UNIVERSIDADE ESTADUAL DE MARINGÁ

Leonardo Oliveira Mendes

Competição, memória e popularidade em jogos: perspectivas da Física de Sistemas Complexos

Maringá

2022

Leonardo Oliveira Mendes

Competição, memória e popularidade em jogos: perspectivas da Física de Sistemas Complexos

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Física da Universidade Estadual de Maringá como requisito parcial para a obtenção do Grau de Doutor em Física. Área de concentração: Sistemas Complexos.

Orientador:
Prof. Dr. Renio dos Santos Mendes

Coorientador:
Prof. Dr. Max Javier Jáuregui Rodríguez

Maringá

2022

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

M538c

Mendes, Leonardo Oliveira

Competição, memória e popularidade em jogos : perspectivas da Física de Sistemas Complexos / Leonardo Oliveira Mendes. -- Maringá, PR, 2022.

183 f.: il. color., figs., tabs.

Orientador: Prof. Dr. Renio dos Santos Mendes.

Coorientador: Prof. Dr. Max Javier Jáuregui Rodríguez.

Tese (Doutorado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Física, Programa de Pós-Graduação em Física, 2022.

1. Video games. 2. Memória coletiva. 3. Jogos de computador. 4. Sistemas complexos - Física. I. Mendes, Renio dos Santos, orient. II. Rodríguez, Max Javier Jáuregui, coorient. III. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Física. Programa de Pós-Graduação em Física. IV. Título.

CDD 23.ed. 530.1

Rosana de Souza Costa de Oliveira - 9/1366

LEONARDO OLIVEIRA MENDES

COMPETIÇÃO, MEMÓRIA E POPULARIDADE EM JOGOS: PERSPECTIVAS DA FÍSICA DE SISTEMAS COMPLEXOS

Tese apresentada à Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de doutor.

Aprovado em: Maringá, 29 de junho de 2022.

BANCA EXAMINADORA

_____
Prof. Dr. Renio dos Santos Mendes
Universidade Estadual de Maringá – UEM

_____
Prof. Dr. Ervin Kaminski Lenzi
Universidade Estadual de Ponta Grossa – UEPG

_____
Prof. Dr. Luciano Rodrigues da Silva
Universidade Federal do Rio Grande do Norte – UFRN

_____
Prof. Dr. Breno Ferraz de Oliveira
Universidade Estadual de Maringá – UEM

_____
Prof. Dr. Sérgio de Picoli Júnior
Universidade Estadual de Maringá – UEM

Resumo

Descrever a permanência de objetos culturais é um passo importante para o entendimento das tendências de uma sociedade. Um objeto cultural relativamente novo é o *video game*, mídia interativa em que o jogador é um colaborador ativo. Nesta tese, investigamos a permanência de *video games* na memória coletiva, usando sua popularidade como métrica e empregando dados baseados na plataforma *Steam* de julho 2012 até dezembro 2020. Alguns de seus objetivos foram: caracterizar o banco de dados; estudar o crescimento do número de jogadores, de jogos e de categorias de jogos; fornecer um modelo para a distribuição de sua popularidade relativa e aplicar esse modelo em três estratos, quais sejam, global, categorias principais e entre categorias. Detectamos tendências lineares de crescimento no número de jogadores e no número de categorias e uma exponencial no número de jogos lançados. Além disso, verificamos que distribuições lognormais, decorrentes de processos multiplicativos, fornecem uma primeira aproximação para a popularidade relativa. Finalmente, propusemos uma melhoria via transformações Box-Cox e, interpretando a magnitude de cada parâmetro Box-Cox como uma medida de efeitos de memória, conseguimos justificar esse modelo aprimorado.

Primeiramente, considerando sua pertinência ao longo da tese, apresentamos aspectos da formulação de Assmann para o conceito de memória, da análise de crescimento e das transformações de potência, com ênfase na transformação Box-Cox. Em segundo lugar, discorremos acerca do material utilizado na tese, isto é, explicamos a plataforma de jogos escolhida, apresentamos os programas desenvolvidos para a aquisição e a análise dos dados e oferecemos uma síntese da base de dados. A seguir, expomos as hipóteses centrais da tese e os resultados relacionados ao primeiro estrato, o global, além de uma descrição quantitativa dos padrões de crescimento. A etapa seguinte é dedicada ao segundo estrato, à investigação da popularidade dos jogos em categorias e a um estudo do número de categorias ao longo do tempo e sua distribuição. Na análise dos resultados, enfatizamos o modelo malthusiano de crescimento e o modelo de memória coletiva. Finalmente, elaboramos parágrafos referentes às conclusões e às perspectivas futuras deste trabalho. Materiais complementares foram disponibilizados como apêndices.

Palavras-chave: Video games; Popularidade; Processos multiplicativos; memória.

Abstract

Describing the permanence of cultural objects is an important step in understanding trends in a society. A relatively new cultural object is the *video game*, being an interactive medium in which the player is an active contributor. This thesis investigated the permanence of *video games* in collective memory using their popularity as a metric, utilizing data based on the *Steam* platform from July 2012 to December 2020. Some of the objectives were to characterize the database, study the growth of players, games, and game categories, provide a model for the relative popularity distribution and apply this model in three strata, global, main categories, and across categories. We detected linear growth trends in the number of players and the number of categories, and exponential growth trends in the number of games released. Furthermore, we found that lognormal distributions, arising from multiplicative processes, provide a first approximation to relative popularity. Finally, we proposed an improvement via Box-Cox transformations and were able to justify this improved model by interpreting the magnitude of each Box-Cox parameter as a measure of memory effects.

Firstly, we present aspects, relevant throughout the thesis, of Assmann's formulation for the concept of memory, of growth analysis, and of power transformations, with emphasis on the Box-Cox transformation. Secondly, we discuss the material used for the thesis, that is, we explain the chosen game platform, we supply the programs developed for the data acquisition and analysis and we offer a synthesis of the database. Next, we expose the central hypotheses of the thesis and the results in the first stratum, the global one, in addition to showing a quantitative description of the growth patterns. We dedicate the fourth stage to the second stratum, investigating the popularity of games in categories and a study of the number of categories over time and their distribution. We also provide an analysis of the results, with emphasis on the Malthusian model of growth and on the model of collective memory. We also offer paragraphs focused on the conclusions and future perspectives of this work. Finally, we present supplementary materials as appendices.

Keywords: Video games; Popularity; Multiplicative processes; Memory.

Lista de Figuras

1 **Esquema de propagação de diferentes modos de memória.** Esquema de propagação de memória adaptado de (SONG et al., 2021). Em primeiro lugar, o sujeito tem uma experiência (em azul na figura) e a relata para outras pessoas (em laranja na figura). Além destas, outras pessoas podem participar indiretamente da experiência, por exemplo, lendo um livro (em amarelo na figura). Note-se que um mesmo sujeito pode participar de dois ou mais modos de memória simultaneamente e esses modos possivelmente terão mecanismos de esquecimento distintos. 23

2 **Taxonomia dos diferentes modos de memória.** A categoria mais ampla, memória coletiva, engloba modos de memória como a familiar ou a social. Seu recorte temporal é longo, podendo até mesmo ser mitológico (a exemplo de mitos de formação de um povo). Já a memória cultural está ligada com a produção de objetos ou lugares que servem para cristalizar as memórias de um grupo ou povo, como museus, livros, filmes, etc. Finalmente, passamos para a memória individual, subjetiva e, muitas vezes, atemporal, para as lembranças e a percepção de si. Nesta última categoria, notamos a consciência da memória, isto é, o sujeito não só guarda memórias passivamente, mas tem consciência do que se recorda. 24

3 **Esquema do modelo malthusiano de recursos e demanda.** A teoria de Malthus identifica um crescimento aritmético (linear) para os recursos, enquanto identifica um crescimento geométrico (exponencial) para a demanda. Utilizamos os termos "recursos" e "demanda" por ser mais gerais, embora a teoria malthusiana tenha sido inicialmente aplicada à dinâmica de populações humanas. Inicialmente, há recursos suficientes para atender a toda a demanda (área verde clara, período de abundância). Com o crescimento da demanda, chega-se a um tempo crítico em que os recursos já não são mais suficientes e há disputa (área laranja, período de escassez). Em nosso contexto, identificamos os recursos com os jogadores e a demanda com os jogos, que competem pela atenção dos usuários. 25

4 **Evolução do número de jogadores.** Evolução temporal mês-a-mês do número médio de jogadores ativos por hora na *Steam*. Os discos roxos representam o dado empírico enquanto o ajuste linear (linha vermelha), obtido por meio do método dos mínimos quadrados, é governado pela equação $N = 441\,618 + 36\,690\,m$. 42

5 **Evolução do número de jogos lançados.** Evolução temporal mês-a-mês dos jogos lançadas na *Steam*. Os discos verdes representam o dado empírico enquanto o ajuste exponencial (linha vermelha na escala log), obtido por meio do método dos mínimos quadrados, é dado por $\ln N = 1{,}90 + 0{,}011\,m$. 43

6 **Médias dos grupos trimestrais.** (**A**) Média do número médio de jogadores por hora em cada mês (x_i) de todos os jogos de cada grupo trimestral ao longo dos meses após o lançamento. (**B**) Usando a variável logarítmica ($y_i = \log x_i$), apresentamos a dependência no tempo das médias correspondentes. Cada um dos trinta e quatro grupos trimestrais tem sua própria cor. 45

7 **Desvios-padrão dos grupos trimestrais.** (**A**) Para esses grupos, temos a evolução temporal de seus desvios-padrão. (**B**) Usando a variável logarítmica ($y_i = \log x_i$), apresentamos a dependência no tempo dos respectivos desvios-padrão. Cada um dos trinta e quatro grupos trimestrais tem sua própria cor. 46

8 **Análise Box-Cox do grupo trimestral 01/2015.** (**A**) Função de distribuição de probabilidade da variável logarítmica normalizada $z_i = (y_i - \langle y_i \rangle)/\sigma_y$ com $y_i = \log x_i$ para o grupo trimestral 01/2015, em que x_i é o número médio de usuários por hora em cada mês do i-ésimo jogo. (**B**) O mesmo que (**A**), mas com o emprego do valor ótimo λ ($\lambda_{opt} = -0{,}07$) da transformação Box-Cox, equação (3.6). As linhas tracejadas referem-se a uma distribuição normal com a média zero e o desvio-padrão unitário. . . 47

9 **Evolução do parâmetro λ ótimo para o grupo trimestral 01/2015.** Evolução temporal do parâmetro ótimo λ, λ_{opt}, do grupo trimestral ao longo dos vinte e três trimestres. Os discos representam os λ_{opt}'s ao longo da evolução do tempo do grupo trimestral 01/2015 e a linha reta é a média de todos os trimestres, $\langle \lambda_{opt} \rangle = -0{,}05$, que se aproxima do valor obtido com λ_{opt} calculado por meio do método de máxima verossimilhança e empregado na figura 8B. As barras de erro referem-se ao intervalo de confiança de 95% dos λ_{opt}'s estimados por meio do método de máxima verossimilhança. . . . 48

10 **Análise Box-Cox para todos os grupos trimestrais.** O parâmetro λ ótimo, λ_{opt}, para cada grupo trimestral (disco) e o valor médio, $\overline{\lambda_{opt}} = -0{,}04$, representado pela linha reta vermelha. As barras de erro indicam intervalos de confiança de 95%. 51

11 **Análise Box-Cox para todos os grupos trimestrais.** (**A**) Para os dados globais normalizados, a PDF da variável logarítmica, $z^{(0)}$. (**B**) O mesmo que em (**A**), mas com o emprego de $\lambda_{opt} = -0{,}05$ na transformação Box-Cox. As linhas tracejadas correspondem à distribuição normal com média zero e desvio-padrão unitário. 51

12 **Evolução do número de jogadores na categoria *Indie*.** O número médio de jogadores por hora em cada mês para a categoria *Indie* (discos roxos) e um ajuste linear (linha reta vermelha), em que o mês um é julho de 2012 e o último mês é dezembro de 2020. Essa linha vermelha corresponde à função $N = 14636 + 4494\,m$. 55

13 **Evolução do número de jogos lançados na *Steam* para a categoria *Indie*.** Durante os cento e quatro meses dos nossos dados, aparece o crescimento do número de jogos lançados nessa categoria (discos verdes) e um ajuste exponencial (na escala logarítmica, linha reta vermelha). A equação que governa esse ajuste é dada por $\log N = 1{,}21 + 0{,}015\,m$. 56

14 **Densidade de probabilidade para a categoria *Indie*.** PDF para a categoria *Indie* obtida por meio da transformação Box-Cox ($\lambda_{opt} = -0{,}049$) é análoga à figura 11B (todos os trimestres de todos os grupos trimestrais juntos, a linha tracejada correspondendo à uma gaussiana de média nula e desvio-padrão unitário). 58

15 **Evolução do parâmetro Box-Cox para a categoria *Indie*.** Parâmetro Box-Cox para jogos *Indie* em cada trimestre de nosso conjunto de dados, análogo à figura 10. As barras referem-se a intervalos de confiança de 95%. A linha reta corresponde à média de todos os parâmetros trimestrais de Box-Cox, $\overline{\lambda_{opt}} = -0{,}06$. 59

16 **Principais categorias de *video games*.** Parâmetro Box-Cox ótimo, λ_{opt}, para as doze maiores categorias de jogos, calculado pelo mesmo procedimento de máxima verossimilhança. Os parâmetros λ_{opt}'s são classificados em ordem decrescente de seu número de jogos por categoria. A linha reta representa a média desses λ_{opt}'s, $\langle\lambda\rangle = -0{,}028$ para as doze categorias principais. 59

17 **Evolução do número de categorias.** Evolução temporal do número de categorias ao longo dos meses de julho 2012 a dezembro 2020. A linha reta vermelha representa um ajuste linear, $N_c = 263{,}09 + 7{,}98m$, do mês 10 até o último mês da nossa base de dados. 60

18 **Distribuição das categorias de *video games*.** PDF para categorias obtidas por meio da transformação Box-Cox ($\lambda_{opt} = -0{,}12$) em função do número de *video games* em cada categoria em dezembro de 2020. A linha tracejada refere-se a uma distribuição gaussiana de média zero e desvio-padrão unitário. 61

19 **Evolução do parâmetro Box-Cox para o número de categorias.** Evolução temporal de julho de 2012 a dezembro de 2020 do parâmetro ótimo, λ_{opt}, para o número de categorias, em que cada disco representa um semestre. As barras indicam 95% intervalos de confiança dos λ_{opt}'s. A linha reta vermelha exibe o comportamento médio dos últimos dois anos. . 62

Lista de Tabelas

1 **Resumo dos dados utilizados neste trabalho**. A tabela resume alguns aspectos dos dados coletados e do banco de dados estudado: número de jogos, número de categorias (*tags*), tempo englobado pela coleta (Δt), tempo inicial t_i, tempo final t_f, total da média de jogadores por mês durante todo o período ρ_{global} e total da média de jogadores por mês durante todo o período para jogos na categoria *Indie*, ρ_{Indie}. 35

2 **Principais categorias empregadas na investigação**. A tabela apresenta as doze categorias considerando a quantidade de jogos. Na primeira coluna, temos o nome da categoria (*tag*) e seu número de jogos entre parênteses (em dezembro de 2020). Na segunda coluna, descrevemos sucintamente o que caracteriza os jogos na categoria. Mencionamos que, à exceção da categoria *Indie* e dos acrônimos RPG e 2D, utilizamos aqui o nome traduzido das categorias para facilitar a compreensão; no entanto, mantivemos o nome em inglês ao longo do texto para ajudar na conferência do conjunto de dados disponibilizado. 53

3 **Parâmetros estatísticos para os dados transformados.** Parâmetros dos dados transformados de Box-Cox (primeira coluna): parâmetro Box-Cox (λ), média (μ), desvio-padrão (σ), assimetria (γ) e curtose (κ); os valores para esses parâmetros log-transformados são mostrados na segunda coluna; todos os dados transformados via Box-Cox (terceira coluna); os dados após a transformação Box-Cox para a categoria *Indie* (*tag*) (quarta coluna) e os dados transformados via Box-Cox para a distribuição do número de jogos por categoria (quinta coluna). Após cada símbolo $\pm$, exibimos o intervalo de confiança de 95% do correspondente parâmetro. 62

Sumário

Introdução **12**

1 Métodos **16**

1.1 Objeto central e contextualização 16

1.2 Estudos de memória e memória cultural 19

1.3 Modelo malthusiano para crescimento 23

1.4 Alguns aspectos estatísticos 25

1.4.1 Análise de crescimento e método dos mínimos quadrados 26

1.4.2 Transformações de potência e máxima verossimilhança 29

2 Materiais **34**

2.1 Obtenção e tratamento dos dados 34

2.2 Apresentação do código 36

3 Análise global dos dados **40**

3.1 Hipóteses 40

3.2 Análise de crescimento 41

3.3 Médias e desvios-padrão 42

3.4 Distribuições de probabilidade 44

3.4.1 Análise de um grupo trimestral 44

3.4.2 Análise global 50

4 Análise categórica dos dados **52**

4.1 Resultados para as maiores categorias 52
4.2 Análise entre as categorias 57

5 Análise dos resultados 63

5.1 Comportamentos de crescimento e modelo malthusiano 63
5.2 Popularidade relativa e modelo *memoryless* 65
5.3 Presença de efeitos de memória 66
5.4 Popularidade relativa nas categorias com maior quantidade de jogos 67
5.5 Distribuição do número de jogos nas categorias 68

Conclusões e perspectivas 70

Referências 74

Apêndice A – Código desenvolvido e utilizado para a obtenção, tratamento e análise dos dados 84

A.1 Código 84
A.1.1 Extração dos dados 84
A.1.2 Extração da *Steamcharts* 88
A.1.3 Laço para todos os jogos 91
A.1.4 Extração das categorias 92
A.1.5 Construção da lista de *tags* 96
A.1.6 Higiene dos dados e construção do *dataframe* 99
A.1.7 Código para obtenção das figuras 4 e 5 102
A.1.8 Código para obtenção das figuras 6 e 7 105
A.1.9 Código para obtenção das figuras 8 e 9 107
A.1.10 Código para obtenção das figuras 10 e 11 110
A.1.11 Código para o caso *Indie* 112
A.1.12 Código para o estudo do número de categorias ao longo do tempo . 118

A.1.13 Código para obtenção das figuras 17, 18 e 19 121

Apêndice B – Função de distribuição de probabilidades dos grupos trimestrais 124

Apêndice C – Evolução temporal dos parâmetros de Box-Cox para os grupos trimestrais 142

Apêndice D – Padrões de crescimento e funções de distribuição das categorias principais 158

Apêndice E – Evolução temporal dos parâmetros de Box-Cox para as categorias principais 175

Introdução

"Eu / um universo de átomos / um átomo do universo" são versos de um poema de Feynman, incluído em seu discurso para a Academia Nacional de Ciências dos Estados Unidos da América (FEYNMAN, 1955). A expressão é sucinta, mas captura o impulso por trás do fazer científico. Desde os primórdios da civilização, conforme as narrativas mitológicas ou as primeiras abstrações filosóficas, a humanidade sempre buscou entender o universo a seu redor e seu papel nele. Não é despropósito afirmar que a física, também na antiga nomenclatura de filosofia natural, tem um papel central nessa busca. Ao longo dos séculos, os filósofos e os físicos desenvolveram ferramentas e técnicas, fazendo florescer o conhecimento do mundo. Além disso, possivelmente tão antiga quanto a própria física, está a questão da posição do homem na ordem natural: se, para um ponto de vista, somos tão integrados ao mundo material que nos desenvolvemos por meio de um longo processo evolutivo, para outro, especialmente o platônico, fazer-se material seria um declínio e o mundo externo, uma sombra do ideal, conjunto de reflexões dos entes perfeitos. Essa oposição, isto é, entre o homem e o mundo externo, permeia todo o desenvolvimento do conhecimento humano, da dialética socrática até a epistemologia contemporânea, passando por nomes como Berkeley (BERKELEY, 1709) e Kant (KANT, 1908), incluindo questões de cunho religioso. Além disso, a própria linguagem cotidiana revela essa contradição já que, por exemplo, quando falamos em 'animais', esse plural usualmente significa todos os animais exceto o homem, a não ser quando explicitado o contrário.

Em consonância com esse contraste, o conhecimento científico e, por extensão, o conhecimento acadêmico dividiu-se em duas grandes áreas, as humanidades e as ciências naturais. A primeira categoria devota-se, grosso modo, ao entendimento da ação humana em contextos de construção humana, sejam de ordem histórica ou socioeconômica, criados a partir da interação do homem consigo mesmo ou até mesmo por uma combinação dos anteriores. A segunda categoria, historicamente, devotou-se à compreensão dos fenômenos naturais externos ao homem. Especialmente nos séculos XVIII e XIX, observamos que a maior parte do programa científico, pelo menos na física, tinha o enfoque da descrição objetiva da realidade, de uma descrição por meio da qual poderíamos prever o futuro do universo caso soubéssemos a posição e a velocidade de cada partícula nele. Apesar do

grande sucesso dessa escola de pensamento em termos da mecânica e do eletromagnetismo clássicos, os avanços do século XX, especialmente no que diz respeito às probabilidades inerentes à física quântica, colocariam em xeque a visão determinística. Com a citação que abre a introdução, deixamos claro que a conexão entre as maneiras pelas quais buscamos explicar os fenômenos naturais e os fenômenos naturais em si é mais sutil e difícil de ignorar do que se costumava estimar.

Com essa mudança de paradigma e com todo o repositório de ferramentas desenvolvidas ao longo dos séculos, uma parcela dos esforços científicos surgidos a partir da segunda metade do século XX destinou-se a abrir novos caminhos entre as humanidades e as ciências exatas, a aumentar a comunicação entre saberes diversos, em suma, a criar áreas de pesquisa que fossem interdisciplinares, que usassem técnicas das ciências naturais para entender e descrever a ação humana. Desses esforços, nasceu o campo intitulado física dos sistemas complexos, que, em um primeiro momento, podem ser descritos como sistemas cujas características principais emergem das interações entre suas partes. Dito de outra forma, seriam sistemas complexos nos quais não é suficiente descrever suas partes para descrever o todo, pois muitos padrões emergentes do todo dificilmente poderiam ser inferidos ou mesmo imaginados do comportamento de apenas uma das partes. Com o tempo, o estudo de sistemas complexos foi evoluindo para subáreas, como a dinâmica de sistemas não-lineares (também chamado de teoria do caos) (ALLIGOOD; SAUER; YORKE, 1996), a econofísica (MANTEGNA; STANLEY, 1999), o estudo do comportamento coletivo de animais (SUMPTER, 2010) e a sociofísica (GALAM, 2012). Ademais, essas investigações interdisciplinares, com base em técnicas e ferramentas de física, encontraram respaldo na comunidade científica, passando a ser largamente publicadas nos mais diversos e conceituados periódicos da física, como a *Physica A* (PHYSICA-A, 2021), o *European Physics Journal B* (JOURNAL-B, 2021), a *Chaos* (CHAOS, 2021) e o *Physical Review E* (E, 2021). Nessa mesma direção, atualmente, vários centros de pesquisa de física contam com grupos de estudo voltados a sistemas complexos. Esse é o caso, por exemplo, do *ComplexLab* na Universidade Estadual de Maringá (LAB, 2012). Nesse grupo, vários pesquisadores têm sido formados: Sergio Picoli (2007), Haroldo V. Ribeiro (2012), Fernando J. Antonio (2013), Magda C. Mantovani (2013), Andréia I. Silva (2015), Edenize S. Santos (2015), Renato R. Guimarães (2016), Leandro S. Costa (2017), Denner S. Vieira (2019) e Higor Y. D. Sigaki (2022).

Outro fenômeno da virada do século passado contribuiu para uma nova mudança de paradigma: a popularização e a globalização da internet. A partir desse fenômeno, os cientistas obtiveram acesso jamais visto a um grande volume de dados das mais diversas

naturezas, como a geolocalização, transações financeiras, uso do transporte público e publicações em rede sociais. Para termos uma noção do volume dessa produção de dados, a estimativa em 2020 era de que produzíamos dois e meio quintilhões de *bytes* por dia (ANALYTICS, 2020). Tamanha abundância permitiu a investigação de novas áreas e questões em uma escala que era previamente impossível, especialmente no caso das ciências sociais. Dessa maneira, é possível afirmar que o casamento entre o esforço interdisciplinar mencionado no parágrafo anterior e os "grandes dados" (*big data*, em inglês) tem sido frutífero e permite uma crescente comunicação entre áreas do conhecimento que estavam usualmente desconectadas umas das outras. Particularmente, esse direcionamento envolve fortemente estatísticos e profissionais da computação.

Nesta tese, a partir de técnicas e ferramentas típicas em física, apresentamos uma análise interdisciplinar da popularidade de um produto relativamente recente de nossa cultura, os *video games*. O objetivo central é investigar e quantificar, por meio de um modelo interpretativo, a conexão entre popularidade e mecanismos de memória. Para tanto, oferecemos uma descrição quantitativa dos dados e investigamos sua popularidade em três distintas camadas: global, maiores categorias e entre categorias.

Inicialmente, investigamos o crescimento do número de jogadores e do número de jogos, apresentando um modelo malthusiano para descrever o crescimento da quantidade de jogadores e de jogos. Propomos também um modelo interpretativo em que associamos a popularidade de um jogo (ou categoria de jogos) com uma medida da memória, conforme distinção feita por J. Assmann entre memória cultural e comunicativa. Há diversos motivos para o estudo da popularidade de jogos. O primeiro, mais imediato, é o crescente impacto socioeconômico desse passatempo: desde 2020 nos Estados Unidos, por exemplo, o mercado de *video games* é maior que o de filmes ou esportes (WITKOWSKI, 2021). O impacto dessa indústria tem aumentado bastante, especialmente no público jovem, desencadeando um interesse científico crescente em correlacionar o uso de *video games* com diferentes métricas educacionais e sociais (BAVELIER et al., 2011; GRANIC; LOBEL; ENGELS, 2014; DANIEL; GARRY, 2018; HAMMAR, 2019). No entanto, não há, no melhor do nosso conhecimento, um estudo quantitativo que utilize uma grande base de dados para entender padrões de popularidade tanto de *video games* quanto em suas categorias.

Em nosso trabalho, buscamos contribuir nessa direção, empregando dados baseados em uma plataforma digital de jogos, a *Steam* (VALVE, 2020a) e provendo, assim, uma novidade metodológica para o estudo do impacto dos jogos. Identificamos padrões de

crescimento, de distribuição de popularidade e de categorias, tanto em seus surgimentos quanto em suas popularidades relativas. Além do modelo interpretativo para a popularidade, propomos um modelo de crescimento malthusiano tanto para o número de jogadores quanto para o número de jogos lançados mensalmente.

A tese está dividida da seguinte maneira: no primeiro capítulo, apresentaremos aspectos da formulação de Assmann, da análise de crescimento e das transformações de potência, com ênfase na transformação Box-Cox; no segundo capítulo, discorreremos acerca do material utilizado para a tese, isto é, explicaremos a plataforma de jogos escolhida, exporemos os programas desenvolvidos para a captura e a análise dos dados e ofereceremos uma síntese da base de dados obtida; no terceiro capítulo, apresentaremos as hipóteses centrais da tese e, também, os resultados relacionados à primeira camada, a global, além de uma descrição quantitativa dos padrões de crescimento encontrados nos dados; o quarto capítulo será dedicado à segunda camada, ou seja, à investigação da popularidade dos jogos divididos em categorias e, também, a um estudo do número de categorias ao longo do tempo e a sua distribuição; o quinto capítulo é direcionado a uma análise dos resultados obtidos para *video games*, com ênfase no modelo malthusiano de crescimento e no modelo de memória coletiva. Dedicamos, também, alguns parágrafos às conclusões e às perspectivas futuras deste trabalho. Os apêndices contêm um conjunto de materiais complementares.

1 Métodos

Neste capítulo, no esforço de manter o trabalho o mais auto-contido possível e considerando que a teoria pertinente serve de estrutura interpretativa para os modelos apresentados nos próximos capítulos, discorreremos acerca do conceito de memória cultural, formulado por Assmann (ASSMANN, 2011), em particular acerca da subdivisão de memória em memória cultural e comunicativa. Além disso, revisaremos alguns aspectos da análise de crescimento e o método dos mínimos quadrados para ajustes lineares. Finalmente, discutiremos as transformações de potência, com ênfase especial na transformação Box-Cox, utilizada ao longo da tese, e faremos uma apresentação do princípio de máxima verossimilhança em sua forma logarítmica (*log-likelihood*), já que foi por meio desse princípio que obtivemos os parâmetros ótimos nas transformações aplicadas. Caso o (a) leitor(a) esteja familiarizado(a) com os conceitos apresentados neste capítulo, ele ou ela pode se sentir à vontade para prosseguir de imediato para a leitura do segundo capítulo. Entretanto, recomendamos a leitura das três primeiras seções, que serão dedicadas à contextualização e aos motivos da escolha de *video games* como objeto de estudo.

1.1 Objeto central e contextualização

Uma das hipóteses da tese é de que a popularidade dos objetos culturais no momento em que são divulgados e nos anos seguintes pode ser um meio para medir a permanência de longo prazo na memória coletiva. Supomos que, se um novo objeto cultural foi recebido positivamente, é plausível que permaneça por mais tempo na memória coletiva. Certamente, esta não é uma maneira exaustiva de se medir essa permanência, visto que uma produção cultural, mesmo não sendo popular, pode impactar profundamente a área em que se insere. Por outro lado, medidas de popularidades entre diferentes espécies de objetos culturais tendem a ser comparáveis, i.e., é possível comparar a recepção de filmes com músicas e *video games* e a comparação não perderia sentido. Outra preocupação pertinente é com a delimitação do que seria "longo prazo". Tomemos um exemplo do

cinema: Carl T. Dreyer filmou *The passion of Joan of Arc* em 1928 e o filme foi recebido imediatamente como uma obra prima (HALL, 1929). Além do impacto cinematográfico, o filme influenciou Érico Veríssimo a escrever *A vida de Joana d'Arc* (VERÍSSIMO, 1960) e, provavelmente, cimentou o caminho para a segunda geração de cineastas escandinavos (entre eles, Ingmar Bergman). Dessa maneira, parece plausível afirmar que o filme de Dreyer é uma obra cultural de repercussão atemporal, mas, como um filme mudo, decaiu em popularidade, em parte, por causa dos avanços tecnológicos da indústria. Somando-se isso ao declínio natural no tempo, podemos dizer que é um filme historicamente importante e um clássico *cult*, mas não popular até os dias atuais. Isso significa que a hipótese mencionada no começo do parágrafo está incorreta? Consideramos que não, mas, quanto maior for o "longo prazo", quanto mais elástico for o intervalo de tempo, mais fraca a hipótese se torna em razão das mudanças tecnológicas, de paradigmas e de sensibilidades.

Na investigação da popularidade dos *video games*, enquadramos o trabalho, pelo menos em parte, no contexto mais amplo da memória coletiva e, mais especificamente, no da memória cultural. Essa última pode ser caracterizada, com base na formulação de Assmann (ASSMANN, 2011), como "[...] uma forma de memória coletiva em que um número de pessoas compartilha a memória cultural e em que ela lhes transmite uma identidade coletiva (ou seja, cultural)". Normalmente, os estudos sobre memória e esquecimento de produtos analisam dados provenientes de consumo, por exemplo, de livros e filmes (LORENZ-SPREEN et al., 2019).

Nesta tese, oferecemos duas novidades nesse debate. Primeira, o foco na popularidade ao longo do tempo de *video games* é indicativo de sua permanência na memória coletiva. A principal métrica empregada para quantificar a popularidade dos *video games* é o número médio de jogadores de cada jogo por mês de julho de 2012 até dezembro de 2020. Outras métricas também empregadas são a data de lançamento e as categorias de cada *video game*. Segunda, a escolha dos estudos de memória como uma estrutura para interpretar nosso modelo quantitativo. Embora estudos de memória já tenham sido empregados na investigação do processo de esquecimento de objetos culturais (CANDIA et al., 2019), este trabalho é o primeiro, até onde sabemos, no contexto da popularidade dos *video games*.

Uma preocupação válida é com o período estudado, visto que a memória cultural está geralmente preocupada com a continuidade intergeracional e intervalos de tempo mais longos. Porém, nesta era digital, novas tendências culturais são criadas, apropriadas, ressignificadas e descartadas de maneira quase contínua, em processos influenciados por interações interpessoais, redes sociais e pela presença (até mesmo onipresença) de publi-

cidade *online*. Assim, em uma sociedade cada vez mais acelerada, parece-nos plausível discutir objetos e memórias culturais em períodos mais curtos e essa discussão torna-se mais relevante em um contexto cultural dominado por pessoas mais jovens (ASSOCIATION; ASSOCIATION et al., 2015; DUGGAN, 2015).

Nesse ponto, uma objeção poderia ser levantada: por que estudar a popularidade dos *video games*? Para responder e na direção de minimizar os problemas mencionados, afirmamos que a escolha por *video games* se dá porque, por um lado, o recorte temporal é bem delimitado e, por outro, porque não houve no período estudado uma mudança técnica comparável à do cinema mudo para o falado[1]. Além desses motivos, devemos observar que o papel socioeconômico dos jogos digitais continua se acentuando, o que se reflete no aumento do interesse científico de diversas áreas de pesquisa como a da educação (TAKATALO et al., 2011; CONNOLLY et al., 2012; KUEIDER et al., 2012; DOMINGUEZ et al., 2013; BOYLE; HAINEY et al., 2016; CHANG; LIN, 2019), da ciência da computação/de rede (BAINBRIDGE, 2007; SZELL; LAMBIOTTE; THURNER, 2010; MERRITT; CLAUSET, 2013; YOKOTA; SOSHI; NARUSE, 2019; LANDFRIED; SLEZAK; MOCSKOS, 2019), da psicologia (PARK; SONG; TENG, 2011; ADACHI; WILLOUGHBY, 2013; TENG et al., 2020), e da saúde (KUSS; GRIFFITHS, 2012; ANGUERA et al., 2013; PETRY et al., 2014; GONG et al., 2015). De particular interesse para este trabalho é a existência de estudos a respeito da popularidade das categorias de *video game* (ELLIOTT et al., 2012; REHBEIN et al., 2016) e dos fatores responsáveis por manter os jogadores investidos nesse passatempo, principalmente o jogo *online* com amigos (fator social), a diversão intrínseca do jogo (fator de imersão) e as conquistas (fator individual) (YEE, 2006; HOFFMAN; NADELSON, 2010; BOYLE; CONNOLLY et al., 2012; BILLIEUX et al., 2013). Por último, além do crescente interesse científico, os *video games* têm um valor próprio como objeto cultural, dado que, ao contrário da maioria dos outros objetos, são uma mídia interativa, o que significa que os jogadores contribuem ativamente para a experiência geral. Em outras palavras, somente entendendo as preferências do usuário é que podemos compreender totalmente os *video games* como fenômenos culturais.

[1]Uma possibilidade seria a dos jogos de realidade virtual (VR, em inglês). Pensamos que, se estes se popularizem e se tornem o padrão de fato da indústria, esta seria uma mudança técnica similar à experimentada pelo cinema. No momento, o equipamento físico necessário ainda é relativamente caro e limitado aos países de primeiro mundo e os jogos ainda não estão se popularizando na mesma medida. O tempo dirá se essa transição ocorrerá. Vale mencionar que esse argumento não implica que não houve avanços técnicos no desenvolvimento de jogos — essa seria uma posição indefensável. O que entendemos é que, apesar dos avanços, a estrutura, o esqueleto dos jogos foram pouco alterados no intervalo de tempo estudado.

Tendo exposto as motivações e justificado a escolha por *video games* como objetos culturais, passamos à apresentação da estrutura interpretativa, qual seja o conceito de memória cultural apresentado por Assmann (ASSMANN, 2011), que se baseou em trabalhos anteriores de Halbwachs e Nietzsche (ASSMANN, 2016).

1.2 Estudos de memória e memória cultural

O conceito de memória é onipresente tanto nas conversas cotidianas quanto nas investigações das ciências sociais. Nossas identidades, pessoais, familiares e nacionais, são moldadas a partir de memórias enraizadas em interações interpessoais, discursos de cima para baixo (por exemplo, se uma nação passa a comemorar um feriado político, mas não outro) e autorreflexão. No entanto, apesar do uso costumeiro, a popularização acadêmica do conceito de memória nas investigações citadas só ocorreu mais recentemente, a partir da década de 1980, inaugurando os chamados estudos de memória (*memory studies* em inglês). Talvez seus proponentes mais famosos, J. e A. Assmann, tenham considerado esses estudos no contexto de uma geração alemã que havia crescido no pós-guerra e ainda havia de lidar com a divisão do país em duas esferas políticas antagônicas. Tal proposta foi fundamentada tanto nos estudos sociológicos da década de 1920 (em particular, a partir de M. Halbwachs (WEBER; PEREIRA, 2010; GUZZI, 2011) quanto na filosofia mais antiga, de F. Nietzsche (ASSMANN, 1988). Como um modelo interpretativo do presente trabalho é baseado nos estudos de memória e há escassez relativa de uma caracterização e de um diálogo com a sociologia nas áreas de física (pelo menos em língua portuguesa), decidimos nos alongar na apresentação de um contorno dessa área tão fértil nas humanidades. Ressaltamos que tal exposição está em acordo com o cuidado de manter este trabalho o mais auto-contido possível.

Nesta seção, apresentaremos uma visão geral dos estudos de memória, incluindo alguns pensadores que influenciaram diretamente em sua formação como campo acadêmico. Além disso, faremos a distinção entre história e memória[2], de maneira que possamos justificar nosso modelo interpretativo como um "estudo sobre memória" dos *video games* em vez de uma "história da popularidade dos *video games*". Para finalizar a seção, discutiremos as distinções entre categorias de memória, como apresentado por Assmann, e como essas distinções podem ser aplicadas ao modelo a ser discutido nos capítulos subsequentes. Embora uma exposição mais abrangente deste último ponto deva aguardar a apresentação

[2] A distinção deve ser entendida, é claro, no contexto deste trabalho. Como se trata de uma distinção complexa, poderíamos devotar muitas páginas a ela, mas nos limitaremos a apresentar uma distinção que seja suficiente para a pesquisa sobre a popularidade de *video games*.

dos dados e do modelo, a explanação neste capítulo servirá para estabelecer nomenclatura e conceitos utilizados ao longo do trabalho.

Começamos por esclarecer que discorreremos sobre os estudos de memória e sobre o conceito de memória como entendido no contexto de Assmann. Não é do nosso interesse, nos capítulos subsequentes, fazermos uma exposição sobre o conceito de memória em geral, visto que nosso enfoque interpretativo foi exatamente a distinção entre os meios de memória fornecida por Assmann.

Observamos, inicialmente, que, apesar de os estudos de memória, da maneira como entendemos e usamos neste trabalho, terem sido propostos apenas em 1980, suas raízes podem ser encontradas já em 1925 na obra de Halbwachs (HALBWACHS, 2010), que cunhou o termo "memória coletiva". Tal desenvolvimento permitiu a elaboração de um conceito de memória de um grupo e, muitas vezes, uma memória não-oficial, não-escrita (podemos pensar nos costumes passados de geração em geração na família ou das tradições orais de povos que não desenvolveram ou não utilizam com frequência a linguagem escrita). Encontramos ainda, nesse autor, a intuição de que a memória coletiva evoluiria de forma diferente das memórias individuais que a compõem, ou seja, o todo difere da soma de suas partes. Indo além, observamos a distinção entre "memória histórica" e "memória coletiva". A primeira forma, grosso modo, uma memória oficial, de recordações e reconstruções dos fatos, conhecimento esse cristalizado em instituições. A memória coletiva, por sua vez, apresenta-se como um fluxo contínuo coletivo que se preocupa em preservar do passado aquilo que ainda se encontra presente, isto é, a preservação daquilo que um grupo ainda percebe como conhecimento útil. Devemos notar que, como o enfoque está dirigido para aquilo que interessa a um grupo, a memória coletiva, para Halbwachs, não ultrapassa os limites desse grupo. Além disso, "os conjuntos de homens que constituem um mesmo grupo em dois períodos sucessivos são como duas barras em contato por suas extremidades opostas, mas que não se juntam de outro modo, e não formam realmente um mesmo corpo" (HALBWACHS, 1990). Por último, mencionaremos que, apesar de ser Jung a pessoa mais associada à ideia de memória coletiva, como ele se dedicou à investigação do espaço interno da memória e do seu impacto na psiquê, não influencia diretamente nossa interpretação do modelo, embora tenha sido também uma influência para Assmann (ASSMANN, 2011).

Evitando o debate ocorrido ao longo do século XX sobre a existência e a importância do estudo da memória não-oficial e também de ouvir grupos marginalizados, passamos agora à exposição da proposta de Assmann (ASSMANN, 2011), já que esta é de interesse central para a interpretação de um modelo proposto em nosso estudo de *video games*. Esse

autor propõe uma subdivisão para o caso da memória coletiva, a saber, uma subdivisão em memória cultural e comunicativa. Como mencionado, Halbwachs não incluiu em seu conceito de memória coletiva a transmissão intergeneracional de tradições e costumes. A contribuição de Assmann está, portanto, não na alteração do conceito de Halbwachs, mas na insistência em tratar esse mecanismo cultural de transferência de conhecimento como memória também (ASSMANN; CZAPLICKA, 1995; ASSMANN, 2011). Não se trata de substituir a ideia de memória coletiva pela de memória cultural e, sim, nas palavras de Assmann, de "[...] distinguir entre dois *modi memorandi*, ou maneiras de relembrar" (ASSMANN, 2011). Essa distinção pode ser abordada da seguinte forma: a memória comunicativa seria compartilhada e transmitida em um grupo social definido por memórias comuns de interação pessoal via meios verbais de comunicação ao longo de um período de 80 a 100 anos (sobre o recorte temporal, ver a tabela e a seção intitulada *Time frames* presentes em (ASSMANN, 2011)). Particularmente, não é um meio de lembrar institucionalizado ou carregado de símbolos. Já a "[...] memória cultural é uma instituição. É exteriorizada, objetificada e armazenada em formas simbólicas que, ao contrário dos sons das palavras ou da aparência dos gestos, são mais estáveis e transcendem a situação. Elas podem ser transferidas de uma situação para outra e transmitidas de uma geração para outra. Ao contrário da memória comunicativa, a memória cultural é desencarnada. Para funcionar como memória, no entanto, suas formas simbólicas devem não apenas ser preservadas, mas também circuladas e reencarnadas em uma sociedade" (ASSMANN, 2011). Além disso, deve ficar claro que a memória cultural é, usualmente, de mais longa duração se comparada à memória comunicativa, podendo existir por séculos em uma sociedade.

Expomos, assim, os conceitos que serão utilizados na interpretação de nosso modelo de popularidade envolvendo memória. Resta-nos esclarecer dois aspectos entre os que mencionamos acima: a motivação para estudos de memória para *video games* e como esses conceitos de memória serão aplicados no modelo. O primeiro ponto pode ser respondido de maneira sucinta, visto que foi parcialmente respondido na introdução deste capítulo. O dado estudado refere-se a usuários jogando de maneira livre e desimpedida, não em um ambiente controlado de laboratório. A escolha por um jogo (ou por uma categoria de jogos) é de responsabilidade do próprio usuário, o que remove graus de subjetividade encontrados em pesquisas realizadas via questionários. Indo além, com o desenvolvimento da internet, os usuários estão em constante contato com amigos e com outros usuários, tendo acesso a jogos que são populares, etc. Dessa forma, a influência do boca-a-boca e da rede de jogadores na totalidade não pode ser ignorada em um modelo que tente

descrever a popularidade dos *video games*. Esse mecanismo de popularidade e memória remete ao conceito de memória comunicativa já descrito. Para responder ao segundo ponto, dependemos de um argumento um pouco abstrato, mas que consideramos válido, qual seja, o de que as categorias de jogos e a escolha, por parte das desenvolvedoras de jogos, de uma espécie de jogo em detrimento de outra representem uma espécie de memória cultural, cristalizada nas instituições-empresas. Dito de outra forma, no nível do usuário, a mudança de popularidade e de opinião é rápida, sendo mais influenciada pelas redes sociais e pelo contato com amigos e com outros usuários (memória comunicativa), enquanto a decisão de desenvolver um estilo de jogo ou de revisitar uma categoria impopular seja tanto uma decisão econômica quanto um reflexo de padrões estéticos ou avanços tecnológicos que mais se assemelham ao conceito de memória cultural[3]. As respostas aos dois pontos podem parecer sucintas, mas serão revisitadas ao longo da discussão do modelo e dos dados. Se o leitor puder guardar apenas uma informação desta seção, que seja a seguinte: podemos inicialmente propor um modelo em que os agentes (usuários, jogadores) atuem de maneira independente e aleatória, mas sabemos, tanto pelo conhecimento de domínio quanto pelo conhecimento geral, que essa não seria uma abstração razoável. Por isso, buscando na sociologia um modelo para nortear nossa base interpretativa e de introdução da interação entre agentes dessa plataforma de *video games*, examinamos tal interação sob a ótica do conceito de memória.

