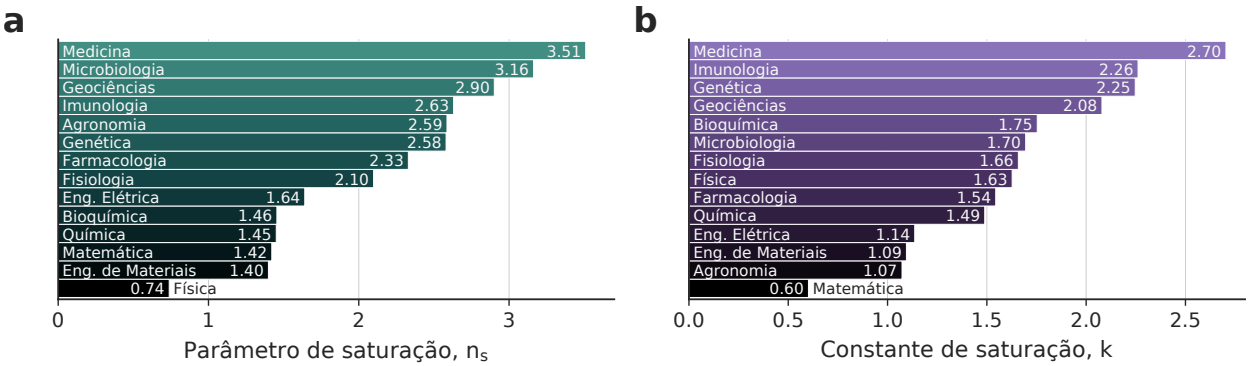

Figura 3.19: Parâmetros do ajuste exponencial para diferentes disciplinas. **(a)** Coeficiente de saturação n_s. **(b)** Constante de saturação k.

Considerações finais

Nessa dissertação, investigamos aspectos da relação entre produtividade e impacto científico por meio da análise de pesquisadores bolsa produtividade do CNPq de diversas disciplinas. Com propósito de caracterizar o comportamento dos pesquisadores, extraímos informações de seus currículos *vitae* disponíveis na Plataforma Lattes. Para as medidas de impacto, utilizamos os indicadores das revistas presentes nas bases de dados da *Web of Science*, o fator de impacto, e *SCOPUS*, o indicador SJR.

Primeiramente, por meio de regressões lineares, verificamos que existe uma tendência de inflação da produtividade de pesquisadores de todas as disciplinas estudadas. Em relação ao fator de impacto, constatamos novamente a existência de um comportamento crescente para todas as áreas. Porém, a situação é diferente quando consideramos o indicador SJR: existem algumas disciplinas com comportamento decrescente.

A partir da padronização das medidas de produtividade e impacto, construímos um plano impacto-produtividade com todas as disciplinas. Esse plano foi dividido em setores que representam o comportamento do pesquisador referente à produtividade e ao impacto científico. Assim, há regiões em que o pesquisador produz acima da média (+) ou abaixo da média (−) em determinado quesito. Além disso, por meio da definição de padronização, conseguimos definir o conceito de *outlier* (++) para o qual o pesquisador é considerado destaque em determinado quesito. Constatamos que o setor mais povoado é aquele em que o pesquisador desempenha abaixo da média em produtividade e impacto. Por outro lado, o setor menos povoado é aquele em que o pesquisador é *outlier* nas duas categorias. Adicionalmente, construímos o plano impacto-produtividade para medidas deflacionadas. Dessa maneira, conseguimos definir os valores médios e os limiares para ser *outlier* em termos de unidades reais de produtividade e impacto.

Em seguida, motivados por evidências de que existe uma dificuldade em ser *outlier* em ambos quesitos do ponto de vista global, exploramos aspectos dos *outliers* num nível indi-

vidual. Em geral, constatamos que pesquisadores que são *outliers* em produtividade não são *outliers* em impacto durante a carreira e vice-versa. Além disso, mesmo aqueles que conseguem esse feito não apresentam uma performance exageradamente acima da média. Verificamos que a fração de anos *outliers* na carreira de um pesquisador é pequena, isto é, anos *outliers* são raros. Por fim, realizamos uma análise logística para verificar o impacto do número de anos *outliers* em produtividade na probabilidade de ser *outlier* em impacto. Constatamos que existem áreas para as quais ser *outlier* em produtividade impossibilita a performance como *outlier* em impacto. Enquanto isso, para outras disciplinas isso é possível com uma probabilidade decrescente com o aumento do número de anos *outliers* em produtividade.

À parte da análise de *outliers*, investigamos o comportamento de pesquisadores não-*outliers* em relação às suas frações nos setores do plano impacto-produtividade durante a carreira. Verificamos que o setor com distribuição mais uniforme é aquele em que há uma performance abaixo da média em ambas as categorias. Portanto, existem frações grandes e pequenas aproximadamente na mesma proporção, com sutil privilégio para os pequenos valores. Os demais setores apresentam distribuições de probabilidade assimétricas em direção aos menores valores de fração. Além disso, por meio da entropia normalizada de Shannon, verificamos que pesquisadores não-*outliers* tendem a transitar entre as seções durante a carreira.

Analisamos também a dinâmica de carreira de pesquisadores de diferentes disciplinas no plano impacto-produtividade. No que concerne aos setores *outliers*, constatamos que existe uma tendência de produzir com altíssimo impacto durante o começo da carreira (com exceção da Matemática), enquanto, ao final, é mais comum se tornar hiperprolífico. Em relação aos setores não-*outliers*, o padrão é específico para cada área. Porém, observamos que existe uma tendência comum para as áreas de Agronomia, Bioquímica, Fisiologia, Genética, Geociências, Imunologia e Química, de migrar de regiões de baixa produtividade no início da carreira para regiões de alta produtividade em estágios posteriores. Para Física e Medicina, existe uma transição da região de alto impacto e baixa produtividade no início da carreira para região de baixo impacto e alta produtividade.

Realizamos uma regressão linear mista para investigar a influência da produtividade no impacto científico. Constatamos que os comportamentos são característicos para cada disciplina. No entanto, existe uma tendência majoritária de que a influência seja positiva, isto é, o impacto científico aumenta com a produtividade para pesquisadores não-*outliers*. Além disso, conseguimos quantificar essa influência a partir da regressão linear mista com medidas deflacionadas.

Finalmente, propomos um modelo exponencial para analisar a variabilidade dos valores do indicador de impacto médio anual como função da produtividade. Conseguimos interpretar o parâmetro n_s do modelo como uma medida de "variedade média de valores do indicador de

impacto". Dessa maneira, verificamos que existe uma clara diferença dessa variedade média para diferentes disciplinas, sendo a Medicina a área com a maior variedade média para o fator de impacto e a Saúde Coletiva para o indicador SJR. A área com menor variedade média é a Física para os dois indicadores. Além disso, interpretamos o parâmetro k como o valor médio da variação máxima do indicador de impacto para determinada disciplina em unidades de impacto. Constatamos que a área com maior variação média é a Medicina para o fator de impacto e a Ecologia para o indicador SJR.

Com esta dissertação, esperamos contribuir com o desenvolvimento da ciência nos âmbitos global e individual. Do ponto de vista global, esperamos que as análises realizadas possam fundamentar as tomadas de decisões pelas instituições de fomento à ciência. Por outro lado, no âmbito individual, esperamos que as informações contidas nessa dissertação possam auxiliar pesquisadores na escolha de estratégias referentes a sua produtividade e impacto científico durante a carreira.

APÊNDICE A

Medidas robustas para *z-score*

A medida *z-score* convencional pode ser definida como

$$(\text{z-score})_i = \frac{x_i - \mu}{\sigma} \ , \tag{A.1}$$

em que x_i é a i-ésima amostra, μ é a média e σ é o desvio padrão da amostra. Se a amostra segue um comportamento normal, o *z-score* corresponde à padronização do conjunto de dados. Em outros termos, podemos dizer que essa quantidade assume a distribuição $\mathcal{N}(0, 1)$. Assim, o termo $(\text{z-score})_i$ corresponde a quantos desvios padrões a i-ésima amostra está acima ou abaixo da média. Essa medida é muito útil quando queremos comparar amostras provenientes de distribuições com parâmetro de localização e escala distintos.

Em nosso trabalho, existe um padrão de inflação temporal tanto para produtividade quanto para o impacto médio das revistas. Além disso, pesquisadores de áreas diferentes apresentam comportamentos diferentes. Por isso, é necessário realizar algum tipo de padronização para possibilitar a comparação mais precisa entre pesquisadores de diferentes períodos e áreas. Porém, conforme discutido na Seção 1.7, as medidas de localização e escala da Eq. (A.1) são sensíveis à presença de *outliers*, que é o caso em nossa base de dados. Por conta disso, propomos a seguinte medida de padronização robusta ou *z-score* robusto

$$(\text{z-score robusto})_i^{a,y} = \frac{x_i - \mu_{a,y}}{\sigma_{a,y}} \ , \tag{A.2}$$

em que o índice a corresponde à área, o índice y corresponde ao ano, e $\mu_{a,y}$ e $\sigma_{a,y}$ são, respectivamente, as medidas de localização e escala de Huber (veja a Seção 1.7) da área a no ano y. Conforme já discutimos, as estatísticas de Huber são robustas à presença de *outliers* e representam bem a tendência central e de dispersão da amostra.

APÊNDICE B

Coeficiente de correlação de Pearson

O coeficiente de correlação amostral de Pearson indica a intensidade da associação linear entre duas variáveis x e y. Essa medida é usualmente definida por [124]

$$r_{xy} = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_i (x_i - \mu_x)^2}\sqrt{\sum_i (y_i - \mu_y)^2}} \ , \tag{B.1}$$

em que x_i é a i-ésima amostra da variável x, y_i é a i-ésima amostra da variável y, μ_x é a média de x e μ_y é a média de y. Podemos reescrever a Eq. (B.1) como

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\mathbb{E}[\sum_i (x_i - \mu_x)(y_i - \mu_y)]}{\sqrt{\mathbb{E}[\sum_i (x_i - \mu_x)^2]}\sqrt{\mathbb{E}[\sum_i (y_i - \mu_y)^2]}} \ , \tag{B.2}$$

em que σ_{xy} é a covariância entre x e y, σ_x é o desvio padrão de x e σ_y é o desvio padrão de y. Dessa forma, o coeficiente pode ser interpretado como uma espécie de covariância normalizada. Por meio da desigualdade de Cauchy-Schwarz podemos mostrar as seguintes propriedades

$$\begin{aligned} &|\sigma_{xy}|^2 \leq |\sigma_x||\sigma_y| \\ &|\sigma_{xy}| \leq \sqrt{|\sigma_x||\sigma_y|} \\ &|r_{xy}| \leq 1 \\ &-1 \leq r_{xy} \leq 1 \ . \end{aligned} \tag{B.3}$$

Notamos que o valor do coeficiente r_{xy} está contido no intervalo $[-1, 1]$. O extremo positivo indica correlação linear positiva perfeita, ou seja, um acréscimo de x equivale a um acréscimo proporcional de y para qualquer observação da amostra. Similarmente, o valor -1 indica

uma correlação negativa perfeita, ou seja, um acréscimo unitário em x implica um decréscimo equivalente em y. Por fim, um valor próximo de zero indica que não existe correlação linear entre as duas variáveis. É importante ressaltar que a noção de correlação não implica causalidade. Porém, ela indica o grau de similaridade linear entre as duas variáveis.

APÊNDICE C

Figuras adicionais

Neste apêndice, apresentamos todas as figuras referentes à análise do indicador SJR.