Finalmente, apresentamos duas figuras esquemáticas sobre os modos de memória. A figura 1, adaptada de (SONG et al., 2021), exemplifica o processo de difusão dos diferentes modos de memória na sociedade a partir da primeira experiência de um dado evento. Notamos que existem pelo menos duas maneiras de se pensar o processo difusivo da informação: preocupando-nos ou com o espalhamento da informação em uma rede ou grupo de pessoas (por exemplo, na difusão de fofoca em redes sociais (LIND et al., 2007)) ou com a retenção e decaimento da informação em um certo indivíduo ou grupo. A figura 2 funciona como uma espécie de taxonomia, embora o nome seja um pouco inadequado, dos diversos modos de memória e os exemplifica com alguns casos usuais. Decidimos apresentar, nessa segunda figura, a consciência de memória, isto é, a percepção que um sujeito tem do que ele se lembra. Embora esta distinção não seja usada ao longo do trabalho, pensamos ser importante salientar, neste capítulo de caráter explanatório, a interrelação entre os tipos de memória e a formação da identidade pessoal.

[3]Aqui reencontramos o "problema" do recorte temporal, já respondido na introdução deste capítulo. Também, como estamos interessados em uma análise quantitativa do fenômeno de popularidade, uma investigação muito extensa temporalmente enfrentaria problemas no que tange a encontrar dados comparáveis com os que utilizamos.

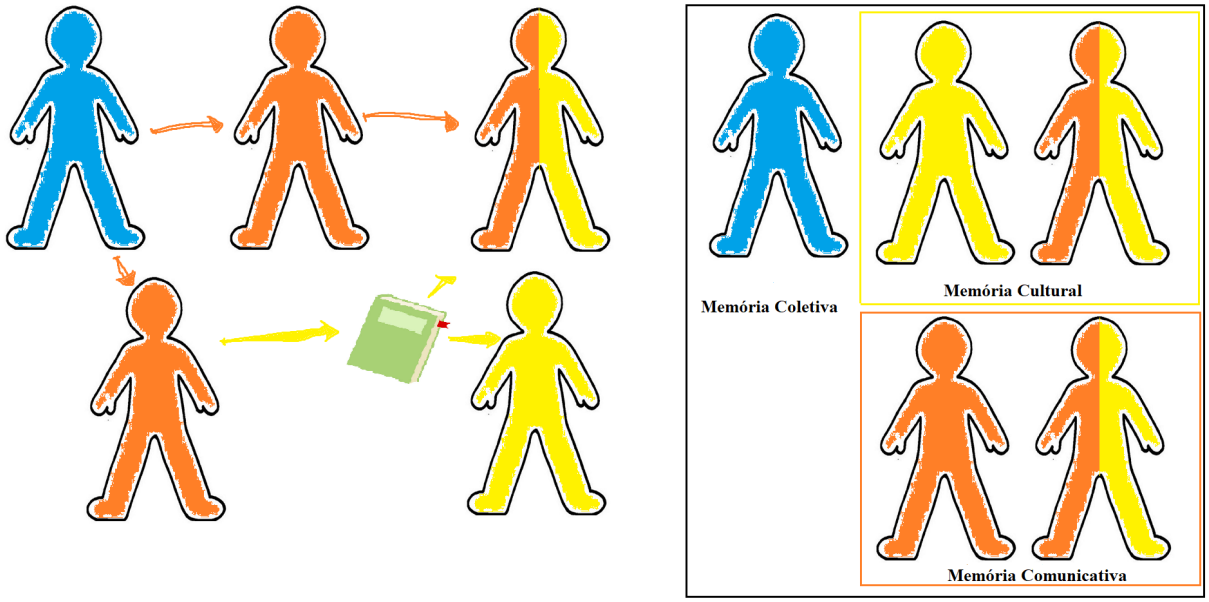


Figura 1: **Esquema de propagação de diferentes modos de memória.** Esquema de propagação de memória adaptado de (SONG et al., 2021). Em primeiro lugar, o sujeito tem uma experiência (em azul na figura) e a relata para outras pessoas (em laranja na figura). Além destas, outras pessoas podem participar indiretamente da experiência, por exemplo, lendo um livro (em amarelo na figura). Note-se que um mesmo sujeito pode participar de dois ou mais modos de memória simultaneamente e esses modos possivelmente terão mecanismos de esquecimento distintos.

1.3 Modelo malthusiano para crescimento

Antes de passarmos à apresentação dos aspectos estatísticos empregados neste trabalho, consideramos útil introduzir um modelo interpretativo além daquele baseado em estudos de memória. Esse outro modelo interpretativo é direcionado à análise de crescimento do número de jogadores e de jogos a ser retomada nos próximos capítulos. Investigar a natureza de crescimento de uma quantia pode ser interessante por si, mas buscar compreender o porquê de uma quantia crescer de uma forma e outra de outra e de se relacionarem entre si também tem valor, seja prático ou teórico. Nesse contexto, Malthus (MALTHUS; WINCH; JAMES, 1992) investigou o crescimento da população comparativamente à evolução na produção de alimentos e identificou, no primeiro caso, um crescimento exponencial e, no segundo, um linear. Ele postulou que, no caso de o crescimento exponencial, em algum momento, sobrepassar o linear, o alimento se tornaria escasso com o passar do tempo e, pela competição e pelo crescimento populacional, atingiria um limite. Embora na sociedade humana, esse fenômeno não tenha sido (ainda) verificado, o modelo serviu de base para inúmeros outros em diversas áreas, notoriamente no estudo de competição entre

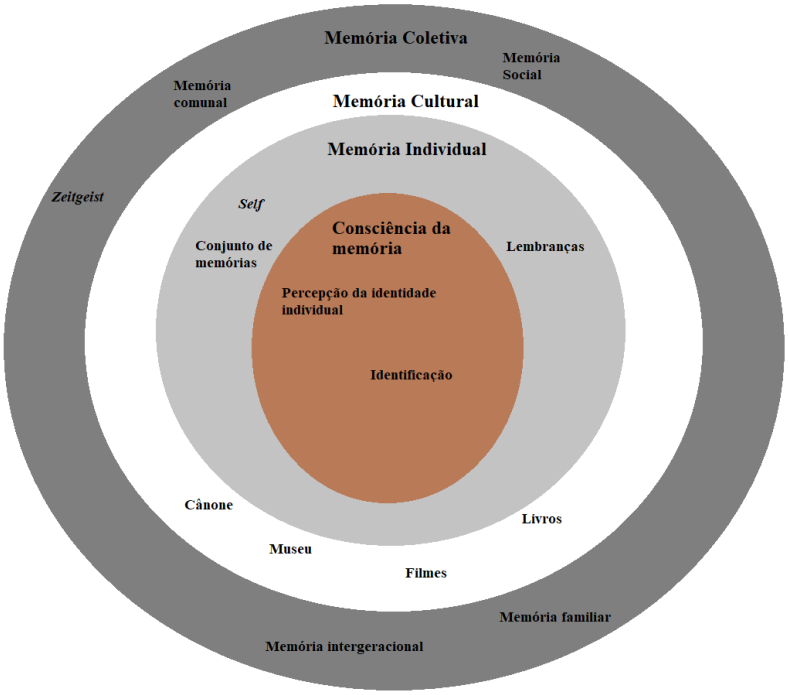


Figura 2: **Taxonomia dos diferentes modos de memória.** A categoria mais ampla, memória coletiva, engloba modos de memória como a familiar ou a social. Seu recorte temporal é longo, podendo até mesmo ser mitológico (a exemplo de mitos de formação de um povo). Já a memória cultural está ligada com a produção de objetos ou lugares que servem para cristalizar as memórias de um grupo ou povo, como museus, livros, filmes, etc. Finalmente, passamos para a memória individual, subjetiva e, muitas vezes, atemporal, para as lembranças e a percepção de si. Nesta última categoria, notamos a consciência da memória, isto é, o sujeito não só guarda memórias passivamente, mas tem consciência do que se recorda.

espécies (NEKOLA et al., 2013; DEMETRIUS; LEGENDRE, 2013) e na investigação de populações (LEE; LOSCHKY, 1987), relações de terra (ANDRÉ; PLATTEAU, 1998) e renda (SHARP; STRULIK; WEISDORF, 2012).

Fazendo um paralelo com o nosso tema, um questionamento nessa direção seria se o comportamento do crescimento de jogos e jogadores é similar e, em caso negativo, se eles podem ser enquadrados qualitativamente nessa interpretação, isto é, a da competição entre jogos em face de um crescimento linear no número de jogadores. Essas estruturas interpretativas poderiam servir de base para uma investigação sobre a condição da indústria em si, sobre a existência de um fluxo constante de jogadores e jogos (e, portanto, espaço para crescer) ou, em contrapartida, de alguma estagnação ou limite desse crescimento.

Essa análise malthusiana, chamemos assim, será revisitada ao longo da nossa discussão. Gostaríamos, entretanto, de reforçar que tais comentários, já no primeiro capítulo,

servem para esclarecer que a investigação realizada sobre o crescimento da base de dados vai além de uma análise exploratória e busca um entendimento da dinâmica que entrelaça jogos e jogadores.

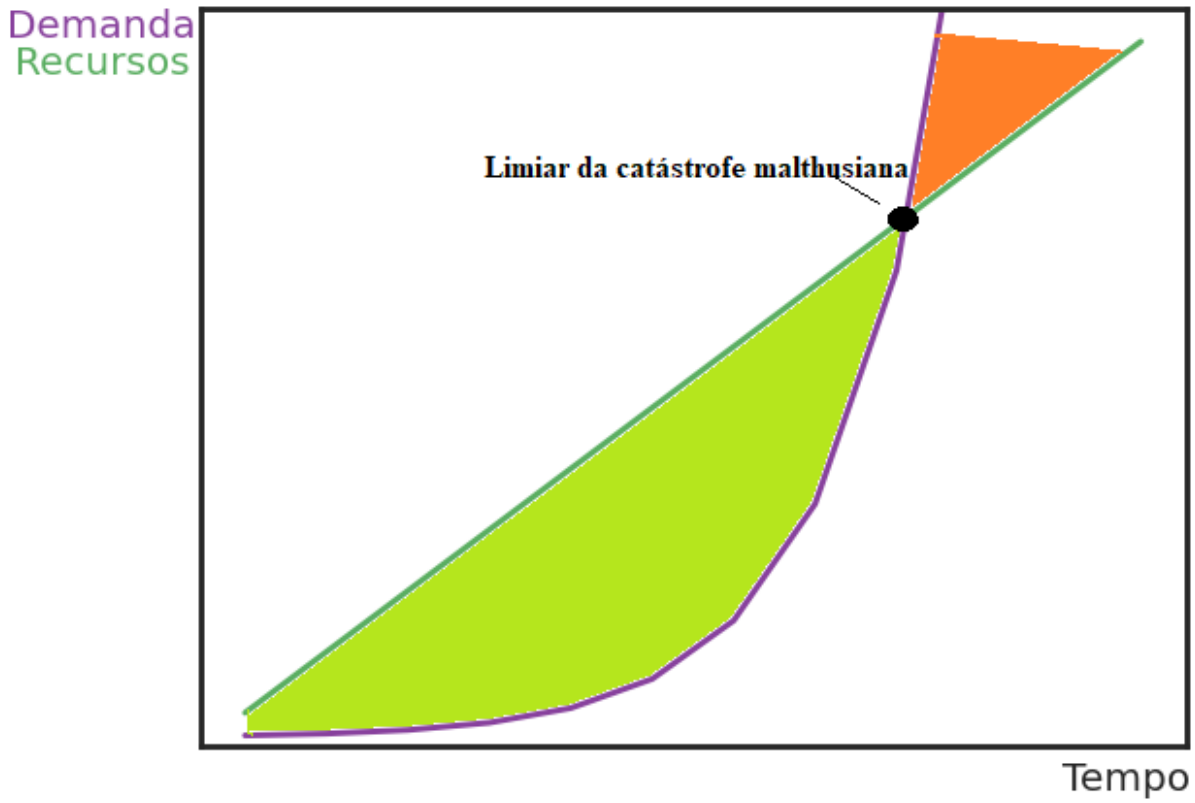


Figura 3: **Esquema do modelo malthusiano de recursos e demanda.** A teoria de Malthus identifica um crescimento aritmético (linear) para os recursos, enquanto identifica um crescimento geométrico (exponencial) para a demanda. Utilizamos os termos "recursos" e "demanda" por ser mais gerais, embora a teoria malthusiana tenha sido inicialmente aplicada à dinâmica de populações humanas. Inicialmente, há recursos suficientes para atender a toda a demanda (área verde clara, período de abundância). Com o crescimento da demanda, chega-se a um tempo crítico em que os recursos já não são mais suficientes e há disputa (área laranja, período de escassez). Em nosso contexto, identificamos os recursos com os jogadores e a demanda com os jogos, que competem pela atenção dos usuários.

1.4 Alguns aspectos estatísticos

Dedicaremos esta seção à introdução e à explanação de alguns conceitos estatísticos e notações utilizadas ao longo do trabalho. Damos particular importância aos métodos dos mínimos quadrados e da máxima verossimilhança, especialmente por seu papel na obtenção dos principais resultados apresentados no presente estudo. Finalmente, estabe-

leceremos em breves palavras uma relação entre esses dois métodos.

1.4.1 Análise de crescimento e método dos mínimos quadrados

Frequentemente, a constatação qualitativa do crescimento de uma grandeza não requer nenhuma análise mais aprofundada. Tomando nosso contexto como exemplo, acreditamos que não seria surpresa afirmar que os *video games* se tornaram mais populares nas duas últimas décadas. No entanto, igualmente frequente é a percepção de que esse tipo de constatação, por si, é pobre em descrever o fenômeno estudado, beirando uma tautologia. São necessários métodos quantitativos por meio dos quais possamos, e.g., estimar o comportamento de um fenômeno nos intervalos de tempo subsequentes. Um desses métodos é colocar em gráfico duas quantias em que estejamos interessados (no contexto desta tese, podemos tomar o número de usuários nos jogos e o número de meses analisados, por exemplo) e tentar identificar como uma quantidade se modifica em função da outra. Em um primeiro momento, podemos já substituir a constatação genérica de crescimento por outra mais útil: a de que "o número de jogadores apresenta um crescimento aproximadamente linear ao longo dos meses analisados".

Não obstante o discutido acima, seria mais útil ainda se conseguíssemos fornecer uma fórmula que melhor ajustasse esses dados e que nos desse realmente uma maneira de quantificar a evolução temporal do fenômeno. Duas coisas são necessárias para isso: conseguir indicar uma curva e sua forma algébrica com base em dados empíricos e estimar o erro no ajuste, de maneira que possamos minimizá-lo, garantindo assim um modelo ótimo[4]. O primeiro ponto, estimar uma curva, depende fortemente do conhecimento de domínio, das hipóteses iniciais e da forma da distribuição dos dados. Assim, é difícil oferecer um guia geral. Já o segundo ponto pode ser detalhado e exemplificado por meio do caso utilizado ao longo deste trabalho, o método dos mínimos quadrados. Nossa exposição seguirá a apresentada na referência (DEKKING et al., 2005), com algumas notas históricas e com ênfase em certos pontos.

A primeira exposição clara do método dos mínimos quadrados se deve a Legendre (LEGENDRE, 1806), o que ocasionou uma rápida adoção do método na área de astronomia. Uma das primeiras demonstrações da força do método veio em 1809, quando Gauss pôde

[4]Vale mencionar que, quando nos referimos a um modelo ótimo, este é condicionado ao grau de correção da hipótese de crescimento subjacente. Dito de outra forma, se supomos que uma quantia y cresce linearmente com x, então a minimização de erros do ajuste linear dos dados empíricos garante-nos um modelo linear ótimo. No entanto, se nossa hipótese de dependência linear estiver incorreta, o modelo estará também. A verificação da hipótese depende de uma miríade de fatores, incluindo conhecimento de domínio, confirmação com dados futuros, mudança na natureza da evolução dos dados, etc.

estimar a posição de um cometa sem a utilização das complicadas equações não-lineares de Kepler. Enunciaremos, sem demonstrar[5], o teorema de Gauss-Markov que asserta que, se nosso modelo for linear e os erros tivessem média nula, variâncias iguais e não estivessem correlacionados, o estimador proveniente desse método é o melhor estimador linear não enviesado[6] dos coeficientes do modelo. Isso implica que, sob hipóteses razoavelmente gerais em relação ao erro, a utilização da abordagem dos mínimos quadrados é indicada.

Passando agora para a descrição matemática do método, consideremos um conjunto de dados de duas variáveis, $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Suponhamos que y_i seja a realização de uma variável aleatória Y_i e que possamos escrever

$$Y_i = \alpha + \beta x_i + \xi_i \qquad i = 1, 2, \cdots, n, \tag{1.1}$$

em que os ξ_i's são variáveis independentes com média zero e variância σ^2. O problema resume-se em estimar os parâmetros α, β e σ^2. Para isso, escrevemos a seguinte quantidade S, conhecida como soma dos quadrados:

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 . \tag{1.2}$$

O método passa por escolher α e β tal que S seja um mínimo, um problema de extremização assim como o método da máxima verossimilhança que será apresentado na próxima seção. A interpretação geométrica dessa soma é que seu i-ésimo termo é o quadrado da distância vertical a partir de (x_i, y_i) até a linha $y = \alpha + \beta x$.

Para obtermos α e β que minimizam S, empregaremos as condições necessárias de mínimo, $\partial S/\partial\alpha = 0$ e $\partial S/\partial\beta = 0$. Diferenciamos, portanto, S em relação a α e β e igualamos a zero:

$$\frac{\partial S}{\partial \alpha} = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \tag{1.3}$$

e

[5] O leitor interessado pode encontrar a demonstração em diversas fontes: no trabalho original de Gauss (GAUSS, 1823); em (WIKIPEDIA, 2021), sendo um tratamento mais moderno, incluindo generalizações; ou um tratamento com viés histórico no capítulo 3 do livro (FISCHER, 2011).

[6] A importância desse ponto depende da análise estatística que buscamos realizar e é falso que um estimador não enviesado seja sempre desejado. Essa discussão foge do nosso escopo de revisão e estabelecimento de notação, mas aconselhamos fortemente o leitor interessado a ler o capítulo 17 do *Probability theory*, escrito por Jaynes (JAYNES, 2003), que pode ser encontrado *online* livremente (https://www-biba.inrialpes.fr/Jaynes/cc17h.pdf).

$$\frac{\partial S}{\partial \beta} = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)\, x_i = 0. \tag{1.4}$$

Distribuindo as somas, podemos reorganizar essas duas equações em um sistema para α e β, a saber:

$$n\alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{1.5}$$

e

$$\alpha \sum_{i=1}^{N} x_i + \beta \sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} x_i y_i. \tag{1.6}$$

Como assumimos que temos um conjunto de dados de (x_i, y_i), podemos resolver diretamente esse sistema de duas equações para as incógnitas α e β. Assim procedendo, verificamos explicitamente as fórmulas para α (coeficiente linear) e β (coeficiente angular):

$$\alpha = \overline{y}_n - \beta \overline{x}_n. \tag{1.7}$$

e

$$\beta = \frac{n \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n \sum {x_i}^2 - \left(\sum x_i\right)^2}. \tag{1.8}$$

Aqui, escrevemos $\sum$ ao invés de $\sum_{i=1}^{n}$ por simplicidade de notação e porque a barra sobre a grandeza indica sua média, e.g., $\overline{y}_n = \left(\sum y_i\right)/n$. É usual, em textos estatísticos, a notação com acento circunflexo para indicar que a quantidade em questão é um estimador, por exemplo, os lados esquerdos das equações (1.8) e (1.7) são escritos como $\hat{\beta}$ e $\hat{\alpha}$, respectivamente.

Para finalizar, no caso de um modelo de regressão linear simples, um estimador não enviesado para o desvio-padrão é dado pela variável aleatória $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \alpha - \beta x_i\right)^2. \tag{1.9}$$

Não justificaremos essa fórmula por não ser necessária ao longo do trabalho (DEKKING

et al., 2005). Decidimos escrevê-la para completar nossa descrição do método dos mínimos quadrados.

1.4.2 Transformações de potência e máxima verossimilhança

A história da transformação de potência remonta, pelo menos, a mais de sessenta anos, tendo sida introduzida por Tukey (TUKEY, 1957) com o intuito de normalizar dados empíricos. Se x_i for uma variável aleatória positiva, introduz-se $\tilde{y}_i^{(\lambda)}$ via

$$\tilde{y}_i^{(\lambda)} = \begin{cases} x_i^{\lambda} & (\text{se } \lambda \neq 0) \\ \log x_i & (\text{se } \lambda = 0) \end{cases} . \tag{1.10}$$

Para facilitar o tratamento matemático e por considerações de homoscedacidade[7], Box e Cox (BOX; COX, 1964) propuseram uma transformação linear da transformação de potência válida para quaisquer $x_i > 0$ e $\lambda \in \mathbb{R}$:

$${y_i}^{(\lambda)} = \begin{cases} \frac{{x_i}^{\lambda}-1}{\lambda} & (\text{se } \lambda \neq 0) \\ \log x_i & (\text{se } \lambda = 0) \end{cases} . \tag{1.11}$$

Notemos que a transformação (1.11), quando comparada com a equação (1.10), promove uma passagem contínua de potência para logaritmo, ao considerar o limite $\lambda \to 0$.

Os autores sugeriram que a transformação aprimora modelos estatísticos porque remove a) não-linearidades, b) heteroscedacidade[8] e c) assimetrias (*skewness*) e a não-normalidade dos erros. Nas décadas seguintes, outras transformações foram propostas, como a de Bickel-Doksum (BICKEL; DOKSUM, 1981) e a de Yeo-Johnson (YEO; JOHNSON, 2000). Como essas transformações não foram utilizadas no trabalho, não nos delongaremos em seus detalhamentos. No entanto, o leitor interessado pode encontrar uma revisão e uma análise matemática em (SOLEYMANI, 2018). Para desenvolver e reobter os resultados centrais da transformação Box-Cox, seguiremos o supracitado trabalho original de Box e Cox e a análise realizada por Spitzer (SPITZER, 1982).

Começamos pela definição de dois conjuntos, $\{x_i\}$ e $\{{y_i}^{(\lambda)}\}$, como os conjuntos de

[7] Um conjunto de variáveis aleatórias é homoscedástica se todas as variáveis aleatórias têm a mesma variância e se ela é finita. Para detalhes, ver (KENNEDY, 2008).

[8] Um conjunto de variáveis aleatórias é heteroscedástica se esse conjunto não observar a hipótese de homoscedacidade, ou seja, se nem todas as variáveis aleatórias tiverem a mesma variância (ou se ela não for finita). Para mais detalhes, o livro (DOUGHERTY, 2011) tem um capítulo dedicado à heteroscedacidade.

variáveis não-transformadas e transformadas, respectivamente. Em seguida, escrevemos a log-função de verossimilhança para o conjunto não-transformado, isto é,

$$l(x_1, \cdots, x_n) = \ln [\tilde{f}(x_1, \cdots, x_n)]. \tag{1.12}$$

Apesar da relativa simplicidade da equação (1.12), precisamos deixar claro que esse é um modelo aditivo, visto que a função original (não logarítmica) de verossimilhança é construída a partir do produtório da função densidade de n observações independentes,

$$\tilde{f}(x_1, \cdots, x_n) = \prod_{j=1}^{n} f(x_j). \tag{1.13}$$

A função $\tilde{f}$ é uma probabilidade (ou uma densidade de probabilidade), um produto das probabilidades $f(x_i)$'s, e indica a probabilidade da distribuição f representar o conjunto de dados $\{x_i\}$. Em geral, $f(x)$ depende de um conjunto de parâmetros que devem ser otimizados para obter a maior probabilidade $\tilde{f}$, a máxima verossimilhança. Um exemplo de distribuição de probabilidade $f(x)$ é a gaussiana (normal),

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{1.14}$$

em que a média μ e o desvio-padrão σ são os parâmetros que podem ser vistos como de otimização. Ressaltamos que é comum empregar $N(\mu, \sigma)$ para indicar essa distribuição de probabilidade.

Quando empregamos a transformação Box-Cox, imaginamos que a distribuição de probabilidade na nova variável ${y_i}^{\lambda}$ seja gaussiana. Assim, segundo o modelo, a função densidade $f_{(\lambda)}$ da variável transformada também será gaussiana para um λ conveniente. Portanto, podemos escrever

$$\begin{aligned} \ln[f_{(\lambda)}(y_1^{(\lambda)}, \cdots, y_n^{(\lambda)})] &= \ln[f({y_1}^{(\lambda)})] + \cdots + \ln[f({y_n})^{(\lambda)}] \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{1.15}$$

Notemos que o papel da média μ é assumido por $\mathbf{X}\boldsymbol{\beta}$. Além disso, a omissão do somatório e a notação matricial antes e a seguir são comuns na literatura estatística. Por isso, decidimos mantê-la, a despeito de sua relativa opacidade.

A relação entre as duas funções de verossimilhança, $\tilde{f}$ e $f_{(\lambda)}$, é

$$\tilde{f}(x_1,\ldots,x_n) = f_{(\lambda)}(y_1^{(\lambda)},\ldots,y_n^{(\lambda)})|\mathbf{J}|, \tag{1.16}$$

em que $|\mathbf{J}|$ é o determinante da matriz jacobiana da transformação Box-Cox. Esse determinante pode ser calculado observando que os elementos diagonais da matriz $\mathbf{J}$ são $\partial y_i^{(\lambda)}/\partial x_i = x_i^{\lambda-1}$ e os não-diagonais são nulos, pela hipótese de independência. Assim, é imediato obter que

$$|\mathbf{J}| = \prod_{i=1}^{n} x_i^{\lambda-1}. \tag{1.17}$$

Temos interesse na função verossimilhança do dado empírico original ou, mais precisamente, na obtenção da função verossimilhança após aplicar a transformação, mas a partir dos dados originais. Dessa maneira, combinando as equações (1.15), (1.16) e (1.17) com a equação (1.12) e denotando $\ln \tilde{f}(x_1,\cdots,x_n)$ por $l(\lambda,\beta,\sigma^2)$, podemos escrever

$$l(\lambda,\boldsymbol{\beta},\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta}) + (\lambda-1)\sum_{i=1}^{n}\log(y_i) \tag{1.18}$$

que é uma função dependente de λ explicitamente.

Como $|\mathbf{J}|$ depende de λ, é evidente que escolhas diferentes desse parâmetro implicarão funções de verossimilhanças distintas. Embora o objetivo seja maximizar a equação (1.18), não fica imediatamente claro como proceder. Uma maneira de contornar o problema é supor que λ seja fixo. Para cada λ, o ajuste do modelo linear resulta em

$$\hat{\beta}(\lambda) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}^{(\lambda)}, \tag{1.19}$$

$$RSS(\lambda) = (\mathbf{Y}^{(\lambda)})^T(\mathbf{I_X})\mathbf{Y}^{(\lambda)} \tag{1.20}$$

e

$$\hat{\sigma}^2(\lambda) = \frac{RSS(\lambda)}{n}, \tag{1.21}$$

em que $\mathbf{I_X}$ representa a matriz identidade e RSS a soma residual de quadrados (*residual*

sum of squares, em inglês). Assim, a função de máxima verossimilhança de perfil[9] para o parâmetro λ é dada por

$$l_{max}(\lambda) = c - \frac{n}{2}\log\left[\frac{RSS(\lambda)}{n}\right] + (\lambda - 1)\sum_{i=1}^{n}\log y_i, \tag{1.22}$$

em que c é uma constante. Tratamos essa função como normalmente tratamos log-funções de verossimilhança: valores de λ próximos ao valor de maximização $\hat{\lambda}$ são consistentes com os dados. A equação (1.22) tem uma forma computacionalmente eficiente (afinal, o cálculo dos resíduos é computacionalmente barato), sendo a função utilizada para a estimativa do parâmetro λ ao longo do trabalho.

Para finalizarmos este capítulo, comentaremos uma relação entre os dois métodos aqui apresentados, seguindo novamente (DEKKING et al., 2005) na nossa exposição. Na aplicação do método dos mínimos quadrados, não é necessário conhecer a distribuição de erros. Por outro lado, se conhecermos essa distribuição, o princípio de máxima verossimilhança se aplica. Considere a situação em que os erros ξ_i na equação (1.1) são independentes com distribuição $N(0, \sigma^2)$. O problema pode ser enunciado como: encontrar as estimativas da máxima verossimilhança para α e β. Nesse caso, os y_i's são independentes e têm uma distribuição normal (1.14) com μ dado por $\alpha + \beta x_i$, $N(\alpha + \beta x_i, \sigma^2)$. A função densidade de probabilidade pode ser escrita como

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(y - \alpha - \beta x_i)^2}{2\sigma^2}\right]. \tag{1.23}$$

Como

$$\ln\left[f(y_i)\right] = -\ln\sigma - \ln\sqrt{2\pi} - \frac{1}{2}\left(\frac{y - \alpha - \beta x_i}{\sigma}\right)^2, \tag{1.24}$$

a verossimilhança logarítmica é dada por

[9] *Profile likelihood* em inglês. É um procedimento por meio do qual reduzimos as dimensões do problema, concentrando a função de verossimilhança para um subconjunto de parâmetros, expressando os parâmetros-empecilhos como funções dos de interesse e substituindo-os na função de verossimilhança (TAKESHI; AMEMIYA, 1985; GOURIEROUX; MONFORT, 1995).

$$\begin{aligned} l(\alpha, \beta, \sigma) &= \ln[f(y_i)] + \cdots + \ln[f(y_n)] \\ &= -n\ln(\sigma) - n\ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 . \end{aligned} \tag{1.25}$$

Podemos perceber que, para um $\sigma > 0$ fixo, essa verossimilhança é máxima exatamente quando a soma

$$\sum_{i=1}^{n}(y - \alpha - \beta x_i)^2 \tag{1.26}$$

for mínima, ou seja, quando a função S, dada na equação (1.2), é mínima. Dessa maneira, no caso em que os erros ξ_i's sejam independentes e obedeçam a tal distribuição, o princípio de máxima verossimilhança e o método dos mínimos quadrados fornecem os mesmos estimadores β e α, dados nas equações (1.8) e (1.7), respectivamente.

Resta-nos apenas encontrar o estimador de máxima verossimilhança para σ. Derivando a equação (1.25) em relação a σ, obtemos

$$\frac{\partial}{\partial\sigma} l(\alpha, \beta, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 \tag{1.27}$$

e, ao empregarmos $\partial l/\partial\sigma = 0$, chegamos a

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 . \tag{1.28}$$

Pelo princípio de invariância, o estimador de máxima verossimilhança é o mesmo da equação (1.9) quando o número de dados n é muito grande. Como mencionado anteriormente, esse estimador tem menor relevância para nosso trabalho, razão pela qual foi apresentado de maneira sucinta.

2 Materiais

Este capítulo é dedicado a uma explanação de alguns aspectos técnicos da nossa investigação, a saber, a plataforma escolhida, a natureza dos dados, o volume de dados estudado e as partes do código desenvolvido ao longo do estudo, bem como o processo necessário para replicar os resultados apresentados nos capítulos 3 e 4. Assim, a exemplo do que afirmamos a respeito do capítulo 1, este também pode não ser lido em uma primeira passagem pela tese ou ser considerado como referência para a leitura dos capítulos subsequentes. Sugerimos, no entanto, que o leitor atente para a primeira seção, mais curta, visto que ela conterá uma descrição resumida, de todo o dado utilizado, fornecendo também suporte para a decisão na escolha de uma plataforma específica. A segunda seção é dedicada a uma apresentação qualitativa do código. Por sua vez, por natureza mais árida e específica, o código na íntegra encontra-se no apêndice A, que pode ser facilmente desconsiderado em uma primeira leitura.

2.1 Obtenção e tratamento dos dados

Em nossa investigação sobre a popularidade dos *video games*, empregamos dados da plataforma digital, a *Steam*. Essa plataforma é uma distribuição de mídia digital desenvolvida pela *Valve Corporation* e uma loja de jogos *online* responsável por gerenciar os direitos autorais dos jogos em seu banco de dados (VALVE, 2020a). A *Steam* atualmente tem mais de noventa milhões de usuários (TAYLOR, 2018), alcançando mais de vinte milhões de usuários *online* simultaneamente (VALVE, 2020b). Infelizmente, em razão do anonimato desses usuários, não podemos determinar seus países e não podemos estimar a população de jogadores por país. No entanto, os dados são uma amostra representativa da população global de jogadores, já que os dados de uso de rede por país são disponibilizados, o que torna possível verificar que há uso de rede em todos os cinco continentes, sendo a China e os Estados Unidos os maiores participantes. Também esclarecemos que a *Steam* já era a maior plataforma de distribuição digital para jogos de computador em 2013 (EDWARDS,

2013) em número de usuários. Hoje, na *Steam*, existem mais de cinquenta mil *video games* (VALVE, 2021).

Como já mencionamos, os dados do usuário são anônimos e não identificáveis. Por exemplo, não é possível saber quem está jogando, apenas existem informações de quantos usuários estão jogando cada jogo a cada hora. Eventualmente, o mesmo usuário pode jogar mais de um jogo simultaneamente. Por exemplo, se um usuário jogar dois jogos distintos com menos de uma hora de diferença, ele será contado como jogador de cada um dos dois jogos. Do número de usuários por jogo por hora, podemos obter dados derivados como o número médio de usuários por hora em cada mês, conforme apresentado no *site* Steamcharts (GRAY, 2020). Como nossa análise é limitada a jogos indexados na Steamcharts, foram investigados cerca de vinte e dois mil jogos. Como os dados da *Steam* usados aqui não são relatados pelos usuários, pode-se esperar uma redução nos aspectos subjetivos da análise. Outra espécie de dado empregado na análise é a das meta-informações de cada jogo (data de lançamento, categorias, etc.) na *Steam* (VALVE, 2020a). A partir desses dados, foi obtido o número de jogos lançados a cada mês e o número de jogos temporalmente distintos por categoria.

Tabela 1: **Resumo dos dados utilizados neste trabalho**. A tabela resume alguns aspectos dos dados coletados e do banco de dados estudado: número de jogos, número de categorias (*tags*), tempo englobado pela coleta (Δt), tempo inicial t_i, tempo final t_f, total da média de jogadores por mês durante todo o período ρ_{global} e total da média de jogadores por mês durante todo o período para jogos na categoria *Indie*, ρ_{Indie}.

Total de jogos: 21752	Número de *tags*: 1044
Δt: 8,5 anos	t_i: Julho de 2012
t_f: Dezembro de 2020	ρ_{global}: $1{,}78 \times 10^8$
ρ_{Indie}: $2{,}46 \times 10^7$	

Todos esses dados correspondem ao período de julho de 2012 até dezembro de 2020. Os jogos lançados antes de 2012 e os jogos sem série temporal foram excluídos. Quanto aos dados do usuário por jogo por hora, a *Steam* os fornece livremente apenas a partir

da hora atual. Portanto, para criar um banco de dados maior, é necessário que haja um acúmulo sistemático de dados ao longo de um determinado período. Embora a plataforma *Steam* tenha começado em setembro de 2003, esse processo de armazenamento de dados só foi disponibilizado livremente pela Steamcharts de 2012 até os dias atuais, o que justifica a data de início para os dados usados no estudo. Por sua vez, os dados foram baixados e analisados usando *scripts* em Python e via a API da *Steam*. Observamos que o método de coleta de dados obedeceu aos termos e condições do *website*. Além disso, o compartilhamento dos dados também está em conformidade com os termos e condições. Para resumir, um esboço dos dados empregados nesse trabalho é fornecido na tabela 1.

2.2 Apresentação do código

Antes de iniciarmos a apresentação propriamente dita, devemos salientar que as subseções mencionadas ao longo desta seção encontram-se no apêndice A. Para a realização deste trabalho, houve necessidade de desenvolver programas para as diversas etapas. Em primeiro lugar, precisávamos de uma lista dos jogos disponíveis na *Steam* indexados de uma maneira única. Percebemos que, para cada jogo, havia um código numérico. Com base nisso, a primeira etapa computacional foi varrer o *site* Steam DB, que contém uma lista com todos os aplicativos da plataforma com seu respectivo código. Dentre esses aplicativos, separamos aqueles marcados como *Game* e os comparamos com a ferramenta de busca da própria *Steam* (VALVE, 2021) para verificar a concordância do número de jogos. O código utilizado para essa extração encontra-se no

apêndice A, *A*.1.1.

Munidos dessa lista, varremos o *site* da Steamcharts (GRAY, 2020) para obter, a partir do código único de cada jogo, a evolução temporal do número médio mensal de jogadores por hora, nossa principal métrica e o dado bruto do trabalho. Dividimos o código em duas partes: a primeira serviu para extrairmos as informações de um *site* da Steamcharts e a segunda foi um simples laço sobre a lista. Dessa forma, o código para extrairmos informações encontra-se no

apêndice A, *A*.1.2

e o laço, no

apêndice A, *A*.1.3.

Infelizmente, nem todo jogo listado estava disponível nesse último *site*. Os motivos são diversos, mas podemos compreender alguns deles: jogos que ainda não foram lançados (acesso antecipado ou pré-venda), jogos listados privadamente (ainda em desenvolvimento ou removidos da loja por algum motivo) ou ainda indexação sem venda (jogo anunciado, mas ainda não disponível para compra). Assim, foram obtidas vinte e um mil setecentos e cinquenta e duas séries temporais distintas.

Além dessas etapas, utilizamos a lista de códigos únicos disponíveis na Steamcharts para acessar a *Steam* e extrair outras informações de cada jogo, como sua data de lançamento, as categorias votadas pelos usuários, as categorias definidas pelos desenvolvedores e alguns outros dados que não foram úteis para a execução desse trabalho. O código para esse acesso pode ser encontrado no

apêndice A, *A*.1.4,

em que *dates2021.txt* é um arquivo de texto que lista os indexadores únicos encontrados no *site steamdb* e *TAGLIST* é uma lista com as categorias de jogos construída a partir do código apresentado no

apêndice A, *A*.1.5.

Nossos dados, então, foram organizados em colunas da seguinte maneira: o indexador único (*AppID*), a data de lançamento (*Date*), as categorias votadas pelos usuários (*Tags*), as categorias definidas pelos desenvolvedores (*devTags*) e a série temporal do jogo em questão (*PlayerTimeSeries*).

Para construir o *dataframe* e verificar a sensatez dos dados e excluir entradas espúrias, utilizamos o código exibido no

apêndice A, *A*.1.6.

Com o intuito de evitar repetição, decidimos importar todas as dependências necessárias para os códigos utilizados na preparação das figuras. Por meio do código acima, removemos jogos lançados antes de 2012 e jogos com série temporal vazias. O número de entradas restantes depois desse processo foi de vinte e um mil setecentos cinquenta e duas, organizados em uma tabela com as entradas descritas. Arquivos nos formatos .ipynb e .csv, os quais contêm o dado que serviu de base para todas nossas análises, estão disponíveis em `https://gitlab.com/tdfb/steam-data-2020/`, além de um passo-a-passo resumido do que já foi descrito nesta seção.

Passando para a análise quantitativa, para obtermos as duas primeiras figuras do trabalho, o código utilizado encontra-se no

apêndice A, *A*.1.7.

Notamos que, para a obtenção dos ajustes linear e exponencial, o algoritmo utilizado baseia-se no método dos mínimos quadrados, descrito no capítulo 1. Para a obtenção das figuras 6 e 7, nenhum ajuste foi necessário, apenas a transformação logarítmica. O código correspondente a essas figuras encontra-se no

apêndice A, *A*.1.8.

A obtenção das figuras 8 e 9 requereu passos mais complexos do que os das figuras anteriores. Para começar, separamos os jogos por seu mês de lançamento e os agrupamos em trimestres (por exemplo, o primeiro trimestre contém jogos lançados em julho de 2012, agosto de 2012 ou setembro de 2012) e escolhemos aleatoriamente um desses trimestres para a análise. Realizamos as transformações logarítmica e de Box-Cox e, por meio do método da máxima verossimilhança, também descrito no capítulo 1, encontramos o valor numérico para o parâmetro λ ótimo. É importante salientar que, antes das transformações, normalizamos o dado, removendo a média e dividindo pelo desvio-padrão, com o intuito de facilitar a comparação com a distribuição normal de média zero e desvio-padrão unitário. Além dessa transformação para todos os dados do trimestre, analisamos a evolução temporal do parâmetro ótimo (e do seu intervalo de confiança por meio do método de *bootstrap*) do grupo trimestral para cada intervalo trimensal. Assim, o primeiro ponto na figura 9 refere-se ao parâmetro ótimo da transformação Box-Cox aplicada ao grupo trimestral 01/2015 nos meses 1, 2 e 3 de 2015; o segundo ponto refere-se ao parâmetro ótimo da transformação aplicada ao mesmo grupo trimestral, mas nos meses subsequentes, meses 4, 5 e 6 e assim sucessivamente. Como mencionamos, o intervalo de confiança foi calculado via *bootstrap*[1] com amostragem de mil elementos. O código para obtenção dessas figuras pode ser verificado no

apêndice A, *A*.1.9.

Continuando, o processo de obtenção da figura 11 é similar ao da figura 8, com a

[1] A ideia fundamental do *bootstrap* é a de que a inferência sobre uma população a partir de dados amostrais pode ser modelada reamostrando os dados e realizando inferências sobre uma amostra a partir desses dados. O leitor interessado pode encontrar uma revisão em (HENDERSON, 2005).

importante diferença de que não depende da separação inicial em grupos trimestrais, isto é, o processo é o mesmo, utilizando todos os dados sem nenhum agrupamento. No entanto, para se obter a figura 10, estimamos o parâmetro ótimo para cada um dos grupos trimestrais, sem a evolução temporal desses. O intervalo de confiança foi também calculado via *bootstrap* com amostragem de mil elementos. O código pertinente para fazer as transformações necessárias e obter as figuras 10 e 11 pode ser encontrado no

apêndice A, *A*.1.10.

No caso do capítulo 4, figuras 12, 13, 14, 15 e 16, o processo é essencialmente o mesmo, mas considerando apenas determinada categoria de jogos. Como exemplo, o código necessário para a reobtenção dos resultados no caso *Indie* encontra-se no

apêndice A, *A*.1.11.

Na obtenção dos dados necessários para a análise do número de categorias ao longo do tempo, foi preciso organizar os jogos por mês de lançamento e, depois, separá-los por categorias. Dessa forma, podemos responder a perguntas como "quantos jogos de ação existiam na *Steam* em agosto de 2013?". O código utilizado está no

apêndice A, *A*.1.12.

De posse desses dados, no ajuste linear da figura 17, foi utilizado o método dos mínimos quadrados, como esperado. Ademais, como o número de categorias é menor que o de jogos, utilizamos intervalos semestrais para evitar o cálculo da função densidade de probabilidade com poucos dados. O código para a obtenção das figuras 17, 18 e 19 está disponível no

apêndice A, *A*.1.13.

3 Análise global dos dados

Começamos este capítulo com algumas hipóteses relacionadas ao nosso estudo sobre *video games*. A seguir, elaboramos uma análise exploratória, cujo objetivo é explicar o conjunto de dados, que caracterizou o aumento da quantidade de jogadores e de jogos temporalmente. Além desses dois aspectos envolvendo a popularidade dos *video games*, fomos motivados a propor um modelo para a popularidade relativa lognormal. Como observamos desvios desse modelo, avançamos para um modelo que refinou a suposição inicial, empregando uma transformação Box-Cox. Nossa investigação, assim como os capítulos subsequentes, segue a referência (MENDES; CUNHA; MENDES, 2022).

3.1 Hipóteses

Tendo em vista a discussão apresentada no capítulo 1, avançaremos na análise considerando as seguintes hipóteses a respeito da popularidade dos jogos e de seu crescimento ao longo do tempo:

H1 **O número de *video games* e de jogadores está crescendo linear ou exponencialmente ao longo do tempo**. Dado que os *video games* se tornaram mais presentes nos últimos anos e que os motores e as ferramentas de desenvolvimento (*middleware*) tornaram mais acessível o desenvolvimento de um jogo, é possível esperar que esses números cresçam e, em uma primeira aproximação, cresçam com padrões simples, como o linear ou exponencial;

H2 **A popularidade relativa dos jogos é descrita por uma distribuição que se assemelha à lognormal**. Essa hipótese foi impulsionada pelo fato de que a distribuição foi detectada e estudada no contexto da análise de popularidade de outros produtos culturais, como citações de patentes, citações científicas, entradas da Wikipedia e memes em redes sociais (RATKIEWICZ et al., 2010; WANG; SONG; BARABÁSI, 2013; HIGHAM et al., 2017; YOOK; KIM, 2020);

H3 **As distribuições discutidas em H2 são estáveis ao longo do tempo**. Dado que, geralmente, um *video game* é mais popular no primeiro mês ou nos primeiros meses após seu lançamento, presumimos ser improvável que um número suficiente de jogos se tornem populares anos após seus lançamentos para perturbar a distribuição de popularidade de maneira significativa;

H4 **As distribuições de popularidade para as categorias principais serão semelhantes entre si e a global**. A natureza da popularidade de *video games*, representada pela distribuição de probabilidade, pode independer de sua categoria. A hipótese é de que o fato de os produtos serem jogos contribui mais para um mesmo comportamento de consumo do que a diferença interna entre suas categorias;

H5 **O número de categorias crescerá mais lentamente em comparação com o número de jogos e de jogadores**. Visto que uma nova categoria implica, muitas vezes, a criação de um novo gênero de *video game*, a hipótese é de que essa criação seja muito mais esporádica do que o influxo de novos jogos e jogadores no banco de dados.

3.2 Análise de crescimento

Como já mencionamos em nossa investigação sobre a popularidade dos video games, focamos, inicialmente, características bastante gerais do conjunto de dados, as exibidas nas figuras 4 e 5. A figura 4 representa a evolução temporal do número médio de jogadores por hora em cada mês na *Steam*. Em uma hora de dezembro de 2019, por exemplo, alguns milhões de usuários conectados à *Steam* estavam em algum jogo. A figura 5 apresenta o número de jogos lançados mensalmente na *Steam*. Exemplificando, setecentos e vinte e sete jogos foram lançados em nosso conjunto de dados em dezembro de 2020. Os dados vão de julho de 2012 a dezembro de 2020 e estão organizados de forma que o primeiro mês (Mês 1) represente julho de 2012, o segundo mês (Mês 2) represente agosto de 2012 e sucessivamente.

Na figura 4 notamos uma tendência de crescimento linear médio, com a taxa de cerca de trinta e sete mil jogadores por hora por mês. Por sua vez, a figura 5 também mostra uma tendência de crescimento, mas com um padrão diferente do linear. Apesar de discrepâncias, um arquétipo que parece ser mais conveniente para o caso da figura 5 é o exponencial. Um ajuste dos jogos lançados na *Steam* por meio desse modelo conduz

a uma taxa de crescimento intrínseco[1] de aproximadamente 0,01/mês. Particularmente, pode-se perceber na figura 5 que há um desvio sensível do comportamento exponencial a partir do mês 80, aproximadamente. Esse desvio, sistematicamente menor que os valores previstos pelo ajuste exponencial, pode ser indício de uma saturação na produção de jogos. Nesse caso, o mercado já estaria produzindo um volume de jogos que tende ao suficiente para atender à demanda dos usuários e a competição entre os lançamentos impediria o aumento na produção. Uma discussão relacionando crescimentos linear e exponencial será retomada no capítulo 5, considerando um modelo malthusiano.

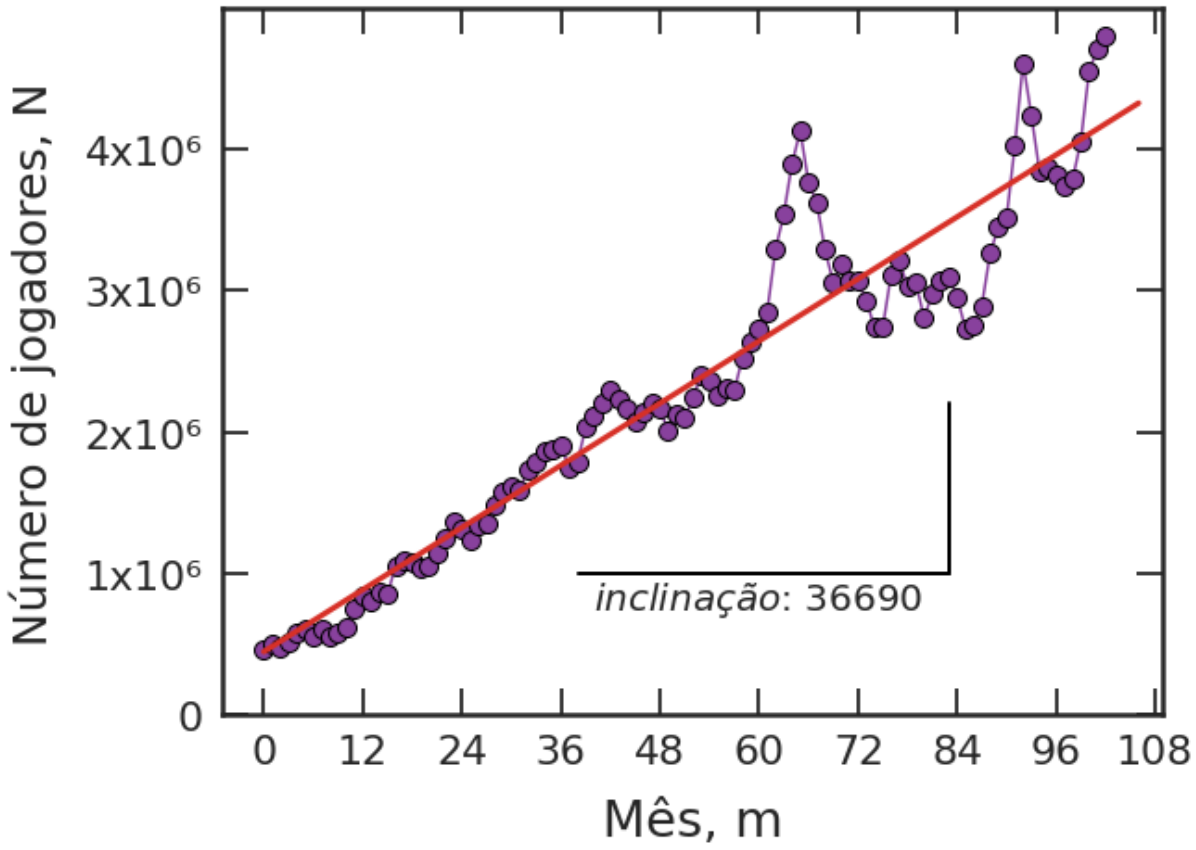


Figura 4: **Evolução do número de jogadores.** Evolução temporal mês-a-mês do número médio de jogadores ativos por hora na *Steam*. Os discos roxos representam o dado empírico enquanto o ajuste linear (linha vermelha), obtido por meio do método dos mínimos quadrados, é governado pela equação $N = 441\,618 + 36\,690\,m$.

3.3 Médias e desvios-padrão

Uma forma de investigar mais profundamente a popularidade dos *video games* é pela divisão dos dados por trimestre de lançamento (por exemplo, jogos lançados em janeiro,

[1] Em um modelo de crescimento exponencial, ditado pela equação $dy/dx = \alpha y$, a taxa de crescimento intrínseco corresponde ao parâmetro α.

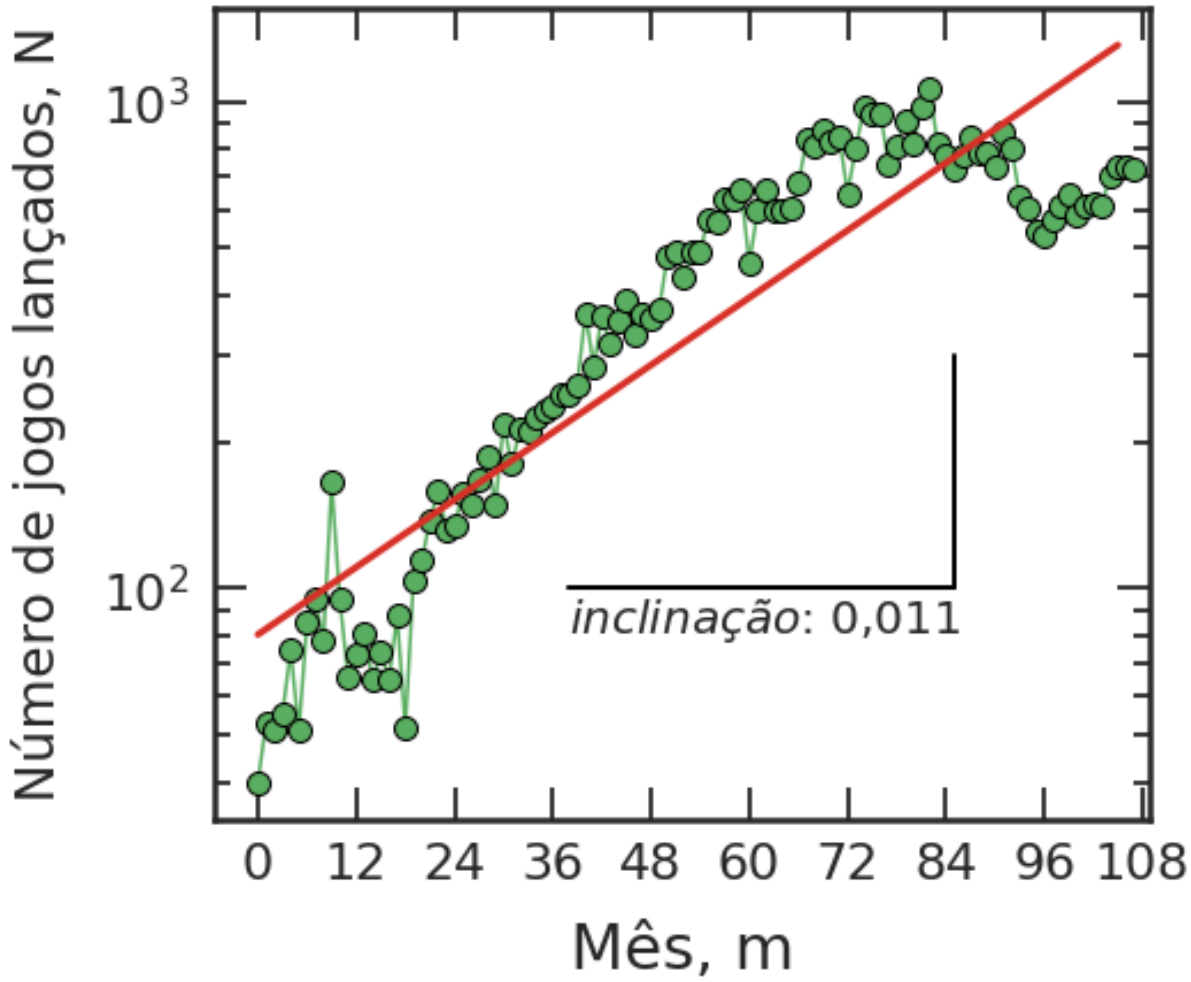


Figura 5: **Evolução do número de jogos lançados.** Evolução temporal mês-a-mês dos jogos lançadas na *Steam*. Os discos verdes representam o dado empírico enquanto o ajuste exponencial (linha vermelha na escala log), obtido por meio do método dos mínimos quadrados, é dado por $\ln N = 1{,}90 + 0{,}011\, m$.

fevereiro ou março do mesmo ano pertenceriam a um mesmo conjunto). Esse critério de coleta de dados coincide com o lançamento sazonal de jogos. Tais períodos de amostragem também correspondem a relatórios trimestrais para investidores e acionistas. Por outro lado, caso tivéssemos escolhido uma janela de agrupamento mensal, teríamos poucos jogos por janela, principalmente nos primeiros meses de nosso banco de dados. Para cada conjunto trimestral de dados, analisamos a evolução temporal de facetas de popularidade nos meses subsequentes. As figuras 6 e 7 mostram a evolução temporal das médias e desvios-padrão, em que cada cor corresponde a um agrupamento trimestral distinto. Usamos o mês um como o primeiro mês de cada grupo, consequentemente, os conjuntos de jogos lançados mais recentemente têm séries temporais mais curtas, com a menor série tendo dois meses. O comportamento médio dos diferentes grupos varia muito em todos os intervalos, o que é marcantemente exemplificado pelas três curvas superiores na figura

6A. Um comportamento semelhante é também observado com o desvio-padrão, mas apenas em uma escala diferente (figura 7A). É pertinente observar que a presença de valores discrepantes (*outliers*) dificulta a identificação de um padrão geral em uma faixa limitada para todas as médias e desvios-padrão.

Um procedimento que pode favorecer a identificação de um padrão universal relacionado às figuras 6A e 7A pode ser obtido por meio da atenuação dos efeitos causados por *outliers*. Nesse sentido, empregaremos uma nova variável que suaviza as contribuições atípicas. Essa variável é definida como

$$y_i = \log x_i \,, \tag{3.1}$$

em que x_i representa o número médio de jogadores por hora em cada mês no i-ésimo jogo e log se refere ao logaritmo na base e. Utilizando essa nova variável, verificamos de fato uma melhora no comportamento temporal das médias e desvios-padrão no sentido de revelar um comportamento mais constante nos meses subsequentes ao lançamento do jogo. Esse fato é ilustrado nas figuras 6B e 7B.

3.4 Distribuições de probabilidade

Para ir além das médias e desvios-padrão, movimentamo-nos na direção das distribuições de probabilidade. Essa abordagem permite-nos investigar a popularidade relativa dos *video games*, visto que a frequência relativa de ocorrências de *video games* com um número diferente de jogadores pode ser colocada em evidência. Com esse objetivo, primeiro determinamos funções de densidade de probabilidade (PDFs) de y_i. Em razão da constância aproximada de cada média e desvio-padrão discutidos anteriormente, esperamos que as PDFs tenham uma tendência à estabilidade. Depois, prosseguindo na investigação dessas PDFs, agrupamos os jogos em trimestres com base em sua data de lançamento, conforme empregado na discussão de médias e desvios-padrão.

3.4.1 Análise de um grupo trimestral

Para realizar uma análise exploratória dos dados, começamos pelos trimestres de data de lançamento separadamente. O grupo escolhido aleatoriamente a partir dos trinta e quatro trimestres em nosso banco de dados foi aquele com jogos lançados no primeiro trimestre de 2015 (01/2015). Motivados pela constância aproximada das médias e desvios-padrão (figuras 6B e 7B), agrupamos a média mensal de jogadores por jogo por hora em cada

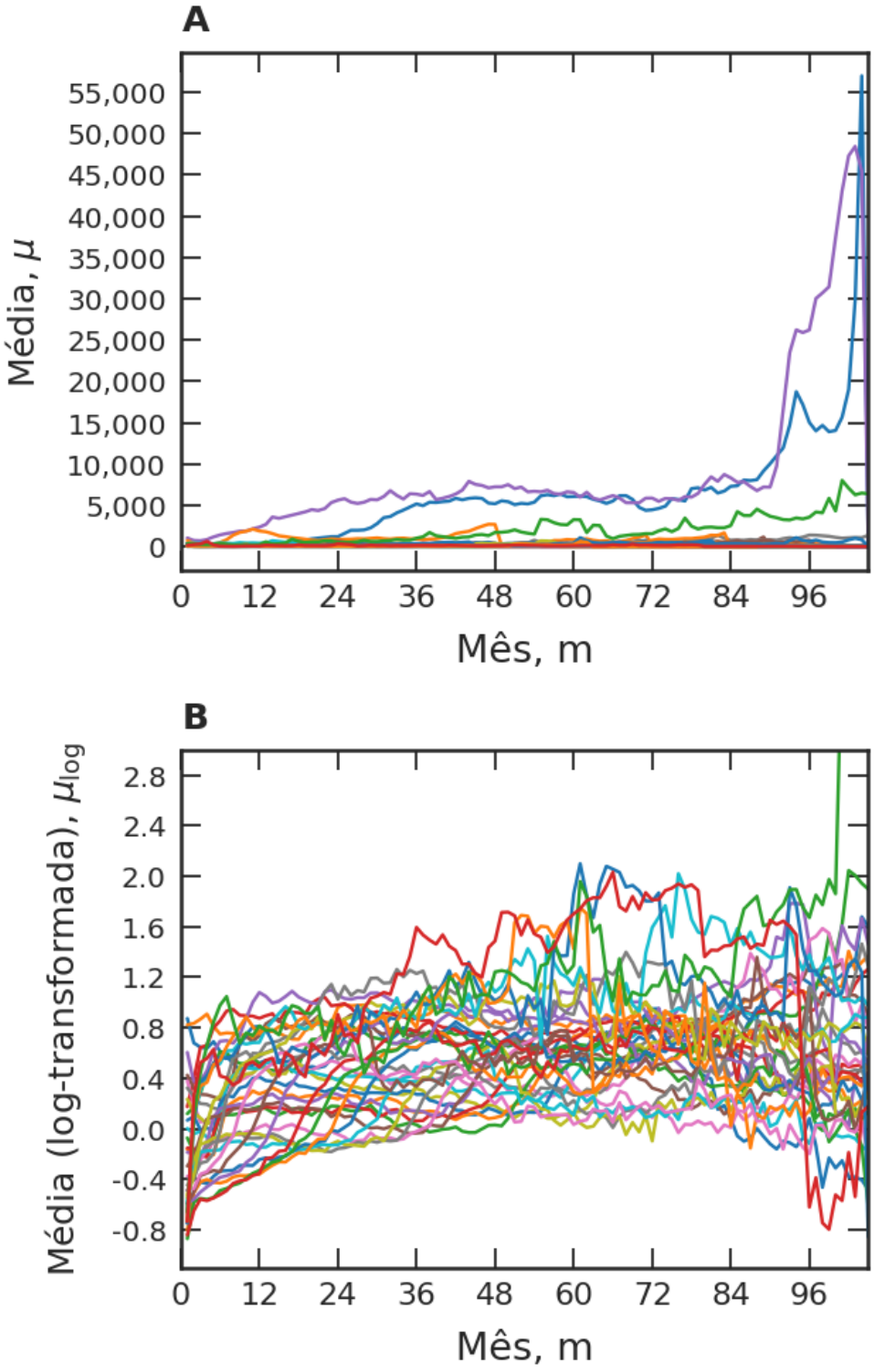


Figura 6: **Médias dos grupos trimestrais.** (**A**) Média do número médio de jogadores por hora em cada mês (x_i) de todos os jogos de cada grupo trimestral ao longo dos meses após o lançamento. (**B**) Usando a variável logarítmica ($y_i = \log x_i$), apresentamos a dependência no tempo das médias correspondentes. Cada um dos trinta e quatro grupos trimestrais tem sua própria cor.

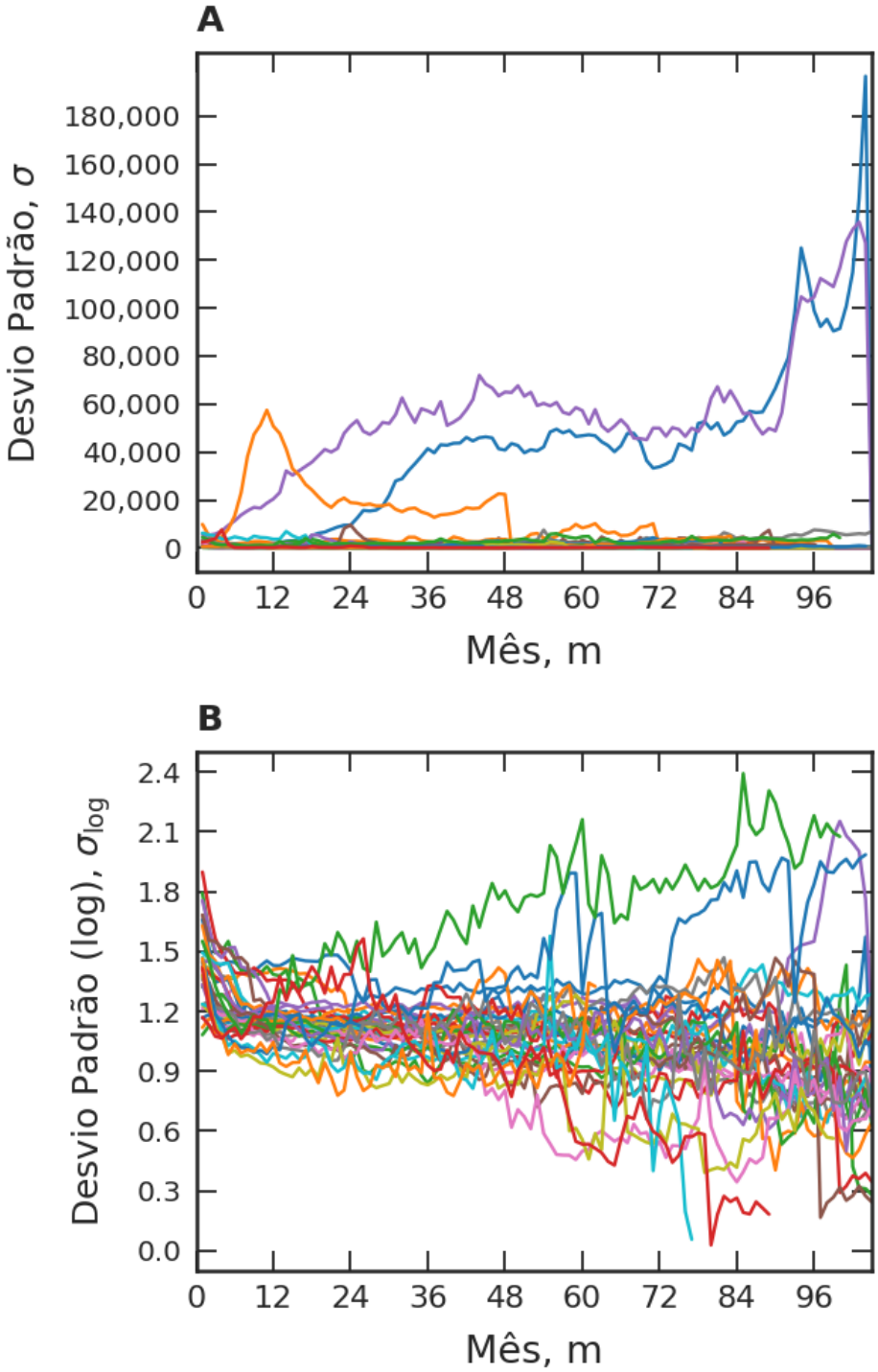


Figura 7: **Desvios-padrão dos grupos trimestrais.** (**A**) Para esses grupos, temos a evolução temporal de seus desvios-padrão. (**B**) Usando a variável logarítmica ($y_i = \log x_i$), apresentamos a dependência no tempo dos respectivos desvios-padrão. Cada um dos trinta e quatro grupos trimestrais tem sua própria cor.

trimestre subsequente até dezembro 2020, o que resultou em vinte e três trimestres de evolução temporal. Como há quatrocentos e vinte jogos no trimestre 01/2015, a quanti-

dade de pontos y_i considerada é 420×24 ($= 10080$). A figura 8A mostra o PDF desses dez mil e oitenta dados na variável z, definida por

$$z_i = \frac{y_i - \langle y_i \rangle}{\sigma_y}, \tag{3.2}$$

com $\langle y_i \rangle$ e σ_y sendo respectivamente a média e o desvio-padrão do conjunto trimestral de y_i's. Segue, dessa definição, que z_i tem média nula e desvio-padrão unitário, isto é, é normalizada.

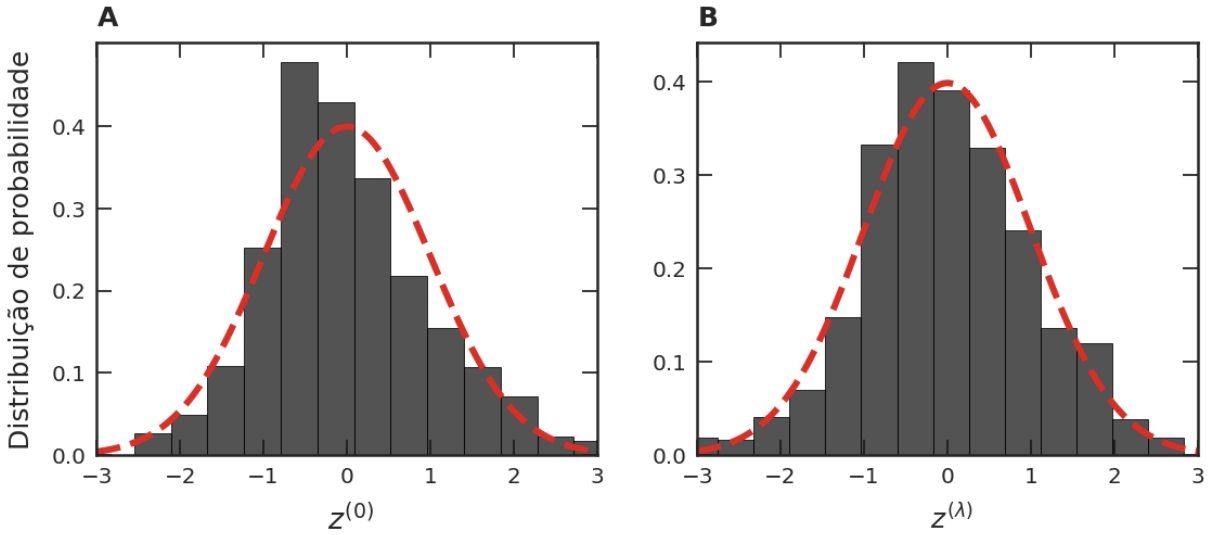


Figura 8: **Análise Box-Cox do grupo trimestral 01/2015.** (**A**) Função de distribuição de probabilidade da variável logarítmica normalizada $z_i = (y_i - \langle y_i \rangle)/\sigma_y$ com $y_i = \log x_i$ para o grupo trimestral 01/2015, em que x_i é o número médio de usuários por hora em cada mês do i-ésimo jogo. (**B**) O mesmo que (**A**), mas com o emprego do valor ótimo λ ($\lambda_{opt} = -0{,}07$) da transformação Box-Cox, equação (3.6). As linhas tracejadas referem-se a uma distribuição normal com a média zero e o desvio-padrão unitário.

Como pode ser visto, o gráfico dos dados da figura 8A se assemelha a uma distribuição gaussiana de média nula e desvio-padrão unitário na variável z:

$$p(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right). \tag{3.3}$$

Essa PDF foi considerada porque a variável z_i, por construção, também tem média nula e desvio-padrão igual a um. Por sua vez, a distribuição de probabilidades na variável original $p(x)$ pode ser obtida a partir da igualdade de probabilidades $p(z)dz = p(x)dx$, ou seja,

$$p(x) = p(z(x)) \frac{dz(x)}{dx}. \tag{3.4}$$

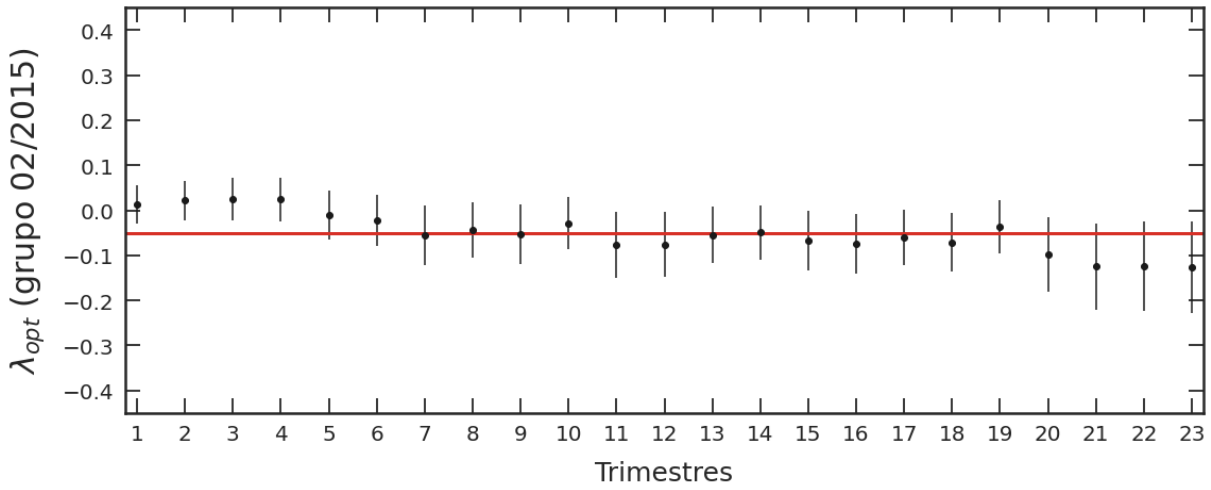


Figura 9: **Evolução do parâmetro λ ótimo para o grupo trimestral 01/2015.** Evolução temporal do parâmetro ótimo λ, λ_{opt}, do grupo trimestral ao longo dos vinte e três trimestres. Os discos representam os λ_{opt}'s ao longo da evolução do tempo do grupo trimestral 01/2015 e a linha reta é a média de todos os trimestres, $\langle\lambda_{opt}\rangle = -0{,}05$, que se aproxima do valor obtido com λ_{opt} calculado por meio do método de máxima verossimilhança e empregado na figura 8B. As barras de erro referem-se ao intervalo de confiança de 95% dos λ_{opt}'s estimados por meio do método de máxima verossimilhança.

Consequentemente, a distribuição na variável original x corresponde a uma lognormal:

$$p(x) = \frac{1}{\sigma_y\sqrt{2\pi}}\frac{1}{x}\,\exp\left[-\frac{(\ln x - \langle y_i\rangle)^2}{2{\sigma_y}^2}\right]. \tag{3.5}$$

Não obstante a semelhança citada, uma assimetria da distribuição dos dados é verificada quando a comparamos à normal de média zero e desvio-padrão unitário. Realmente, os dados estão um pouco para o lado esquerdo da gaussiana na figura 8A. Esse fato sugere a necessidade de uma transformação de variável mais refinada do que o logaritmo [equação (3.1)]. Em casos como esse, a transformação Box-Cox é frequentemente empregada (BOX; COX, 1964; WILSON; NICHOLAS, 2008; TESTI; IVALDI, 2009; ASSIS BOMFIM et al., 2018; YANG et al., 2019). Como pontuamos no capítulo 1, subseção 1.4.2, essa transformação [equação (1.11)] é dada por

$${y_i}^{(\lambda)} = \begin{cases} \frac{{x_i}^{\lambda}-1}{\lambda} & (\text{se } \lambda \neq 0) \\ \log x_i & (\text{se } \lambda = 0) \end{cases}. \tag{3.6}$$

Ressaltamos novamente que, no limite $\lambda \to 0$, o comportamento logarítmico é recuperado e o valor do parâmetro λ indica o grau de desvio desse comportamento.

É oportuno mencionar que, mais de vinte anos após a proposta inicial de Box e Cox, a transformação (3.6) foi considerada de maneira independente no seio da mecânica

estatística generalizada de Tsallis (TSALLIS, 2009). Nesse contexto, foi proposta uma generalização da função logaritmo, a $q-$logaritmo definida como

$$\ln_q(x) = \frac{x^{1-q} - 1}{1 - q}, \tag{3.7}$$

em que q é um parâmetro real. Esse parâmetro indica o grau de desvio da entropia de Shannon-Gibbs, que rotula a entropia de Tsallis, usualmente denotada por S_q. No limite $q \to 1$, a entropia de Shannon-Gibbs é recuperada. Nesse mesmo limite, $\ln_q(x)$ recai no logaritmo usual. Geralmente, o $q-$logaritmo é utilizado em aplicações que requerem um grau de modificação do logaritmo usual, governada pelo parâmetro q. Uma comparação da transformação Box-Cox (3.6) e do $q-$logaritmo (3.7) mostra que ambos são idênticos se $\lambda = 1 - q$. Dessa maneira, em nossa investigação, pensaremos na transformação Box-Cox como uma maneira controlada de deformar a função logaritmo, promovendo um ajuste ótimo aos dados empíricos.

Na subseção 1.4.2, vimos também que um procedimento padrão para chegar a um parâmetro λ ótimo é o método de máxima verossimilhança. O resultado desse procedimento, aplicado aos mesmos dados da figura 8A, pode ser visto na figura 8B. Nessa última figura, usamos a variável normalizada

$$z_i{}^{(\lambda)} = \frac{y_i{}^{(\lambda)} - \langle y_i^{(\lambda)} \rangle}{\sigma_{y^{(\lambda)}}}, \tag{3.8}$$

em que $\langle y_i^{(\lambda)} \rangle$ é o valor médio μ de mais de dez mil $y_i^{(\lambda)}$'s e $\sigma_{y^{(\lambda)}}$ é o desvio-padrão correspondente. É perceptível que o grau de simetria da distribuição dos dados na figura 8B é maior que o da figura 8A. Esse tipo de ganho de simetria foi verificado, em maior ou menor grau, para todos os trinta e quatro grupos trimestrais de nosso banco de dados (ver apêndice B). Observemos também que, em termos da variável original, empregando a transformação Box-Cox (3.6) e as equações (3.8) e (3.4), a distribuição de popularidade é representada por

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x^{1-\lambda}} \exp\left[-\frac{\left(\frac{x^\lambda - 1}{\lambda} - \mu \right)^2}{2\,\sigma^2} \right], \tag{3.9}$$

em que $x \geq 0$. Nessa expressão, $\left(x^\lambda - 1\right)/\lambda$ deve ser substituído por $\log x$ quando $\lambda = 0$, reduzindo $p(x)$ à distribuição lognormal (3.5), e, quando $\lambda = 1$, temos uma PDF gaussiana. De maneira geral, a transformação Box-Cox proporciona a identificação de uma

opção interpolante entre gaussiana e lognormal que mais se adapte aos dados. Além disso, com a possibilidade de a PDF dos dados se aproximar de uma gaussiana na variável $z^{(\lambda)}$ com uma escolha conveniente de λ, a transformação Box-Cox é frequentemente recomendada como uma maneira de promover uma gaussianização dos dados (ZHANG; LEI; KOYAMADA, 2018; QIN; YU; ZHANG, 2020). Nesse contexto de gaussianização e procedendo de maneira semelhante com outros conjuntos trimestrais de dados, reforçamos que o padrão ligeiramente assimétrico também ocorre nos demais trinta e três trimestres de datas de lançamento restantes (ver apêndices B e C).

Por último, devemos mencionar que, como apresentado na figura 9, a evolução temporal do parâmetro ótimo, λ_{opt}, para esse grupo trimestral variou pouco ao longo do tempo e, mesmo considerando as barras de erro, permaneceu negativo, resultados que apontam para um cenário robusto. A evolução temporal do parâmetro ótimo para os demais trimestres encontra-se no apêndice C.

Consistentemente com nosso agrupamento dos vinte e três trimestres de evolução temporal para os dados de 01/2015, observamos a possibilidade de aplicar a transformação Box-Cox separadamente para cada um desses trimestres. Os resultados dessas aplicações são PDFs comparáveis ao apresentado na figura 8B e estão apresentados no apêndice B. Além disso, o λ ótimo para cada um desses PDFs está próximo do λ_{opt} considerado na figura 8B, conforme apresentado na figura 9.

3.4.2 Análise global

Quando um procedimento análogo ao discutido em conexão com a figura 8B é aplicado a cada um dos trinta e quatro grupos trimestrais, uma média de λ_{opt}'s, $\overline{\lambda_{opt}}$, é obtida. Esses valores de λ_{opt} são apresentados na figura 10. Entre todos os λ_{opt}, o menor λ_{opt} é $-0{,}07$, o maior é $-0{,}02$ e a média é $-0{,}04$. Esses fatos apontam para uma semelhança das PDFs dos dados trimestrais normalizados por meio da transformação dada na equação (3.8). Essa robustez também sugere a investigação do comportamento de todos os dados trimestrais padronizados como um conjunto de dados exclusivo de uma única PDF.

Na figura 11A ilustramos como seria a distribuição em termos da variável $z^{(0)}$ para um conjunto que inclui todos os dados e a compara com uma distribuição lognormal. Observemos que a assimetria dessa distribuição de dados é consistente com a do conjunto de dados apresentado na figura 8A. Por sua vez, quando a variável $z^{(\lambda)}$ com a máxima verossimilhança é empregada, obtemos o parâmetro global ótimo igual a $\lambda_{opt} = -0{,}05$. Observemos que esse valor está próximo do comportamento dos diferentes λ_{opt}'s mostrados

como pontos na figura 10 e também próximo do valor médio de todos λ_{opt}'s, $\overline{\lambda_{opt}}$, exibido como linha contínua. Nas figuras 11A e 11B, também verificamos que a semelhança com o gaussiano é acentuada quando λ_{opt} é usado em vez de $\lambda = 0$, tal como ocorre com os dados usados nas figuras 8A e 8B. Esses fatos apontam para uma PDF padrão unificadora da variável normalizada, incluindo todos os trinta e quatro trimestres, bem como para sua evolução no tempo.

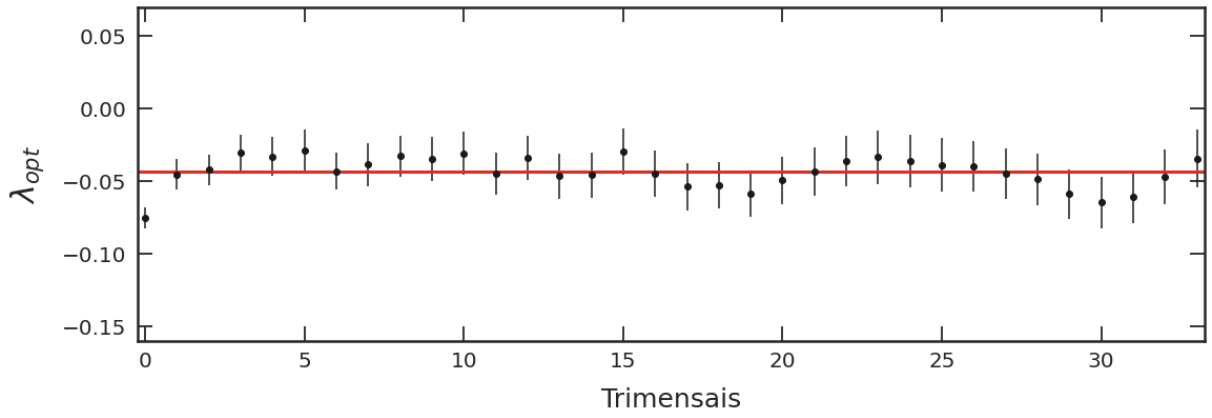


Figura 10: **Análise Box-Cox para todos os grupos trimestrais.** O parâmetro λ ótimo, λ_{opt}, para cada grupo trimestral (disco) e o valor médio, $\overline{\lambda_{opt}} = -0{,}04$, representado pela linha reta vermelha. As barras de erro indicam intervalos de confiança de 95%.

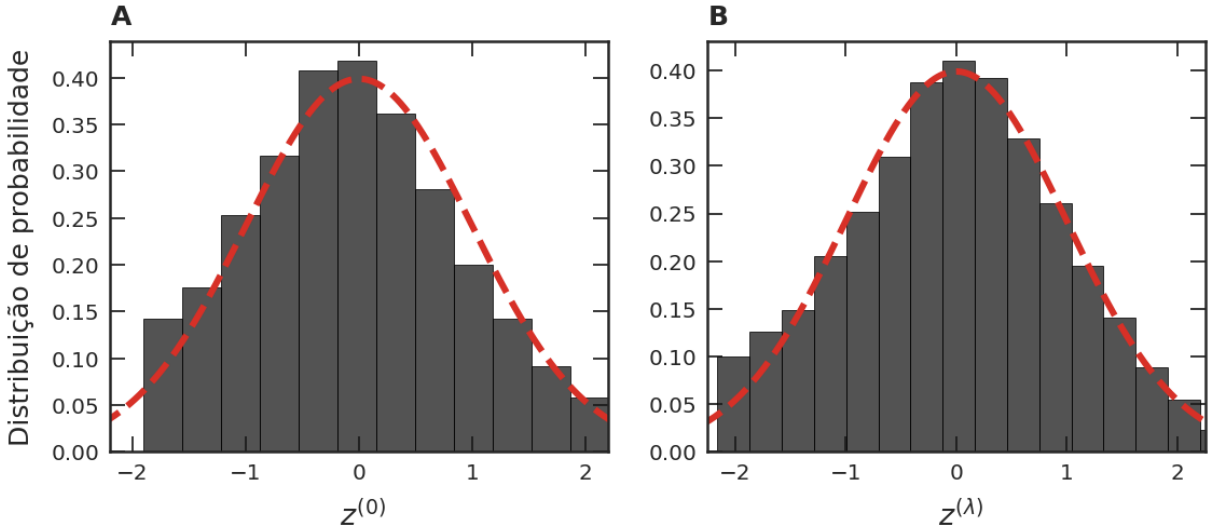


Figura 11: **Análise Box-Cox para todos os grupos trimestrais.** (**A**) Para os dados globais normalizados, a PDF da variável logarítmica, $z^{(0)}$. (**B**) O mesmo que em (**A**), mas com o emprego de $\lambda_{opt} = -0{,}05$ na transformação Box-Cox. As linhas tracejadas correspondem à distribuição normal com média zero e desvio-padrão unitário.

4 Análise categórica dos dados

Neste capítulo, estenderemos a análise e os resultados anteriores para outra forma de focar a popularidade dos *video games*, a saber, pela comparação de jogos com outros da mesma categoria (também chamada de *tag* na *Steam*). Isso significa que precisamos primeiro decidir uma maneira de definir a categoria de um jogo. No caso da plataforma considerada, a categoria de um *video game* é indicada tanto pelos próprios desenvolvedores como pelos usuários. Um usuário pode, por exemplo, definir a categoria de "Ação" para um jogo e outros usuários podem decidir se essa categoria descreve bem o jogo em questão e votar nela. Dessa maneira, a lista de categorias definidas pelos usuários é democrática no sentido de ser decidida pelo "voto popular" dos jogadores. Como os objetivos centrais de nossa investigação são exatamente descrever e compreender a popularidade dos *video games*, escolhemos como categoria(s) aquela(s) com mais votos dos jogadores.

4.1 Resultados para as maiores categorias

Considerando a definição proposta de categoria, listamos na tabela 2 as doze com o maior número de jogos e uma explicação sucinta do estilo de jogo na categoria. Notamos também uma sobreposição de categorias[1]; por exemplo, um jogo pode pertencer tanto à categoria *Indie* quanto à categoria Quebra-cabeças. Ressaltamos que a listagem mencionada é resultado da contagem de vezes que usuários definiram um jogo com essa categoria. Posteriormente, conforme a introdução deste capítulo, tomaremos apenas a mais votada como a definição de categoria do jogo.

[1] Todo jogo tem uma lista de categorias vinculadas a ele. Concretamente, o jogo TAL tem uma lista de categorias [A, B, C...], ordenada da mais votada para a menos votada. Assim, esse jogo pertence a todas essas categorias.

Table 2: **Principais categorias empregadas na investigação**. A tabela apresenta as doze categorias considerando a quantidade de jogos. Na primeira coluna, temos o nome da categoria (*tag*) e seu número de jogos entre parênteses (em dezembro de 2020). Na segunda coluna, descrevemos sucintamente o que caracteriza os jogos na categoria. Mencionamos que, à exceção da categoria *Indie* e dos acrônimos RPG e 2D, utilizamos aqui o nome traduzido das categorias para facilitar a compreensão; no entanto, mantivemos o nome em inglês ao longo do texto para ajudar na conferência do conjunto de dados disponibilizado.

Categoria (Núm. de jogos)	Descrição
Indie (15348)	São criados sem apoio de grande publicadora ou os desenvolvedores têm independência criativa.
Casual (9201)	São direcionados a um público de mercado de massa e geralmente envolvem regras simples e sessões curtas.
Ação (8705)	Enfatizam desafios físicos, incluindo coordenação mão-olho e tempo de reação.
Aventura (8601)	O jogador assume o papel de protagonista em uma história interativa impulsionada pela exploração e/ou resolução de quebra-cabeças. Pelo foco na narrativa, são os jogos mais semelhantes a filmes ou à literatura.
Único jogador (6835)	Jogados apenas por um jogador.
Estratégia (5063)	Enfatizam o pensamento e o planejamento para alcançar a vitória ou finalizar o jogo.
RPG (4027)	Sigla de *Roleplaying Game*, jogos nos quais os usuários assumem papéis de personagens em um cenário fictício.
2D (3451)	Jogos que não tentam criar a ilusão de profundidade, sendo restritos ao plano bidimensional.
Quebra-cabeças (2753)	Testam a engenhosidade ou o conhecimento do jogador em resolver problemas.

Continua na próxima página ...