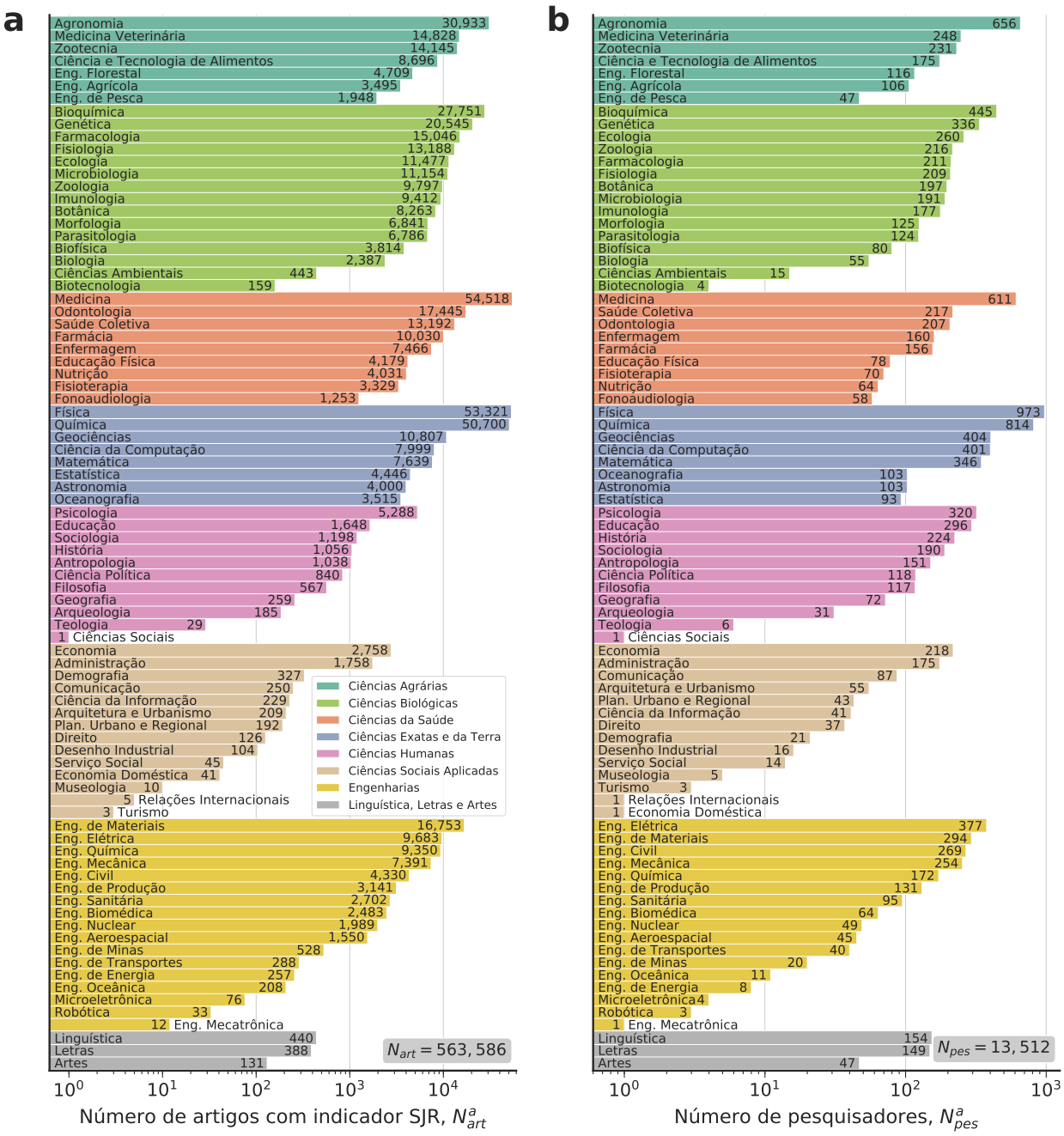

Figura C.1: Caracterização das áreas em relação aos artigos com indicador SJR. **(a)** Número de artigos com indicador SJR em cada área, sendo o total de artigos $N_a = 563\ 586$. **(b)** Número de pesquisadores em cada área, sendo o total de pesquisadores $N_p = 13\ 512$. As cores representam as diferentes grandes áreas conforme indicado pela legenda.

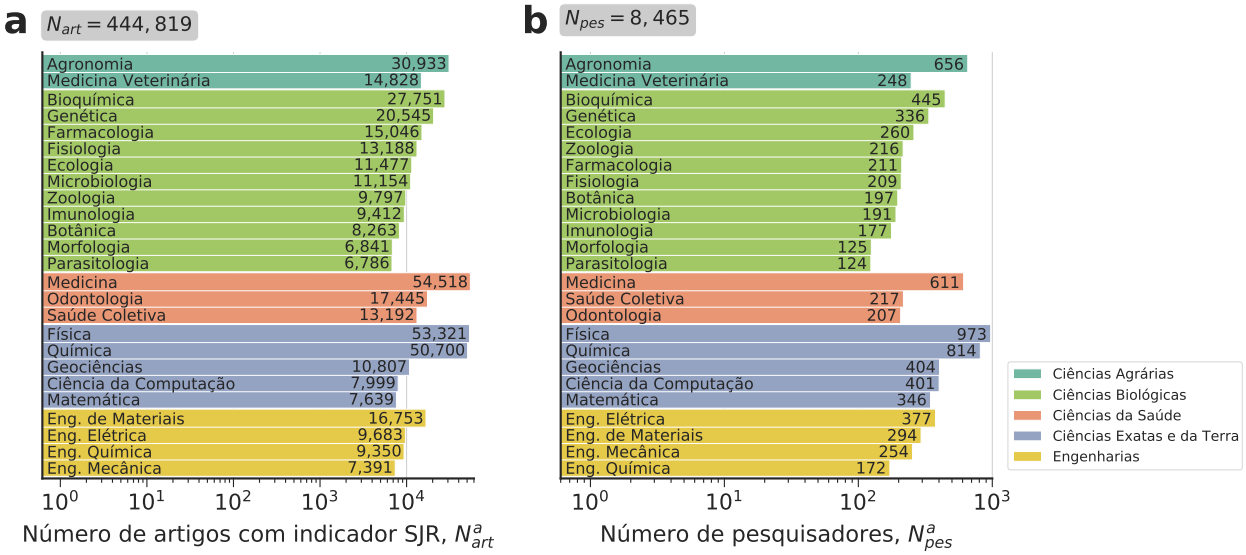

Figura C.2: Caracterização das 25 áreas selecionadas após o filtro em relação aos artigos com indicador SJR. (a) Número de artigos com indicador SJR em cada área, sendo o total de artigos $N_a = 444\,819$. **(b)** Número de pesquisadores em cada área, sendo o total de pesquisadores $N_p = 8\,465$. As cores representam as diferentes grandes áreas conforme indicado pela legenda.

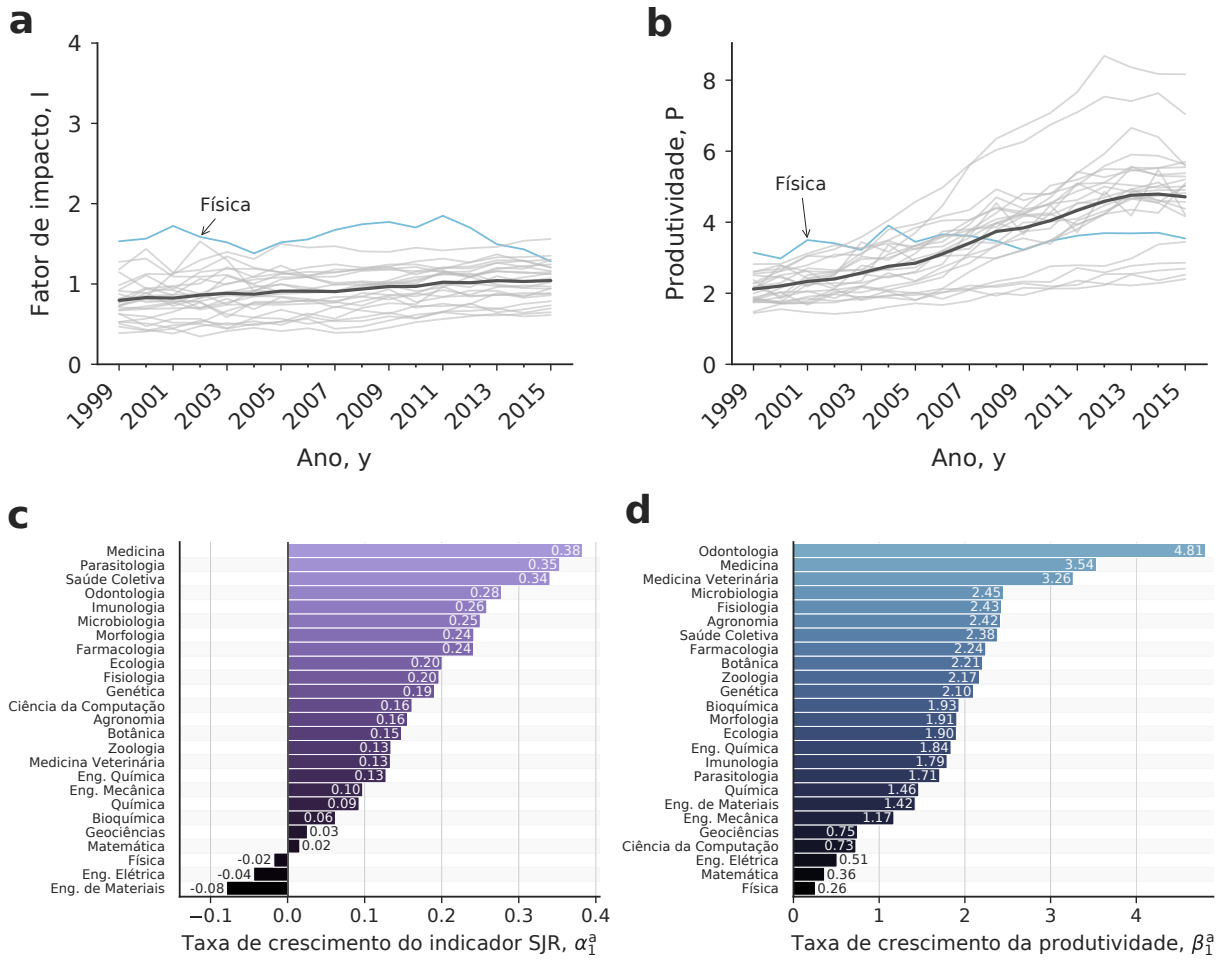

Figura C.3: Inflação do indicador SJR e da produtividade entre as diferentes áreas do conhecimento. **(a)** Padrão de crescimento do indicador SJR com o decorrer do tempo. A linha escura representa o comportamento global de todas as disciplinas. As linhas claras representam o comportamento individual de cada área. A linha azul representa o comportamento da Física. **(b)** Padrão de crescimento da produtividade com o decorrer do tempo. **(c)** Taxas de crescimento por década do indicador SJR para diferentes áreas. **(d)** Taxas de crescimento por década da produtividade referente ao conjunto de dados do indicador SJR para diferentes áreas.

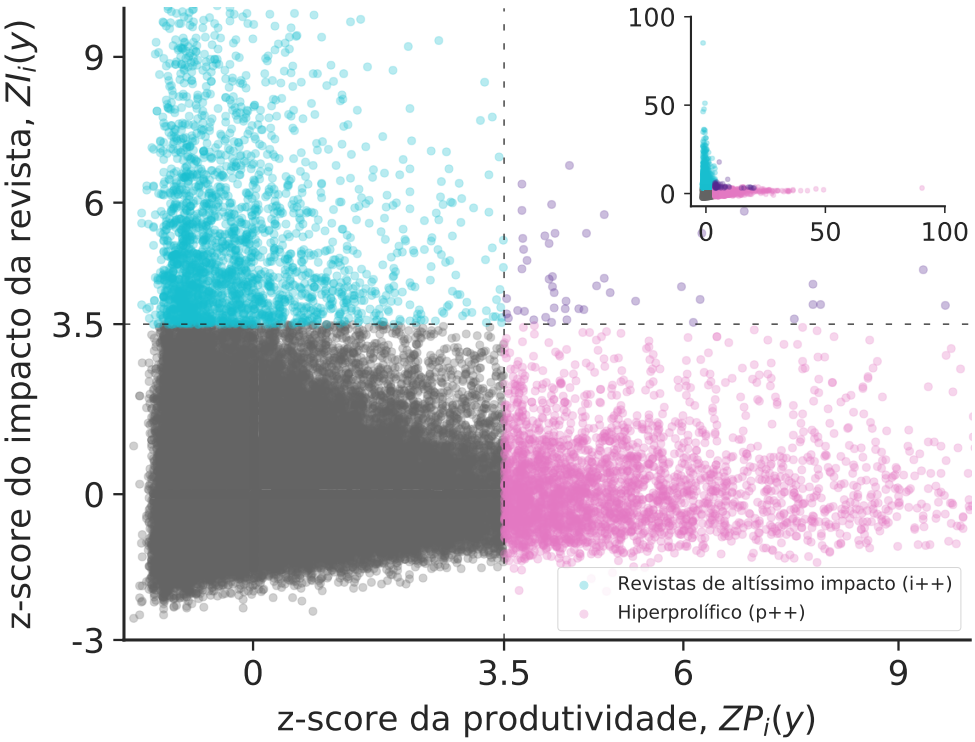

Figura C.4: Plano impacto-produtividade agregando todas as áreas do conhecimento. A partir da interpretação do *z-score*, dividimos o plano impacto-produtividade em setores. Valores acima de 3.5 indicam que o pesquisador foi *outlier* naquele quesito. Os pontos azuis indicam anos de pesquisadores hiperprolíficos. Enquanto isso, os pontos rosas indicam que o pesquisador publicou apenas em revistas de grande indicador SJR para sua área naquele ano. Valores positivos (negativos) indicam que o pesquisador esteve acima (abaixo) da média naquele ano, em sua área e em determinado quesito. O gráfico mostra valores de *z-score* até 10. O *inset* mostra o plano inteiro.