Table 2: **Principais categorias empregadas na investigação**. A tabela apresenta as doze categorias considerando a quantidade de jogos. Na primeira coluna, temos o nome da categoria (*tag*) e seu número de jogos entre parênteses (em dezembro de 2020). Na segunda coluna, descrevemos sucintamente o que caracteriza os jogos na categoria. Mencionamos que, à exceção da categoria *Indie* e dos acrônimos RPG e 2D, utilizamos aqui o nome traduzido das categorias para facilitar a compreensão; no entanto, mantivemos o nome em inglês ao longo do texto para ajudar na conferência do conjunto de dados disponibilizado. (Continuação)

Categoria (Núm. de jogos)	Descrição
Atmosférico (2479)	Jogos que se preocupam em criar um clima ou sentimento (por exemplo, jogos de terror).
Simulação (5514)	Preocupam-se em criar uma experiência simulada (por exemplo, pilotar um avião).
Multijogador (2540)	Jogados por vários jogadores ou *online*.

Neste capítulo, como procedemos no anterior, iniciamos a análise das categorias considerando o análogo da figura 4. Na figura 12, mostramos a evolução temporal do número médio de jogadores por hora em cada mês (de julho de 2012 até dezembro de 2020) na maior categoria, *Indie*. Observamos que existe uma tendência média de crescimento linear que coincide com o comportamento identificado no caso geral mostrado na figura 4. As tendências temporais de crescimento linear do número médio de jogadores por hora também foram verificadas na maioria das doze maiores categorias. Outras categorias, contendo menos de 150 jogos, apresentam muitas flutuações estatísticas em sua evolução temporal, dificultando a identificação de um padrão geral. Por outro lado, como no caso geral (figura 5), o número de jogos lançados mensalmente exibe uma tendência de crescimento exponencial aproximado (pelo menos nos meses intermediários) para a maioria das doze principais categorias de jogos. Esse comportamento é ilustrado com o caso *Indie* na figura 13. Para cada uma das outras categorias, o número de jogos lançados por mês é menor e a possível identificação de um comportamento exponencial torna-se menos clara.

Via de regra, há uma evolução no número de jogadores nos últimos meses da nossa base de dados para as categorias principais. Essa evolução pode ser explicada por meio do crescimento natural da plataforma, isto é, do aumento de usuários da *Steam*. Por outro

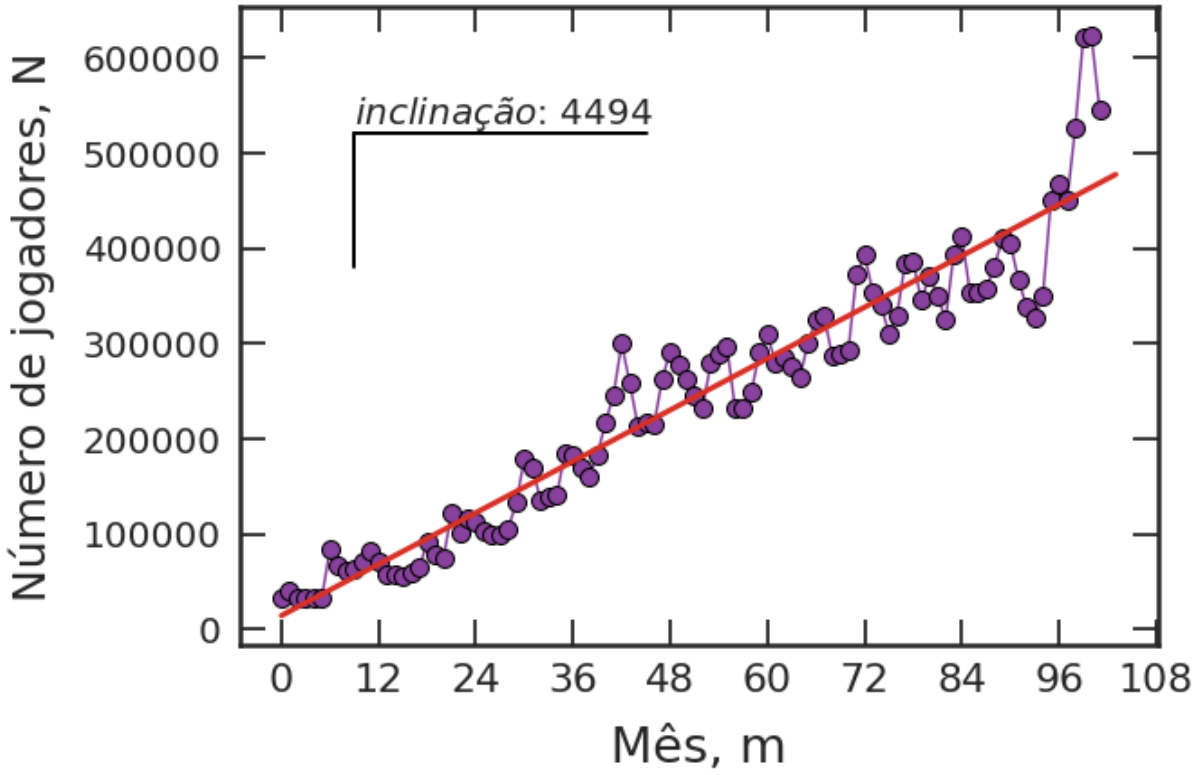


Figura 12: **Evolução do número de jogadores na categoria *Indie*.** O número médio de jogadores por hora em cada mês para a categoria *Indie* (discos roxos) e um ajuste linear (linha reta vermelha), em que o mês um é julho de 2012 e o último mês é dezembro de 2020. Essa linha vermelha corresponde à função $N = 14636 + 4494\,m$.

lado, no caso específico da categoria *Indie*, o crescimento parece maior que o esperado no processo de fragmentação de novos usuários. Dessa maneira, ainda que seja demasiado cedo para apontar uma explicação definitiva, três hipóteses surgem: a primeira, de viés econômico, é que os jogos *Indie* são usualmente mais baratos que os chamados AAA[2]. Considerando o período da pandemia, é possível que os usuários tenham buscado alternativas acessíveis. Outra hipótese, de cunho estético, seria a busca por jogos diferentes, o que implicaria uma certa exaustão por parte dos usuários com a repetição das fórmulas presentes nos jogos já famosos. A última hipótese é puramente acidental: o pico se explica pela presença de alguns jogos lançados naquele trimestre, os quais foram mais populares do que se esperava (no trimestre em questão, destacam-se *Fall Guys*[3], *Phasmophobia*[4] e

[2]Termo emprestado das qualificações de títulos da indústria de crédito, em que os títulos AAA representavam a oportunidade com maior probabilidade de atingir as metas financeiras. Tal termo refere-se a jogos de empresas de grande capital ou já muito famosos, como *FIFA* (também conhecido como *FIFA Football*, uma série de *video games* de futebol) ou *Call of Duty* (série de jogos de tiro em primeira pessoa, usualmente com elementos históricos.)

[3]Jogo da categoria de batalha real, que é uma competição para ser o último a sobreviver em uma zona de jogo que encolhe gradativamente.

[4]Jogo de terror independente em que até quatro jogadores assumem o papel de caça-fantasmas.

Hades[5] como jogos com mais de quinze mil jogadores, pertencentes à categoria *Indie*.). Como mencionamos, fincar uma hipótese seria prematuro, mas novos dados apontarão na direção da explicação correta para esse aumento de popularidade.

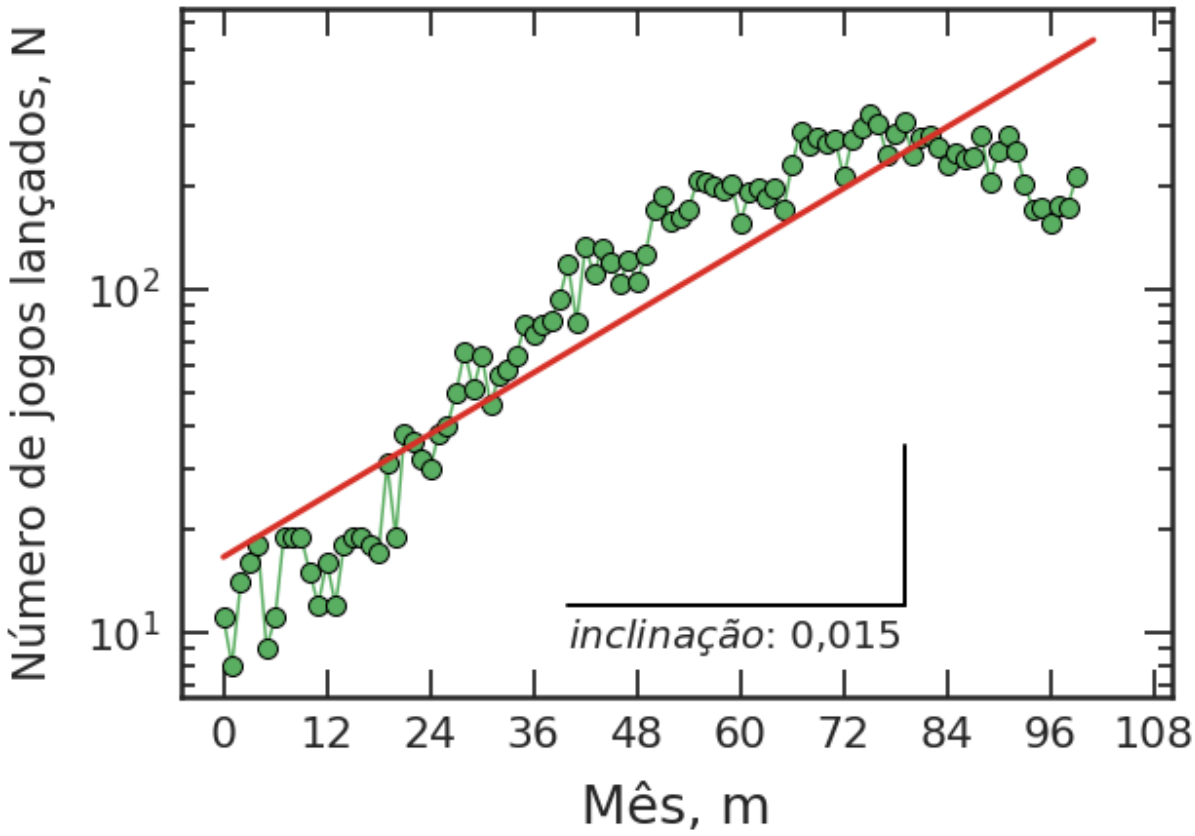


Figura 13: **Evolução do número de jogos lançados na *Steam* para a categoria *Indie*.** Durante os cento e quatro meses dos nossos dados, aparece o crescimento do número de jogos lançados nessa categoria (discos verdes) e um ajuste exponencial (na escala logarítmica, linha reta vermelha). A equação que governa esse ajuste é dada por $\log N = 1{,}21 + 0{,}015\, m$.

Em adição, podemos perceber na figura 13, assim como na figura 5, que o desvio do comportamento exponencial ocorre marcantemente para meses posteriores ao 80. Como explicamos no caso geral, esse comportamento pode ser indício de uma saturação na produção de jogos, isto é, os desenvolvedores produziriam um volume de jogos suficiente para atender à demanda dos usuários. Além disso, como podemos argumentar que um declínio na produção de jogos em geral não necessariamente implicaria um declínio na produção de jogos em uma categoria específica, o decaimento na categoria *Indie* poderia resultar no desenvolvimento acentuado em outra. Nesse sentido, podemos argumentar que,

[5]Jogo da categoria *roguelike*, isto é, jogos em que o jogador deve finalizar o jogo inteiro em um dado tempo ou número de vidas. Caso o tempo ou o número de vidas se esgotem, o jogador deve recomeçar da zona inicial. Especificamente, o objetivo de *Hades* é escapar do submundo da mitologia grega e os jogadores assumem o papel do deus órfico Zagreu.

dentre os usuários da *Steam*, todos os que teriam interesse em jogos *Indie* são jogadores de algum jogo dessa categoria e os usuários que nunca consumiram um jogo que contenha essa *tag* provavelmente continuarão sem consumir, o que justifica a percepção de um menor retorno no investimento para o desenvolvimento e, portanto, explicaria a redução nos lançamentos.

Continuando com a análise da popularidade dos *video games* categorizados, a figura 14 mostra o análogo da figura 11B (todos os trimestres de todos os grupos trimestrais juntos) com foco na distribuição do número de usuários para *video games Indie*. Nesse caso, temos uma transformação Box-Cox com $\lambda_{opt} = -0{,}049$. Esse valor é consistente com o λ_{opt} obtido para cada dado trimestral da categoria *Indie*, mostrado na figura 15 (o análogo da figura 10). Comportamentos semelhantes aos mostrados nas figuras 14 e 15 também foram obtidos para as outras onze categorias principais de jogos, cujos λ_{opt}'s são apresentados na figura 16 (ver também apêndices D e E). Por sua vez, a média desses doze valores médios de λ_{opt} é igual a $-0{,}028$, um resultado próximo à média global mostrada na figura 10 (linha contínua). O cálculo desses λ's ótimos para as categorias restantes também está próximo de $-0{,}03$ na maioria dos casos, embora o número de jogos envolvidos em cada categoria seja pequeno.

4.2 Análise entre as categorias

Neste último estrato de nossa análise, examinaremos outros aspectos da popularidade das categorias de *video games*. O primeiro é o crescimento do número de categorias distintas de *video games* ao longo do tempo. Conforme mostrado na figura 17, identificamos uma tendência de crescimento linear, cuja taxa de aumento é de aproximadamente 8 categorias/mês.

Quanto à ainvestigação da popularidade relativa entre as categorias, a abordagem é semelhante à das figuras 14 e 15. Nesse caso, concentramos nossa atenção na distribuição de popularidade relativa das categorias como uma função de seu número de *video games*. Empregamos a variável $n^{(\lambda)}$ dada por

$$n_i{}^{(\lambda)} = \frac{\nu_i{}^{(\lambda)} - \langle \nu_i{}^{(\lambda)} \rangle}{\sigma_{n^{(\lambda)}}}, \tag{4.1}$$

em que $\nu_i{}^{(\lambda)}$ é a transformação Box-Cox, equação (3.6), do número de jogos na categoria i, $\langle \nu_i{}^{(\lambda)} \rangle$ é o valor médio e $\sigma_{n^{(\lambda)}}$ é o desvio-padrão da variável $\nu_i{}^{(\lambda)}$. Usando uma transfor-

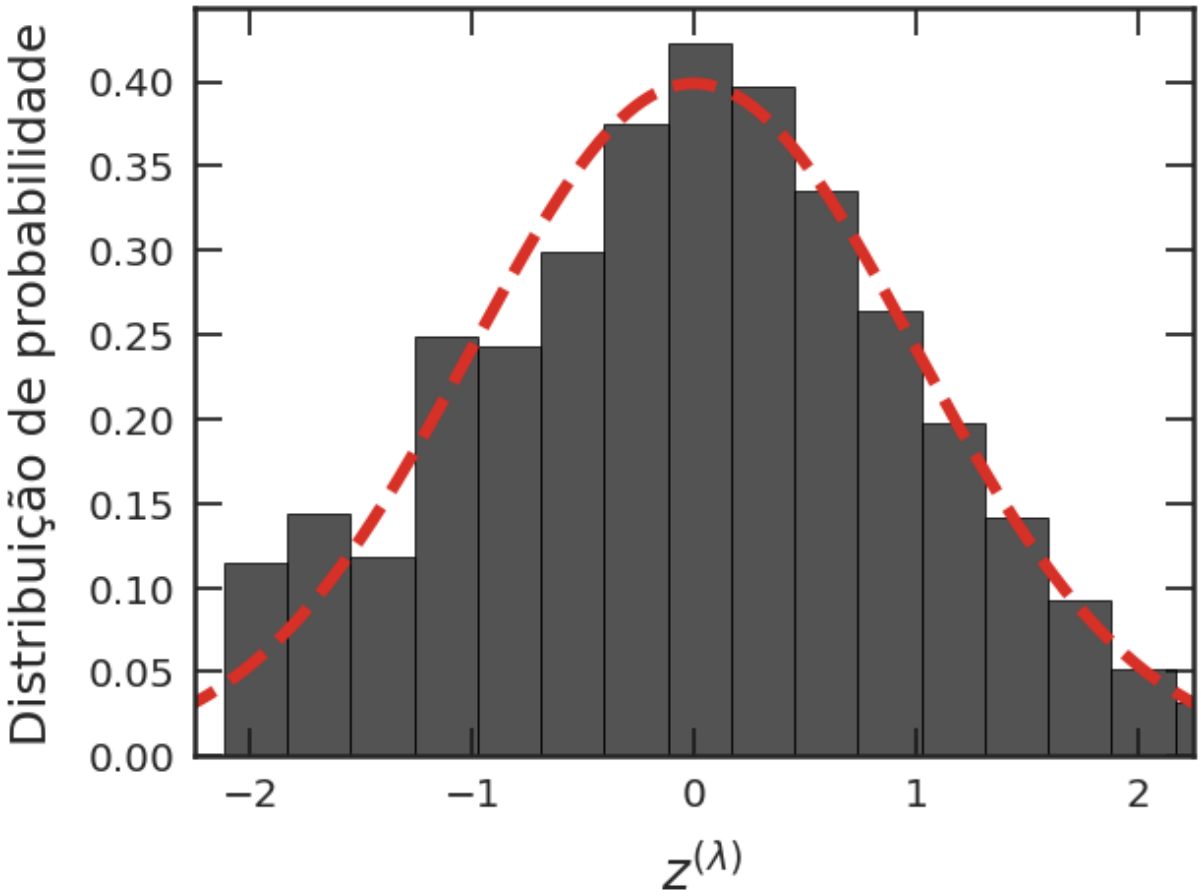


Figura 14: **Densidade de probabilidade para a categoria *Indie*.** PDF para a categoria *Indie* obtida por meio da transformação Box-Cox ($\lambda_{opt} = -0{,}049$) é análoga à figura 11B (todos os trimestres de todos os grupos trimestrais juntos, a linha tracejada correspondendo à uma gaussiana de média nula e desvio-padrão unitário).

mação Box-Cox ótima, essa distribuição em dezembro de 2020 é exibida na figura 18. Em dezembro de 2020, havia um mil e quarenta e quatro categorias, abrangendo vinte e um mil setecentos e cinquenta e dois jogos. Verificamos que existem algumas categorias com apenas um jogo e não as consideramos em nossa análise. Optamos por esse procedimento porque a nomenclatura da categoria pertinente pode ser bastante incerta, dado que há um número muito limitado de votos para essas categorias, o que leva a incertezas estatísticas.

Como podemos ver na figura 18, os dados transformados via Box-Cox podem ser ajustados por uma distribuição normal de média zero e desvio-padrão unitário, mas de maneira menos precisa do que em nossos estudos anteriores (figuras 8B, 11B, e 14). No entanto, o λ ótimo, $\lambda_{opt} = -0{,}12$, está próximo dos encontrados anteriormente, indicando uma visão unificada para nosso estudo da popularidade dos *video games*. Esse resultado é reforçado pela estabilidade da evolução temporal de λ_{opt} nos últimos anos, mostrada na

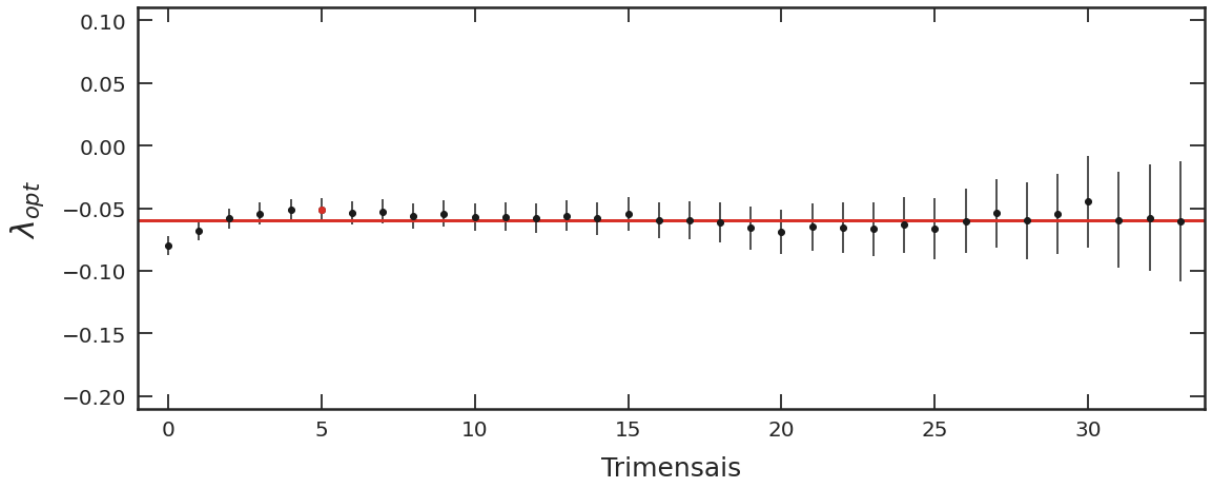


Figura 15: **Evolução do parâmetro Box-Cox para a categoria *Indie*.** Parâmetro Box-Cox para jogos *Indie* em cada trimestre de nosso conjunto de dados, análogo à figura 10. As barras referem-se a intervalos de confiança de 95%. A linha reta corresponde à média de todos os parâmetros trimestrais de Box-Cox, $\overline{\lambda_{opt}} = -0{,}06$.

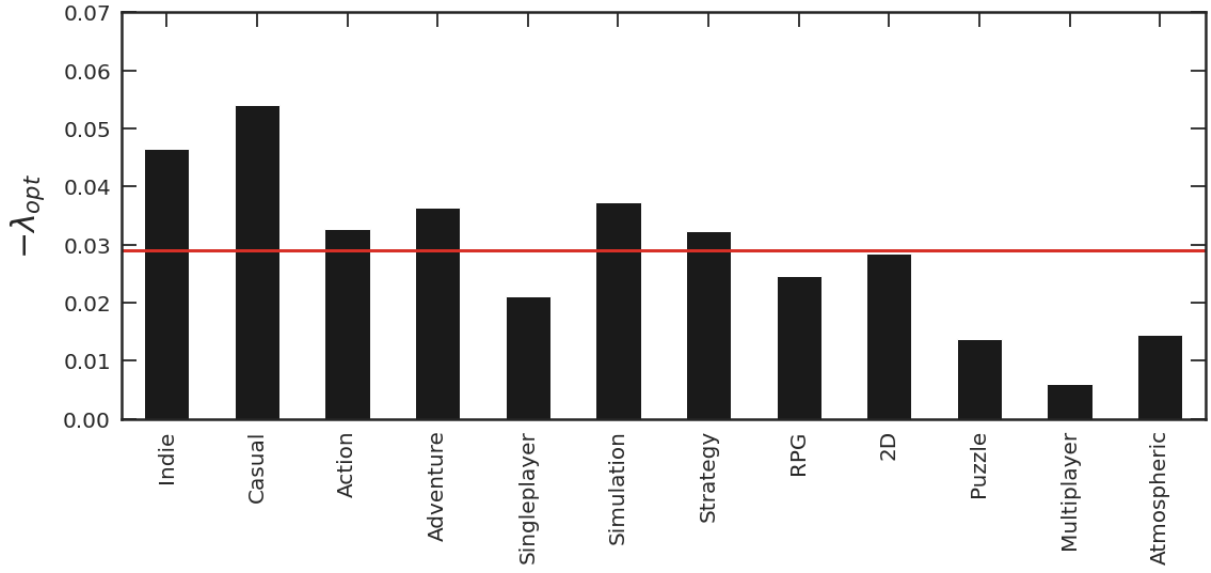


Figura 16: **Principais categorias de *video games*.** Parâmetro Box-Cox ótimo, λ_{opt}, para as doze maiores categorias de jogos, calculado pelo mesmo procedimento de máxima verossimilhança. Os parâmetros λ_{opt}'s são classificados em ordem decrescente de seu número de jogos por categoria. A linha reta representa a média desses λ_{opt}'s, $\langle\lambda\rangle = -0{,}028$ para as doze categorias principais.

figura 19. Nos primeiros anos de nosso banco de dados, havia poucas categorias e poucos jogos, o que piorou o ajuste dos dados por uma distribuição Box-Cox, levando a um valor λ_{opt} pouco claro, evidenciado por intervalos de confiança mais amplos. Como nas análises anteriores, visto que λ_{opt} está próximo de zero, os dados como os da figura 18 podem ser

representados por uma distribuição lognormal em uma primeira aproximação.

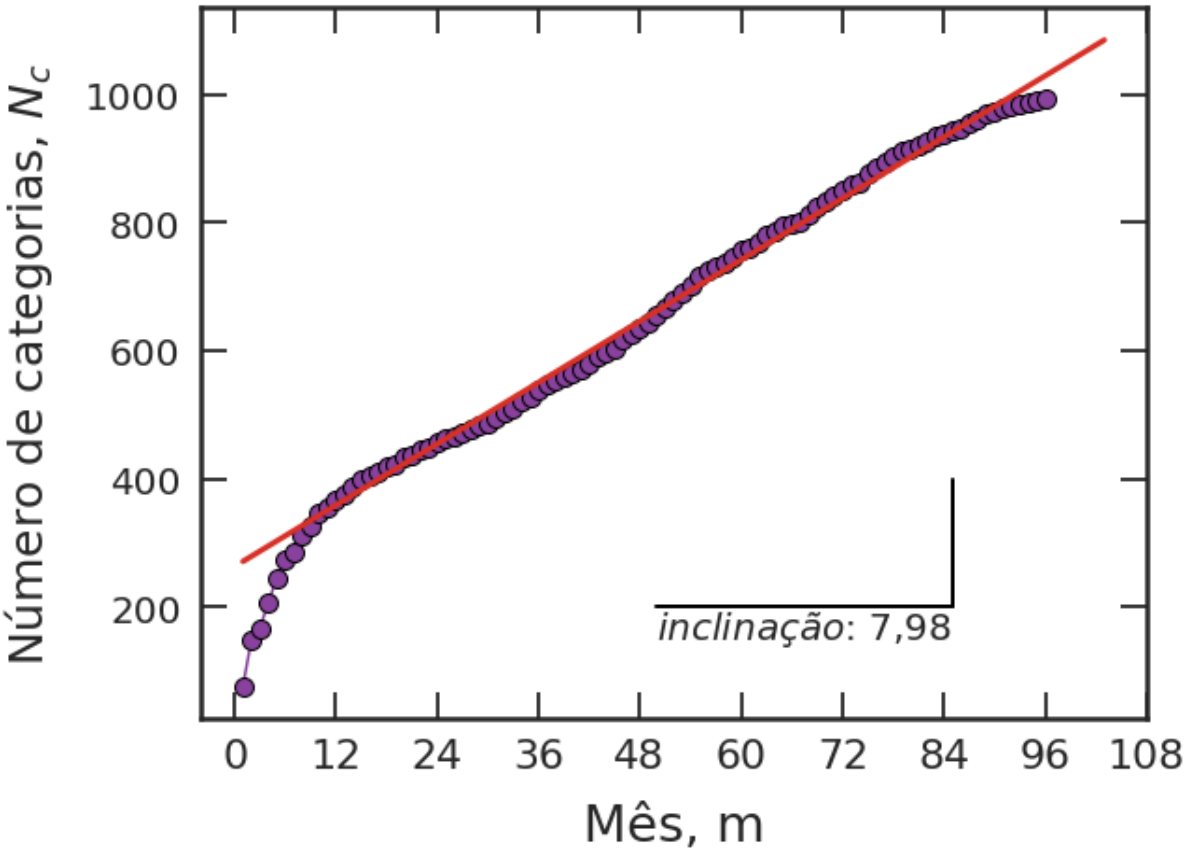


Figura 17: **Evolução do número de categorias.** Evolução temporal do número de categorias ao longo dos meses de julho 2012 a dezembro 2020. A linha reta vermelha representa um ajuste linear, $N_c = 263{,}09 + 7{,}98m$, do mês 10 até o último mês da nossa base de dados.

Para resumir nossos resultados quantitativos sobre a popularidade comparativa dos *video games*, apresentamos na tabela 3 valores típicos de λ ótimos, bem como a média (μ), desvio-padrão (σ), assimetria (γ) e curtose (κ) da variável transformada via Box-Cox, para a distribuição de popularidade de *video games* globalmente, da categoria *Indie* e por categoria. Essas duas últimas quantidades são definidas por

$$\gamma = \frac{\langle (x_i - \langle x_i \rangle)^3 \rangle}{\sigma^3} \tag{4.2}$$

e

$$\kappa = \frac{\langle (x_i - \langle x_i \rangle)^4 \rangle}{\sigma^4} - 3, \tag{4.3}$$

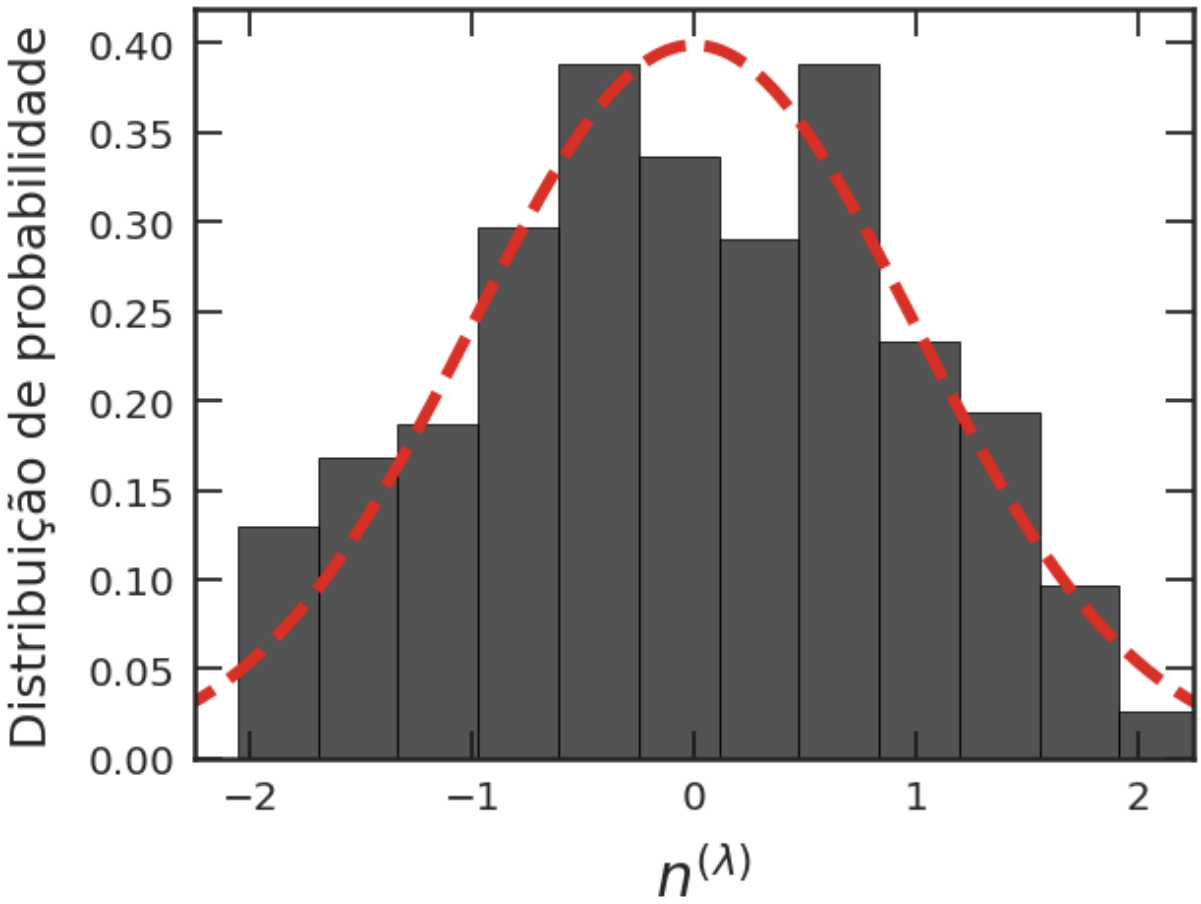


Figura 18: **Distribuição das categorias de *video games*.** PDF para categorias obtidas por meio da transformação Box-Cox ($\lambda_{opt} = -0{,}12$) em função do número de *video games* em cada categoria em dezembro de 2020. A linha tracejada refere-se a uma distribuição gaussiana de média zero e desvio-padrão unitário.

em que x_i é uma variável aleatória e $\langle \cdots \rangle$ representa o valor médio. Se a PDF, considerando uma transformação de variável conveniente, fosse gaussiana, seria simétrica em relação a μ (portanto, teríamos $\gamma = 0$) e haveria um equilíbrio entre a cauda da distribuição e seu pico, correspondendo a $\kappa = 0$ (DEGROOT; SCHERVISH, 2012). Consequentemente, os desvios dos valores nulos de γ e κ indicam discrepâncias da normalidade. Por exemplo, se uma variável logarítmica ($\lambda = 0$) for usada e se γ e κ estiverem próximos de zero, esses valores de γ e κ indicam, em uma primeira aproximação, que os dados podem ser vistos como lognormalmente distribuídos (usualmente, esta aproximação pode ser melhorada empregando um λ ótimo). Em termos das variáveis originais, a distribuição de popularidade otimizada para cada estrato referido na tabela 3 é obtida pelo uso dos valores correspondentes para λ, μ e σ na equação (3.9). Particularmente, deve ser notado que os valores da assimetria γ são bastante próximos de zero quando o λ ótimo é

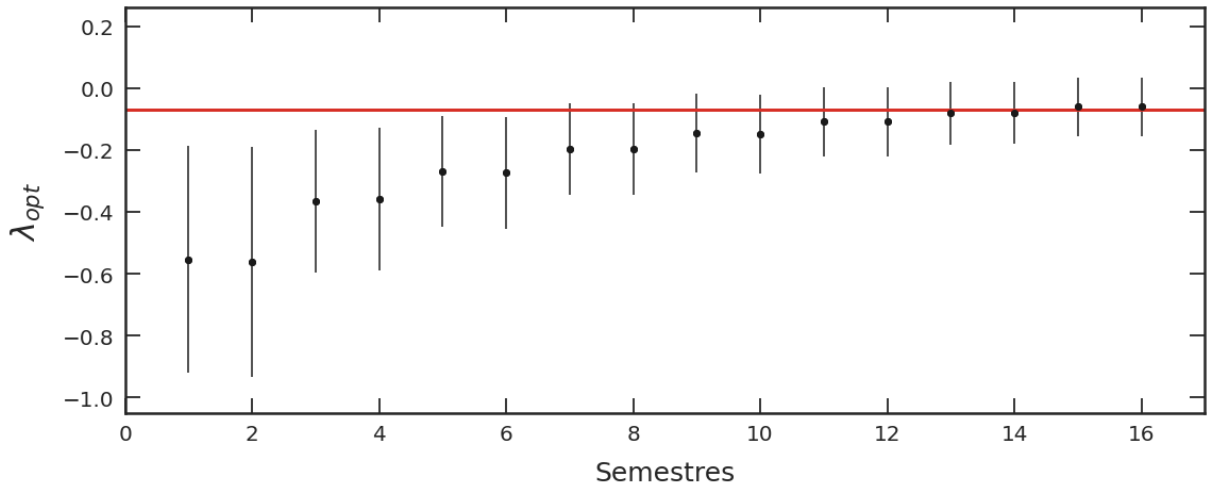


Figura 19: **Evolução do parâmetro Box-Cox para o número de categorias.** Evolução temporal de julho de 2012 a dezembro de 2020 do parâmetro ótimo, λ_{opt}, para o número de categorias, em que cada disco representa um semestre. As barras indicam 95% intervalos de confiança dos λ_{opt}'s. A linha reta vermelha exibe o comportamento médio dos últimos dois anos.

empregado.

Tabela 3: **Parâmetros estatísticos para os dados transformados.** Parâmetros dos dados transformados de Box-Cox (primeira coluna): parâmetro Box-Cox (λ), média (μ), desvio-padrão (σ), assimetria (γ) e curtose (κ); os valores para esses parâmetros log-transformados são mostrados na segunda coluna; todos os dados transformados via Box-Cox (terceira coluna); os dados após a transformação Box-Cox para a categoria *Indie* (*tag*) (quarta coluna) e os dados transformados via Box-Cox para a distribuição do número de jogos por categoria (quinta coluna). Após cada símbolo $\pm$, exibimos o intervalo de confiança de 95% do correspondente parâmetro.

	Lognormal (todos os dados)	Box-Cox (todos os dados)	Box-Cox (*Tag Indie*)	Box-Cox (jogos por *tags*)
λ	0,00	$-0{,}053 \pm 0{,}001$	$-0{,}049 \pm 0{,}002$	$-0{,}12 \pm 0{,}05$
μ	$0{,}27 \pm 0{,}01$	$0{,}43 \pm 0{,}02$	$1{,}60 \pm 0{,}02$	$3{,}72 \pm 0{,}25$
σ	$1{,}2 \pm 0{,}1$	$2{,}6 \pm 0{,}2$	$2{,}6 \pm 0{,}2$	$0{,}74 \pm 0{,}15$
γ	$0{,}41 \pm 0{,}02$	$0{,}01 \pm 0{,}01$	$0{,}01 \pm 0{,}01$	$0{,}03 \pm 0{,}05$
κ	$0{,}15 \pm 0{,}07$	$-0{,}28 \pm 0{,}02$	$-0{,}34 \pm 0{,}03$	$-0{,}5 \pm 0{,}6$

5 Análise dos resultados

Este capítulo é dedicado à análise dos resultados obtidos no terceiro e no quarto capítulos. Em particular, discutiremos os modelos empregados na interpretação qualitativa desses resultados prévios. Tais modelos, a saber, o modelo malthusiano de crescimento, o modelo lognormal (*memoryless*) e o modelo Box-Cox, os dois últimos para a popularidade relativa, foram aplicados em três diferentes níveis hierárquicos, o global, o das categorias maiores e o entre categorias. Dessa maneira, a análise será dedicada, também, à investigação dos resultados nesses três níveis, já que os resultados empíricos sobre a popularidade dos *video games*, baseados em uma das principais plataformas de jogos (*Steam*), sugerem um cenário robusto.

5.1 Comportamentos de crescimento e modelo malthusiano

Em primeiro lugar, os resultados são consistentes com a noção de um número crescente de jogadores de *video game* ao redor do mundo, bem como do aumento do número de *video games* lançados. Quantitativamente, o número médio de jogadores por hora ao longo dos meses apresenta um crescimento mais ou menos linear, tanto do ponto de vista global (figura 4) quanto do das categorias de jogos (figura 12), especialmente as principais. Para o número de categorias distintas, o padrão de crescimento semelhante (isto é, linear) ao apresentado nas figuras 4 e 12 também foi verificado na figura 17. Ademais, podemos compreender qualitativamente os fenômenos observados na figura 4. Por um lado, no intervalo entre 20 e 80 meses, o ajuste parece melhor e o crescimento, uniforme e mais ou menos linear. Por outro, a parte final passa por um fenômeno de *outlier*, no qual há o lançamento de um jogo (ou jogos) muito mais populares que a média (nesse caso, o jogo em questão é o *PlayerUnknown's Battlegrounds*[1]). Quanto à evolução do número de jogos lançados, foi identificada uma tendência de crescimento aproximadamente exponencial no nível global (figura 5), assim como para as maiores categorias (figura 13). Dessa forma,

[1]Fenômeno do já citado gênero de batalha real, mas que, diferentemente do mencionado *Fall Guys*, é embasado na realidade de uma zona de guerra, com armas baseadas em modelos reais.

verificamos a hipótese **H1** apresentada na primeira seção do capítulo 3.

O comportamento de crescimento linear (aritmético) de recursos e demanda exponencial (geométrica) já foi explorado por Malthus (MALTHUS; WINCH; JAMES, 1992) e revisto brevemente no capítulo 1 desta tese. Em nosso estudo, se interpretarmos os jogos como competindo entre si pela atenção dos jogadores (ver figura 3), estes ocupariam o papel de recursos e os jogos, o papel de demanda na visão malthusiana. Nesse caso, o aumento exponencial dominaria drasticamente o linear no longo prazo. Além das figuras 4 e 5, observamos também, ao analisarmos os dados em categorias (figuras 12 e 13), um padrão malthusiano de crescimento, mas menos robusto. Esses estudos apontam para padrões comuns de crescimento nos três níveis sob investigação: dados globais, categorias principais e entre categorias.

O comportamento malthusiano é comum na dinâmica populacional de sistemas biológicos (MURRAY, 2004) e sociais (BACAËR, 2011) quando não há limitações, ocorrendo tipicamente nos estágios iniciais da evolução desses sistemas. Por outro lado, as restrições a esse regime de crescimento inicial comumente se manifestam ao longo do tempo, indicando alguma saturação populacional (RICKLEFS, 2008). Nesse sentido, como último aspecto em relação ao número de jogos, é justo dizer que um ajuste exponencial não é completamente satisfatório ao longo de todo o intervalo, apesar de ser bem mais apropriado no intervalo entre vinte e quatro meses e oitenta meses. Apesar desse resultado parcialmente negativo, é interessante observar o desvio do comportamento exponencial na parte final da figura 5. Isso pode indicar uma mudança na natureza do desenvolvimento de jogos, pois o desvio parece ser sistemático e não uma flutuação estatística. Como mencionamos, essa desaceleração pode prenunciar uma saturação no mercado, seja porque é atingido um certo limite temporário[2] seja porque ocorre aumento nos custos de desenvolvimento e de marketing, refletido em menos jogos lançados anualmente[3]. No entanto, seria prematuro afirmar que há de fato tal saturação. Trabalhos futuros poderiam retornar a esse problema, beneficiando-se da maior disponibilidade de dados. Uma comparação com outras indústrias de entretenimento (a exemplo da evolução do custo da produção cinematográfica) também poderia ser bastante pertinente. Além disso, uma primeira tentativa de desvio do padrão exponencial que considere um processo de saturação é o modelo logístico de crescimento, tal como foi discutido em dinâmica de populações

[2]Uma maneira mnemônica de enunciar esse argumento é que as pessoas que jogam *video game* hoje são as pessoas que jogariam *video game*, isto é, não existe um grande público ainda a ser "capturado" por essa indústria.

[3]Outra alternativa seria de saturação da própria *Steam*. Infelizmente, a única maneira de contornar essa ambiguidade, em nossa percepção, seria comparar os dados dessa base com outras, como a da *Sony* ou *Nintendo*.

de sistemas biológicos (MURRAY, 2004). Nesse contexto populacional, também são bastante investigados modelos cíclicos que envolvem competição, como o de Lotka-Volterra (SZOLNOKI; OLIVEIRA; BAZEIA, 2020). Assim, outra vertente, menos clara, para estudar padrões de crescimento relacionados a *video games* seria ponderar a possibilidade de algum tipo de modelo cíclico.

5.2 Popularidade relativa e modelo *memoryless*

Outro aspecto investigado neste trabalho é o da popularidade relativa dos jogos por meio de distribuições de probabilidade. Os resultados nos conduziram inicialmente à distribuição lognormal, considerando a estabilidade das médias e desvios-padrão log-transformados (figuras 6 e 7). No entanto, enfatizamos outro ponto de vista. Como ficará claro na discussão que segue, essa outra abordagem permite relacionar nossos resultados com o conceito de memória na teoria de Assmann, delineada no capítulo 1. Nesse sentido, de uma perspectiva qualitativa, podemos considerar que cada lançamento de *video game* fragmenta (particiona) a base de jogadores em uma determinada proporção, pois certa quantidade de jogadores é realocada para o novo jogo. Denominaremos por w_1 esse fator de proporção da fragmentação. Por sua vez, lançamentos sucessivos levam a fragmentações (partições) subsequentes da base de jogadores, proporcionando consecutivos w_i's. A hipótese mais simples é de que os jogos são desconhecidos dos jogadores no momento de seu lançamento; como consequência, podemos supor, em uma primeira aproximação, que essas fragmentações sucessivas ocorrem inteiramente ao acaso e independentes umas das outras, ou seja, ocorrem de uma maneira sem memória (*memoryless*). Dessa maneira, ao assumirmos essa hipótese, estabelecemos que o conjunto $\{w_i\}$ é composto de variáveis positivas aleatórias idêntica e independentemente distribuídas. Como estamos interessados em sucessivas fragmentações, a proporção final, W, após N fragmentações, é igual ao produto das sucessivas etapas, isto é,

$$W = \prod_{i=1}^{N} w_i. \tag{5.1}$$

Portanto, temos um processo multiplicativo para a variável W.

Tomando o logaritmo dos dois lados da igualdade 5.1, obtemos uma variável Y dada por

$$Y = \log W = \sum_{i=1}^{N} \log w_i. \tag{5.2}$$

No lado direito desta última equação, percebemos a soma de variáveis aleatórias indepen-

dentes, que, suporemos, apresentarem valor médio e variância bem definidos. Nesse caso, podemos aplicar o famoso teorema do limite central (CHEBYSHEV, 1900), que indica que a distribuição de probabilidade para a variável soma Y, se N for suficientemente grande, é bem descrita por uma distribuição gaussiana, equação (1.14), quando as variáveis integrantes da soma são suficientemente aleatórias, independentes e identicamente distribuídas. Dessa maneira, chega-se a uma distribuição normal para Y e, consequentemente, à distribuição lognormal, equação (3.5), quando se retorna à variável original W. Assim, a hipótese simples, de fragmentação sem memória, é condição suficiente para a obtenção da distribuição lognormal (AITCHISON; BROWN, 1957; LIMPERT; STAHEL; ABBT, 2001) empregada no capítulo 3. Dessa maneira, verificam-se as hipóteses **H2** e **H3** apresentadas no início do capítulo 3.

Um resultado importante deste estudo foi detectar padrões lognormais aproximados semelhantes nos três estratos diferentes. Não é um exagero mencionar que, apesar da aparente simplicidade da hipótese de ausência de memória (aleatoriedade), a distribuição lognormal tem sido empregada com sucesso nos mais diversos contextos, como na investigação de notícias e memes (WU; HUBERMAN, 2007; YOOK; KIM, 2020), em eleições proporcionais (FORTUNATO; CASTELLANO, 2007) ou na análise de rede de dispersão de sementes (ASSIS BOMFIM et al., 2018).

5.3 Presença de efeitos de memória

Como discutimos na seção prévia, a distribuição lognormal é uma primeira aproximação da distribuição empírica para a popularidade relativa de *video games*. Os desvios dessa distribuição estão relacionados à violação de uma ou mais hipóteses empregadas, ou seja, alguns fatores pensados para conduzir à distribuição lognormal não são totalmente verdadeiros. Particularmente, não é razoável supor que os jogos sejam completamente desconhecidos dos jogadores no momento de seu lançamento ou que a ignorância permaneça com o tempo. Por exemplo, cada jogo tem suas características definidoras, suas peculiaridades, seu público-alvo e seu modelo de negócio. Além disso, existe a memória cultural produzida por anúncios e franquias de jogos bem estabelecidos. A memória comunicativa entre os jogadores, como ocorre nas redes sociais, a antecipação do lançamento de um jogo e o acúmulo de fama também alteram a percepção dos jogos. Assim, consideramos prudente investigar a possibilidade de desvios da lognormal.

Visto que todos esses fatores mencionados perturbam a hipótese de ausência de me-

mória (aleatoriedade), podem contribuir para explicar a assimetria mostrada nas figuras 8A e 11A. Quantitativamente, o parâmetro λ_{opt} fornece uma medida do afastamento da hipótese inicial, isto é, da distribuição lognormal. Assim, interpretamos a magnitude de λ_{opt} como uma medida da presença de memórias culturais e comunicativas na distribuição de popularidade, em que $\lambda_{opt} = 0$ corresponde à hipótese da ausência de memória. Essa descoberta é consistente ao longo do tempo (figura 9), aparecendo em diferentes grupos de jogos (figura 10) e nas categorias de jogos (figura 15), levando à verificação da hipótese **H4**, exibida no início do capítulo 3. Portanto, apesar de pequenas variações em torno da média (evidenciando que a popularidade de alguns jogos é mais afetada pelos efeitos de memória do que a de outros), o efeito das memórias está presente em todos os estratos investigados.

Além disso, mesmo considerando as barras de erro, os parâmetros λ_{opt}'s são consistentemente negativos, além de serem de magnitudes similares. Esse ponto deve ser salientado, visto que, apesar de pertencerem à mesma base de dados, a natureza dos dados é distinta. Embora fosse possível esperar que o parâmetro de uma das categorias principais concordasse com o parâmetro global, considerando que ela é uma amostra bastante representativa do todo, não segue que, na análise do número de categorias, isso se apresente como verdade. Enquanto dois dos aspectos estudados, global e por categoria, são baseados no número de usuários em cada jogo, o último, a evolução do número de categorias, não o é. Essa concordância indica, como mencionamos, um cenário robusto, mas também parece apontar para uma forte interrelação entre número de jogos, de categorias e de jogadores. Dito de outra forma, o crescimento em popularidade dos jogos incentiva a indústria a desenvolver, além de novos jogos, novos estilos de jogo, como uma maneira seja de se diferenciar da concorrente, seja de buscar novas maneiras de pensar uma mídia interativa e, frequentemente, conectada, inserida em uma rede. Retornando às barras de erro, estas usualmente diminuem com o tempo, nos intervalos em que há mais dados, o que reforça a robustez do resultado, já que é coerente com a nossa intuição sobre o comportamento estatístico.

5.4 Popularidade relativa nas categorias com maior quantidade de jogos

Algumas particularidades das categorias de jogos podem estar concretamente relacionadas às memórias culturais e comunicativas. Como mostra a figura 16, algumas categorias têm a magnitude λ_{opt} notavelmente maior do que outras. Particularmente, as popularidades

das categorias *Indie* e *Casual* (representadas pelas PDFs) são mais afetadas pelos efeitos de memória, visto que apresentam maiores magnitudes de λ.

Uma explicação consistente com a nossa interpretação é de que tal fato é resultante do modelo de negócios adotado pelos desenvolvedores dessas categorias de jogos. Por exemplo, os jogos *Indie* seriam produções independentes, muitas vezes pequenas, no mercado de jogos. Frequentemente, esses jogos são desconhecidos do público e o objetivo desses desenvolvedores é vender o máximo possível de jogos a um preço justo. Dessa forma, os jogos que sobrevivem e alcançam sucesso devem-se ao boca a boca entre os jogadores, o que caracteriza uma elevada memória comunicativa. Na mesma linha, os jogos da categoria *Casual* possuem características semelhantes, ou seja, não costumam envolver produções caras, são considerados um passatempo de poucas horas e seu modelo de negócio é semelhante ao do *Indie*. Em contraste, o modelo de negócios adotado pelos desenvolvedores de jogos multijogador é bem diferente. Muitos desses jogos são *freemium*, o que significa que são gratuitos ou acessíveis, e seu objetivo é manter as pessoas jogando o máximo possível e gastando dinheiro em customizações. Como esses jogos estão disponíveis para todos os usuários, a decisão de continuar jogando ou parar de jogar é de inteira responsabilidade da pessoa, favorecendo a hipótese da aleatoriedade e, consequentemente, a pequena magnitude de λ_{opt}.

Apesar das variações, todos os λ_{opt} encontrados foram negativos. Quanto maior a magnitude do λ_{opt}, mais assimétrica é a distribuição empírica (ver figuras 8A e 8B), favorecendo a presença de *outliers*. Além disso, o fato de os parâmetros serem consistentemente negativos implica uma assimetria na direção de jogos muito pequenos, ou seja, a presença de muitos jogos impopulares. Dessa maneira, os mecanismos de memória contribuem para aumentar a desigualdade da popularidade entre os *video games* quando comparada com o padrão lognormal de quantidade de jogadores por jogos. Isso parece indicar que muitos jogadores procuram jogos que têm alguma recomendação, seja de amigos, de redes sociais, de *marketing*, seja, no caso de franquias já estabelecidas, da própria lembrança de jogos anteriores.

5.5 Distribuição do número de jogos nas categorias

A desigualdade ocasionada pelos mecanismos de memória foi também investigada em relação à distribuição das categorias de *video games* (figura 18) ao longo do tempo (figura 19). Curiosamente, os padrões descobertos são consistentes com os encontrados nas figuras

8B e 11B, e nas figuras 9 e 10, respectivamente. Dizemos "curiosamente" porque essa consistência não seria necessariamente esperada a priori, pois a natureza dos dados é distinta, apesar de estes serem associados com a popularidade de *video games*. Essa distinção advém do fato de que a métrica estudada nos dois primeiros estratos foi o número médio mensal de jogadores, enquanto no último estrato investigamos o número de jogos apenas, desacoplado de qualquer popularidade representada pelo número de usuários. Realmente, parece-nos um fruto inesperado de nossa pesquisa que os resultados obtidos nas três camadas distintas sejam quantitativamente similares, mesmo sendo provenientes de métricas distintas (no caso do modelo malthusiano, devemos comparar a figura 17 com as figuras 4 e 12 dos outros dois estratos).

Inúmeras razões podem ser levantadas para essa semelhança nos resultados similares obtidos a partir de métricas distintas, mas uma possível razão pode ser que alguns desenvolvedores de *video games* tentem desenvolver jogos pertencentes a categorias populares, dado que é difícil avaliar a recepção popular de um gênero desconhecido, havendo um risco maior de fracasso comercial. Dessa maneira, pela aversão natural a riscos, os desenvolvedores aglutinariam os jogos em um número limitado de categorias.

Finalmente, retornando à figura 17, verificamos a hipótese **H5** apresentada no capítulo 3. De um lado, a taxa de crescimento de categorias por mês é pequena (o que pode ser constatado pela inclinação do ajuste linear) e, de outro, a escala do eixo y, do número de categorias, não ultrapassa a ordem de 10^3, comparada à escala de 10^6 e de 10^5 das figuras 4 e 12, respectivamente. Devemos ressaltar, no entanto, que há um desvio do padrão linear nos últimos meses, o que, como aconteceu no caso do ajuste exponencial para o número de jogos lançados, poderia indicar um processo de saturação: conforme o mercado se desenvolve e mais jogos estão disponíveis, torna-se cada vez mais difícil pensar em um estilo novo de jogo ou em alguma categoria inovadora.

Conclusões e perspectivas

Neste trabalho, apresentamos uma análise quantitativa da popularidade de *video games* de uma plataforma digital, a *Steam*, com base em dois modelos interpretativos, o malthusiano de crescimento e um modelo de memória envolvendo a distribuição lognormal e a transformação Box-Cox. Esse último, deve-se mencionar, permite um diálogo com outras áreas que investigam o impacto do uso de *video games* na sociedade, a exemplo da psicologia e da sociologia. Ressaltamos esse nexo, pois, assim nos parece, a interconexão entre a física e certas áreas das humanidades é ainda relativamente rara. Em suma, conseguimos, por meio das técnicas e modelos apresentados ao longo deste trabalho, identificar e descrever padrões e comportamentos coletivos sobre popularidade de *video games*.

Nos capítulos 1 e 2, apresentamos os métodos e materiais utilizados ao longo da tese. No capítulo 3, por sua vez, oferecemos uma análise de crescimento ao longo do tempo para o conjunto de dados global e identificamos um padrão de crescimento linear para o número de jogadores por hora e um padrão aproximadamente exponencial para o número de jogos lançados mensalmente. Além disso, a análise de médias e desvios-padrão conduziu-nos à transformação logarítmica dos dados e à verificação de que a distribuição lognormal é um bom ajuste para a popularidade relativa de jogos tanto no caso de um trimestre escolhido aleatoriamente quanto para o dado global. No entanto, como a distribuição empírica apresenta uma pequena assimetria que não é capturada por esse modelo, aplicamos a transformação Box-Cox como um meio de deformar continuamente a lognormal e propiciar uma melhora no ajuste. Interpretamos o modelo lognormal como emergente do produto de fragmentações aleatórias dos jogadores da plataforma em jogos distintos e, consistentemente com isso, que os jogadores são agentes sem memória (modelo *memoryless*). Por outro lado, identificamos a magnitude do parâmetro da transformação Box-Cox, isto é, a divergência entre a lognormal e a transformação Box-Cox ótima, como uma medida dos fenômenos de memória (modelo de memórias cultural e comunicativa). No capítulo 4, oferecemos o mesmo tipo de análise feita no capítulo 3, porém para uma categoria com maior número de jogos, a categoria *Indie*. Identificamos padrões similares em relação à assimetria na distribuição empírica e ao ajuste ótimo de Box-Cox. Esses padrões são verificados também em outras grandes categorias de *video games*, como mos-

tram os apêndices D e E. Além disso, identificamos os mesmos padrões de crescimento obtidos no caso global. Por último, ainda no capítulo 4, expandimos nossa análise para uma métrica distinta, isto é, deixamos de analisar o número de jogadores em um jogo para analisar o número de jogos em uma categoria. Ao realizarmos as mesmas análises para esse terceiro estrato, de maneira surpreendente, considerando uma métrica distinta, os resultados obtidos foram similares aos verificados tanto no caso global quanto no caso *Indie*. Concluindo, identificamos padrões robustos e similares entre si em três diferentes níveis hierárquicos (global, maiores categorias e entre categorias) a partir de duas métricas distintas (número médio de jogadores por hora e número de jogos em uma dada categoria).

Como mostram os resultados apresentados, a transformação Box-Cox foi suficiente para os objetivos delineados neste trabalho, permitindo identificar os efeitos da memória e descrever aspectos proeminentes das distribuições de popularidade relativa. Sem dúvida, outras famílias de distribuição poderiam ser empregadas na tentativa de refinar ainda mais a descrição dos dados empíricos.

Reconhecemos que uma limitação do presente estudo é que não conseguimos separar de maneira categórica as memórias culturais e comunicativas e, portanto, nossas interpretações da magnitude do parâmetro λ dependem do conhecimento de domínio. Como uma próxima etapa no aprimoramento do modelo, pode-se, em uma direção semelhante à proposta de (CANDIA et al., 2019), tentar desembaraçar os fenômenos de memória por meio de um modelo com parâmetros extras (por exemplo, uma transformação Box-Cox multivariada poderia ser potencialmente empregada e cada um dos parâmetros seria interpretado como um mecanismo de memória distinto). Provavelmente, esse novo modelo hipotético provavelmente exigiria outros tipos de dados para quantificar corretamente os diversos mecanismos de memória (por exemplo, o número de *tweets* dos usuários distintos ao se referir a um jogo nas semanas após seu lançamento seria uma forma de avaliar a memória comunicativa). Além disso, valeria a pena investigar novamente o crescimento linear do número de jogadores e o número exponencial de jogos lançados para ampliar a cobertura da análise realizada e para confirmar se a competição malthusiana continua válida. Esse último ponto seria de especial valor se fosse possível combinar vários bancos de dados de jogos, dado que, se as tendências de crescimento forem semelhantes em bases diferentes, isso poderia indicar a existência de um padrão universal no consumo de *video game*.

Em termos de perspectivas futuras, acreditamos que o modelo proposto para descrever

a popularidade relativa poderia enriquecer os debates que envolvem aspectos da psicologia e da sociologia. Por exemplo, uma questão perene em discussões acadêmicas (TENG et al., 2020; SHOSHANI; BRAVERMAN; MEIROW, 2021) e cotidianas (GUARDIAN, 2020) é se o uso de *video games* contribui para o comportamento violento dos jogadores. Uma possível aplicação futura desta tese seria, primeiro, correlacionar a popularidade das categorias de jogos violentos (por exemplo, na *Steam* há jogos com categorias como *Gore* e *Violent*) com indicadores socioeconômicos no mesmo período. Se a popularidade relativa dessas categorias for melhor descrita por uma distribuição transformada de Box-Cox ao invés de uma lognormal, poderia sugerir um efeito de memória comunicativa, ou seja, uma preferência coletiva por jogos violentos guiada pelos usuários. Essa descrição poderia então ser utilizada para melhor compreender esse padrão de consumo e para apoiar as ciências sociais na formulação de hipóteses relacionadas à questão acima mencionada.

Atualmente, existem muitas pesquisas envolvendo *video games* e o impacto social ocasionado. Porém, no melhor do nosso conhecimento, há poucos estudos que se utilizem de grandes bases de dados, como fizemos nesta tese. Assim, várias pesquisas relativas a *video games* podem ser pensadas da perspectiva de "grandes dados". Por exemplo, uma análise detalhada do papel de grandes jogos como influenciadores da criação de novos jogos poderia indicar se um comportamento seria da indústria de entretenimento em geral. Por exemplo, pode-se escolher um jogo dentre aqueles com maior número de jogadores por hora e, a partir de sua data de lançamento, verificar quantos novos jogos com as mesmas *tags* foram lançados desde aquela data. Uma comparação do número de jogos lançados com essas *tags* dois anos antes e dois anos depois do jogo popular poderia potencialmente mostrar o impacto nos desenvolvedores que este jogo popular teve. Indo além, como se tem um conjunto de *tags* para cada jogo, poder-se-ia criar uma medida de parentesco entre jogos, isto é, dois que concordem em 80% das *tags* podem ser pensados como bem próximos (ou concorrentes) enquanto jogos que tenham 50% das mesmas *tags*, por exemplo, podem ser vistos como relacionados ou o mais antigo, caso seja popular, teria influenciado indiretamente o jogo mais recente, etc. A possibilidade dessa análise setorial com graus variáveis de relação poderia servir de base para uma comparação entre um modelo de mercado em que um jogo age independentemente da existência de outros jogos e outro que inclua uma competição mais direta entre os jogos.

Outra perspectiva de investigação seria empregar redes complexas, por exemplo, a investigação sobre redes de corrupção (RIBEIRO et al., 2018). Uma possibilidade seria partir de uma rede em que os nós seriam os jogos e, se um usuário jogasse dois jogos quaisquer, criaríamos um *link* entre esses jogos. Entre outras possibilidades, a densidade

de *links* serviria de métrica para analisar a proximidade entre os jogos, sem necessariamente passar por uma classificação a priori de categoria. Construir esse tipo de rede e, com base na identificação dessa outra forma de categoria, tornaria possível refazer a análise categórica apresentada ao longo desta tese no intuito de verificar se a robustez do resultado permanece e se novas características emergem dessa classificação dinâmica.

De qualquer forma, esperamos ter deixado claro, com esses exemplos, que a investigação sobre *video games* e, mais especificamente, sobre a popularidade de *video games* é um campo fértil para pesquisas científicas, sejam elas multidisciplinares ou limitadas à análise quantitativa de dados. Além disso, oferecemos dois modelos interpretativos, o malthusiano e o de efeitos de memória, que pensamos não ser utilizados com frequência na área da física e, assim, podem ser úteis na investigação de outros dados de frequência ou popularidade, como *tweets*, filmes ou categorias musicais.

Referências

ADACHI, Paul JC; WILLOUGHBY, Teena. More than just fun and games: The longitudinal relationships between strategic video games, self-reported problem solving skills, and academic grades. **Journal of Youth and Adolescence**, Springer, v. 42, n. 7, p. 1041–1052, 2013.

AITCHISON, John; BROWN, James AC. The lognormal distribution with special reference to its uses in economics. **Journal of the Royal Statistical Society. Series A (General)**, Cambridge Univ. Press, v. 120, n. 4, p. 481, 1957.

ALLIGOOD, Kathleen; SAUER, Tim; YORKE, James. **Chaos: An Introduction to Dynamical Systems**. [S. l.]: Springer-Verlag, 1996.

ANALYTICS, SG. **2.5 quintillion bytes of data generated everyday**. [S. l.: s. n.], 2020. Disponível em: <https://us.sganalytics.com/blog/2-5-quintillion-bytes-of-data-generated-everyday-top-data-science-trends-2020/>.

ANDRÉ, Catherine; PLATTEAU, Jean-Philippe. Land relations under unbearable stress: Rwanda caught in the Malthusian trap. **Journal of Economic Behavior & Organization**, Elsevier, v. 34, n. 1, p. 1–47, 1998.

ANGUERA, Joaquin A et al. Video game training enhances cognitive control in older adults. **Nature**, Nature Publishing Group, v. 501, n. 7465, p. 97, 2013.

ASSIS BOMFIM, Jamille de et al. Local extinctions of obligate frugivores and patch size reduction disrupt the structure of seed dispersal networks. **Ecography**, Wiley Online Library, v. 41, n. 11, p. 1899–1909, 2018.

ASSMANN, Jan. Communicative and cultural memory. In: CULTURAL memories. [S. l.]: Springer, 2011. p. 15–27.

_____. Kollektives Gedächtnis und kulturelle Identität, 1988.

_____. Memória comunicativa e memória cultural. **História Oral**, v. 19, n. 1, p. 115–128, 2016.

ASSMANN, Jan; CZAPLICKA, John. Collective memory and cultural identity. **New german critique**, JSTOR, n. 65, p. 125–133, 1995.

ASSOCIATION, Entertainment Software; ASSOCIATION, Entertainment Software et al. Essential facts about the computer and video game industry: 2015 sales, demographic and usage data. **Washington, DC: Entertainment Software Association**, 2015.

BACAËR, Nicolas. **A short history of mathematical population dynamics**. [S. l.]: Springer Science & Business Media, 2011.

BAINBRIDGE, William Sims. The scientific research potential of virtual worlds. **Science**, American Association for the Advancement of Science, v. 317, n. 5837, p. 472–476, 2007.

BAVELIER, Daphne et al. Brains on video games. **Nature reviews neuroscience**, Nature Publishing Group, v. 12, n. 12, p. 763–768, 2011.

BERKELEY, George. **An essay towards a new theory of vision**. [S. l.]: Aaron Rhames, 1709.

BICKEL, Peter J; DOKSUM, Kjell A. An analysis of transformations revisited. **Journal of the American Statistical Association**, Taylor & Francis, v. 76, n. 374, p. 296–311, 1981.

BILLIEUX, Joël et al. Why do you play World of Warcraft? An in-depth exploration of self-reported motivations to play online and in-game behaviours in the virtual world of Azeroth. **Computers in Human Behavior**, Elsevier, v. 29, n. 1, p. 103–109, 2013.

BOX, George EP; COX, David R. An analysis of transformations. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 26, n. 2, p. 211–243, 1964.

BOYLE, Elizabeth; CONNOLLY, Thomas et al. Engagement in digital entertainment games: A systematic review. **Computers in Human Behavior**, Elsevier, v. 28, n. 3, p. 771–780, 2012.

BOYLE, Elizabeth; HAINEY, Thomas et al. An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. **Computers & Education**, Elsevier BV, v. 94, p. 178–192, 2016.

CANDIA, Cristian et al. The universal decay of collective memory and attention. **Nature Human Behaviour**, Nature Publishing Group, v. 3, n. 1, p. 82–91, 2019.

CHANG, Shan-mei; LIN, Sunny SJ. Online gaming motive profiles in late adolescence and the related longitudinal development of stress, depression, and problematic internet use. **Computers & Education**, Elsevier, v. 135, p. 123–137, 2019.

CHAOS. **About**. [S. l.: s. n.], 2021. Disponível em: <https://aip.scitation.org/cha/info/focus>.

CHEBYSHEV, P. Sur deux théorèmes relatifs aux probabilités. **Acta Mathematica**, Institut Mittag-Leffler, v. 14, p. 305–315, 1900.

CONNOLLY, Thomas et al. A systematic literature review of empirical evidence on computer games and serious games. **Computers & Education**, Elsevier, v. 59, n. 2, p. 661–686, 2012.

DANIEL, Muriel; GARRY, Crawford. **Video games as culture: considering the role and importance of video games in contemporary society**. [S. l.]: Routledge, 2018.

DEGROOT, M.H.; SCHERVISH, M.J. **Probability and statistics**. [S. l.]: Addison-Wesley, 2012.

DEKKING, Frederik Michel et al. **A Modern Introduction to Probability and Statistics: Understanding why and how**. [S. l.]: Springer Science & Business Media, 2005.

DEMETRIUS, Lloyd; LEGENDRE, Stéphane. Evolutionary entropy predicts the outcome of selection: competition for resources that vary in abundance and diversity. **Theoretical Population Biology**, Elsevier, v. 83, p. 39–54, 2013.

DOMINGUEZ, Adrian et al. Gamifying learning experiences: Practical implications and outcomes. **Computers & Education**, Elsevier, v. 63, p. 380–392, 2013.

DOUGHERTY, Christopher. **Introduction to econometrics**. [S. l.]: Oxford University Press, 2011.

DUGGAN, Maeve. **Gaming and Gamers**. [S. l.: s. n.], 2015. Disponível em: <http://www.pewinternet.org/2015/12/15/gaming-and-gamers/>. Acesso em: 22 fev. 2019.

E, Physical Review. **About**. [S. l.: s. n.], 2021. Disponível em: <https://journals.aps.org/pre/about>.

EDWARDS, Cliff. **Valve Lines Up Console Partners in Challenge to Microsoft, Sony**. [S. l.: s. n.], 2013. Disponível em: <https://www.bloomberg.com/news/articles/2013-11-04/valve-lines-up-console-partners-in-challenge-to-microsoft-sony>. Acesso em: 27 abr. 2021.

ELLIOTT, Luther et al. The contribution of game genre and other use patterns to problem video game play among adult video gamers. **International Journal of Mental Health and Addiction**, Springer, v. 10, p. 948–969, 2012.

FEYNMAN, Richard P. The value of science. **Engineering and Science**, California Institute of Technology, v. 19, n. 3, p. 13–15, 1955.

FISCHER, Hans. **A history of the central limit theorem: From classical to modern probability theory**. [S. l.]: Springer, 2011.

FORTUNATO, Santo; CASTELLANO, Claudio. Scaling and universality in proportional elections. **Physical Review Letters**, APS, v. 99, n. 13, p. 138701, 2007.

GALAM, Serge. **Sociophysics: A physicist's modeling of psycho-political phenomena**. [S. l.]: Springer Science e Business Media, 2012.

GAUSS, Carl Friedrich. **Theoria combinationis observationum erroribus minimis obnoxiae**. [S. l.]: H. Dieterich, 1823. v. 2.

GONG, Diankun et al. Enhanced functional connectivity and increased gray matter volume of insula related to action video game playing. **Scientific Reports**, Nature Publishing Group, v. 5, p. 9763, 2015.

GOURIEROUX, Christian; MONFORT, Alain. **Statistics and econometric models**. [S. l.]: Cambridge University Press, 1995. v. 1, p. 170–175.

GRANIC, Isabela; LOBEL, Adam; ENGELS, Rutger CME. The benefits of playing video games. **American psychologist**, American Psychological Association, v. 69, n. 1, p. 66, 2014.

GRAY, James. **Steam Charts**. [S. l.: s. n.], 2020. Disponível em: <https://steamcharts.com>. Acesso em: 22 fev. 2020.

GUARDIAN, The. **Playing video games doesn't lead to violent behaviour**. [S. l.: s. n.], 2020. Disponível em: <https://www.theguardian.com/games/2020/jul/22/playing-video-games-doesnt-lead-to-violent-behaviour-study-shows>. Acesso em: 22 jul. 2020.

GUZZI, Diego. Per una definizione di memoria pubblica. Halbwachs, Ricoeur, Assmann, Margalit. **Scienza & Politica. Per una storia delle dottrine**, v. 23, n. 44, 2011.

HALBWACHS, Maurice. **A memória coletiva**. [S. l.]: Vértice, 1990.

_____. **Les cadres sociaux de la memoire**. [S. l.]: Walter de Gruyter, 2010. v. 5.

HALL, Mordaunt. **Poignant French Film**. [S. l.: s. n.], 1929. Disponível em: <https://www.nytimes.com/1929/03/31/archives/poignant-french-film-maria-falconetti-gives-unequaled-performance-a.html>.

HAMMAR, Emil Lundedal. The political economy of cultural memory in the videogames industry. **Digital Culture & Society**, Transcript Verlag, v. 5, n. 1, p. 61–84, 2019.

HENDERSON, A Ralph. The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. **Clinica Chimica Acta**, Elsevier, v. 359, n. 1-2, p. 1–26, 2005.

HIGHAM, Kyle W et al. Fame and obsolescence: Disentangling growth and aging dynamics of patent citations. **Physical Review E**, APS, v. 95, n. 4, p. 042309, 2017.

HOFFMAN, Bobby; NADELSON, Louis. Motivational engagement and video gaming: A mixed methods study. **Educational Technology Research and Development**, Springer, v. 58, n. 3, p. 245–270, 2010.

JAYNES, Edwin T. **Probability theory: The logic of science**. [S. l.]: Cambridge university press, 2003.

JOURNAL-B, European Physics. **Recent articles**. [S. l.: s. n.], 2021. Disponível em: <https://epjb.epj.org/>.

KANT, Immanuel. Critique of pure reason. 1781. **Modern Classical Philosophers, Cambridge, MA: Houghton Mifflin**, p. 370–456, 1908.

KENNEDY, Peter. **A guide to econometrics**. [S. l.]: John Wiley & Sons, 2008.

KUEIDER, Alexandra M et al. Computerized cognitive training with older adults: A systematic review. **PLOS ONE**, Public Library of Science, v. 7, n. 7, e40588, 2012.

KUSS, Daria Joanna; GRIFFITHS, Mark D. Internet gaming addiction: A systematic review of empirical research. **International Journal of Mental Health and Addiction**, Springer, v. 10, n. 2, p. 278–296, 2012.

LAB, Complex. **Complex Lab at State University of Maringá**. [S. l.: s. n.], 2012. Disponível em: <http://complex.pfi.uem.br/>.

LANDFRIED, Gustavo; SLEZAK, Diego Fernández; MOCSKOS, Esteban. Faithfulness-boost effect: Loyal teammate selection correlates with skill acquisition improvement in online games. **PLOS ONE**, Public Library of Science, v. 14, n. 3, e0211014, 2019.

LEE, Maw Lin; LOSCHKY, David. Malthusian population oscillations. **The Economic Journal**, JSTOR, v. 97, n. 387, p. 727–739, 1987.

LEGENDRE, Adrien Marie. **Nouvelles méthodes pour la détermination des orbites des comètes**. [S. l.]: Firmin Didot, 1806.

LIMPERT, Eckhard; STAHEL, Werner A; ABBT, Markus. Log-normal distributions across the sciences: keys and clues: on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: that is the question. **BioScience**, American Institute of Biological Sciences, v. 51, n. 5, p. 341–352, 2001.

LIND, Pedro G et al. Spreading gossip in social networks. **Physical Review E**, APS, v. 76, n. 3, p. 036117, 2007.

LORENZ-SPREEN, Philipp et al. Accelerating dynamics of collective attention. **Nature communications**, Nature Publishing Group, v. 10, n. 1, p. 1–9, 2019.

MALTHUS, Thomas Robert; WINCH, Donald; JAMES, Patricia. **Malthus:'An Essay on the Principle of Population'**. [S. l.]: Cambridge University Press, 1992.

MANTEGNA, Rosario N; STANLEY, H Eugene. **Introduction to econophysics: correlations and complexity in finance**. [S. l.]: Cambridge university press, 1999.

MENDES, Leonardo O; CUNHA, Leonardo R; MENDES, Renio S. Popularity of Video Games and Collective Memory. **Entropy**, 24(7), p. 860, 2022. DOI: https://doi.org/10.3390/e24070860.

MERRITT, Sears; CLAUSET, Aaron. Environmental structure and competitive scoring advantages in team competitions. **Scientific Reports**, Nature Publishing Group, v. 3, p. 3067, 2013.

MURRAY, James D. **Mathematical biology, interdisciplinary applied mathematics**. [S. l.]: Springer-Verlag, 2004.

NEKOLA, Jeffrey C et al. The Malthusian–Darwinian dynamic and the trajectory of civilization. **Trends in ecology & evolution**, Elsevier, v. 28, n. 3, p. 127–130, 2013.

PARK, Jowon; SONG, Yosep; TENG, Ching-I. Exploring the links between personality traits and motivations to play online games. **Cyberpsychology, Behavior, and Social Networking**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 14, n. 12, p. 747–751, 2011.

PETRY, Nancy M et al. An international consensus for assessing internet gaming disorder using the new DSM-5 approach. **Addiction**, Wiley Online Library, v. 109, n. 9, p. 1399–1406, 2014.

PHYSICA-A. **Aims and scope**. [S. l.: s. n.], 2021. Disponível em: <https://www.sciencedirect.com/journal/physica-a-statistical-mechanics-and-its-applications/about/aims-and-scope>.

QIN, Jian; YU, Yu; ZHANG, Pengjie. The copula of the cosmological matter density field is non-Gaussian. **The Astrophysical Journal**, IOP Publishing, v. 897, n. 1, p. 105, 2020.

RATKIEWICZ, Jacob et al. Characterizing and modeling the dynamics of online popularity. **Physical Review Letters**, APS, v. 105, n. 15, p. 158701, 2010.

REHBEIN, Florian et al. Video game playing in the general adult population of Germany: Can higher gaming time of males be explained by gender specific genre preferences? **Computers in Human Behavior**, Elsevier, v. 55, p. 729–735, 2016.

RIBEIRO, Haroldo V et al. The dynamical structure of political corruption networks. **Journal of Complex Networks**, Oxford University Press, v. 6, n. 6, p. 989–1003, 2018.

RICKLEFS, Robert E. **The economy of nature**. [S. l.]: Macmillan, 2008.

SHARP, Paul; STRULIK, Holger; WEISDORF, Jacob. The determinants of income in a Malthusian equilibrium. **Journal of Development Economics**, Elsevier, v. 97, n. 1, p. 112–117, 2012.

SHOSHANI, Anat; BRAVERMAN, Shahar; MEIROW, Galya. Video games and close relations: Attachment and empathy as predictors of children's and adolescents' video game social play and socio-emotional functioning. **Computers in Human Behavior**, Elsevier, v. 114, p. 106578, 2021.

SOLEYMANI, Samira. Exact Box-Cox Analysis - TESE CHECAR ABNT COMO CITAR, 2018.

SONG, Shuang et al. Improving representation of collective memory in socio-hydrological models and new insights into flood risk management. **Journal of Flood Risk Management**, Wiley Online Library, v. 14, n. 1, e12679, 2021.

SPITZER, John J. A primer on Box-Cox estimation. **The Review of Economics and Statistics**, JSTOR, p. 307–313, 1982.

SUMPTER, David JT. **Collective Animal Behavior**. [S. l.]: Princeton University Press, 2010.

SZELL, Michael; LAMBIOTTE, Renaud; THURNER, Stefan. Multirelational organization of large-scale social networks in an online world. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 107, n. 31, p. 13636–13641, 2010.

SZOLNOKI, A; OLIVEIRA, BF de; BAZEIA, D. Pattern formations driven by cyclic interactions: A brief review of recent developments. **EPL (Europhysics Letters)**, IOP Publishing, v. 131, n. 6, p. 68001, 2020.

TAKATALO, Jari et al. User experience in digital games: Differences between laboratory and home. **Simulation & Gaming**, SAGE Publications Sage CA: Los Angeles, CA, v. 42, n. 5, p. 656–673, 2011.

TAKESHI, Amemiya; AMEMIYA, TAKESHI AUTOR. **Advanced econometrics**. [S. l.]: Harvard University Press, 1985. p. 125–127.

TAYLOR, Haydn. **Steam reaches 90m monthly active users**. [S. l.: s. n.], 2018. Disponível em: <https://www.gamesindustry.biz/articles/2018-10-24-steam-reaches-90m-monthly-active-users>. Acesso em: 27 abr. 2021.

TENG, Zhaojun et al. Violent video game exposure and (Cyber) bullying perpetration among Chinese youth: The moderating role of trait aggression and moral identity. **Computers in Human Behavior**, Elsevier, v. 104, p. 106193, 2020.

TESTI, Angela; IVALDI, Enrico. Material versus social deprivation and health: A case study of an urban area. **The European Journal of Health Economics**, Springer, v. 10, n. 3, p. 323, 2009.

TSALLIS, Constantino. Introduction to nonextensive statistical mechanics: approaching a complex world. **Springer**, **Springer**, v. 1, n. 1, p. 2–1, 2009.

TUKEY, John W. On the comparative anatomy of transformations. **The Annals of Mathematical Statistics**, JSTOR, p. 602–632, 1957.

VALVE. **Steam**. [S. l.: s. n.], 2020. Disponível em: <https://store.steampowered.com/>. Acesso em: 22 fev. 2020.

_____. **Steam Search**. [S. l.: s. n.], 2021. Disponível em: <https://store.steampowered.com/search?category1=998>. Acesso em: 27 abr. 2021.

_____. **Steam Stats**. [S. l.: s. n.], 2020. Disponível em: <https://store.steampowered.com/stats>. Acesso em: 22 fev. 2020.

VERÍSSIMO, Érico. **Obras: A vida de Joana d'Arc**. [S. l.]: Editora Globo, 1960.

WANG, Dashun; SONG, Chaoming; BARABÁSI, Albert-László. Quantifying long-term scientific impact. **Science**, American Association for the Advancement of Science, v. 342, n. 6154, p. 127–132, 2013.

WEBER, Regina; PEREIRA, Elenita Malta. Halbwachs e a Memória: contribuição à História Cultural. **Revista Territórios e Fronteiras**, v. 3, n. 1, p. 104–126, 2010.

WIKIPEDIA. **Teorema de Gauss-Markov**. [S. l.: s. n.]. Disponível em: <https://en.wikipedia.org/wiki/Gauss%5C%E2%5C%80%5C%93Markov_theorem>. Acesso em: 12 out. 2021.

WILSON, Marcella; NICHOLAS, Charles. Topological analysis of an online social network for older adults. In: ACM. PROCEEDINGS of the 2008 ACM Workshop on Search in Social Media. [S. l.: s. n.], 2008. p. 51–58.

WITKOWSKI, Wallace. **Videogames are a bigger industry than movies and North American sports combined, thanks to the pandemic**. [S. l.: s. n.], 2021. Disponível em: <https://www.marketwatch.com/story/videogames-are-a-bigger-industry-than-sports-and-movies-combined-thanks-to-the-pandemic-11608654990/>.

WU, Fang; HUBERMAN, Bernardo A. Novelty and collective attention. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 104, n. 45, p. 17599–17601, 2007.

YANG, Linchuan et al. Does bus accessibility affect property prices? **Cities**, v. 84, p. 56–65, 2019. ISSN 0264-2751.

YEE, Nick. Motivations for play in online games. **CyberPsychology & Behavior**, Rapid Communications, v. 9, n. 6, p. 772–775, 2006.

YEO, In-Kwon; JOHNSON, Richard A. A new family of power transformations to improve normality or symmetry. **Biometrika**, Oxford University Press, v. 87, n. 4, p. 954–959, 2000.

YOKOTA, Yusuke; SOSHI, Takahiro; NARUSE, Yasushi. Error-related negativity predicts failure in competitive dual-player video games. **PLOS ONE**, Public Library of Science, v. 14, n. 2, e0212483, 2019.

YOOK, Soon-Hyung; KIM, Yup. Origin of the log-normal popularity distribution of trending memes in social networks. **Physical Review E**, APS, v. 101, n. 1, p. 012312, 2020.

ZHANG, Chi; LEI, Pu-wen; KOYAMADA, Koji. Calculation of Extreme Precipitation Threshold by Percentile Method Based on Box-Cox Transformation. In: SPRINGER. ASIAN Simulation Conference. [S. l.: s. n.], 2018. p. 286–295.

APÊNDICE A – Código desenvolvido e utilizado para a obtenção, tratamento e análise dos dados

A.1 Código

Neste apêndice, listamos o código em *Python* utilizado ao longo da presente pesquisa. Esse código foi separado em treze partes, seguindo a disposição apresentada no capítulo 2.