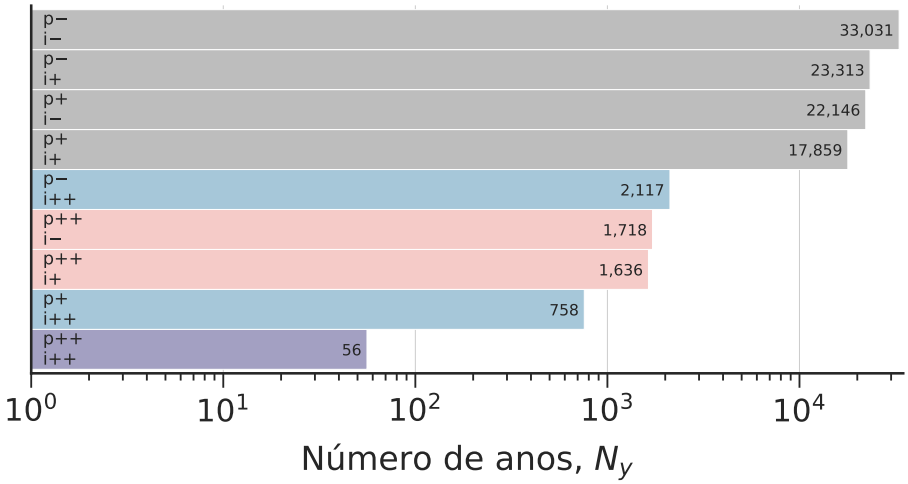

Figura C.5: Quantidade de anos da carreira dos pesquisadores em cada setor do plano impacto-produtividade para o indicador SJR.

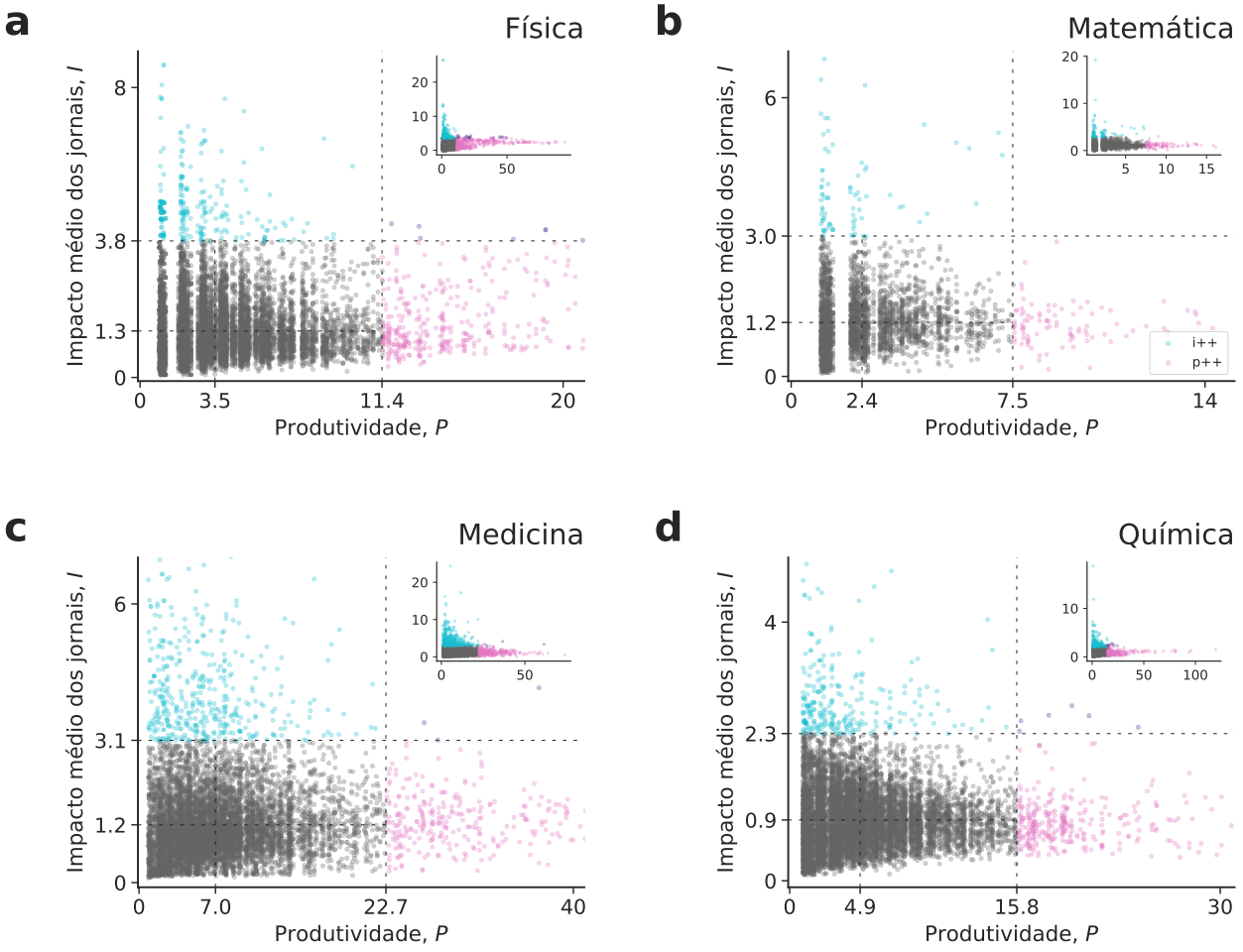

Figura C.6: Plano impacto-produtividade deflacionado para diferentes áreas do conhecimento. Os painéis mostram o plano impacto-produtividade deflacionado para as disciplinas de **(a)** Física, **(b)** Matemática, **(c)** Medicina e **(d)** Química. Os valores médios de cada área foram calculados por meio da medida de localização de Huber e os limiares *outliers* são os valores deflacionados correspondentes ao valor do *z-score* igual a 3.5.

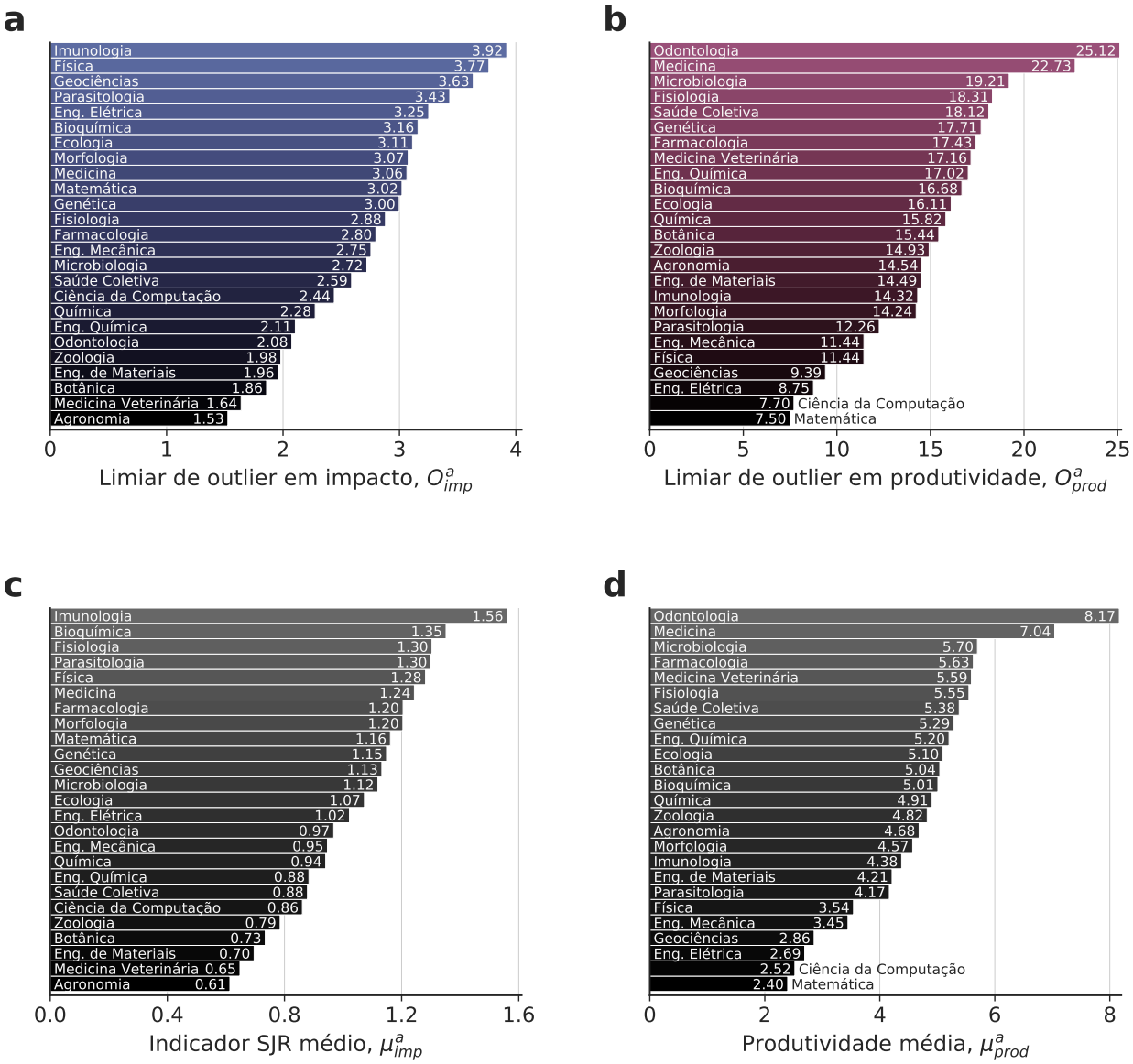

Figura C.7: Caracterização das disciplinas de acordo com as medidas deflacionadas. Como as áreas apresentam comportamentos diferentes, os limiares de *outlier* para impacto e produtividade diferem. Os painéis mostram os limiares de *outlier* para **(a)** o indicador SJR e **(b)** a produtividade, além dos valores médios de Huber para **(c)** o indicador SJR e **(d)** a produtividade para cada área a.

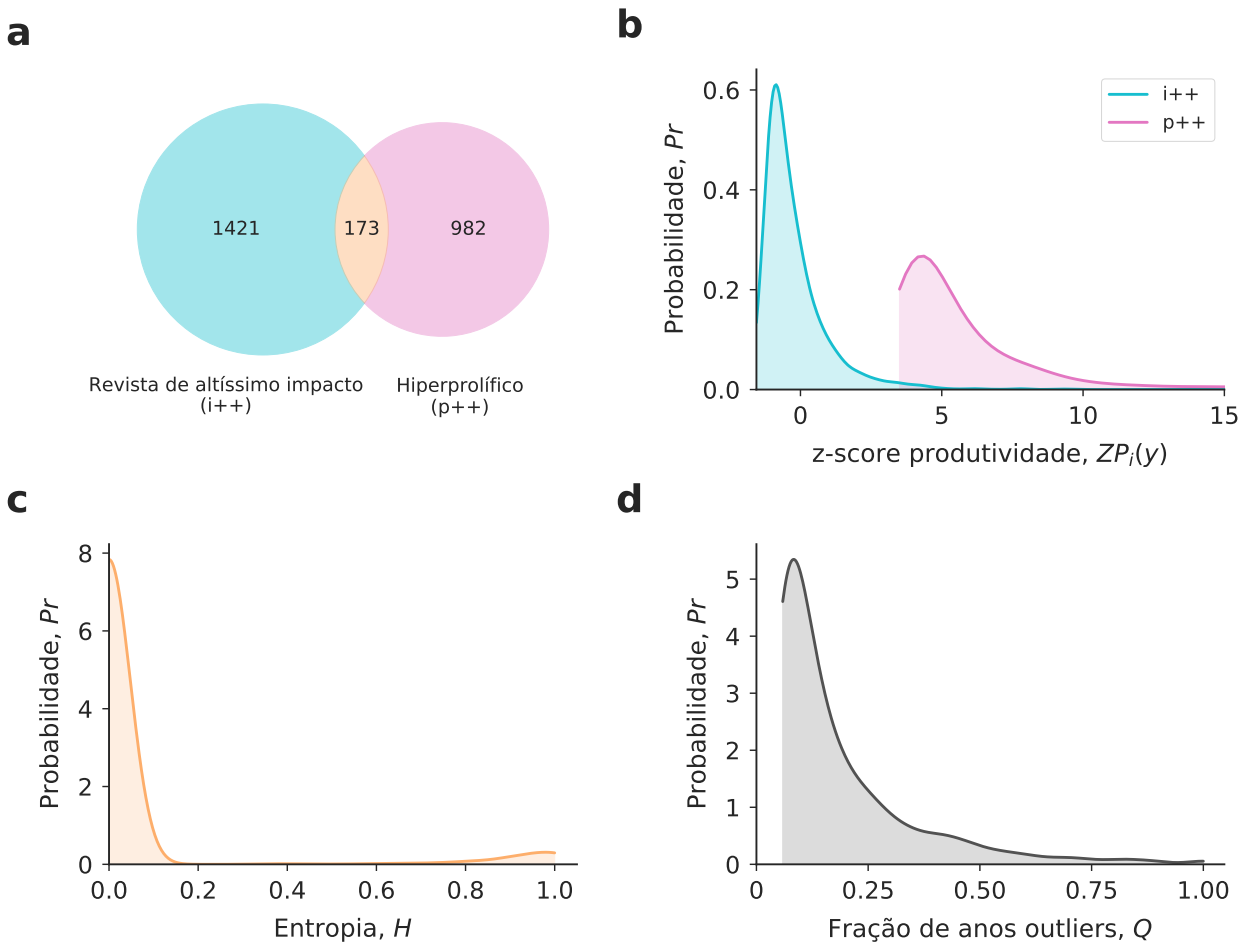

Figura C.8: Análise dos setores *outliers*. **(a)** Diagrama de Venn indicando quantos pesquisadores estão presentes em cada setor. A intersecção do diagrama significa que, no decorrer de sua carreira, o pesquisador esteve presente no setor *outlier* produtividade e no setor *outlier* impacto ao menos um ano. A maioria dos pesquisadores *outliers* está presente exclusivamente em um setor. **(b)** Distribuição de probabilidade do *z-score* produtividade para as duas categorias *outliers*. O gráfico mostra que a intersecção das distribuições de probabilidade é muito pequena. Além disso, notamos que a maioria dos anos *outliers* impacto apresenta produtividade baixa. Os *outliers* de ambos os quesitos não possuem produtividade com valor tão acentuado, ultrapassando o limiar com apenas algumas unidades de desvio padrão. **(c)** Distribuição de probabilidade da entropia normalizada dos pesquisadores que, no decorrer da carreira, foram *outliers* em ambos os quesitos. Como a maioria da densidade se concentra em valores pequenos de entropia, os pesquisadores apresentam um comportamento persistente ao longo das carreiras. **(d)** Distribuição de probabilidade da fração de anos *outliers* por pesquisador. Os anos *outliers* geralmente representam uma pequena parcela da carreira dos pesquisadores.

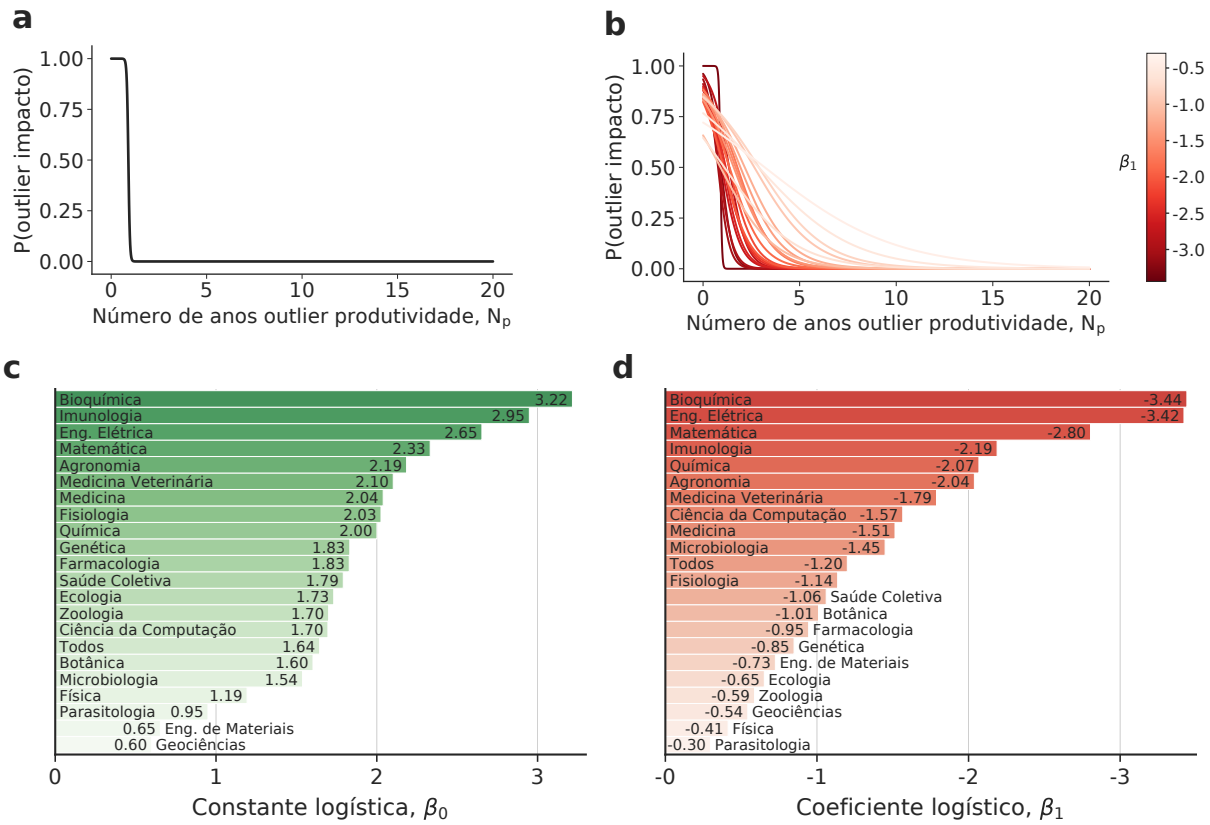

Figura C.9: Regressão logística para análise de *outliers*. **(a)** Regressão logística global agregando todas as áreas do conhecimento. **(b)** Regressão logística por área do conhecimento. A escala de cor vermelha representa diferentes valores de β_1. **(c)** Valores da constante β_0 para as diferentes áreas. **(d)** Valores do coeficiente β_1 para as diferentes áreas.

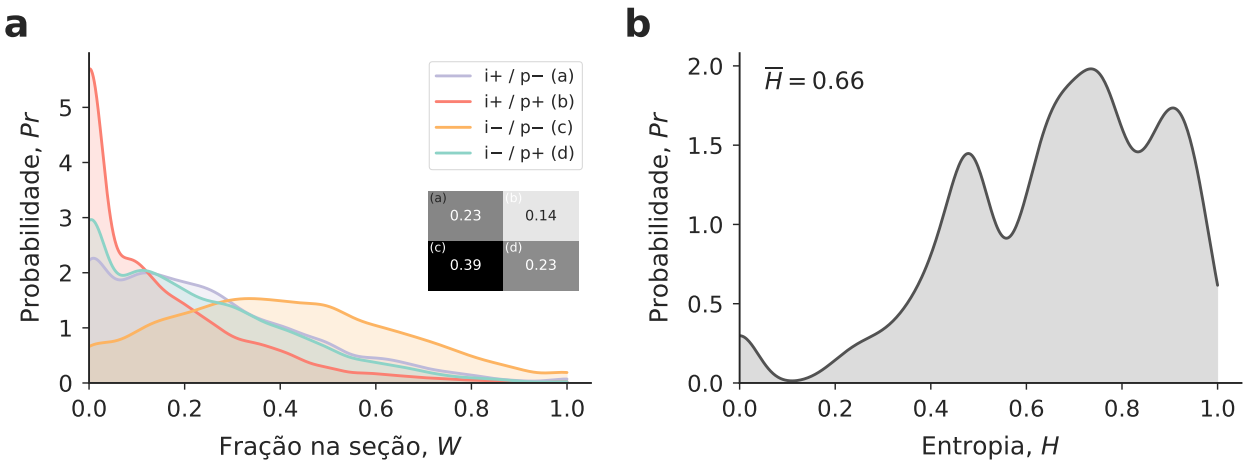

Figura C.10: Análise dos setores não-*outliers*. **(a)** Distribuição de probabilidade dos setores não-*outliers*. A letra i na legenda representa o impacto dos jornais em um ano e o símbolo p a produtividade. O símbolo $+$ indica que o pesquisador publicou acima da média naquele quesito. Por outro lado, o símbolo $-$ indica que o pesquisador publicou abaixo da média em determinada categoria. O *inset* mostra a fração dos anos em cada seção correspondente. **(b)** Distribuição de probabilidade da entropia normalizada da distribuição dos anos dos pesquisadores nas seções não-*outliers*. Os valores elevados indicam que os anos da carreira não estão concentrados em apenas uma seção, isto é, pesquisadores tendem a transitar entre as seções no decorrer de suas carreiras.