A.1.1 Extração dos dados

```
#small script to grab HTML tables from https://steamdb.info/apps/

from bs4 import BeautifulSoup as BS
import urllib.request as urllib2
import csv
import httplib
import sys
import time

import stem
import stem.connection

from stem import Signal
from stem.control import Controller

### ALTERNATIVE COOKIES

#cookies = {
#    'timezoneOffset': '-10800,0',
#    '_ga': 'GA1.2.487916651.1614090293',
```

```
#    'sessionid': '6025779bbbdd62aa046209a1',
#    'steamCountry': 'BR%7C50b7098e97fc05484363f55ab12cf154',
#    'app_impressions': '1382330@2_100300_300_',
#    '_gid': 'GA1.2.1996107189.1614275655',
#    'steamLoginSecure':
    '76561197983234798%7C%7C867DD3A9B23F7E82A9E304B67B59B30FDE826C0C',
#    'steamMachineAuth76561197983234798': 'D1B4620F1D50556F622FEB16C3CB6728A06D157C',
#    'steamRememberLogin': '76561197983234798%7C%7C0505d01f4be2b7e6e03beb2608ae14fd',
#    'browserid': '2432478171016300114',
#}

#headers = {
#    'User-Agent': 'Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:85.0) Gecko/20100101 Firefox/85.0',
#    'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8',
#    'Accept-Language': 'en-US,en;q=0.5',
#   'Referer': 'https://steamdb.info/',
#    'Content-Type': 'application/x-www-form-urlencoded',
#    'Origin': 'https://steamdb.info',
#    'Connection': 'keep-alive',
#    'Upgrade-Insecure-Requests': '1',
#}

#data = {
# 'openid.identity': 'http://specs.openid.net/auth/2.0/identifier_select',
# 'openid.claimed_id': 'http://specs.openid.net/auth/2.0/identifier_select',
# 'openid.ns': 'http://specs.openid.net/auth/2.0',
# 'openid.mode': 'checkid_setup',
# 'openid.realm': 'https://steamdb.info/',
# 'openid.return_to': 'https://steamdb.info/login/'
#}

###


cookies = {
    '__cfduid': 'd383e5e72ada9ddf19058502b638c90ee1614275612',
    '__cf_bm':
        '26ceb1e8b1042b38f05d5402a3b0d76351d72051-1614275614-1800-ASNRVuZlrmEe/mvYX9hlUbK4Xey9qI-
    '_ga': 'GA1.2.466569414.1614275615',
    '_gid': 'GA1.2.2115297284.1614275615',
    'cf_chl_2': '6019e7c56bdac43',
    'cf_chl_prog': 'x19',
    'cf_clearance': '23d38f07625c7f89ef2c8907541b7b05ae408ffe-1614275652-0-150',
```

```
}

headers = {
    'User-Agent': 'Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:85.0) Gecko/20100101 Firefox/85.0',
    'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8',
    'Accept-Language': 'en-US,en;q=0.5',
    'Referer': 'https://steamcommunity.com/openid/login',
    'Connection': 'keep-alive',
    'Upgrade-Insecure-Requests': '1',
    'TE': 'Trailers',
}

params = (
    ('openid.ns', 'http://specs.openid.net/auth/2.0'),
    ('openid.mode', 'id_res'),
    ('openid.op_endpoint', 'https://steamcommunity.com/openid/login'),
    ('openid.claimed_id', 'https://steamcommunity.com/openid/id/76561197983234798'),
    ('openid.identity', 'https://steamcommunity.com/openid/id/76561197983234798'),
    ('openid.return_to', 'https://steamdb.info/login/'),
    ('openid.response_nonce', '2021-02-25T18:02:27Z73eVE3ELEIFk6P0yWSn2VVwhxWM='),
    ('openid.assoc_handle', '1234567890'),
    ('openid.signed',
        'signed,op_endpoint,claimed_id,identity,return_to,response_nonce,assoc_handle'),
    ('openid.sig', 'uoc0QPM9qXybn75YNRmjdij1Tuo='),
)

game_list = []

def url_ok(url):
    def _set_urlproxy():
        proxy_support = urllib2.ProxyHandler({"http" : "127.0.0.1:8118"})
        opener = urllib2.build_opener(proxy_support)
        urllib2.install_opener(opener)

    _set_urlproxy()
    s = requests.Session()
    request = s.get('https://steamdb.info/login/', headers=headers, params=params,
        cookies=cookies)
    con = urllib2.urlopen(request)
    return urllib2.urlopen(request), con.code == 200

'''
USE THIS ONE FOR SMALLER SCRAPING
```

```
'''
#def url_ok(url):
    #blocks simple requests, need an user-agent header
# try:
#     req = urllib2.Request(url, headers={'User-Agent' : "Magic Browser"})
#     con = urllib2.urlopen(req)
# except urllib2.HTTPError, e:
#     if e.code == 429:
#         time.sleep(5)
#         url_ok(url)
# return req, con.code == 200

def extractor(myurl, output):
    if type(output) != str:
        print('IO Error: expecting a name.')
        return 1
    try:
        req, ok_flag = url_ok(myurl)
        soup = BS(req.read())
        table = soup.find('table', {'class': 'table-products table-hover text-left'})
    except:
        f = open(output,'a')
        for game in game_list:
            f.write('app/' + game + '\n')
        f.close()
        sys.exit(1)
    td_list = table.findAll('td')
    i = 1
    j = 1
    for td in td_list:
        try:
            if 'Game' in td_list[i+1].text.encode('ascii', 'ignore'):
                game_appid = td_list[i].text.encode('ascii', 'ignore')
                game_list.append(game_appid)
            i=i+4
        except IndexError: break
    return 0

def main():
    if len(sys.argv) != 2:
        print('Usage: extractor.py output_file.txt')
        return 1
    tot = TOT_OF_GAMES
```

```
    output = sys.argv[1]
    try:
        for i in xrange(0, tot):
            print(i, tot)
            myurl = 'https://steamdb.info/apps/page' + str(i)
            extractor(myurl, output)
    except:
        f = open(output, 'a')
        for game in game_list:
            f.write('app/' + game + '\n')
        f.close()
        sys.exit(1)
    f = open(output, 'a')
    for game in game_list:
        f.write(game + '\n')
    f.close()
    return 0

if __name__ == "__main__":
    main()
```

A.1.2 Extração da *Steamcharts*

```
#small script to grab HTML tables from http://steamcharts.com

from bs4 import BeautifulSoup as BS
import urllib.request as urllib2
import requests
from datetime import datetime as dt
from datetime import timedelta as td
import time
import sys
import re

import stem
import stem.connection

from stem import Signal
from stem.control import Controller
```

```
LIMIT = 300.0
headers={'User-Agent': 'Magic Browser'}
DIR = myDir
SECONDSBETWCHECKS = 5
oldIP = "0.0.0.0"
newIP = "0.0.0.0"

def renew_connection():
    with Controller.from_port(port = 9051) as controller:
        controller.authenticate(password = 'my_password')
        controller.signal(Signal.NEWNYM)
        controller.close()

def url_ok(url):
    def _set_urlproxy():
        proxy_support = urllib2.ProxyHandler({'http' : '127.0.0.1:8118'})
        opener = urllib2.build_opener(proxy_support)
        urllib2.install_opener(opener)

    _set_urlproxy()
    try:
        con = requests.get(url, timeout=5)
    except urllib2.HTTPError as e:
        print(e)
        if e.code == 429:
            if newIP == "0.0.0.0":
                renew_connection()
                newIP = request('http://icanhazip.com')
            else:
                oldIP = newIP
                renew_connection()
                newIP = request('http://icanhazip.com/')

            while oldIP == newIP:
                time.sleep(SECONDSBETWCHECKS)
                newIP = request('http://icanhazip.com/')
                url_ok(url)
```

```
        elif e.code == 502:
            print ('error 502')
            return url_ok(url)
        elif e.code == 302:
            raise
        else:
            f = open('errorlog.txt', 'a')
            f.write(url + '\n')
            f.close()
            raise
    except urllib2.URLError as e:
        f = open('errorlog.txt', 'a')
        f.write(url + '\n')
        f.close()
        raise

    return con, con.status_code == 200

def extractor(myurl, output):
    names = []
    nums = []
    tmp = []
    id_list = []
    if type(output) != str:
        raise ValueError
    html, ok_flag = url_ok(myurl)
    if not ok_flag:
        raise ValueError
    soup = BS(html.content, "lxml")
    table = soup.find('table', {'class': 'common-table'})
    td_list = table.findChildren('td')
    time_series_means = []
    for i in range(0, len(td_list), 5):
        try:
            meanplayers = td_list[i+1].text
            maxplayers = td_list[i+4].text
            time_series_means.append( (meanplayers, maxplayers) )
        except IndexError:
```

```
            pass
    with open(output, 'a') as f:
        for entry in time_series_means:
            f.write( entry[0] + ',' + entry[1] + '\n')
    return 0

def main():
    if len(sys.argv) != 3:
        print('Usage: SCE_top.py url output_file.csv')
        return 1
    myurl = sys.argv[1]
    output = sys.argv[2]

    extractor(myurl, output)
    return 0

if __name__ == "__main__":
    main()
```

A.1.3 Laço para todos os jogos

```
#simple script to loop in steam charts over all app IDs extracted from steamdb

import SCE
import urllib.request as urllib2

DIR = myDirDatadump

def line_by_line(fname):
    with open(fname) as f:
        content = f.readlines()
    content = [x.strip() for x in content]
    return content

def main():
    id_list = line_by_line("index.txt")
    tot_len = len(id_list) - STOP
    i = 1
    for id_num in id_list[START:]:
```

```
            print(str(i) + "/" + str(tot_len))
            try:
                if 'b' in id_num:
                    id_num = (str(id_num)).split('b')[0]
                    name = DIR + id_num + ".csv"
                    print("http://steamcharts.com/bundle/" + id_num)
                    SCE.extractor("https://steamcharts.com/app/" + id_num, name)
                else:
                    name = DIR + str(id_num) + ".csv"
                    print("http://steamcharts.com/app/" + str(id_num))
                    SCE.extractor("https://steamcharts.com/app/" + str(id_num), name)
            except urllib2.HTTPError:
                pass
            except ValueError:
                pass
            i += 1
        return 0

if __name__ == "__main__":
    main()
```

A.1.4 Extração das categorias

```
from bs4 import BeautifulSoup as BS
import urllib.request as urllib2
import urllib
import sys
import re
import requests
from operator import itemgetter
from collections import OrderedDict

import stem
import stem.connection

from stem import Signal
from stem.control import Controller

__TAGLIST = [...]

headers={'User-Agent': 'Magic Browser'}
```

```
DIR = myDir
SECONDSBETWCHECKS = 15
_TRIES = 10
oldIP = "0.0.0.0"
newIP = "0.0.0.0"
tag_list = []
tag_dict = {}
cookies = { 'birthtime': '283993201', 'mature_content': '1' }
bad_chars = ['<', 'l', 'i', '>', '<', 's', 't', 'r', 'o', 'n', 'g', '>', 'S', 't', 'o', 'r', 'a',
    'g', 'e', ':', '<', '/', 's', 't', 'r', 'o', 'n', 'g', '>', 'a', 'v', 'a', 'i', 'l', 'a', 'b',
    'l', 'e', ' ', 's', 'p', 'a', 'c', 'e']

def data_dump(output):
    f = open(output, 'a')
    fstring = '('
    _fstring = 'dev('
    sorted_dict = OrderedDict(sorted(tag_dict.items(), key=itemgetter(1), reverse=True))
    for entry in sorted_dict:
        for _tag in sorted_dict[entry][3]:
            fstring += _tag.decode('ascii') + ','
        for _genre in sorted_dict[entry][4]:
            _fstring += _genre + ','
        f.write(entry + ';' + sorted_dict[entry][0] + ';' + sorted_dict[entry][1] + ';' +
            sorted_dict[entry][2] + ';' + fstring + ')' + ';' + _fstring + ')' + '\n')
        fstring = '('
        _fstring = 'dev('
    f.close()
    return 0

def renew_connection():
    with Controller.from_port(port = 9051) as controller:
        controller.authenticate(password = 'my_password')
        controller.signal(Signal.NEWNYM)
        controller.close()

def url_ok(url, output):
    i = 0
    def _set_urlproxy():
        proxy_support = urllib2.ProxyHandler({'http' : '127.0.0.1:8118'})
        opener = urllib2.build_opener(proxy_support)
        urllib2.install_opener(opener)

    _set_urlproxy()
```

```
    try:
        r = requests.get(url, cookies=cookies)
    except urllib2.HTTPError as e:
        if e.code == 429:
            if newIP == "0.0.0.0":
                renew_connection()
                newIP = request('http://icanhazip.com')
            else:
                oldIP = newIP
                renew_connection()
                newIP = request('http://icanhazip.com/')

            while oldIP == newIP:
                if i < _TRIES:
                    time.sleep(SECONDSBETWCHECKS)
                    newIP = request('http://icanhazip.com/')
                    i += 1
                else:
                    data_dump(output)
                    raise
            url_ok(url)
        elif e.code == 502:
            print ('error 502')
            return url_ok(url, output)
        elif e.code == 302:
            data_dump(output)
            raise
        else:
            data_dump(output)
            raise
    except urllib2.URLError as e:
        data_dump(output)
        raise
    except:
        data_dump(output)
        raise
    return r.text, r.status_code == 200

def line_by_line(fname):
    with open(fname) as f:
            content = f.readlines()
    content = [x.strip() for x in content]
    return content
```

```
def extractor(myurl, output):
    html, ok_flag = url_ok(myurl, output)
    genrs = []
    try:
        soup = BS(html, "lxml")
        hrefs = soup.findAll('a', {'class': 'app_tag'})
        lis = soup.findAll(text=re.compile('Storage'))
        if len(lis) == 0:
            lis = soup.findAll(text=re.compile('Hard Drive'))
        try:
            _par = [li.parent for li in lis]
            _text = [str(par.parent).strip() for par in _par]
            for c in bad_chars: _text[0] = _text[0].replace(c, '')
            for div in soup.findAll('div', {'class': 'details_block'}):
                for a in div.findAll('a'):
                    _b = a.text.encode('ascii', 'ignore')
                    _b = _b.decode('ascii')
                    if _b in __TAGLIST:
                        genrs.append(_b)
        except IndexError:
            _text = []
            _text.append('0MB')
        try:
            price = (soup.find('div', {'class': 'game_purchase_price price'})).text.encode('ascii',
                'ignore')
            price = price.decode('ascii')
            if 'Free To Play' in price:
                price = str(0)
            price = price.replace('R$ ', '')
            price = price.replace(',', '.')
        except AttributeError:
            try:
                price = (soup.find('div', {'class':
                    'discount_original_price'})).text.encode('ascii', 'ignore')
                price = price.decode('ascii')
                if 'Free To Play' in price:
                    price = str(0)
                price = price.replace('R$ ', '')
                price = price.replace(',', '.')
            except AttributeError: #game from the future
                price = 'N/A'
            except:
```

```
                data_dump(output)
                raise
    except:
        data_dump(output)
        raise
    return hrefs, 'N/A', genrs, price

def main():
    output = sys.argv[1]
    id_list = line_by_line('dates2021.txt')
    i = 1
    tot = len(id_list)
    tag_list = []
    genrs = []
    for id_num in id_list:
        strnum1, strnum2 = str(i), str(tot)
        fprint = strnum1 + '/' + strnum2
        print(fprint)
        i += 1
        urlslice = str(id_num.split(',')[0])
        appurl = 'https://store.steampowered.com/app/' + urlslice
        date = id_num.split(',')[1]
        _day = date.split('-')[2]
        year, month, day = date.split('-')[0], date.split('-')[1], _day.split(' ')[0]
        try:
            hrefs, size, genrs, price = extractor(appurl, output)
        except (urllib2.HTTPError, urllib2.URLError) as e:
            continue
        for href in hrefs:
            _gametag = href.text.encode('ascii', 'ignore')
            tag_list.append(_gametag)
        tag_dict[urlslice] = [year + '/' + month + '/' + day, price, size, tag_list, genrs]
        tag_list = []

    data_dump(output)
    return 0

if __name__ == "__main__":
    main()
```

A.1.5 Construção da lista de *tags*

```
from BeautifulSoup import BeautifulSoup as BS
import urllib2
import sys
from operator import itemgetter
from collections import OrderedDict

import stem
import stem.connection

from stem import Signal
from stem.control import Controller

headers={'User-Agent': 'Magic Browser'}
DIR = myDir
SECONDSBETWCHECKS = 5
oldIP = "0.0.0.0"
newIP = "0.0.0.0"
popular_tags = {}

def renew_connection():
    with Controller.from_port(port = 9051) as controller:
        controller.authenticate(password = 'my_password')
        controller.signal(Signal.NEWNYM)
        controller.close()

def url_ok(url, output):
    def _set_urlproxy():
        proxy_support = urllib2.ProxyHandler({'http' : '127.0.0.1:8118'})
        opener = urllib2.build_opener(proxy_support)
        urllib2.install_opener(opener)

    _set_urlproxy()
    try:
        request = urllib2.Request(url, None, headers)
        con = urllib2.urlopen(request)
    except urllib2.HTTPError, e:
        if e.code == 429:
            if newIP == "0.0.0.0":
                renew_connection()
                newIP = request('http://icanhazip.com')
            else:
                oldIP = newIP
```

```
                renew_connection()
                newIP = request('http://icanhazip.com/')

            while oldIP == newIP:
                    time.sleep(SECONDSBETWCHECKS)
                newIP = request('http://icanhazip.com/')
            url_ok(url)
        elif e.code == 502:
            print ('error 502')
            return url_ok(url, output)
        elif e.code == 302:
            raise
        else:
            f = open(output, 'a')
            sorted_dict = OrderedDict(sorted(popular_tags.items(), key=itemgetter(1), reverse=True))
            for entry in sorted_dict:
                f.write(entry + ',' + str(sorted_dict[entry]) + '\n')
            f.close()
            raise
    except urllib2.URLError, e:
        f = open(output, 'a')
        sorted_dict = OrderedDict(sorted(popular_tags.items(), key=itemgetter(1), reverse=True))
        for entry in sorted_dict:
            f.write(entry + ',' + str(sorted_dict[entry]) + '\n')
        f.close()
        raise

    return con, con.code == 200

def line_by_line(fname):
    with open(fname) as f:
            content = f.readlines()
    content = [x.strip() for x in content]
    return content

def extractor(myurl, output):
    html, ok_flag = url_ok(myurl, output)
    try:
        if ok_flag:
            soup = BS(html)
            hrefs = soup.findAll('a', {'class': 'app_tag'})
    except urllib2.HTTPError, e:
        raise
```

```
    return hrefs

def main():
    output = sys.argv[1]

    ###loop
    id_list = line_by_line('index.txt')
    i = 1
    tot = len(id_list)
    tag_list = []
    for id_num in id_list:
        percentage = (float(i)/tot)*100
        print '\%.2f' \% percentage + '\%'
        i += 1
        print id_num.split(",")[0]
        appurl = 'https://store.steampowered.com/' + str(id_num.split(",")[0])
        appid = appurl.split("/")[-1]
        try:
            hrefs = extractor(appurl, output)
        except urllib2.HTTPError, e:
            continue
        for href in hrefs:
            tag = href.text.encode('ascii', 'ignore')
            if tag in tag_list:
                popular_tags[tag] += 1
            else:
                popular_tags[tag] = 1
            tag_list.append(tag)

    f = open(output, 'a')
    sorted_dict = OrderedDict(sorted(popular_tags.items(), key=itemgetter(1), reverse=True))
    for entry in sorted_dict:
        f.write(entry + ',' + str(sorted_dict[entry]) + '\n')
    f.close()
    return 0

if __name__ == "__main__":
    main()
```

A.1.6 Higiene dos dados e construção do *dataframe*

```
from ipykernel import kernelapp as app
from matplotlib.ticker import LogFormatterSciNotation
from scipy import stats
from math import log
from os import listdir
from os.path import isfile , join
from pyhvr import *
from statistics import mean, stdev
from pylab import polyfit
from datetime import datetime
from mpltools import annotation
import numpy as np
import matplotlib.pyplot as plt
import scipy
import scipy.stats
import pylab
import pandas as pd
import glob
import string
import matplotlib as mpl
import seaborn as sns
import os
import matplotlib.pylab as pl
import matplotlib.gridspec as gridspec
import warnings
warnings.filterwarnings('ignore')

#creates and cleans the major dataframe: one including release dates, appids, and tags.

header_list = ["AppID", "Date", "Price", "Size", "Tags", "DevTags"]
df = pd.read_csv('formatted_content_final.txt', sep=';', names=header_list)
repl = []
repl_dev = []
repl_date = []
_e = []
_edev = []
for entry in df.Tags.values:
    try:
        entry = entry.replace('(', '')
        entry = entry.replace(')', '')
    except AttributeError:
        entry = []
    except:
```

```
            print(entry)
            raise
        if entry:
            _e = entry.split(',')
        repl.append(_e[:-1])
for entry in df.DevTags.values:
    try:
        entry = entry.replace('(', '')
        entry = entry.replace(')', '')
        entry = entry.replace('dev', '')
    except AttributeError:
        entry = []
    except:
        print(entry)
        raise
    if entry:
        _edev = entry.split(',')
    repl_dev.append(_edev[:-1])
for entry in df.Date.values:
    repl_date.append(datetime.strptime(entry, '%Y/%m/%d'))
df.Tags = repl
df.DevTags = repl_dev
df.Date = repl_date
ender = glob.glob(myDir + '*.csv')
for entry in df.AppID.values:
    if len(df[df['AppID'] == entry]) > 1:
        index_to_drop = (df[df['AppID'] == entry].index.values[-1])
        df = df.drop(index=index_to_drop)
df = df.set_index('AppID')

#merge time series with tags and release dates

ender = glob.glob(myDir + '*.csv')
tseries_list = []
i = 0
for entry in df.index.values:
    means = []
    file = myDir + str(entry) + '.csv'
    try:
        with open(file) as f:
            lines = f.readlines()
            for line in lines:
                mean = float((line.split(',')[0]))
```

```
                means.append(mean)
            tseries_list.append(means[::-1])
        except:
            tseries_list.append([])
            pass
df['PlayerTimeSeries'] = tseries_list

players_mask = []
for entry in df.PlayerTimeSeries.values:
    if len(entry):
        players_mask.append(1)
    else:
        players_mask.append(0)
df['PlayersMask'] = players_mask

mask = (df['Date'].dt.year > 2011) & (df['Date'].dt.year < 2021)
df_masked = df[mask]
years = df_masked.Date.dt.year.astype(int)
trimes = (df_masked.Date.dt.month.astype(int) - 1) // 3
grouped = df_masked.groupby([years, trimes])

cols = ['AppID', 'Date', 'Tags', 'PlayerTimeSeries']
test = (df_masked[['Date', 'Tags', 'PlayerTimeSeries']])
first_sem_12 = test[test.Date.dt.year == 2012][test.Date.dt.month < 7].index.values
df_list = []
for i in range(len(test.PlayerTimeSeries.values)):
    if len(test.PlayerTimeSeries.values[i]) > 0:
        if test.index[i] not in first_sem_12:
            temp = [test.index[i]]
            for entry in test.iloc[i].values:
                temp.append(entry)
            df_list.append(temp)
cleaned_df = pd.DataFrame(df_list, columns=cols)
```

A.1.7 Código para obtenção das figuras 4 e 5

```
# Dados para figura 1a

ender = glob.glob(myDir + '*.csv')
means = []
for file in ender:
```

```
    app_id = int((file.split('/')[-1]).split('.')[0])
    i = 0
    with open(file) as f:
        lines = f.readlines()
        for line in lines:
            mean = float(line.split(',')[0])
            try:
                means[i] += mean
            except IndexError:
                if i < 103:
                    means.append(mean)
                else:
                    continue
            i += 1
means = means[::-1]

#Mascara para figura 1b

mask = (df['Date'].dt.year > 2011) & (df['Date'].dt.year < 2021)
df_masked = df[mask]
fig1b_dado1_y =
    (df_masked['Date'].groupby(df_masked.Date.dt.to_period("M")).agg('count')).values
fig1b_dado1_x = [i for i in range(len(fig1b_dado1_y))]

XLIM=109
fig1 = pl.figure( figsize =stdfigsize(nx=1,ny=1))
widths=[1]
heights=[0.8]
spec1 = fig1.add_gridspec(1, 1, hspace=0.4, wspace=0.3, width_ratios=widths,
                          height_ratios=heights)

ax = fig1.add_subplot(spec1[0, 0])

ax.plot([i for i in range(len(means))], means, '-o', \
        markeredgewidth= 1, markeredgecolor= 'black', markersize=10, color= '#88419d')
ax.set_ylabel('Numero de jogadores, N', fontsize= 25)
ax.set_yticks([i for i in range(0,4100000,1000000)])

fig1a_eixo_y= ['0']
for i in range(1,5,1):
    fig1a_eixo_y.append('{:,.0f}x10^6'.format(i))

ax.set_yticklabels(fig1a_eixo_y, fontsize= 22)
```

```
ax.set_xlabel('Mes, m', fontsize= 27)
ax.set_xticks([i for i in range(0,XLIM,12)])
ax.set_xticklabels([str(i) for i in range(0,XLIM,12)], fontsize= 22)
ax.set_xlim(right=XLIM)
ax.set_ylim(top=5000000)

m1,b1 = polyfit([i for i in range(len(means))], means, 1)
x= np.linspace(0, XLIM-3, 100)
ax.plot(x, m1*x+b1, '-', color= '#d73027', linewidth=3)

ax.plot ([83,83], [1000000,2200000], color= '#000000', linewidth= 2)
ax.plot ([38,83], [1000000,1000000], color= '#000000', linewidth= 2)
ax.annotate(r'$inclinacao$' + r': ' + r'$36690$', (40,750000), fontsize=20)

fig1 = pl.figure( figsize =stdfigsize(nx=1,ny=1))
widths=[12, 1]
heights=[0.8]
spec1 = fig1.add_gridspec(1, 2, hspace=0.4, wspace=0.3, width_ratios=widths,
                          height_ratios=heights)

ax = fig1.add_subplot(spec1[0, 0])
ax.semilogy(fig1b_dado1_x, fig1b_dado1_y, '-o', markeredgewidth= 1, markeredgecolor= 'black',
    markersize=10, color='#5aae61')
ax.set_ylabel('Numero de jogos lancados, N', fontsize= 23)
ax.set_yscale('log')
ax.set_yticks([100, 1000])
ax.set_yticklabels(['100', '1000'], fontsize= 22)
ax.yaxis.set_major_formatter(LogFormatterSciNotation())
ax.set_xlabel('Mes, m', fontsize= 27)
ax.set_xticks([i for i in range(0,XLIM,12)])
ax.set_xticklabels([str(i) for i in range(0,XLIM,12)], fontsize= 22)
ax.set_xlim(right=XLIM)

m2,b2 = polyfit(fig1b_dado1_x, np.log10(fig1b_dado1_y), 1)
x= np.linspace(0, 105, 100)
ax.semilogy(x, 10**(b2+(m2*x)), '-', color= '#d73027', linewidth=3)
ax.plot ([85,85], [100,300], color= '#000000', linewidth= 2)
ax.plot ([38,85], [100,100], color= '#000000', linewidth= 2)
ax.annotate(r'$inclinacao$: $0.011$', (38,81), fontsize=20)

pl.gcf().subplots_adjust(bottom=0.18)
```

A.1.8 Código para obtenção das figuras 6 e 7

```
mask = (df['Date'].dt.year > 2011) & (df['Date'].dt.year < 2021)
df_masked = df[mask]
fig2_quarter_tseries = []
years = df_masked.Date.dt.year.astype(int)
semes = (df_masked.Date.dt.month.astype(int) - 1) // 3
grouped = df_masked.groupby([years, semes])
for y in range(2012, 2021):
    for q in range(0, 4):
        if y == 2012 and (q == 0 or q == 1):
            continue
        else:
            quarter_tseries = []
            for entry in (grouped.get_group((y, q)))['PlayerTimeSeries'].values:
                if entry:
                    for i in range(len(entry)):
                        try:
                            quarter_tseries[i].append(entry[i])
                        except IndexError:
                            quarter_tseries.append([entry[i]])
            if quarter_tseries:
                fig2_quarter_tseries.append(quarter_tseries)

linear_means = []
linear_std = []
log_means = []
log_std = []
for q in fig2_quarter_tseries:
    q_m = []
    q_s = []
    qlog_m = []
    qlog_s = []
    for m in q:
        m = np.asarray(m)
        q_m.append(np.mean(m))
        q_s.append(np.std(m))
        logm = (np.log10(m + 0.01))
        qlog_m.append(np.mean(logm))
        qlog_s.append(np.std(logm))
    linear_means.append(q_m)
    linear_std.append(q_s)
    log_means.append(qlog_m)
```

```
    log_std.append(qlog_s)

#Figura 3
f,axs = plt.subplots(nrows=2,ncols=1,figsize=stdfigsize(nx=1,ny=2))
XLIM = 108
XSTEP = 12

for i in range(len(linear_means)):
    fig2_x_NOVA= []
    for j in range(len(linear_means[i])):
        fig2_x_NOVA.append(j+1)
    axs[0].set_xlim(0,105)
    axs[0].plot(fig2_x_NOVA, linear_means[i], linewidth=2.4)
axs[0].set_ylabel(r'Media, $\mu$', fontsize= 25)
axs[0].set_yticks([i for i in range(0,60000,5000)])
axs[0].set_title(r'$\mathbf{A}$', loc='left')
axs[0].set_yticklabels(['{:,.0f}'.format(i) for i in range(0,60000,5000)], fontsize = 20)
axs[0].set_xlabel('Mes, m', fontsize = 27)
axs[0].set_xticks([i for i in range(0,XLIM,XSTEP)])
axs[0].set_xticklabels([str(i) for i in range(0,XLIM,XSTEP)], fontsize= 22)

for i in range(len(log_means)):
    fig2_x_NOVA= []
    for j in range(len(log_means[i])):
        axs[1].set_xlim(0,105)
        fig2_x_NOVA.append(j+1)
    axs[1].plot(fig2_x_NOVA, log_means[i], linewidth=2.4)
axs[1].set_yticks([i for i in np.arange(-0.8,3.2,0.4)])
axs[1].set_yticklabels(['{:1.1f}'.format(i) for i in np.arange(-0.8,3.2,0.4)], fontsize = 20)
axs[1].set_ylabel(r'Media (log-transformada), $\mu_{\log}$', fontsize= 25)
axs[1].set_xlabel('Mes, m', fontsize = 27)
axs[1].set_title(r'$\mathbf{B}$', loc='left')
axs[1].set_xticks([i for i in range(0,XLIM,XSTEP)])
axs[1].set_xticklabels([str(i) for i in range(0,XLIM,XSTEP)], fontsize= 22)
axs[1].set_ylim(-1.1,3.0)

f.tight_layout(pad=1, h_pad=1)

#Figura 4
f,axs = plt.subplots(nrows=2,ncols=1,figsize=stdfigsize(nx=1,ny=2))
XLIM = 108
XSTEP = 12
```

```
linear_std_array = []
for month in linear_std:
    linear_std_array.append(np.asarray(month))
for i in range(len(linear_std_array)):
    fig2_x_NOVA= []
    for j in range(len((linear_std_array[i])[linear_std_array[i] > 0])):
        axs[0].set_xlim(0,105)
        fig2_x_NOVA.append(j+1)
    axs[0].plot(fig2_x_NOVA, (linear_std_array[i])[linear_std_array[i] > 0], linewidth=2.4)
axs[0].set_ylabel(r'Desvio Padro , $\sigma$', fontsize = 25)
axs[0].set_yticks([i for i in range(0,200000,20000)])
axs[0].set_title(r'$\mathbf{A}$', loc='left')
axs[0].set_yticklabels(['{:,.0f}'.format(i) for i in range(0,200000,20000)], fontsize = 20)
axs[0].set_xlabel('Mes, m', fontsize = 27)
axs[0].set_xticks([i for i in range(0,XLIM,XSTEP)])
axs[0].set_xticklabels([str(i) for i in range(0,XLIM,XSTEP)], fontsize= 22)

log_std_array = []
for month in log_std:
    log_std_array.append(np.asarray(month))
for i in range(len(log_std)):
    fig2_x_NOVA= []
    for j in range(len((log_std_array[i])[log_std_array[i] > 0])):
        axs[1].set_xlim(0,105)
        fig2_x_NOVA.append(j+1)
    axs[1].plot(fig2_x_NOVA, (log_std_array[i])[log_std_array[i] > 0], linewidth=2.4)
axs[1].set_yticks([i for i in np.arange(0,2.5,0.3) ])
axs[1].set_yticklabels(['{:1.1f}'.format(i) for i in np.arange(0,2.5,0.3) ], fontsize = 20)
axs[1].set_ylabel(r'Desvio Padro (log), $\sigma_{\log}$', fontsize= 25)
axs[1].set_xlabel('Mes, m', fontsize = 27)
axs[1].set_title(r'$\mathbf{B}$', loc='left')
axs[1].set_xticks([i for i in range(0,XLIM,XSTEP)])
axs[1].set_xticklabels([str(i) for i in range(0,XLIM,XSTEP)], fontsize= 22)
axs[1].set_ylim(-0.1,2.5)


f.tight_layout(pad=1, h_pad=1)
```

A.1.9 Código para obtenção das figuras 8 e 9

```
#formatting for figure 5
```

```
nonzeroes = []
for entry in (grouped.get_group((2015, 1)))['PlayerTimeSeries'].values:
    if entry:
        nonzeroes.append(entry)

wrt_str = '{'
fstr = ''
i = 0
for entry in nonzeroes:
    fstr = str(entry)
    fstr = fstr.replace('[', '{')
    fstr = fstr.replace(']', '}')
    fstr = fstr.replace(' ', '')
    if i == len(nonzeroes) - 1:
        wrt_str += fstr
    else:
        wrt_str += fstr + ','
    i += 1
wrt_str += '}'
with open('x13_4pdf.txt', 'w') as f:
    f.write(wrt_str)

trans_nonzeroes = [[] for i in range(207)]

for i in range(len(nonzeroes)):
    j = 0
    while len(nonzeroes[i][j:]) > 0:
        trans_nonzeroes[j].append(nonzeroes[i][j])
        j += 1

tri_nonzeroes = [[] for i in range(207//3)]

k = 0
for i in range(0, len(trans_nonzeroes), 3):
    for j in range(0, 3):
        for entry in trans_nonzeroes[i+j]:
            if entry > 0:
                tri_nonzeroes[k].append(entry)
    k += 1

lista_boxcox_lambda = []
lista_boxcox_ci = []
```

```
for entry in tri_nonzeroes[:24]:
    _, lambda_param, ci = scipy.stats.boxcox(entry, alpha=0.05)
    lista_boxcox_lambda.append(lambda_param)
    lista_boxcox_ci.append(abs(ci[1] - ci[0]))
lista_boxcox_lambda = np.asarray(lista_boxcox_lambda)

flat_list_tri = [item for sublist in fig2_quarter_tseries[5] for item in sublist]
flat_list_all = [item for sublist in fig2_quarter_tseries for subsublist in sublist for item in
    subsublist]
flat_list_all_grouped_tri = [item for sublist in fig2_quarter_tseries for item in sublist]

fig3 = []
for num in flat_list_tri:
    if num > 0:
        fig3.append(num)
fig3 = np.array(fig3)
fig3a_NOVA = (np.log10(fig3) - np.log10(fig3).mean())/np.log10(fig3).std()
fig3b_NOVA, maxlog = scipy.stats.boxcox(fig3)
fig3b_NOVA = (fig3b_NOVA - np.mean(fig3b_NOVA))/np.std(fig3b_NOVA

#Figura 5
gs = gridspec.GridSpec(1, 2, hspace=0.4)
pl.figure( figsize =stdfigsize(nx=2,ny=1))

ax = pl.subplot(gs[0, 0])
ax.set_title(r'$\mathbf{A}$', loc='left')
pl.hist(fig3a_NOVA, bins=14, density=True, histtype= 'bar', rwidth=1.0, color= '#1a1a1a',
    edgecolor='#000000', alpha= 0.75)
x = np.linspace(-3,3,1000)
y = scipy.stats.norm.pdf(x, loc=0,scale=1)
pl.plot(x,y, '--', color= '#d73027',linewidth=6)
pl.ylabel('Distribuicao de probabilidade', fontsize= 25)
pl.yticks([i for i in np.arange(0,0.45,0.1)])
pl.xlabel(r' $z^{(0)}$', fontsize= 27)
pl.xticks([i for i in range(-3,4,1)])
pl.xlim(-3,3)

ax = pl.subplot(gs[0, 1])
ax.set_title(r'$\mathbf{B}$', loc='left')
pl.hist(fig3b_NOVA, bins=14, density=True, histtype= 'bar', rwidth=1.0, color= '#1a1a1a',
    edgecolor='#000000', alpha= 0.75)
plt.xlim(-3,3)
pl.plot(x, y, '--', color= '#d73027', linewidth=6)
```

```
pl.yticks([i for i in np.arange(0,0.45,0.1)])
pl.xlabel(r' $z^{(\lambda)}$', fontsize= 25)
pl.xticks([i for i in range(-3,4,1)])
pl.xlim(-3,3)

#Figura 6
gs = gridspec.GridSpec(1, 2, hspace=0.4)
pl.figure( figsize =stdfigsize(nx=2,ny=1))

ax = pl.subplot(gs[0, :])
lambda_xrange = [i for i in range(len(lista_boxcox_lambda))]
lambda_xrange2 = [i for i in range(-2, len(lista_boxcox_lambda) + 2)]
mean_list = [np.mean(lista_boxcox_lambda) for i in range(len(lista_boxcox_lambda) + 4)]
pl.plot(lambda_xrange[5], lista_boxcox_lambda[5], color='#d73027', zorder=200)
pl.errorbar(lambda_xrange, lista_boxcox_lambda, yerr=lista_boxcox_ci, fmt='o', color="#1a1a1a")
pl.plot(lambda_xrange2, mean_list, color='#d73027', linewidth=3)
pl.ylabel(r'$\lambda_{opt}$', fontsize=30)
pl.xlabel(r'Trimestres', fontsize=25)
pl.xticks(np.arange(len(lista_boxcox_lambda)), np.arange(0, len(lista_boxcox_lambda)))
pl.ylim(-0.31,0.16)
pl.xlim(-0.25, 23.25)
```

A.1.10 Código para obtenção das figuras 10 e 11

```
# Similar a figura 3c: a lista original eh #jogos e sublista eh #meses de cada jogo.
# Precisa transpor primeiro para ficar na forma #meses de jogos e sublista ser #jogos (para alinhar
    as series temporais)
# Depois agrupa de 3 em 3 meses (poderia ser outro) e faz um loop para obter os lambdas de cada
    trimestre
# Notar que a serie temporal eh de tras pra frente, entao [:34] sao os 34 trimestres de 06/2012 ate
    12/2020.
trans_flat_grouped_tri = [[] for i in range(1155)]

for i in range(len(flat_list_all_grouped_tri)):
    j = 0
    while len(flat_list_all_grouped_tri[i][j:]) > 0:
        trans_flat_grouped_tri[j].append(flat_list_all_grouped_tri[i][j])
        j += 1

tri_trans_flat = [[] for i in range(1155//3)]
```

```
k = 0
for i in range(0, len(trans_flat_grouped_tri), 3):
    for j in range(0, 3):
        for entry in trans_flat_grouped_tri[i+j]:
            if entry > 0:
                tri_trans_flat[k].append(entry)
    k += 1

lista_boxcox_lambda_all = []
lista_boxcox_all_ci = []
for entry in tri_trans_flat [:34]:
    _, lambda_param_all, ci = scipy.stats.boxcox(entry, alpha=0.05)
    lista_boxcox_lambda_all.append(lambda_param_all)
    lista_boxcox_all_ci.append(abs(ci[1] - ci[0]))
lista_boxcox_lambda_all = np.asarray(lista_boxcox_lambda_all)
####

#### Figura 4b e 4c
fig4 = []
for num in flat_list_all:
    if num > 0:
        fig4.append(num)
fig4 = np.array(fig4)
fig4b_NOVA = (np.log10(fig4) - (np.log10(fig4)).mean())/(np.log10(fig4)).std()
fig4c_OLD_NOVA, maxlog, maxci = scipy.stats.boxcox(fig4, alpha=0.05)
ci_interval = abs(maxci[1] - maxci[0])
fig4c_NOVA = (fig4c_OLD_NOVA - np.mean(fig4c_OLD_NOVA))/np.std(fig4c_OLD_NOVA)
####

#Figura 7

gs = gridspec.GridSpec(2, 1, hspace=0.4)
pl.figure( figsize =stdfigsize(nx=2,ny=2))
pl.rcParams.update({'font.size': 22})

ax = pl.subplot(gs[0, :])

lambda_xrange = [i for i in range(len(lista_boxcox_lambda_all))]
lambda_xrange2 = [i for i in range(-2, len(lista_boxcox_lambda_all) + 2)]
mean_list = [np.mean(lista_boxcox_lambda_all) for i in range(len(lista_boxcox_lambda_all) + 4)]
pl.errorbar(lambda_xrange, lista_boxcox_lambda_all, yerr=lista_boxcox_all_ci, fmt='o',
    color="#1a1a1a")
```

```
pl.plot(lambda_xrange2, mean_list, color='#d73027', linewidth=3)
pl.ylabel(r'$\lambda_{opt}$', fontsize= 30)
pl.xlabel(r'Trimensais', fontsize=25)
pl.ylim(-0.16,0.07)
pl.xlim(-0.25, 33.25)

#Figura 8

gs = gridspec.GridSpec(1, 2, hspace=0.4)
pl.figure( figsize =stdfigsize(nx=2,ny=1))
pl.rcParams.update({'font.size': 22})

ax = pl.subplot(gs[0, 0])
ax.set_title(r'$\mathbf{A}$', loc='left')
pl.hist(fig4b_NOVA, bins=20, density=True, histtype= 'bar', rwidth=1.0, color= '#1a1a1a',
    edgecolor='#000000', alpha= 0.75)
x = np.linspace(-3.2,3.2,1000)
y = scipy.stats.norm.pdf(x, loc=0,scale=1)
pl.xlim(-2.2,2.2)
pl.xlabel(r' $z^{(0)}$', fontsize= 27)
pl.ylabel('Distribuicao de probabilidade', fontsize= 25)
pl.plot(x,y, '--', color= '#d73027',linewidth=6)

ax = pl.subplot(gs[0, 1])
ax.set_title(r'$\mathbf{B}$', loc='left')
pl.hist(fig4c_NOVA, bins=20, density=True, histtype= 'bar', rwidth=1.0, color= '#1a1a1a',
    edgecolor='#000000', alpha= 0.75)
x = np.linspace(-3.2,3.2,1000)
y = scipy.stats.norm.pdf(x, loc=0,scale=1)
pl.xlim(-2.25,2.25)
pl.xlabel(r' $z^{(\lambda)}$', fontsize= 27)
pl.plot(x,y, '--', color= '#d73027',linewidth=6)
```

A.1.11 Código para o caso *Indie*

```
df_players = df[df['PlayersMask'] == True]
#Create mask for game type
indie_app_ids = []
i = 0
for entry in df_players.Tags.values:
    if "Indie" in entry:
```

```
            indie_app_ids.append(df_players.index.values[i])
        else:
            indie_app_ids.append(0)
        i += 1
indie_df = (df_players[df_players.index == indie_app_ids])
indie_array = indie_df.index.values

#Aggregate means for figure 5a
means_indie = []
for appid in indie_array:
    try:
        with open(myDir + str(appid) + '.csv') as f:
            app_id = int((file.split('/')[-1]).split('.')[0])
            i = 0
            lines = f.readlines()
            for line in lines:
                mean = float(line.split(',')[0])
                try:
                    means_indie[i] += mean
                except IndexError:
                    if i < 110:
                        means_indie.append(mean)
                    else:
                        continue
                i += 1
    except FileNotFoundError:
        continue
means_indie = means_indie[::-1]

#Data for figure 5b

indie_mask = (indie_df['Date'].dt.year > 2011) & (indie_df['Date'].dt.year < 2021)
indie_df_masked = indie_df[indie_mask]
fig5b_dado1_y =
    (indie_df_masked['Date'].groupby(indie_df_masked.Date.dt.to_period("M")).agg('count')).values
fig5b_dado1_x = [i for i in range(len(fig5b_dado1_y))]

tot_players_big_tags = 0.0
tag_game_num = []
# figure 6a
flat_indie_list = []
for game in indie_df_masked.PlayerTimeSeries.values:
    if game:
```

```
        for players in game:
            if players > 0:
                flat_indie_list.append(players)

flat_indie_list = np.array(flat_indie_list)
flat_indie_boxcox, lambda_indie, ci_interval = scipy.stats.boxcox(flat_indie_list, alpha=0.05)
flat_indie_boxcox = (flat_indie_boxcox - np.mean(flat_indie_boxcox))/(np.std(flat_indie_boxcox))
#

# figure 6b
indie_players_list = []
max_len_indie = 0
for game in indie_df_masked.PlayerTimeSeries.values:
    if game:
        indie_players_list.append(game)
        if len(game) > max_len_indie:
            max_len_indie = len(game)

trans_indie_nonzeroes = [[] for i in range(max_len_indie)]

for i in range(len(indie_players_list)):
    j = 0
    while len(indie_players_list[i][j:]) > 0:
        trans_indie_nonzeroes[j].append(indie_players_list[i][j])
        j += 1


tri_indie_nonzeroes = [[] for i in range(max_len_indie//3)]

k = 0
for i in range(0, len(trans_indie_nonzeroes), 3):
    for j in range(0, 3):
        for entry in trans_indie_nonzeroes[i+j]:
            if entry > 0:
                tri_indie_nonzeroes[k].append(entry)
    k += 1

lista_indie_boxcox_lambda = []
lista_indie_boxcox_ci = []
for entry in tri_indie_nonzeroes[:34]:
    _, lambda_indie_param, ci = scipy.stats.boxcox(entry, alpha=0.05)
    lista_indie_boxcox_lambda.append(lambda_indie_param)
    lista_indie_boxcox_ci.append(abs(ci[1] - ci[0]))
```

```
lista_indie_boxcox_lambda = np.asarray(lista_indie_boxcox_lambda)
#


# figure 6d
tag_dic = {}
for game in df_players.Tags.values:
    if game:
        for tag in game:
            try:
                tag_dic[tag] += 1
            except KeyError:
                tag_dic[tag] = 1

sorted_tag_dic = {k: v for k, v in reversed(sorted(tag_dic.items(), key=lambda item: item[1]))}
big_tags = list(sorted_tag_dic.keys())[:12]

rank_dist_tags = []
i = 0
sorted_tag_games = (list(sorted_tag_dic.values()))
for game_num in sorted_tag_games:
    if i == 0:
        rank_dist_tags.append( (game_num, i+1) )
        i += 1
    else:
        if game_num == rank_dist_tags[-1][0]:
            rank_dist_tags.append( (game_num, i+1) )
        else:
            rank_dist_tags.append( (game_num, i+1) )
            i += 1
#

# figure 6c
lambdas_tags = []
for tag in big_tags:
    tag_app_ids = []
    flat_tag_list = []
    i = 0
    for entry in df_players.Tags.values:
        if tag in entry:
            tag_app_ids.append(df_players.index.values[i])
        else:
            tag_app_ids.append(0)
```

```
            i += 1
        tag_df = (df_players[df_players.index == tag_app_ids])
        for game in tag_df.PlayerTimeSeries.values:
            if game:
                for players in game:
                    if players > 0:
                        flat_tag_list.append(players)
        _, lambda_tag = scipy.stats.boxcox(flat_tag_list)
        tag_game_num.append(len(tag_df))
        lambdas_tags.append(lambda_tag)
        tot_players_big_tags += sum(np.asarray(flat_tag_list))
#

XLIM = 109
XSTEP = 12

fig5 = pl.figure( figsize =stdfigsize(nx=1,ny=1))
widths=[1]
heights=[0.8]
spec5 = fig5.add_gridspec(1, 1, hspace=0.8, wspace=0.4, width_ratios=widths,
                          height_ratios=heights)

ax = fig5.add_subplot(spec5[0, 0])
xrange1 = [i for i in range(len(means_indie[:-8]))]
ax.plot(xrange1, means_indie[:-8], '-o', markeredgewidth= 1, markeredgecolor= 'black',
    markersize=10, color= '#88419d')
ax.set_ylabel('Numero de jogadores, N', fontsize= 25)

fig1a_eixo_y= ['0']
for i in range(1,8,1):
    fig1a_eixo_y.append('{:,.0f}x10^6'.format(i))
y_axis0 = [0] + fig1a_eixo_y
ax.set_xlabel('Mes, m', fontsize= 27)
ax.set_xticks([i for i in range(0,XLIM,XSTEP)])
ax.set_xticklabels([str(i) for i in range(0,XLIM,XSTEP)], fontsize= 22)

m1,b1 = polyfit(xrange1, means_indie[:-8], 1)
x= np.linspace(0, XLIM-6, XLIM-6)
ax.plot(x, m1*x+b1, '-', color= '#d73027', linewidth=3)

ax.plot ([9,9], [380000,520000], color= '#000000', linewidth= 2)
ax.plot ([9,45], [520000,520000], color= '#000000', linewidth= 2)
ax.annotate(r'$inclinacao$: $4494$', (9,530000), fontsize =20)
```

```
pl.gcf().subplots_adjust(bottom=0.18)

fig5 = pl.figure( figsize =stdfigsize(nx=1,ny=1))
widths=[1]
heights=[0.8]
spec5 = fig5.add_gridspec(1, 1, hspace=0.8, wspace=0.4, width_ratios=widths,
                          height_ratios=heights)

ax = fig5.add_subplot(spec5[0, 0])
xrange2 = [i for i in range(len(fig5b_dado1_y))]
ax.semilogy(xrange2[:-8], fig5b_dado1_y[:-8], '-o', markeredgewidth= 1, markeredgecolor= 'black',
    markersize=10, color='#5aae61')
ax.set_ylabel('Numero de jogos lancados, N', fontsize= 23)
ax.set_yscale('log')
ax.set_yticks([100, 1000])
ax.set_yticklabels(['100', '1000'], fontsize= 22)
ax.yaxis.set_major_formatter(LogFormatterSciNotation())
ax.set_xlabel('Mes, m', fontsize= 27)
ax.set_xticks([i for i in range(0,XLIM,XSTEP)])
ax.set_xticklabels([str(i) for i in range(0,XLIM,XSTEP)], fontsize= 22)
ax.set_xlim(right=110)

m2,b2 = polyfit(xrange2[:-8], np.log10(fig5b_dado1_y[:-8]), 1)
x= np.linspace(0, XLIM-8, XLIM-10)
ax.semilogy(x, 10**(b2+(m2*x)), '-', color= '#d73027', linewidth=3)
ax.plot ([79,79], [12,35], color= '#000000', linewidth= 2)
ax.plot ([40,79], [12,12], color= '#000000', linewidth= 2)
ax.annotate(r'$inclinacao$: $%.3f $' % m2, (40,9), fontsize=20)

pl.gcf().subplots_adjust(bottom=0.18)

gs = gridspec.GridSpec(1, 1, hspace=0.4, wspace=0.3)
pl.figure( figsize =stdfigsize(nx=1,ny=1))

ax = pl.subplot(gs[0, 0])
# figure 6a
pl.hist(flat_indie_boxcox, bins=20, density=True, histtype= 'bar', rwidth=1.0, color= '#1a1a1a',
    edgecolor='#000000', alpha= 0.75)
x = np.linspace(-3.2,3.2,1000)
y = scipy.stats.norm.pdf(x, loc=0,scale=1)
pl.xlim(-2.25,2.25)
pl.xlabel(r' $z^{(\lambda)}$', fontsize= 27)
```

```
pl.ylabel('Distribuicao de probabilidade', fontsize= 25)
pl.plot(x,y, '--', color= '#d73027',linewidth=6)

gs = gridspec.GridSpec(1, 1, hspace=0.4, wspace=0.3)
pl.figure( figsize =stdfigsize(nx=2,ny=1))

ax = pl.subplot(gs[0, 0])
# figure 6b
lambda_xrange = [i for i in range(len(lista_indie_boxcox_lambda))]
lambda_xrange2 = [i for i in range(-2, len(lista_indie_boxcox_lambda) + 2)]
mean_list = [np.mean(lista_indie_boxcox_lambda) for i in range(len(lista_indie_boxcox_lambda)
    + 4)]
pl.errorbar(lambda_xrange, lista_indie_boxcox_lambda, yerr=lista_indie_boxcox_ci, fmt='o',
    color="#1a1a1a")
pl.plot(lambda_xrange2, mean_list, color='#d73027', linewidth=3)
pl.ylabel(r'$ \lambda_{opt}$', fontsize= 30)
pl.xlabel('Trimensais')
pl.plot(lambda_xrange[5], lista_indie_boxcox_lambda[5], 'o', color='#d73027', zorder=5)
pl.ylim(-0.21,0.11)
pl.xlim(-1, 33.8)

#lambda categorias
gs = gridspec.GridSpec(1, 1, hspace=0.4, wspace=0.3)
pl.figure( figsize =stdfigsize(nx=2,ny=1))

ax = pl.subplot(gs[0, 0])
# figure 6b
freq_series = pd.Series(abs(np.asarray(lambdas_tags)))
xrang = [i-1 for i in range(len(freq_series.values)+2)]
xax = freq_series.plot(kind='bar', color='#1a1a1a')
pl.plot(xrang, [(np.mean(freq_series.values)) for i in range(len(freq_series.values) + 2)],
    color='#d73027', linewidth=3)
xax.set_xticklabels(big_tags)
rects = xax.patches
pl.ylabel(r'$- \lambda_{opt}$', fontsize= 30)
pl.ylim(0,0.07)
```