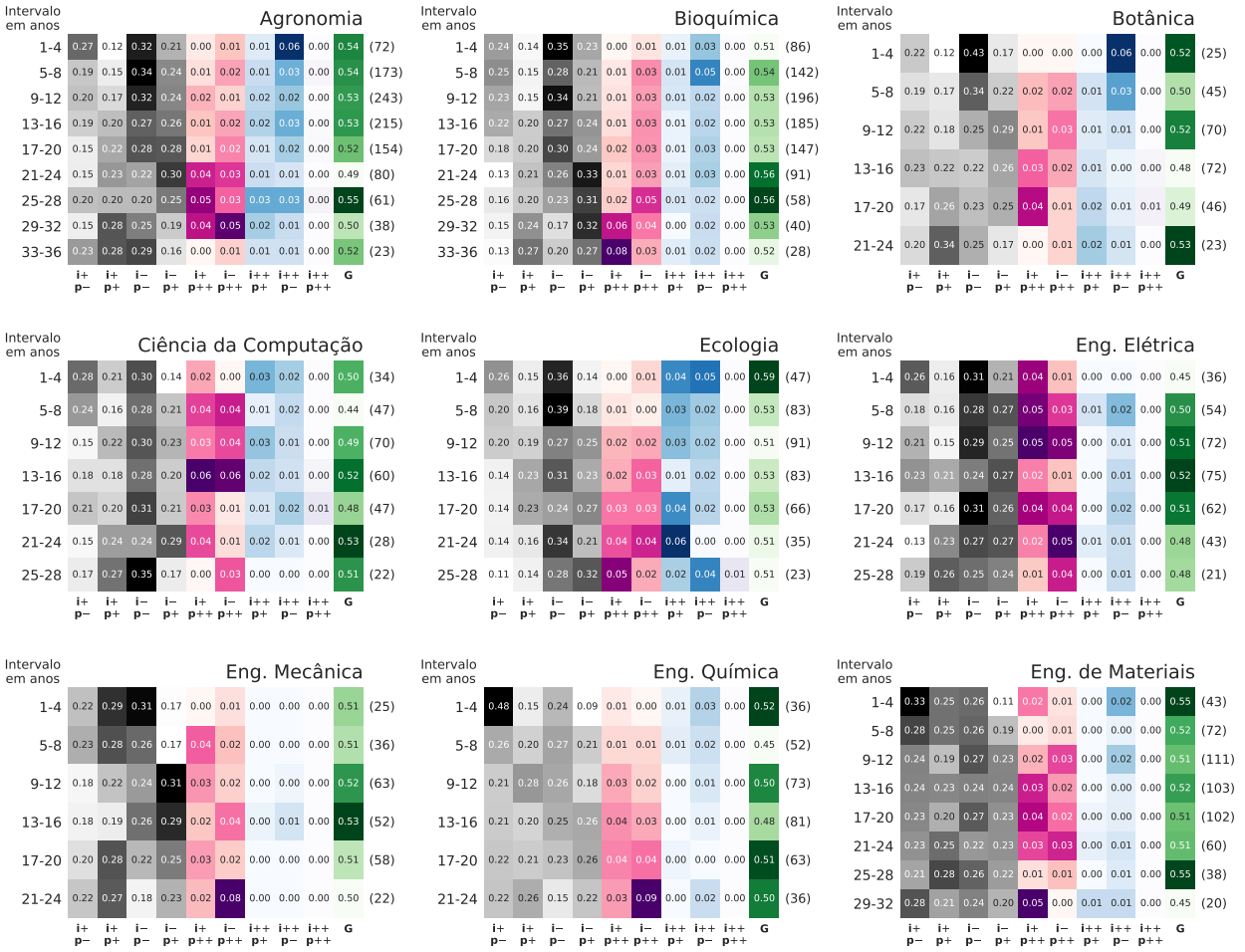

Figura C.11: Análise das frações dos setores impacto-produtividade ao longo das carreiras de pesquisadores de diferentes áreas (Parte 1). Dividimos a carreira dos pesquisadores em janelas de quatro anos, contando como primeiro ano a data de obtenção do título de doutor do pesquisador e, assim, calculamos a fração média de anos em cada seção para cada janela de tempo. As linhas representam os períodos da carreira do pesquisador em determinada área. As nove primeiras colunas representam frações médias em cada uma das seções e a última coluna é o coeficiente de Gini dos setores não-*outliers*. O número de pesquisadores em cada janela temporal é indicado entre parênteses ao final das linhas. As janelas temporais varrem um intervalo de tempo que é superior ao disponível em nossa base de dados (17 anos, 1999-2015) porque existe uma variedade de pesquisadores em épocas diferentes de suas carreiras.

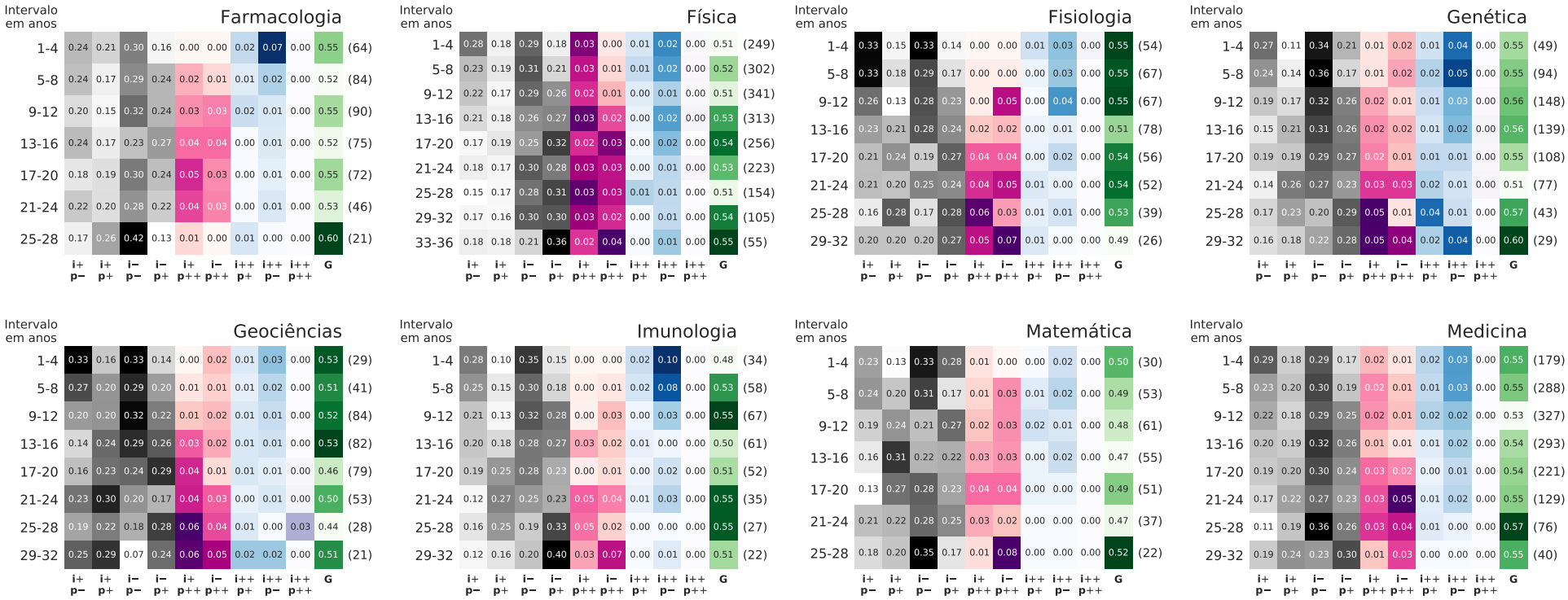

Figura C.12: Análise das frações dos setores impacto-produtividade ao longo das carreiras de pesquisadores de diferentes áreas (Parte 2). Dividimos a carreira dos pesquisadores em janelas de quatro anos, contando como primeiro ano a data de obtenção do título de doutor do pesquisador e, assim, calculamos a fração média de anos em cada seção para cada janela de tempo. As linhas representam os períodos da carreira do pesquisador em determinada área. As nove primeiras colunas representam frações médias em cada uma das seções e a última coluna é o coeficiente de Gini dos setores não-*outliers*. O número de pesquisadores em cada janela temporal é indicado entre parênteses ao final das linhas. As janelas temporais varrem um intervalo de tempo que é superior ao disponível em nossa base de dados (17 anos, 1999-2015) porque existe uma variedade de pesquisadores em épocas diferentes de suas carreiras.

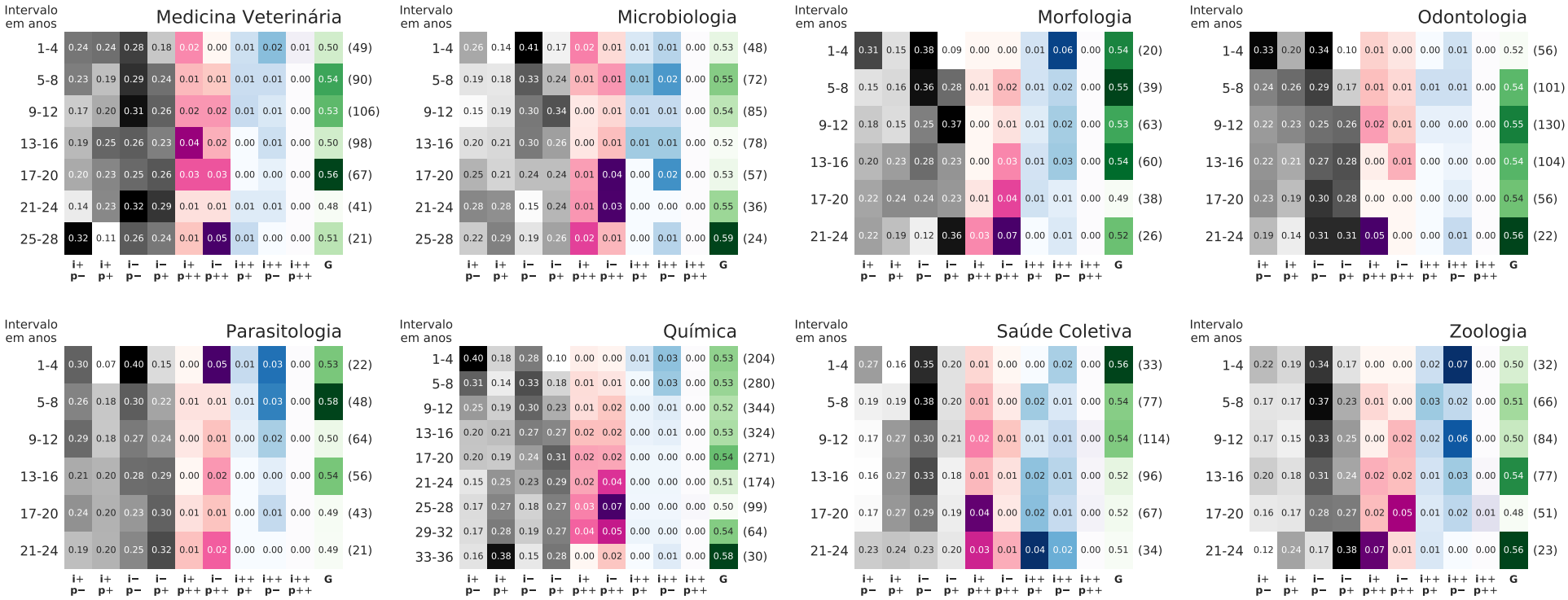

Figura C.13: Análise das frações dos setores impacto-produtividade ao longo das carreiras de pesquisadores de diferentes áreas (Parte 3). Dividimos a carreira dos pesquisadores em janelas de quatro anos, contando como primeiro ano a data de obtenção do título de doutor do pesquisador e, assim, calculamos a fração média de anos em cada seção para cada janela de tempo. As linhas representam os períodos da carreira do pesquisador em determinada área. As nove primeiras colunas representam frações médias em cada uma das seções e a última coluna é o coeficiente de Gini dos setores não-*outliers*. O número de pesquisadores em cada janela temporal é indicado entre parênteses ao final das linhas. As janelas temporais varrem um intervalo de tempo que é superior ao disponível em nossa base de dados (17 anos, 1999-2015) porque existe uma variedade de pesquisadores em épocas diferentes de suas carreiras.

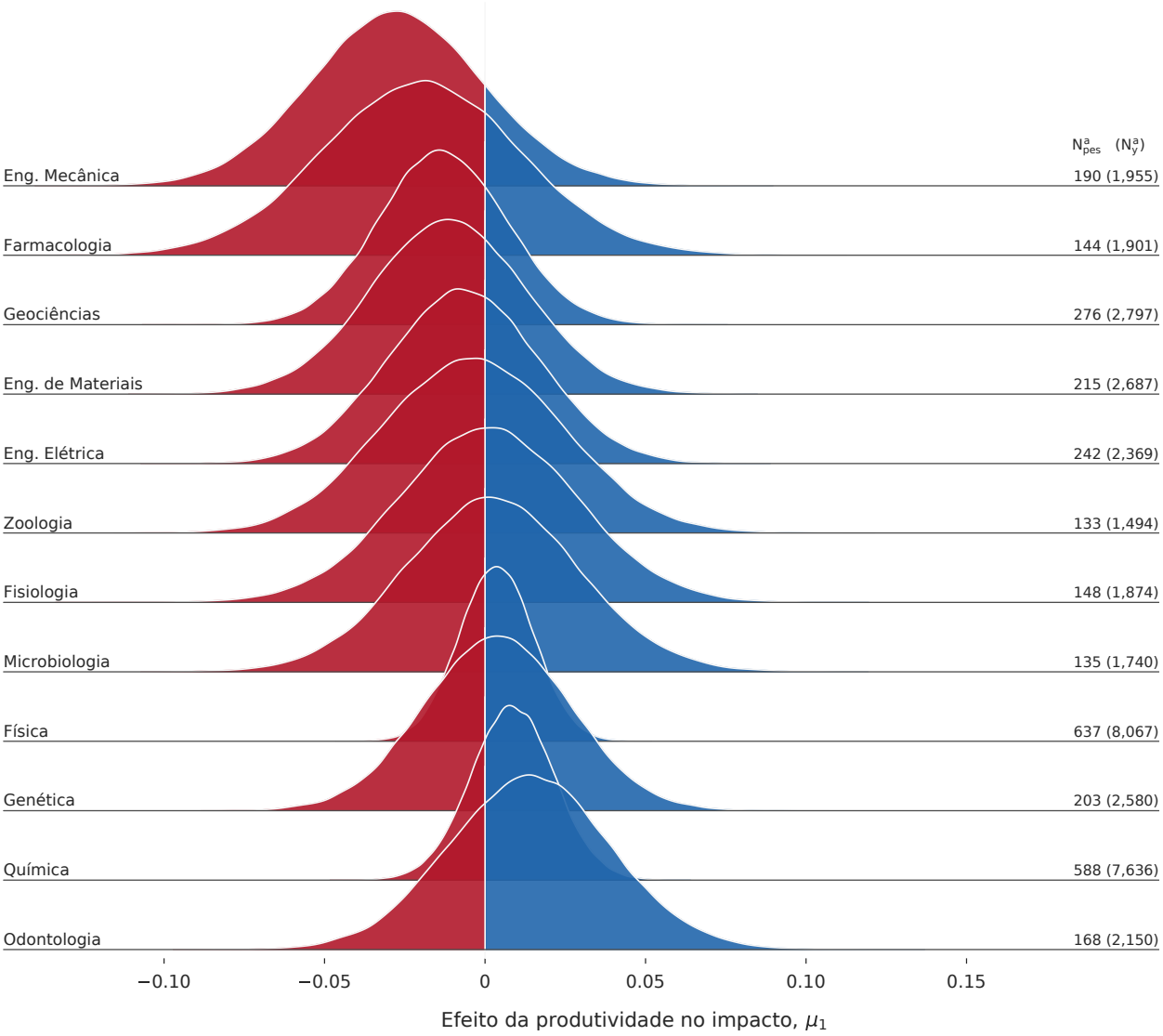

Figura C.14: Distribuições de probabilidade do parâmetro β_1 do modelo linear misto padronizado para cada disciplina (Parte 1). As distribuições de probabilidade marginais são obtidas integrando a distribuição a *posteriori* em relação aos demais parâmetros. À direita de cada distribuição, estão especificados o número de pesquisadores (N^a_{pes}) e, entre parênteses, o número de pontos (N^a_y) utilizados na realização da regressão para cada área a.

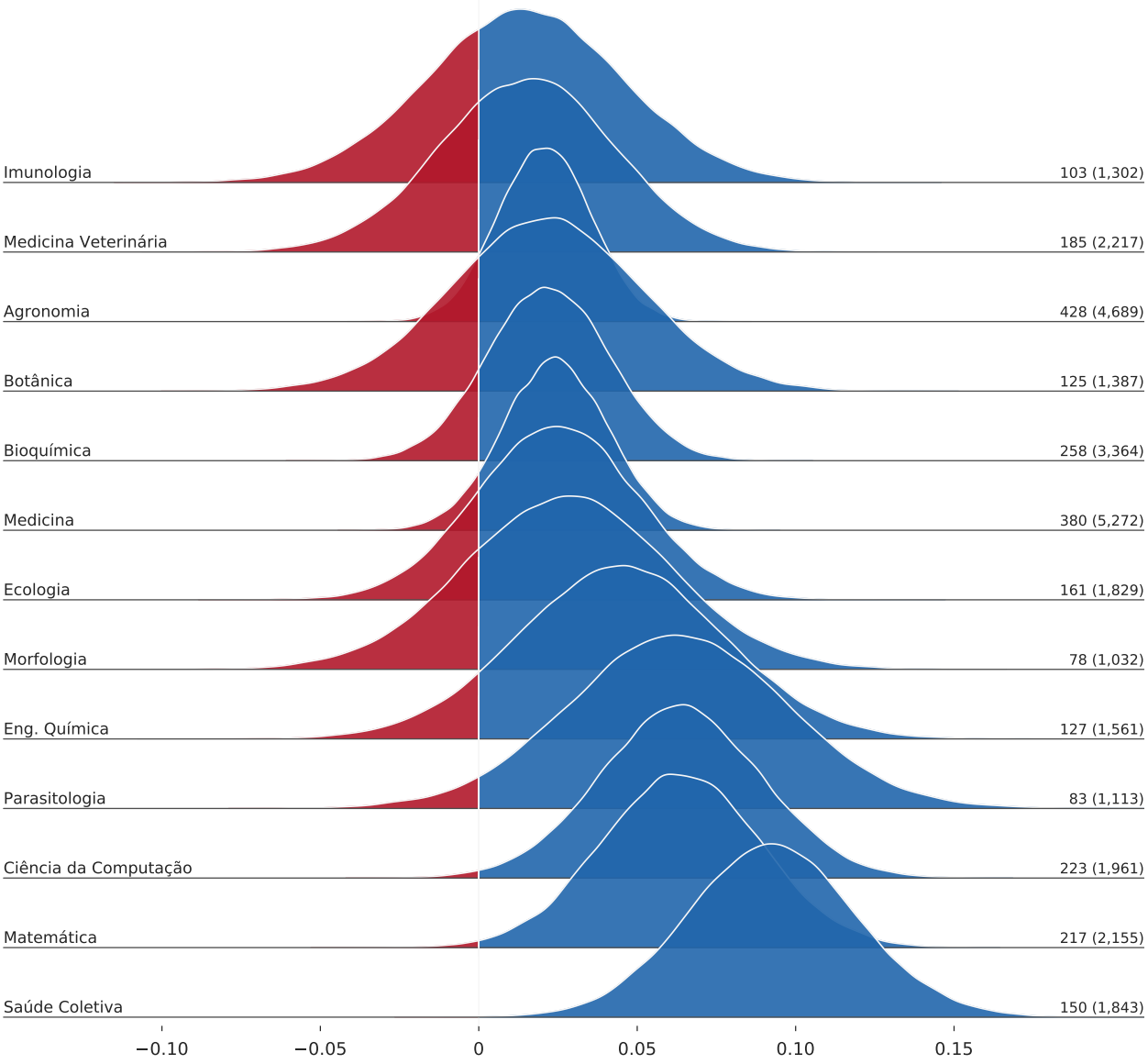

Figura C.15: Distribuições de probabilidade do parâmetro β_1 do modelo linear misto padronizado para cada disciplina (Parte 2). As distribuições de probabilidade marginais são obtidas integrando a distribuição a *posteriori* em relação aos demais parâmetros. À direita de cada distribuição, estão especificados o número de pesquisadores (N^a_{pes}) e, entre parênteses, o número de pontos (N^a_y) utilizados na realização da regressão para cada área a.

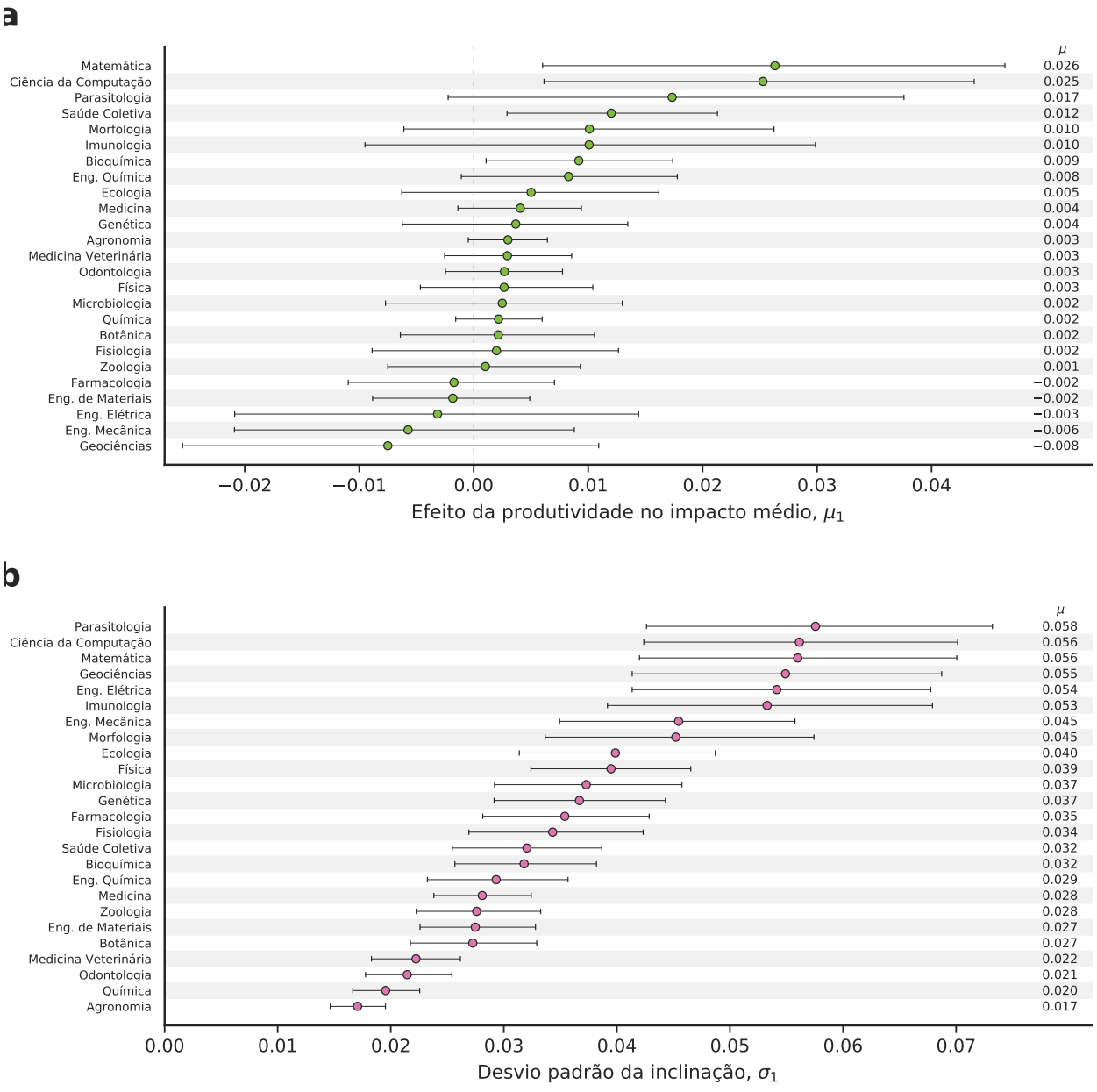

Figura C.16: Estimativa bayesiana dos parâmetros de localização e de escala da distribuição das inclinações β_1. Os painéis apresentam valores médios para **(a)** o valor médio μ_1 e **(b)** o valor do desvio padrão σ_1 do efeito da produtividade no impacto. As barras de erro ilustram a região de maior densidade da distribuição a *posteriori*.