A.1.12 Código para o estudo do número de categorias ao longo do tempo

```
#figure 7b
```

```
def plaw_investigation(year, month, dataframe=df_players):
    #year and month to exclude from dataset!
    time_mask = (dataframe['Date'].dt.year > 2011) & (dataframe['Date'].dt.year < year)
    df_players_masked = dataframe[time_mask]
    month_mask = ((df_players_masked['Date'].dt.year == year-1) &
        (df_players_masked['Date'].dt.month == month))
    exclude_ids = df_players_masked[month_mask].index.values
    a = df_players_masked[~df_players_masked.index.isin(exclude_ids)]
    return a

lim = 100
coef_list = []
new_tag_dic = {}
for year in range(2013, 2021):
    for month in range(1, 13):
        a = plaw_investigation(year, month)
        new_df = a.copy()
        for game in new_df.Tags.values:
            if game:
                for tag in game:
                    try:
                        new_tag_dic[tag] += 1
                    except KeyError:
                        new_tag_dic[tag] = 1

        sorted_new_tag_dic = {k: v for k, v in reversed(sorted(new_tag_dic.items(), key=lambda
            item: item[1]))}
        new_big_tags = list(sorted_new_tag_dic.keys())[:12]

        new_rank_dist_tags = []
        i = 0
        sorted_new_tag_games = (list(sorted_new_tag_dic.values()))
        for game_num in sorted_new_tag_games:
            if i == 0:
                new_rank_dist_tags.append( (game_num, i+1) )
                i += 1
            else:
                if game_num == new_rank_dist_tags[-1][0]:
                    new_rank_dist_tags.append( (game_num, i+1) )
                else:
                    new_rank_dist_tags.append( (game_num, i+1) )
                    i += 1
```

```
        y_rank = []
        x_rank = []
        for entry in new_rank_dist_tags:
            y_rank.append(entry[1])
            x_rank.append(entry[0])

        xr = np.asarray(x_rank[:lim])
        yr = np.asarray(y_rank[:lim])
        logy = np.log10(yr)
        logx = np.log10(xr)
        m,b = np.polyfit(logx, logy, 1)
        coef_list.append(m-1)

#figure 7a
a = plaw_investigation(year, month)
new_df = a.copy()
for game in new_df.Tags.values:
    if game:
        for tag in game:
            try:
                new_tag_dic[tag] += 1
            except KeyError:
                new_tag_dic[tag] = 1

sorted_new_tag_dic = {k: v for k, v in reversed(sorted(new_tag_dic.items(), key=lambda item:
    item[1]))}
new_big_tags = list(sorted_new_tag_dic.keys())[:12]

new_rank_dist_tags = []
i = 0
sorted_new_tag_games = (list(sorted_new_tag_dic.values()))
for game_num in sorted_new_tag_games:
    if i == 0:
        new_rank_dist_tags.append( (game_num, i+1) )
        i += 1
    else:
        if game_num == new_rank_dist_tags[-1][0]:
            new_rank_dist_tags.append( (game_num, i+1) )
        else:
            new_rank_dist_tags.append( (game_num, i+1) )
            i += 1

y_rank = []
```

```
x_rank = []
for entry in new_rank_dist_tags:
    y_rank.append(entry[1])
    x_rank.append(entry[0])

xr = np.asarray(x_rank[:lim])
yr = np.asarray(y_rank[:lim])
logy = np.log10(yr)
logx = np.log10(xr)



n_sets = 1000
n_samples = len(xr)
bootstrap_cutoff = 0.025
top_cutoff = 100

def generate_samples(dataset, n):
    return  list (np.random.choice(dataset, size=n, replace=True))

boot_samples = [generate_samples(xr, n_samples) for _ in range(n_sets)]

boot_coef_list = []
for sample in boot_samples:
    sample = np.asarray(sample)
    sample_slice = list(reversed(sorted(sample)))
    sample_slice = sample_slice[:100]
    y_rank = [i+1 for i in range(len(sample_slice))]
    ylen = len(sample[sample > 200])
    yr = np.asarray(y_rank)
    logx_cutoff = np.log10(sample_slice)
    logy_cutoff = np.log10(yr)
    m_cutoff,b_cutoff = np.polyfit(logx_cutoff, logy_cutoff, 1)
    boot_coef_list.append(m_cutoff)

sorted_bcf = sorted(boot_coef_list)
sliced = sorted_bcf[int(n_sets*bootstrap_cutoff):int(n_sets - n_sets*bootstrap_cutoff)]
mi, mx = min(sliced), max(sliced)
boot_err_list = [(abs(mi-mx)) for _ in range(len(sample_slice))]
```

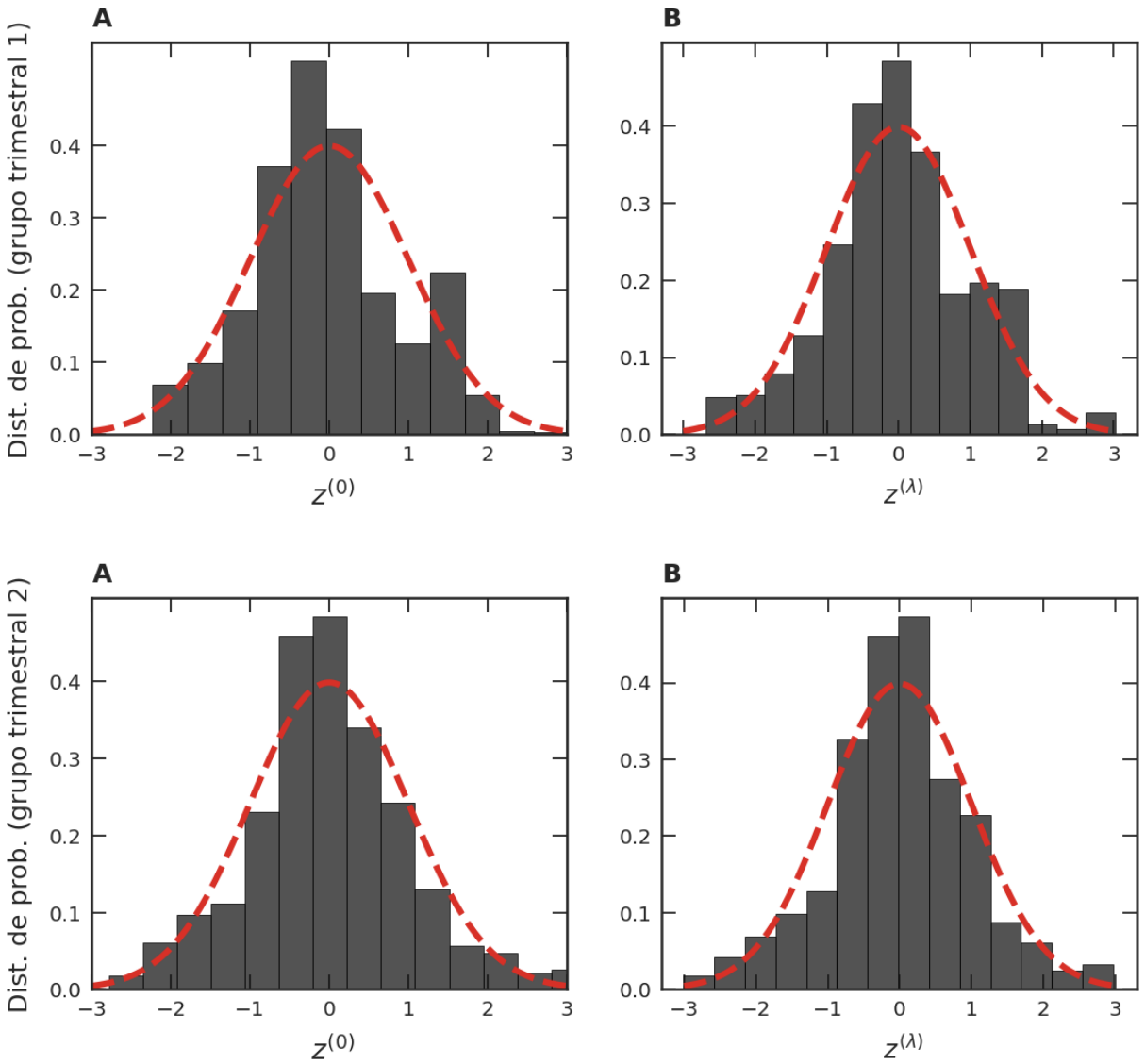
A.1.13 Código para obtenção das figuras 17, 18 e 19

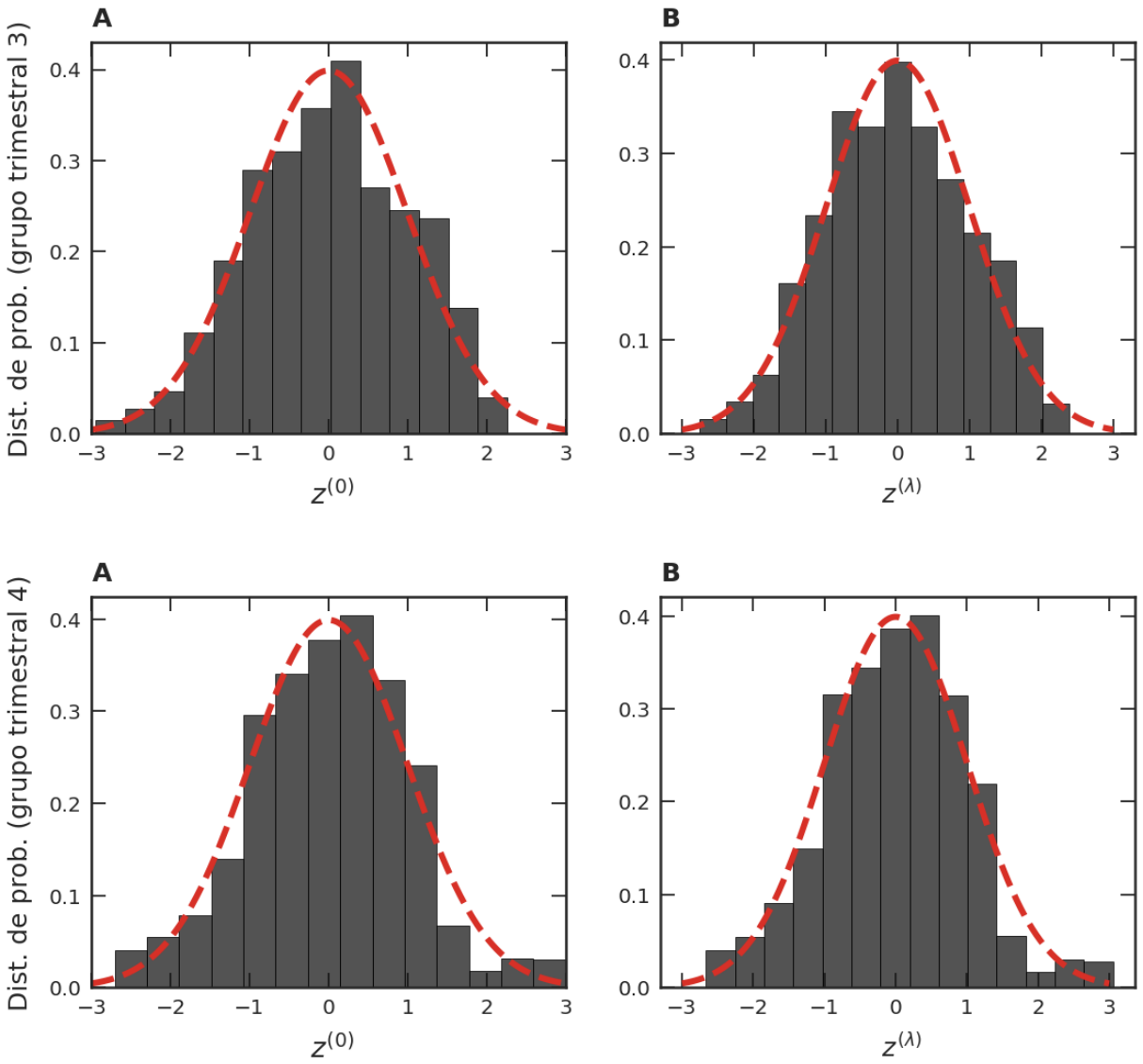
```
gs = gridspec.GridSpec(1, 1, hspace=0.4, wspace=0.3)
pl.figure( figsize =stdfigsize(nx=1,ny=1))

XLIM=109
XSTEP=12

ax = pl.subplot(gs[0, 0])
xrange = [i+1 for i in range(len(tag_evol_2))]
ax.plot(xrange, tag_evol_2, '-o', markeredgewidth= 1, markeredgecolor= 'black', markersize=10,
    color= '#88419d')
ax.set_ylabel(r'Numero de categorias, $N_c$', fontsize= 25)

ax.set_xlabel('Mes, m', fontsize= 27)
ax.set_xticks([i for i in range(0,XLIM,XSTEP)])
ax.set_xticklabels([str(i) for i in range(0,XLIM,XSTEP)], fontsize= 22)

m1,b1 = polyfit(xrange, tag_evol_2, 1)
x= np.linspace(1, XLIM-6, XLIM-7)

m2,b2 = polyfit(xrange[10:], tag_evol_2[10:], 1)
x2= np.linspace(1, XLIM-6, XLIM-7)
ax.plot(x2, m2*x+b2, color='#d73027', linewidth=3)
ax.plot ([85,85], [200,400], color= '#000000', linewidth= 2)
ax.plot ([50,85], [200,200], color= '#000000', linewidth= 2)
ax.annotate(r'$inclinacao$: $%.2f $' % m2, (50,150), fontsize=20)

gs = gridspec.GridSpec(1, 1, hspace=0.4, wspace=0.3)
pl.figure( figsize =stdfigsize(nx=1,ny=1))

XLIM=109
XSTEP=12

ax = pl.subplot(gs[0])
a = np.asarray(list(tag_dic.values()))
a = a[a > 12]
bc_dist, lambda_param, ci_interval = scipy.stats.boxcox(a, alpha=0.05)
bc_old = bc_dist

bc_dist = (bc_dist - np.mean(bc_dist))/np.std(bc_dist)

lambda_list = []
bc_lists = []
err_list_b = []
```

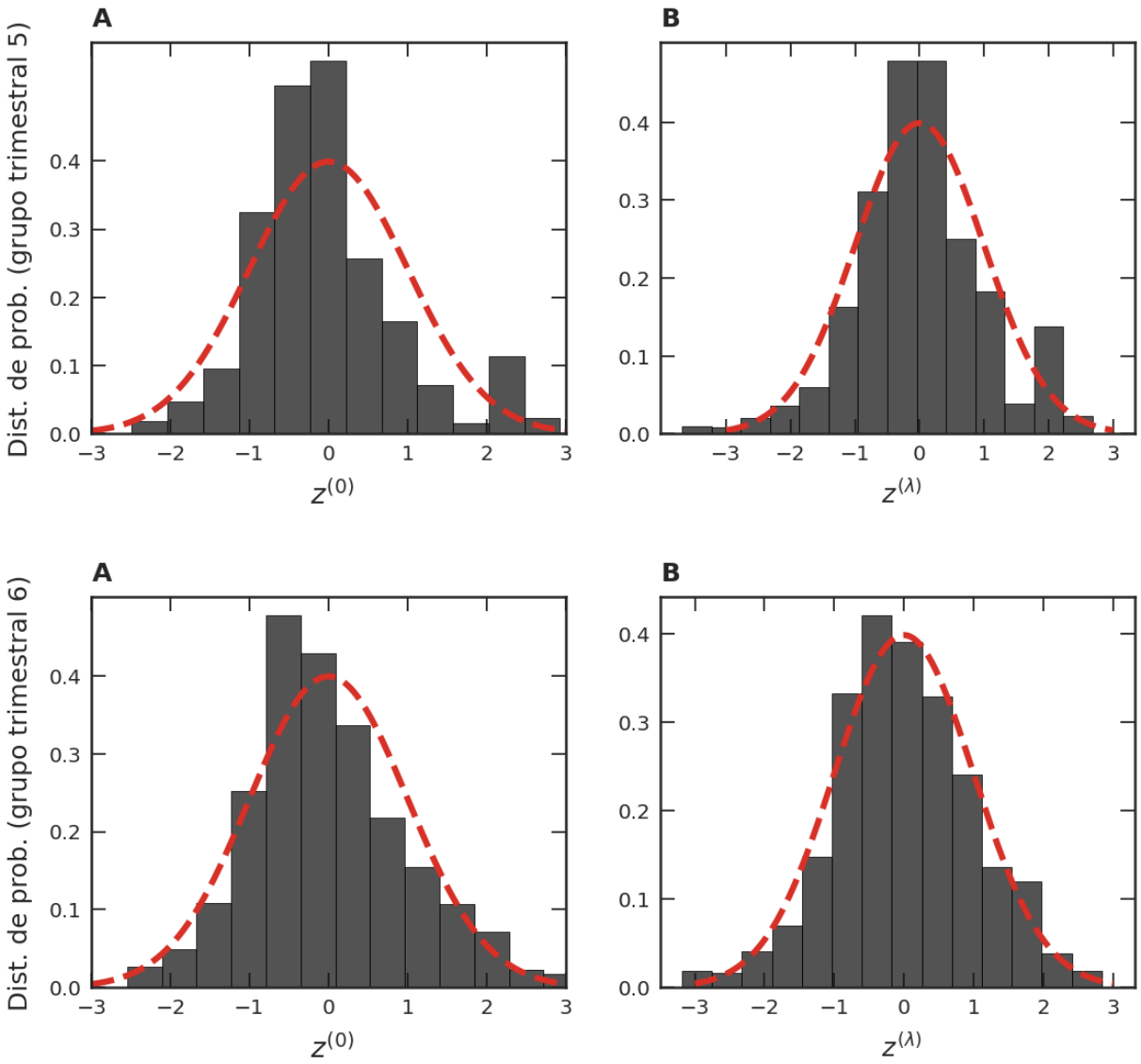
```
for dic in sem_dic_list:
    dic_tri_values = np.asarray(list(dic.values()))
    dic_tri_values = dic_tri_values[dic_tri_values > 20]
    bc_list, lambda_param, ci = scipy.stats.boxcox(dic_tri_values, alpha=0.05)
    bc_list = (bc_list - np.mean(bc_list))/(np.std(bc_list))
    bc_lists.append(bc_list)
    lambda_list.append(lambda_param)
    err_list_b.append(abs(ci[1] - ci[0]) )
lambda_xrange = [i+1 for i in range(len(lambda_list))]

four_years_mean = [np.mean(lambda_list[-4:]) for i in range(len(lambda_list)+3)]

# figure 6a
ax.hist(bc_dist, bins='auto', density=True, histtype= 'bar', rwidth=1.0, color= '#1a1a1a',
    edgecolor='#000000', alpha= 0.75)
x = np.linspace(-3.2,3.2,1000)
y = scipy.stats.norm.pdf(x, loc=0,scale=1)
ax.set_xlim(-2.25,2.25)
ax.set_xlabel(r' $n^{(\lambda)}$', fontsize= 30)
ax.set_ylabel('Distribuicao de probabilidade', fontsize= 25)
ax.plot(x,y, '--', color= '#d73027',linewidth=6)

gs = gridspec.GridSpec(1, 2, hspace=0.4, wspace=0.3)
pl.figure( figsize =stdfigsize(nx=2,ny=1))

XLIM=109
XSTEP=12

ax = pl.subplot(gs[0, :])
# figure 6b
ax.plot(lambda_xrange, lambda_list, 'o', color= '#1a1a1a')
ax.errorbar(lambda_xrange, lambda_list, yerr=err_list_b, fmt='o', color="#1a1a1a")
ax.plot([i-1 for i in range(len(four_years_mean))], four_years_mean, color='#d73027',
    linewidth=3)
ax.set_ylabel(r'$ \lambda_{opt}$', fontsize= 30)
ax.set_xlabel('Semestres')
ax.set_ylim(-1.05,0.26)
ax.set_xlim(0, 17)
#
```

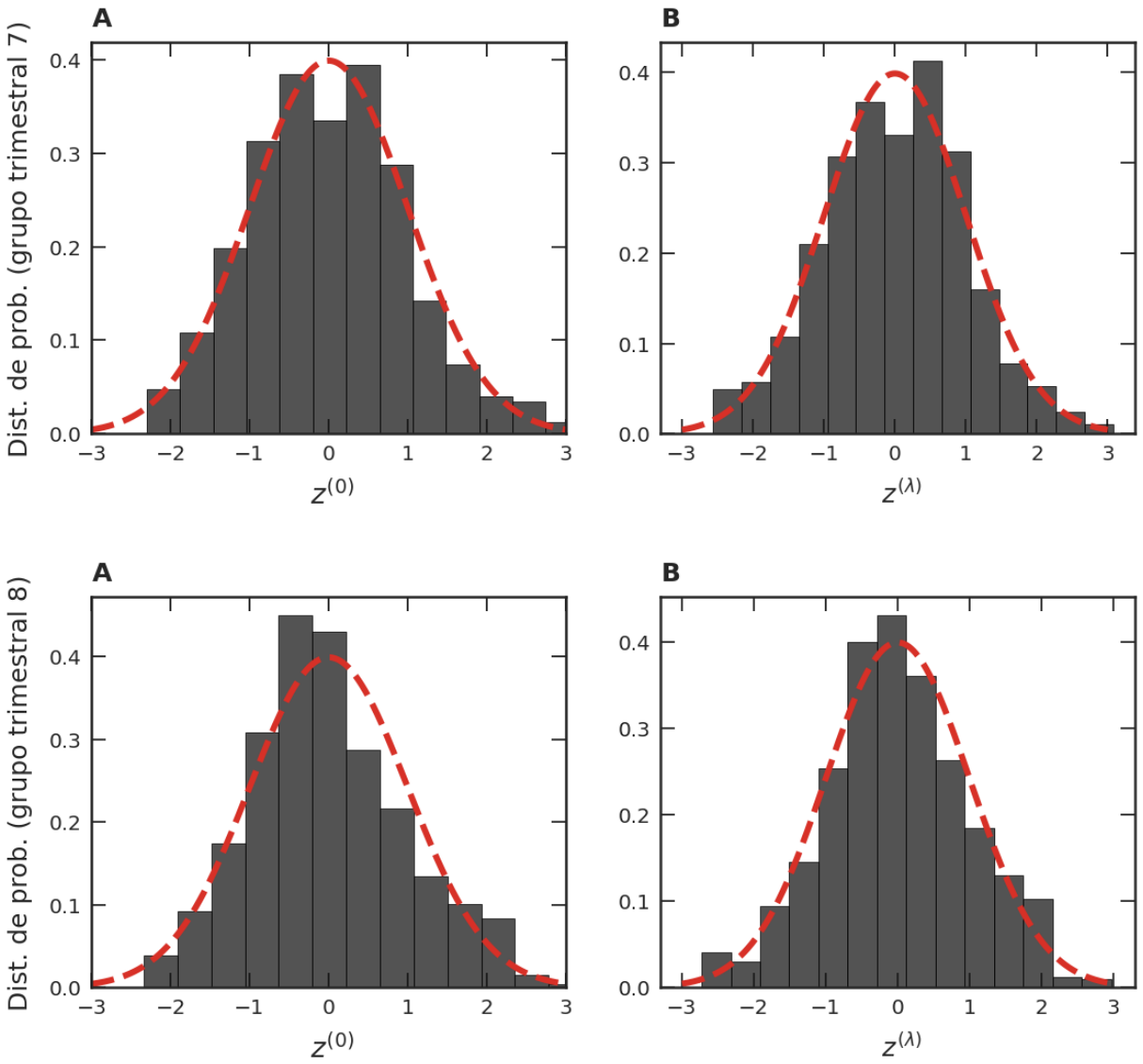
APÊNDICE B – Função de distribuição de probabilidades dos grupos trimestrais

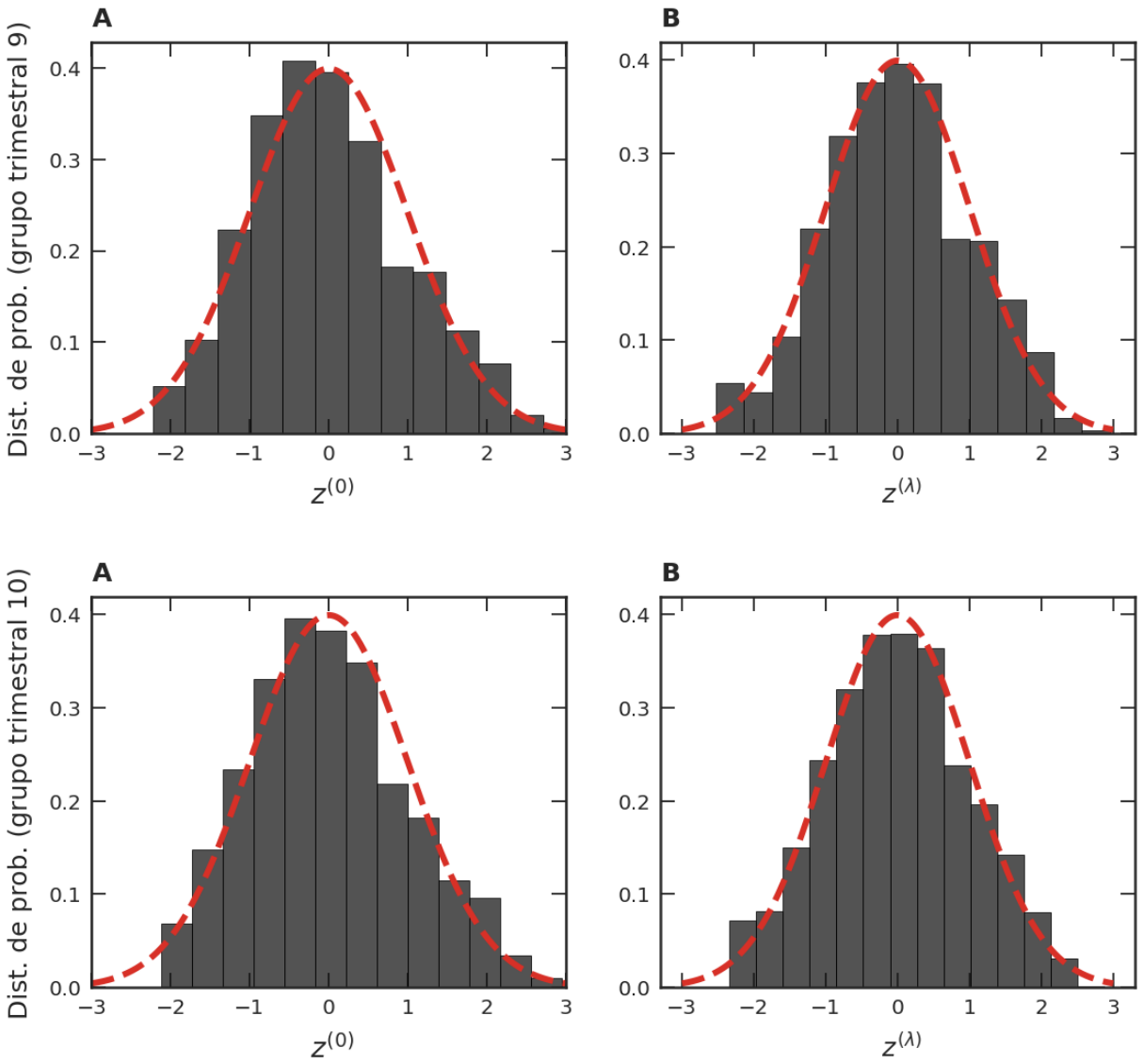
Este apêndice apresenta comparações entre o ajuste lognormal e o ajuste via transformada de Box-Cox para diferentes trimestres, começando no terceiro trimestre de 2012 e indo até o último trimestre de 2020. Portanto, o trimestre 1 refere-se ao terceiro trimestre de 2012 e assim por diante.

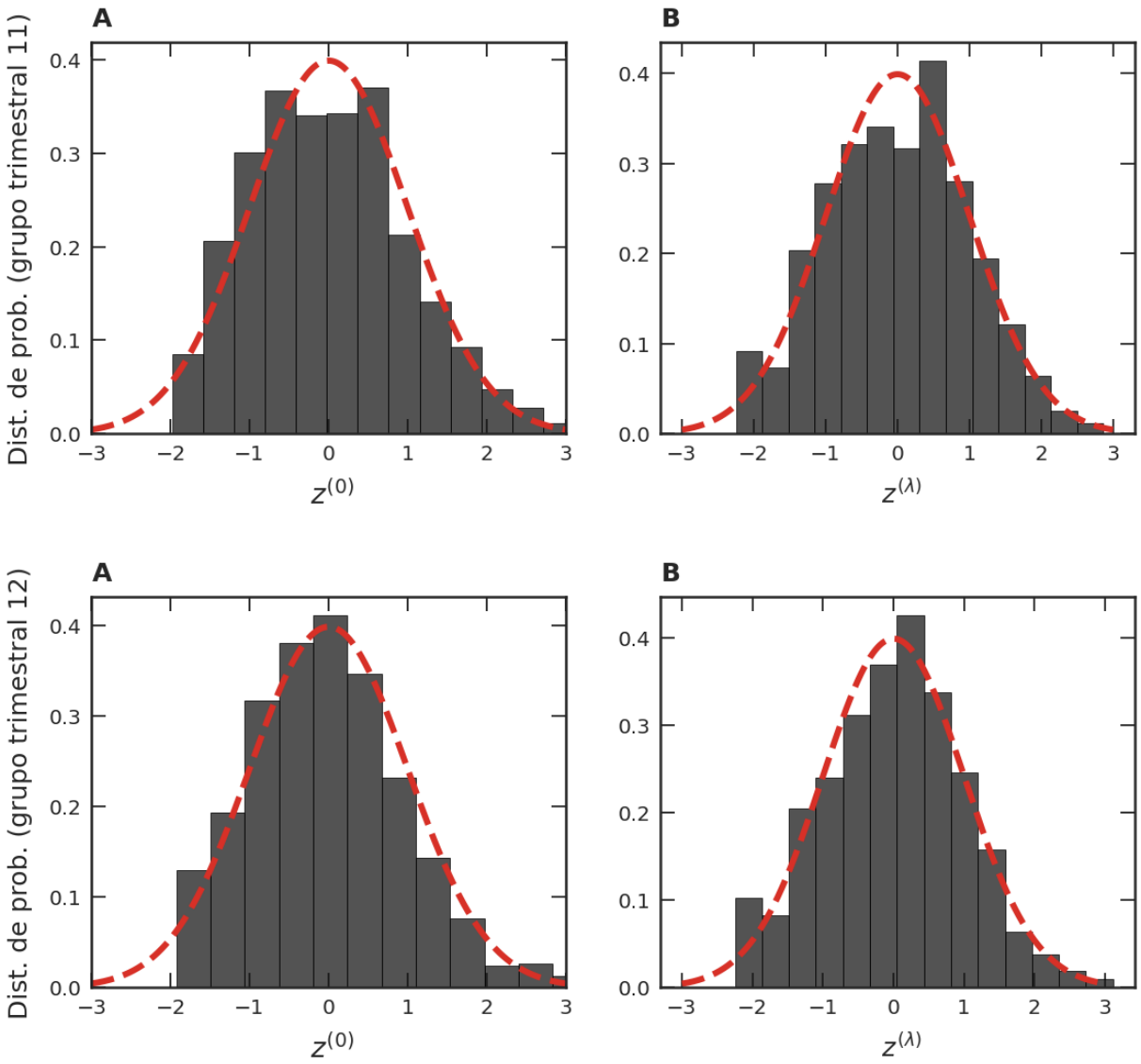

A
B
A
B
Dist. de prob. (grupo trimestral 1)
Dist. de prob. (grupo trimestral 2)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
$z^{(0)}$
$z^{(\lambda)}$

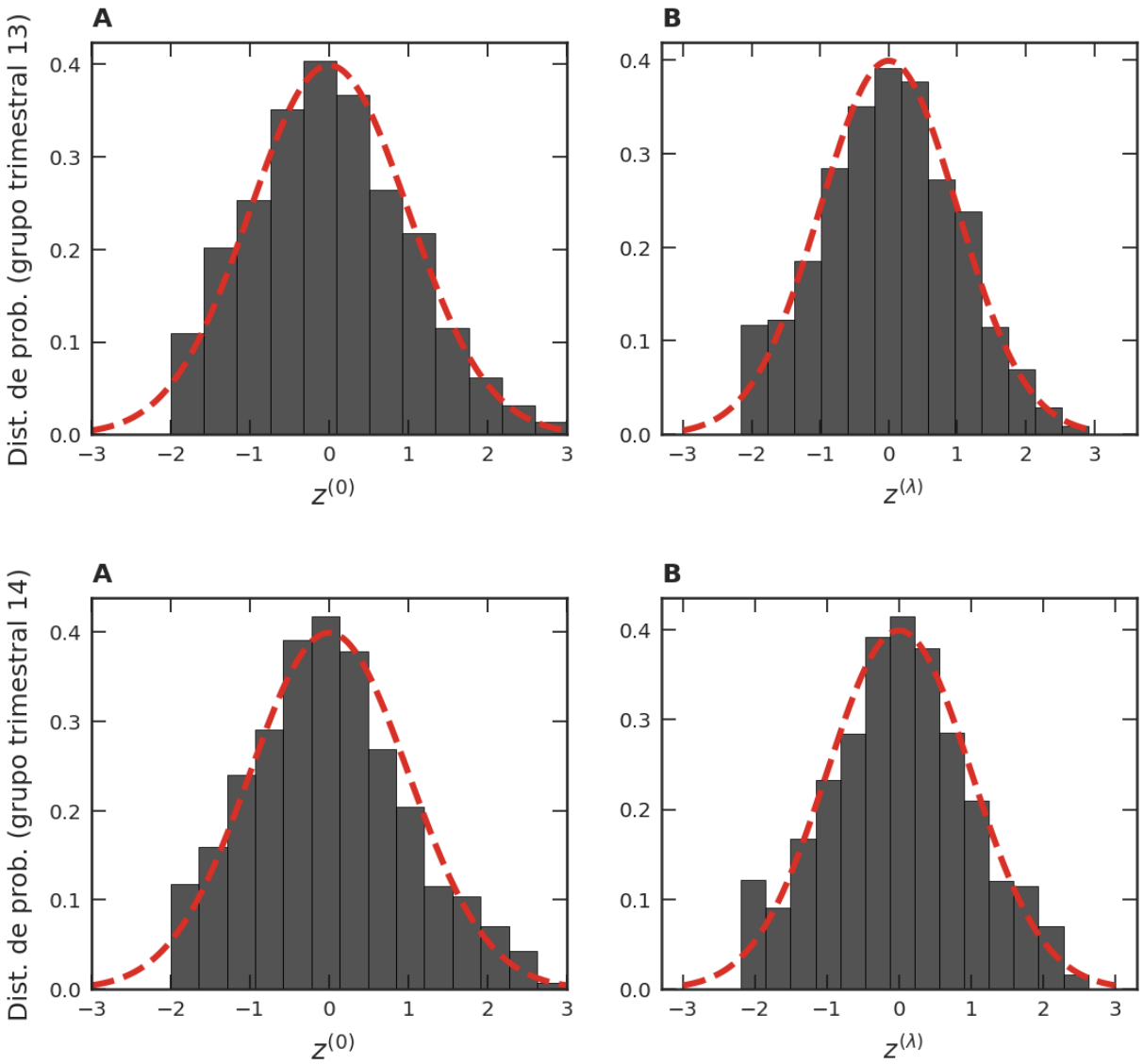

A
B
A
B
Dist. de prob. (grupo trimestral 3)
Dist. de prob. (grupo trimestral 4)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
$z^{(0)}$
$z^{(\lambda)}$

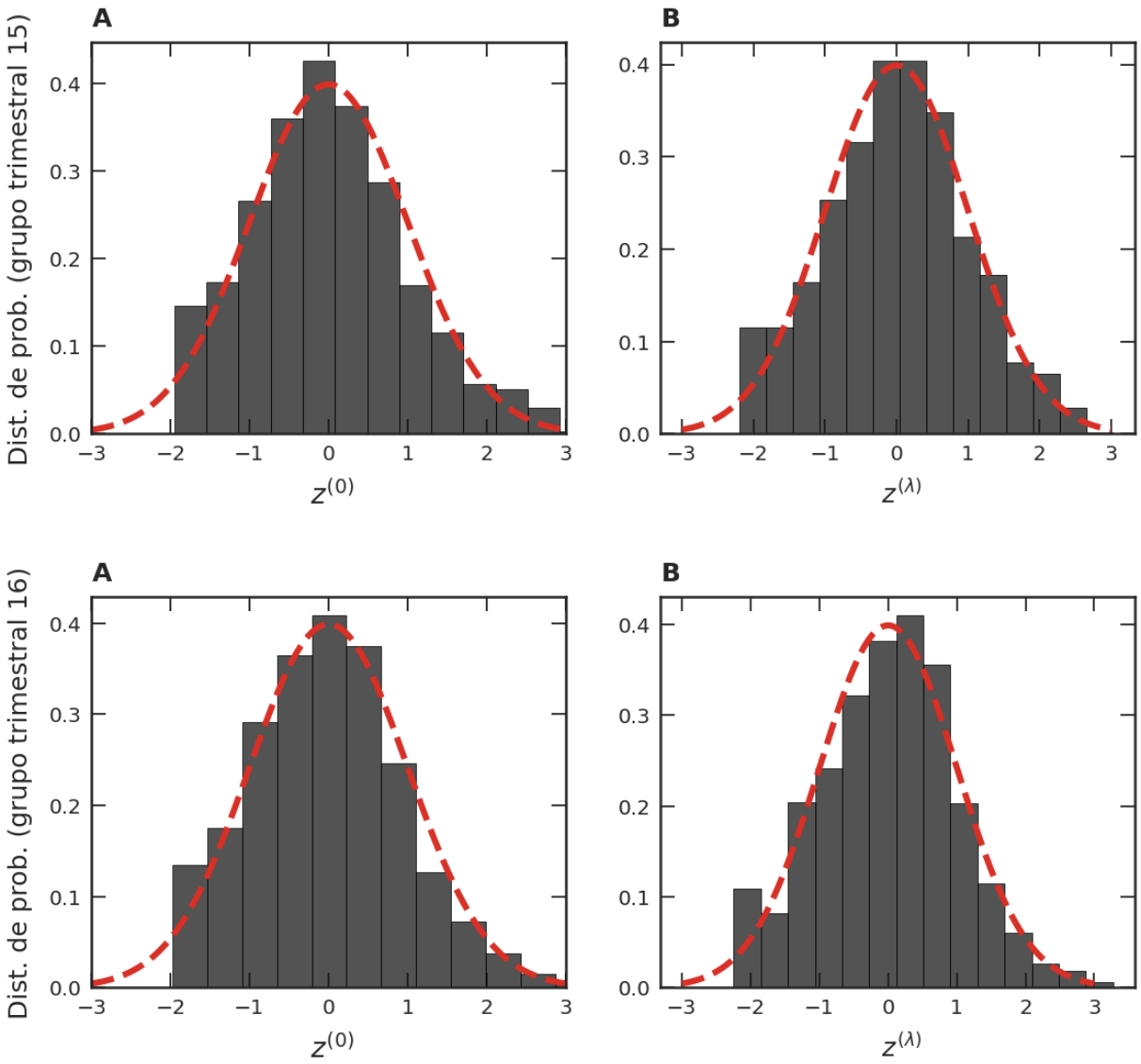

A
B
A
B
Dist. de prob. (grupo trimestral 5)
Dist. de prob. (grupo trimestral 6)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
$z^{(0)}$
$z^{(\lambda)}$