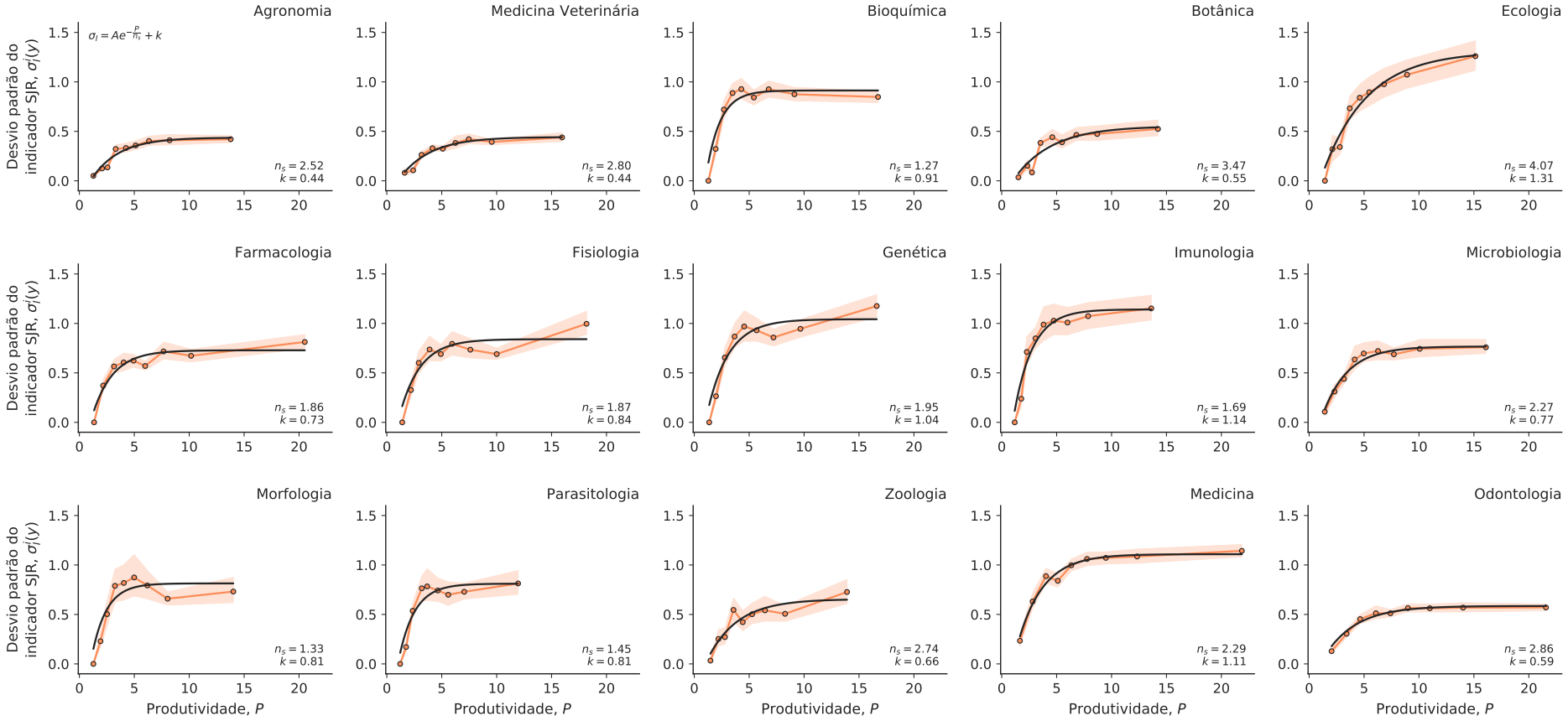

Figura C.17: Variabilidade de valores de indicador SJR (Parte 1) Realizamos um ajuste exponencial para entender a relação média entre o desvio padrão anual do indicador SJR $\sigma_I^i(y)$ e a produtividade P do pesquisador. O coeficiente de saturação n_s indica o quão rápido o desvio padrão satura com o aumento da produtividade. A constante de saturação k indica qual é o valor de saturação para grandes valores de produtividade.

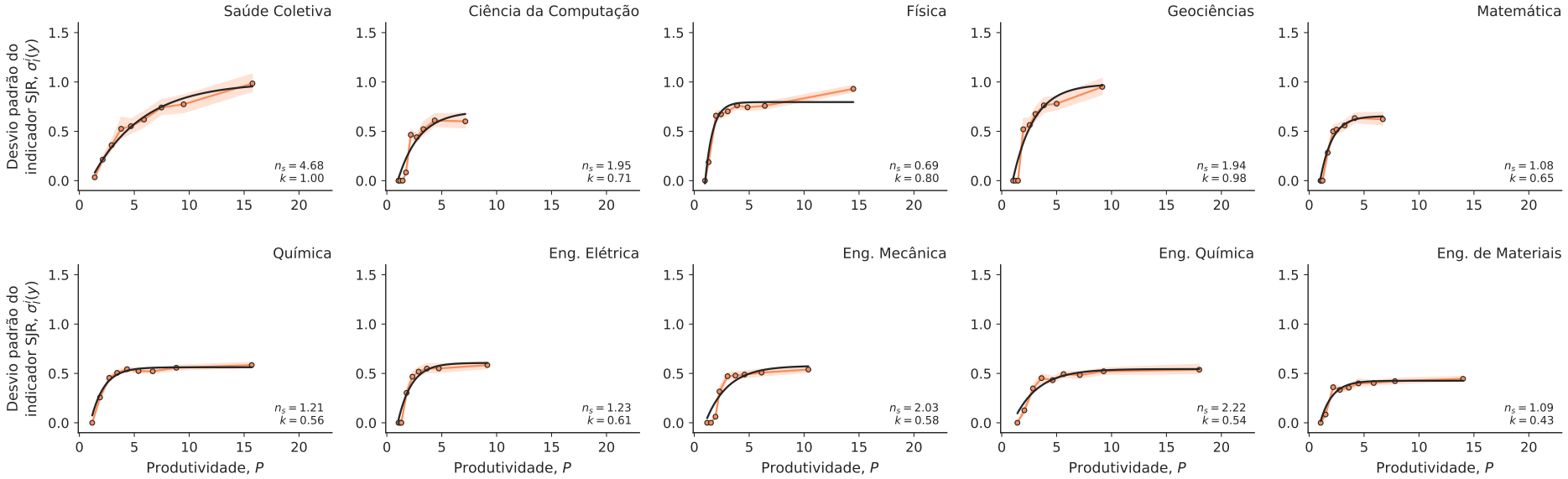

Figura C.18: Variabilidade de valores de indicador SJR (Parte 2) Realizamos um ajuste exponencial para entender a relação média entre o desvio padrão anual do indicador SJR $\sigma_I^i(y)$ e a produtividade P do pesquisador. O coeficiente de saturação n_s indica o quão rápido o desvio padrão satura com o aumento da produtividade. A constante de saturação k indica qual é o valor de saturação para grandes valores de produtividade.

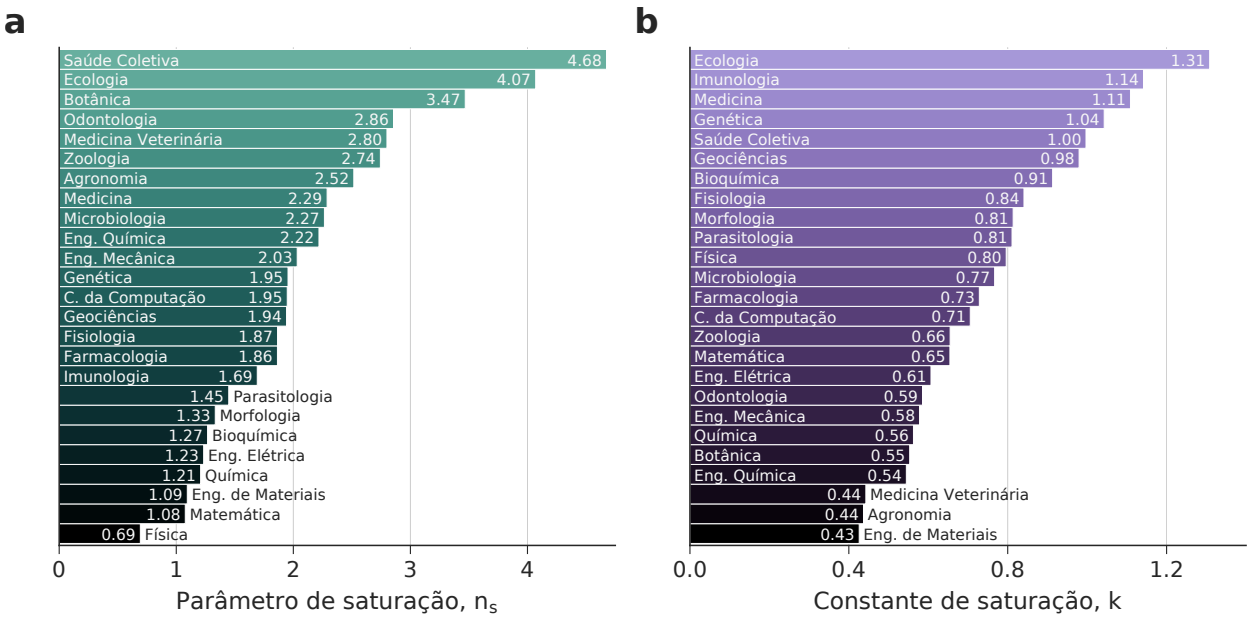

Figura C.19: Parâmetros do ajuste exponencial para diferentes disciplinas. **(a)** Coeficiente de saturação n_s. **(b)** Constante de saturação k.

Referências Bibliográficas

[1] Price, D. What is immunology? https://www.immunology.org/public-information/bitesized-immunology/special-topics/what-is-immunology. [Último acesso 08 de abril de 2019].

[2] Wheeler, W. M. The ant-colony as an organism. *Journal of Morphology* **22**, 307–325 (1911).

[3] Mitchell, M. *Complexity: A Guided Tour* (Oxford University Press, 2009).

[4] Flake, G. W. *The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems, and Adaptation* (MIT Press, 1998).

[5] Levin, S. A. Ecosystems and the biosphere as complex adaptive systems. *Ecosystems* **1**, 431–436 (1998).

[6] Sigaki, H. Y. D., de Souza, R. F., de Souza, R. T., Zola, R. S. & Ribeiro, H. V. Estimating physical properties from liquid crystal textures via machine learning and complexity-entropy methods. *Physical Review E* **99**, 013311 (2019).

[7] Sumpter, D. J. T. *Collective Animal Behavior* (Princeton University Press, 2010).

[8] Couzin, I. D. Collective cognition in animal groups. *Trends in Cognitive Sciences* **13**, 36–43 (2009).

[9] Alves, L. G. A., Ribeiro, H. V., Lenzi, E. K. & Mendes, R. S. Distance to the scaling law: A useful approach for unveiling relationships between crime and urban metrics. *PLOS ONE* **8**, e69580 (2013).

[10] Alves, L. G. A., Ribeiro, H. V. & Mendes, R. S. Scaling laws in the dynamics of crime growth rate. *Physica A: Statistical Mechanics and its Applications* **392**, 2672–2679 (2013).

[11] Sigaki, H. Y. D., Perc, M. & Ribeiro, H. V. History of art paintings through the lens of entropy and complexity. *Proceedings of the National Academy of Sciences* **115**, E8585–E8594 (2018).

[12] Mantegna, R. N. & Stanley., H. E. *Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, 1999).

[13] Sigaki, H. Y. D., Perc, M. & Ribeiro, H. V. Clustering patterns in efficiency and the coming-of-age of the cryptocurrency market. *Scientific Reports* **9**, 1440 (2019).

[14] Altmann, E. G. & Gerlach, M. Statistical Laws in Linguistics. In Esposti, M. D., Altmann, E. G. & Pachet, F. (eds.) *Creativity and Universality in Language*, 7–26 (Springer, 2016).

[15] Vieira, D. S., Picoli, S. & Mendes, R. S. Robustness of sentence length measures in written texts. *Physica A: Statistical Mechanics and its Applications* **506**, 749–754 (2018).

[16] Fortunato, S. *et al.* Science of science. *Science* **359**, eaao0185 (2018).

[17] Sinatra, R., Deville, P., Szell, M., Wang, D. & Barabási, A.-L. A century of physics. *Nature Physics* **11**, 791–796 (2015).

[18] de Solla Price, D. J. *Little Science, Big Science* (Columbia University Press, 1963).

[19] Milojević, S. Quantifying the cognitive extent of science. *Journal of Informetrics* **9**, 962–973 (2015).

[20] Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).

[21] Council, N., Education, D., Board on Behavioral, C., Science, C., Hilton, M. & Cooke, N. *Enhancing the Effectiveness of Team Science* (National Academies Press, 2015).

[22] Shwed, U. & Bearman, P. S. The temporal structure of scientific consensus formation. *American Sociological Review* **75**, 817–840 (2010).

[23] Bruggeman, J., Traag, V. A. & Uitermark, J. Detecting communities through network data. *American Sociological Review* **77**, 1050–1063 (2012).

[24] Shi, F., Foster, J. G. & Evans, J. A. Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks* **43**, 73–85 (2015).

[25] Foster, J. G., Rzhetsky, A. & Evans, J. A. Tradition and innovation in scientists' research strategies. *American Sociological Review* **80**, 875–908 (2015).

[26] Wu, L., Wang, D. & Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature* **566**, 378 (2019).

[27] Bromham, L., Dinnage, R. & Hua, X. Interdisciplinary research has consistently lower funding success. *Nature* **534**, 684–687 (2016).

[28] Yegros-Yegros, A., Rafols, I. & D'Este, P. Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity. *PLOS ONE* **10**, 1–21 (2015).

[29] Boudreau, K. J., Guinan, E. C., Lakhani, K. R. & Riedl, C. Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science* **62**, 2765–2783 (2016).

[30] Azoulay, P., Zivin, J. S. G. & Manso, G. Incentives and creativity: Evidence from the academic life sciences. *National Bureau of Economic Research* (2009).

[31] Petersen, A. M., Riccaboni, M., Stanley, H. E. & Pammolli, F. Persistence and uncertainty in the academic career. *Proceedings of the National Academy of Sciences* **109**, 5213–5218 (2012).

[32] Morgan, A. C., Economou, D. J., Way, S. F. & Clauset, A. Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Science* **7**, 40 (2018).

[33] Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C. R. Bibliometrics: Global gender disparities in science. *Nature News* **504**, 211 (2013).

[34] Duch, J., Zeng, X. H. T., Sales-Pardo, M., Radicchi, F., Otis, S., Woodruff, T. K. & Amaral, L. A. N. The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLOS ONE* **7**, e51332 (2012).

[35] West, J. D., Jacquet, J., King, M. M., Correll, S. J. & Bergstrom, C. T. The role of gender in scholarly authorship. *PLOS ONE* **8**, e66212 (2013).

[36] Zeng, X. H. T., Duch, J., Sales-Pardo, M., Moreira, J. A. G., Radicchi, F., Ribeiro, H. V., Woodruff, T. K. & Amaral, L. A. N. Differences in collaboration patterns across discipline, career stage, and gender. *PLOS Biology* **14**, e1002573 (2016).

[37] Ley, T. J. & Hamilton, B. H. The gender gap in NIH grant applications. *Science* **322**, 1472–1474 (2008).

[38] Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J. & Handelsman, J. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences* **109**, 16474–16479 (2012).

[39] Schellekens, M. H., Holstege, F. & Yasseri, T. Female scholars need to achieve more for equal public recognition. *arXiv: 1904.06310* (2019).

[40] de Solla Price, D. J. Networks of scientific papers. *Science* **149**, 510–515 (1965).

[41] de Solla Price, D. J. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* **27**, 292–306 (1976).

[42] Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

[43] Perc, M. The Matthew effect in empirical data. *Journal of The Royal Society Interface* **11**, 20140378 (2014).

[44] Petersen, A. M., Fortunato, S., Pan, R. K., Kaski, K., Penner, O., Rungi, A., Riccaboni, M., Stanley, H. E. & Pammolli, F. Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences* **111**, 15316–15321 (2014).

[45] Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K. & Fortunato, S. Attention decay in science. *Journal of Informetrics* **9**, 734–745 (2015).

[46] Eom, Y.-H. & Fortunato, S. Characterizing and modeling citation dynamics. *PLOS ONE* **6**, e24926 (2011).

[47] Wang, D., Song, C. & Barabási, A.-L. Quantifying long-term scientific impact. *Science* **342**, 127–132 (2013).

[48] van Raan, A. F. J. Sleeping Beauties in science. *Scientometrics* **59**, 467–472 (2004).

[49] Ke, Q., Ferrara, E., Radicchi, F. & Flammini, A. Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences* **112**, 7426–7431 (2015).

[50] Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. Quantifying the evolution of individual scientific impact. *Science* **354**, aaf5239 (2016).

[51] Janosov, M., Battiston, F. & Sinatra, R. Success and luck in creative careers. *arXiv: 1909.07956* (2019).

[52] Kutner, M. H. *Applied Linear Statistical Models* (McGraw-Hill Irwin, 2005).

[53] Rencher, A. C. & Schaalje, G. B. *Linear Models in Statistics* (Wiley, 2008).

[54] Myung, I. J. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology* **47**, 90–100 (2003).

[55] Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A. & Smith, G. M. *Mixed Effects Models and Extensions in Ecology with R* (Springer, 2011).

[56] Agresti, A. *Categorical Data Analysis* (Wiley, 2003).

[57] Unpingco, J. *Python for Probability, Statistics, and Machine Learning* (Springer, 2016).

[58] Hosmer, D. W. & Lemeshow, S. *Applied Logistic Regression* (Wiley, 2004).

[59] Seabold, S. & Perktold, J. statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference* (2010).

[60] OECD. Gender wage gap. https://www.oecd-ilibrary.org/content/data/7cee77aa-en. [Último acesso 19 de fevereiro de 2020].

[61] Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics* **38**, 963–974 (1982).

[62] Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., Robinson, B. S., Hodgson, D. J. & Inger, R. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* **6**, e4794 (2018).

[63] Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).

[64] Pinheiro, J. & Bates, D. *Mixed-Effects Models in S and S-PLUS* (Springer-Verlag, 2000).

[65] Lambert, B. *A Student's Guide to Bayesian Statistics* (SAGE, 2018).

[66] Downey, A. B. *Think Bayes: Bayesian Statistics in Python* (O'Reilly, 2013).

[67] Amrhein, V., Greenland, S. & McShane, B. Scientists rise up against statistical significance. *Nature* **567**, 305–307 (2019).

[68] Robert, C. & Casella, G. A short history of Markov Chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science* **26**, 102–115 (2011).

[69] Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* **2**, e55 (2016).

[70] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092 (1953).

[71] Geman, S. & Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 721–741 (1984).

[72] OpenBUGS. http://www.openbugs.net/. [Último acesso 15 de janeiro de 2020].

[73] Betancourt, M. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv: 1701.02434* (2017).

[74] Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. Hybrid Monte Carlo. *Physics Letters B* **195**, 216–222 (1987).

[75] Neal, R. M. MCMC using Hamiltonian dynamics. *arXiv: 1206.1901* (2012).

[76] Monnahan, C. C., Thorson, J. T. & Branch, T. A. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution* **8**, 339–348 (2017).

[77] Eastwood, J. W. & Hockney, R. W. *Computer Simulation Using Particles* (A. Hilger, 1988).

[78] Homan, M. D. & Gelman, A. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593–1623 (2014).

[79] Pereyra, M., Schniter, P., Chouzenoux, E., Pesquet, J., Tourneret, J., Hero, A. O. & McLaughlin, S. A survey of stochastic simulation and optimization methods in signal processing. *IEEE Journal of Selected Topics in Signal Processing* **10**, 224–241 (2016).

[80] Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472 (1992).

[81] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. *Bayesian Data Analysis* (Taylor & Francis, 2013).

[82] Murphy, K. P. *Machine Learning: A Probabilistic Perspective* (MIT Press, 2012).

[83] Huber, P. J. *Robust Statistics* (Wiley, 2004).

[84] Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* (Springer, 2003).

[85] Huber, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35**, 73–101 (1964).

[86] Staudte, R. G. & Sheather, S. J. *Robust Estimation and Testing* (Wiley, 1990).

[87] Süli, E. & Mayers, D. F. *An Introduction to Numerical Analysis* (Cambridge University Press, 2003).

[88] Weinberg, B. H. Indexes and religion: Reflections on research in the history of indexes. *The Indexer* **21**, 8 (1999).

[89] Garfield, E. Citation indexes for science: A new dimension in documentation through association of ideas. *Science* **122**, 108–111 (1955).

[90] Smith, D. R. Impact factors, scientometrics and the history of citation-based research. *Scientometrics* **92**, 419–427 (2012).

[91] Garfield, E. "Science Citation Index" - A new dimension in indexing. *Science* **144**, 649–654 (1964).

[92] Web of Science: Science Citation Index Expanded. https://clarivate.com/webofsciencegroup/solutions/webofscience-scie/. [Último acesso 12 de dezembro de 2019].

[93] Garfield, E. Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science* **178**, 471–479 (1972).

[94] Garfield, E. The history and meaning of the journal impact factor. *Journal of the American Medical Association* **295**, 90–93 (2006).

[95] Web of Science: Overview. https://clarivate.com/webofsciencegroup/solutions/web-of-science/. [Último acesso 12 de dezembro de 2019].

[96] Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank citation ranking: Bringing order to the web. *Stanford InfoLab* (1999).

[97] Wang, K., Shen, Z., Huang, C., Wu, C.-H., Eide, D., Dong, Y., Qian, J., Kanakia, A., Chen, A. & Rogahn, R. A review of Microsoft Academic Services for science of science studies. *Frontiers in Big Data* **2**, 45 (2019).

[98] Guerrero-Bote, V. P. & Moya-Anegón, F. A further step forward in measuring journals' scientific prestige: The SJR2 indicator. *Journal of Informetrics* **6**, 674–688 (2012).

[99] Martín-Martín, A., Orduna-Malea, E., Thelwall, M. & López-Cózar, E. D. Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics* **12**, 1160–1177 (2018).

[100] Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., june Paul Hsu, B. & Wang, K. An overview of Microsoft Academic Service and applications. In *Proceedings of the 24th International Conference on World Wide Web*, 243–246 (2015).

[101] Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R. & Karageorgopoulos, D. E. Comparison of SCImago journal rank indicator with journal impact factor. *The FASEB Journal* **22**, 2623–2628 (2008).

[102] Harzing, A.-W. & Alakangas, S. Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics* **106**, 787–804 (2016).

[103] Mongeon, P. & Paul-Hus, A. The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics* **106**, 213–228 (2016).

[104] Seglen, P. O. Why the impact factor of journals should not be used for evaluating research. *BMJ* **314**, 497 (1997).

[105] PLoS Medicine Editors. The impact factor game. *PLOS Medicine* **3**, e291 (2006).

[106] San Francisco Declaration on Research Assessment. https://sfdora.org/read/. [Último acesso 15 de dezembro de 2019].

[107] Larivière, V. & Sugimoto, C. R. *The Journal Impact Factor: A Brief History, Critique, and Discussion of Adverse Effects*, 3–24 (Springer, 2019).

[108] Kim, L., Portenoy, J. H., West, J. D. & Stovel, K. W. Scientific journals still matter in the era of academic search engines and preprint archives. *Journal of the Association for Information Science and Technology* (2019).

[109] Traag, V. A. Inferring the causal effect of journals on citations. *arXiv: 1912.08648* (2019).

[110] Abramo, G., D'Angelo, C. A. & Di Costa, F. Citations versus journal impact factor as proxy of quality: could the latter ever be preferable? *Scientometrics* **84**, 821–833 (2010).

[111] Waltman, L. & Traag, V. A. Use of the journal impact factor for assessing individual articles need not be wrong. *arXiv: 1703.02334* (2017).

[112] Plataforma Lattes. http://lattes.cnpq.br/. [Último acesso 12 de dezembro de 2019].

[113] Araújo, E. B., Moreira, A. A., Furtado, V., Pequeno, T. H. & Andrade Jr, J. S. Collaboration networks from a large CV database: Dynamics, topology and bonus impact. *PLOS ONE* **9**, e90537 (2014).

[114] Araújo, E. B., Araújo, N. A., Moreira, A. A., Herrmann, H. J. & Andrade Jr, J. S. Gender differences in scientific collaborations: Women are more egalitarian than men. *PLOS ONE* **12**, e0176791 (2017).

[115] Way, S. F., Morgan, A. C., Clauset, A. & Larremore, D. B. The misleading narrative of the canonical faculty productivity trajectory. *Proceedings of the National Academy of Sciences* **114**, E9216–E9223 (2017).

[116] Chamada CNPq N° 06/2019 - Bolsas de Produtividade em Pesquisa. http://memoria.cnpq.br/chamadas-publicas?p_p_id=resultadosportlet_WAR_resultadoscnpqportlet_INSTANCE_0ZaM&filtro=encerradas&detalha=chamadaDivulgada&idDivulgacao=8722. [Último acesso 12 de dezembro de 2019].

[117] The ElementTree XML API. https://docs.python.org/3/library/xml.etree.elementtree.html. [Último acesso 18 de fevereiro de 2020].

[118] Althouse, B. M., West, J. D., Bergstrom, C. T. & Bergstrom, T. Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology* **60**, 27–34 (2009).

[119] Iglewicz, B. & Hoaglin, D. C. *How to Detect and Handle Outliers* (ASQC Quality Press, 1993).

[120] Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423 (1948).

[121] Gini, C. Measurement of inequality of incomes. *The Economic Journal* **31**, 124–126 (1921).

[122] StatsDirect. Gini Coefficient of Inequality. https://www.statsdirect.com/help/default.htm#nonparametric_methods/gini.htm. [Último acesso 14 de fevereiro de 2020].

[123] Gelman, A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 1–19 (2006).

[124] Hogg, R. V., McKean, J. W. & Craig, A. T. *Introduction to Mathematical Statistics* (Pearson, 2005).