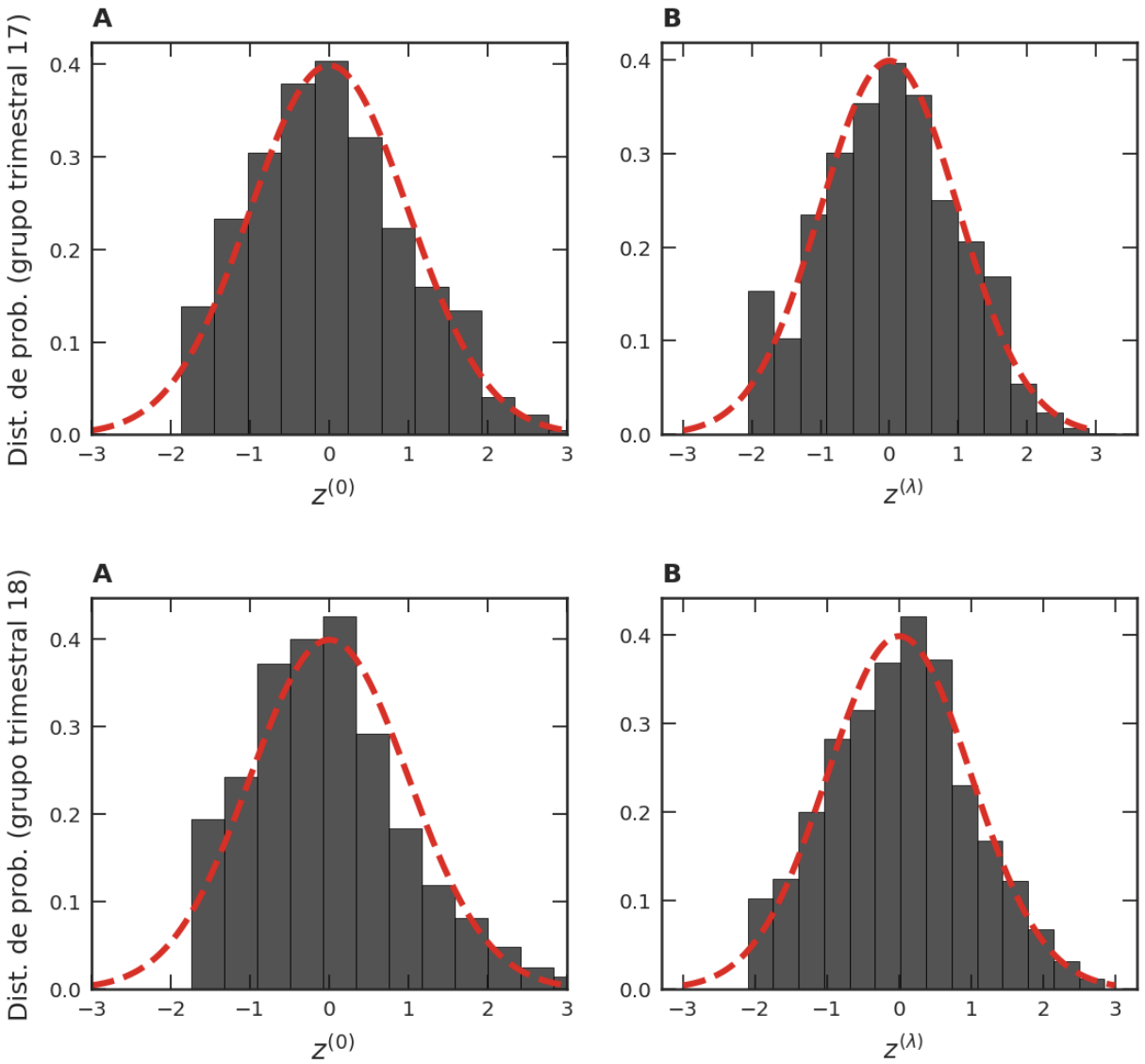

A
B
A
B
Dist. de prob. (grupo trimestral 7)
Dist. de prob. (grupo trimestral 8)
$z^{(0)}$
$z^{(\lambda)}$
$z^{(0)}$
$z^{(\lambda)}$
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3

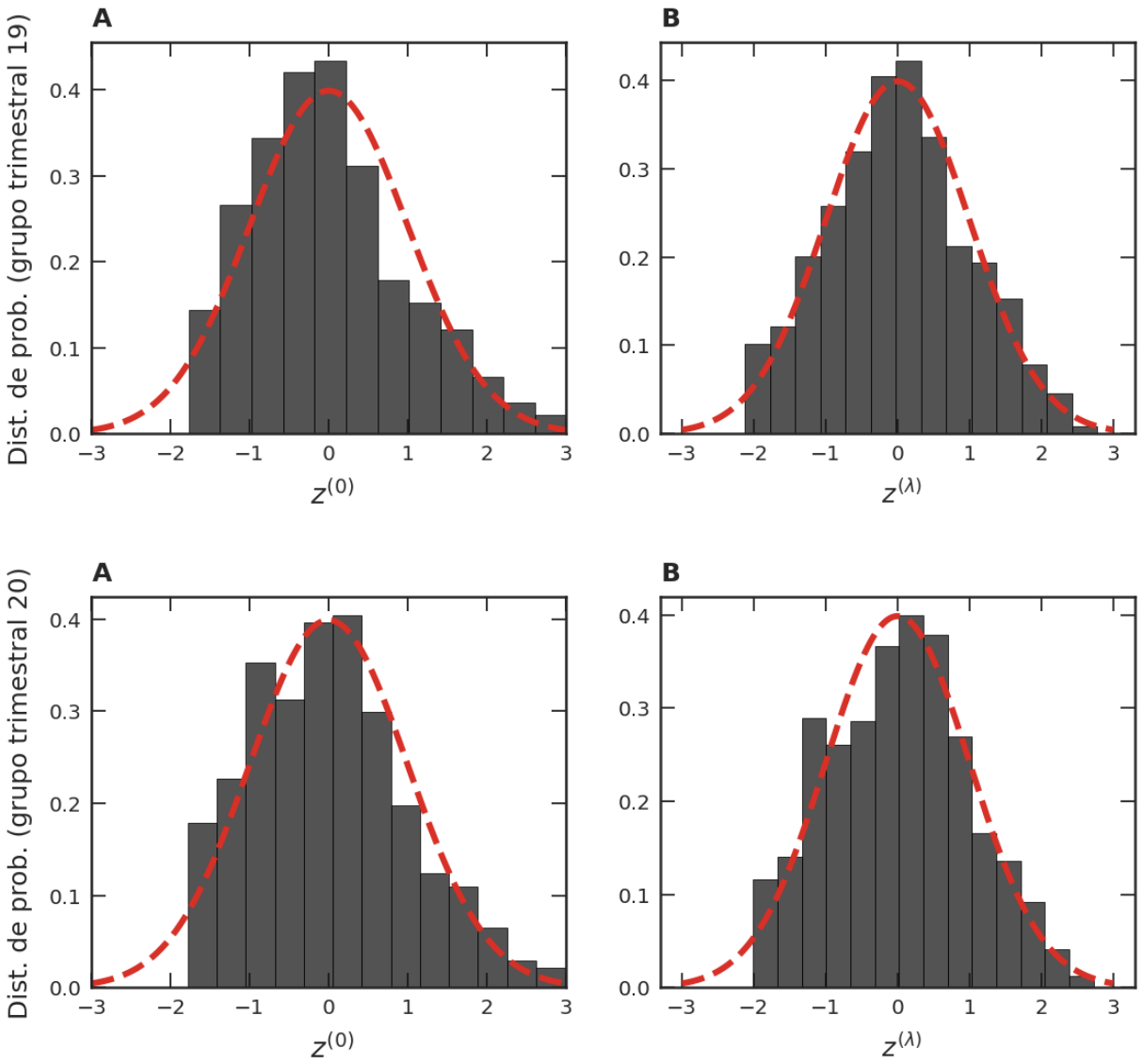

A
B
A
B
Dist. de prob. (grupo trimestral 9)
Dist. de prob. (grupo trimestral 10)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
$z^{(0)}$
$z^{(\lambda)}$

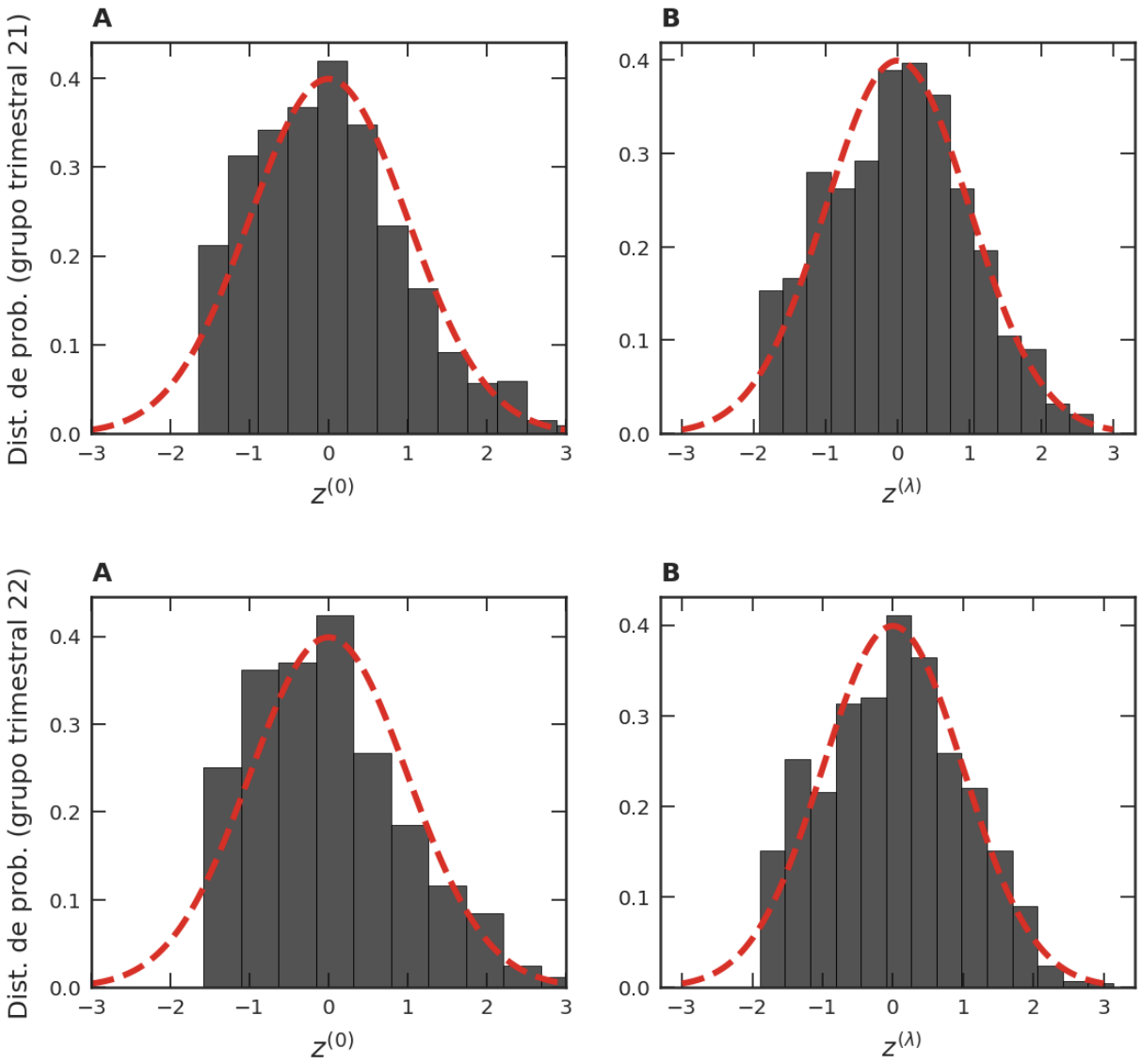

A
B
A
B
Dist. de prob. (grupo trimestral 11)
Dist. de prob. (grupo trimestral 12)
$z^{(0)}$
$z^{(\lambda)}$
$z^{(0)}$
$z^{(\lambda)}$
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3

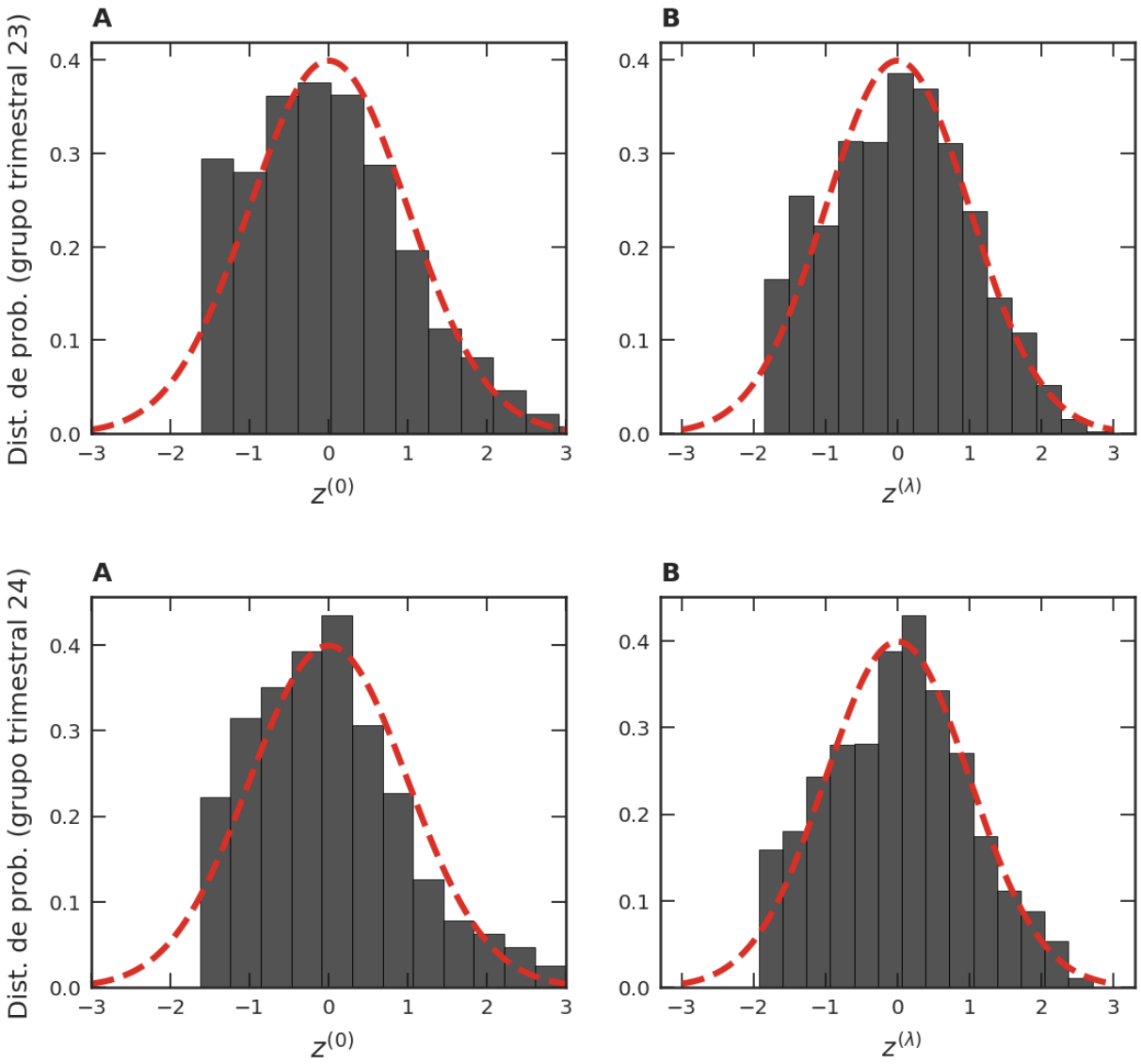

A
B
A
B
Dist. de prob. (grupo trimestral 13)
Dist. de prob. (grupo trimestral 14)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
$z^{(0)}$
$z^{(\lambda)}$

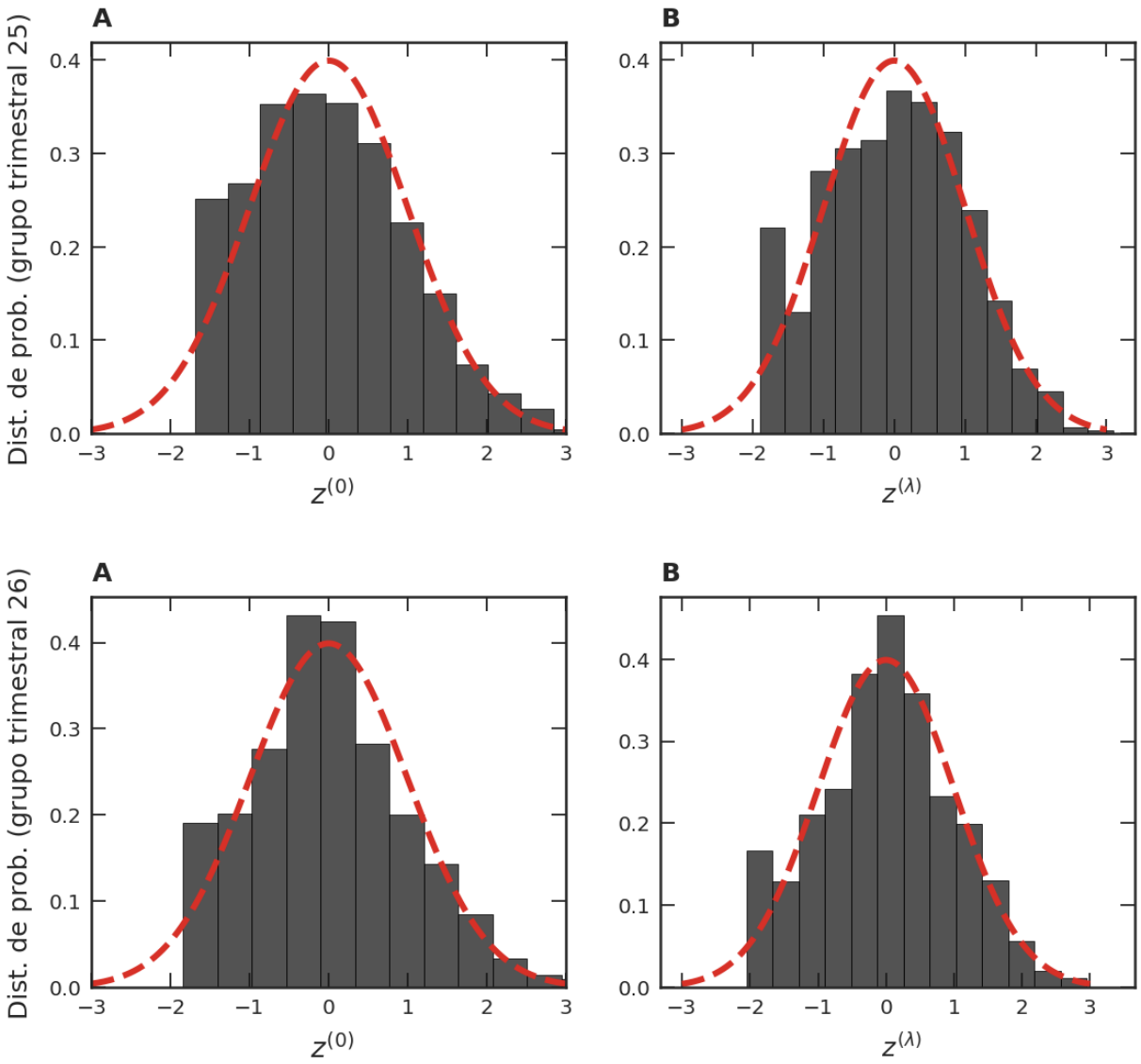

A
B
A
B
Dist. de prob. (grupo trimestral 15)
Dist. de prob. (grupo trimestral 16)
$z^{(0)}$
$z^{(\lambda)}$
$z^{(0)}$
$z^{(\lambda)}$
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3

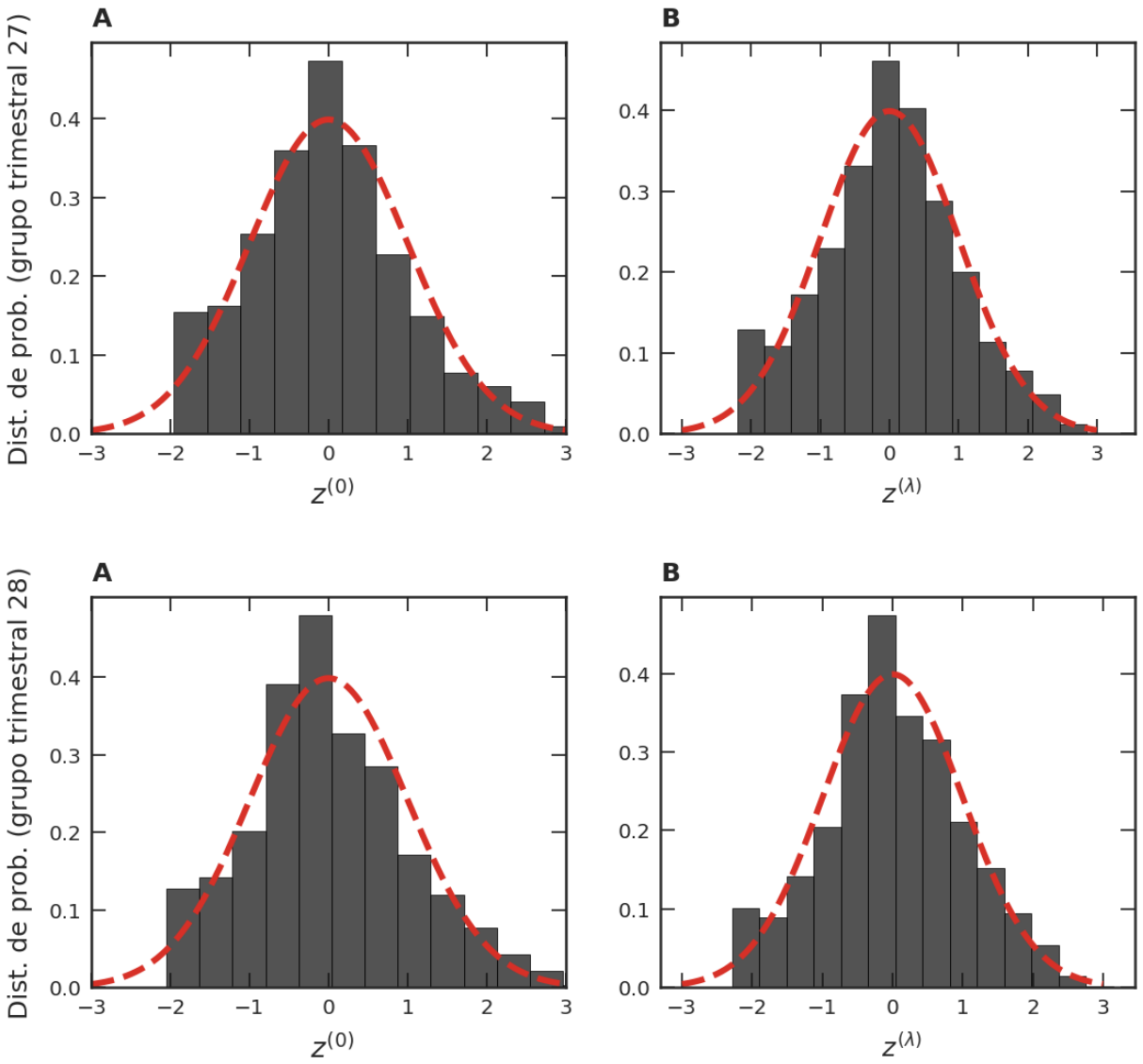

A
B
A
B
Dist. de prob. (grupo trimestral 17)
Dist. de prob. (grupo trimestral 18)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
$z^{(0)}$
$z^{(\lambda)}$

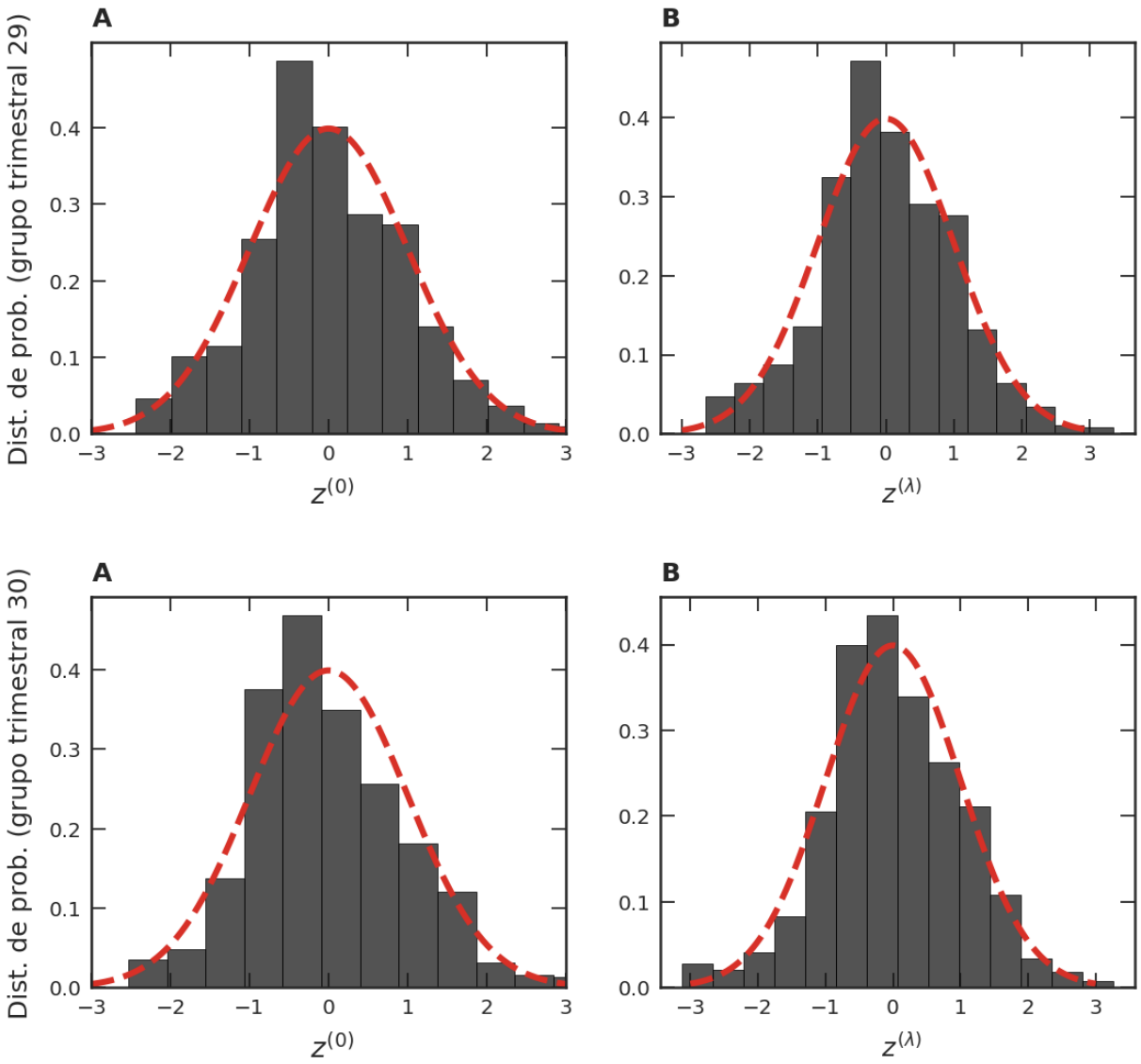

A
B
Dist. de prob. (grupo trimestral 19)
$z^{(0)}$
$z^{(\lambda)}$
A
B
Dist. de prob. (grupo trimestral 20)
$z^{(0)}$
$z^{(\lambda)}$

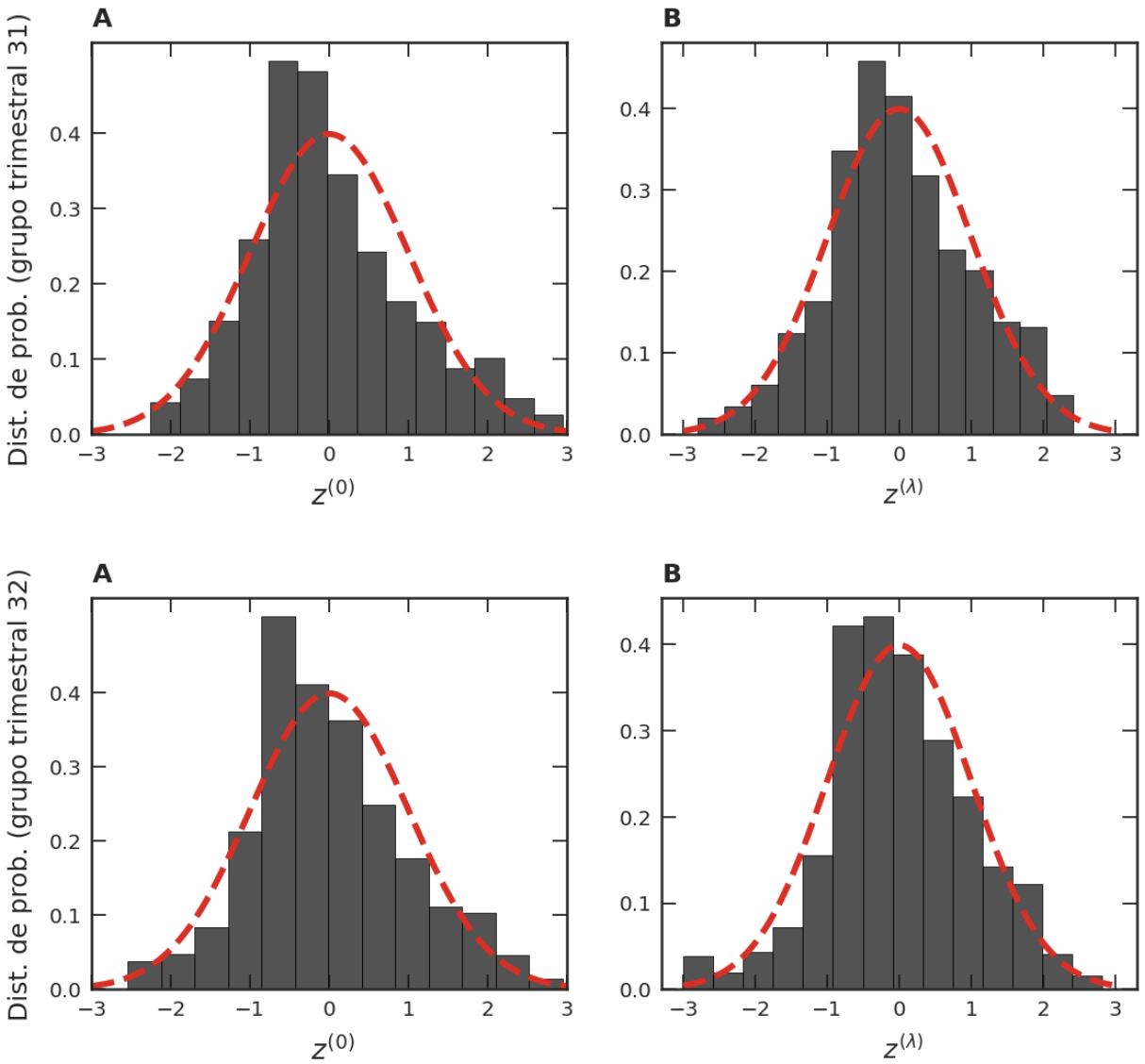

A
B
A
B
Dist. de prob. (grupo trimestral 21)
Dist. de prob. (grupo trimestral 22)
$z^{(0)}$
$z^{(\lambda)}$
$z^{(0)}$
$z^{(\lambda)}$
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3

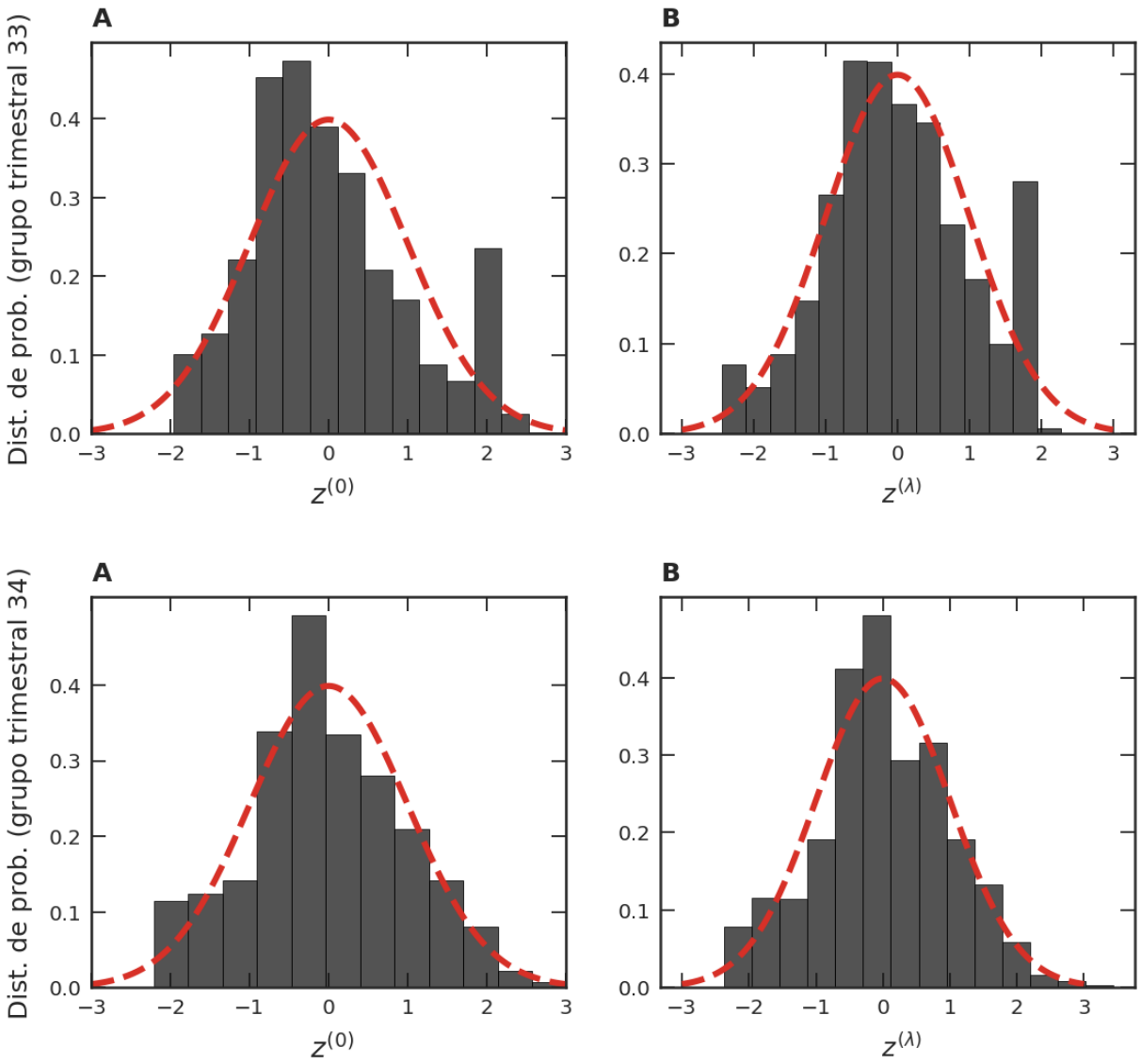

A
B
A
B
Dist. de prob. (grupo trimestral 23)
Dist. de prob. (grupo trimestral 24)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
$z^{(0)}$
$z^{(\lambda)}$


A
B
Dist. de prob. (grupo trimestral 25)
Dist. de prob. (grupo trimestral 26)
$z^{(0)}$
$z^{(\lambda)}$
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3


A
B
Dist. de prob. (grupo trimestral 27)
$z^{(0)}$
$z^{(\lambda)}$
A
B
Dist. de prob. (grupo trimestral 28)
$z^{(0)}$
$z^{(\lambda)}$

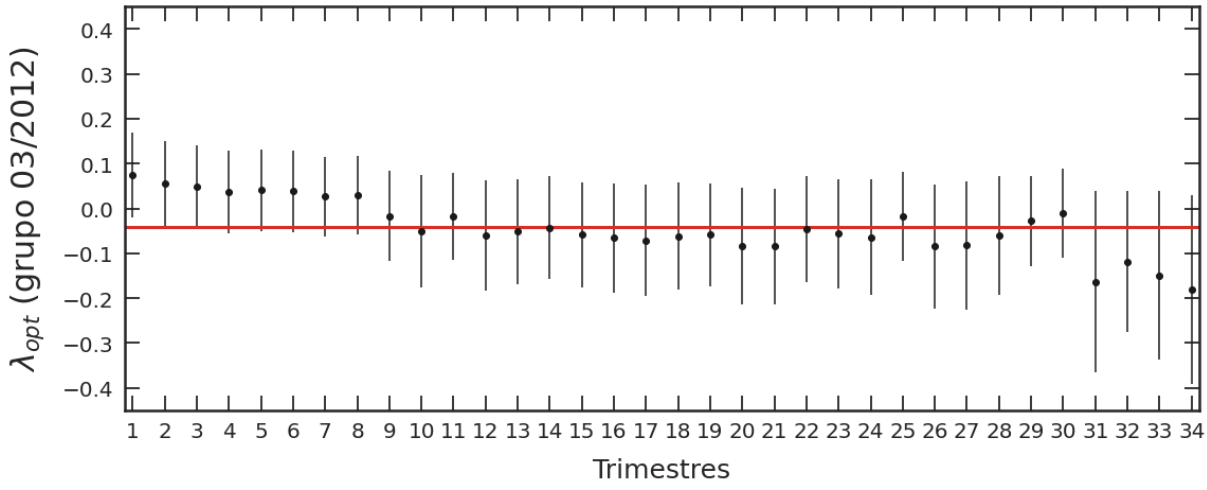

A
B
Dist. de prob. (grupo trimestral 29)
$z^{(0)}$
$z^{(\lambda)}$
A
B
Dist. de prob. (grupo trimestral 30)
$z^{(0)}$
$z^{(\lambda)}$

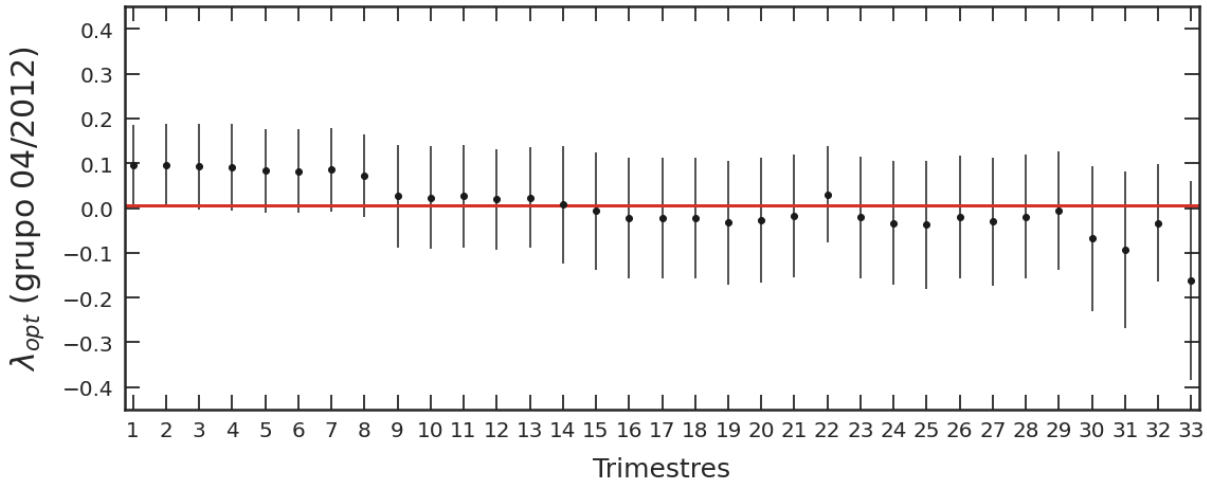

A
B
A
B
Dist. de prob. (grupo trimestral 31)
Dist. de prob. (grupo trimestral 32)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
$z^{(0)}$
$z^{(\lambda)}$

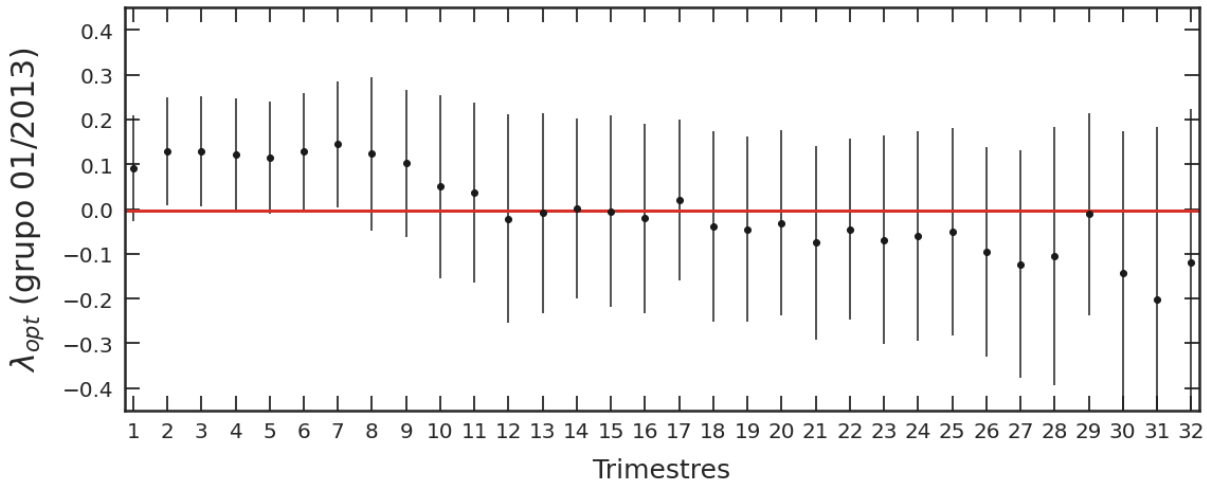

A
B
Dist. de prob. (grupo trimestral 33)
Dist. de prob. (grupo trimestral 34)
$z^{(0)}$
$z^{(\lambda)}$
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3

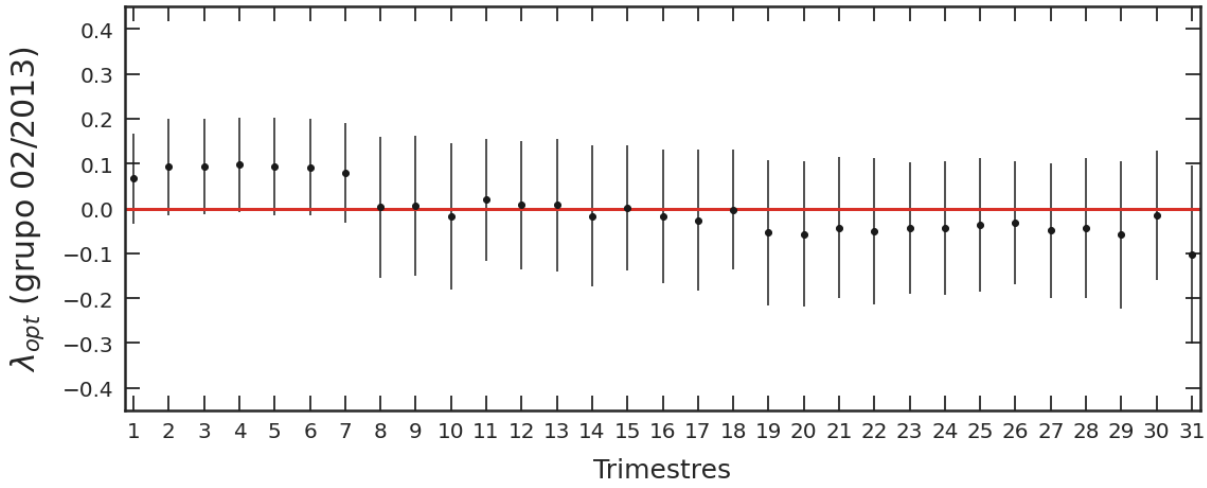
APÊNDICE C – Evolução temporal dos parâmetros de Box-Cox para os grupos trimestrais

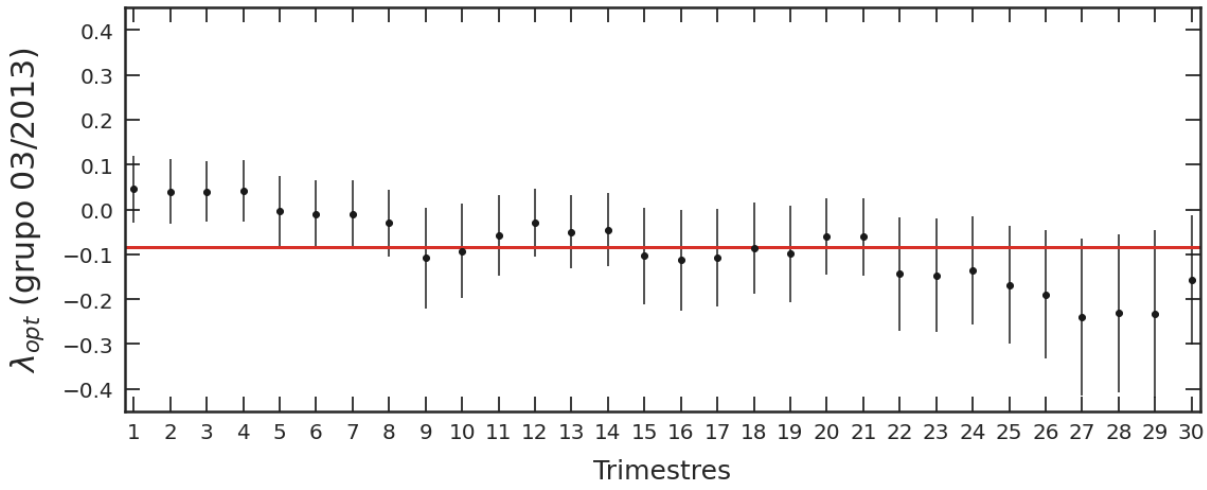
Neste apêndice, apresentamos a evolução temporal do parâmetro ótimo para cada um dos grupos trimestrais. Notamos que decidimos por excluir os grupos 03 e 04 de 2020, já que eles teriam poucos pontos (apenas dois e um, respectivamente).

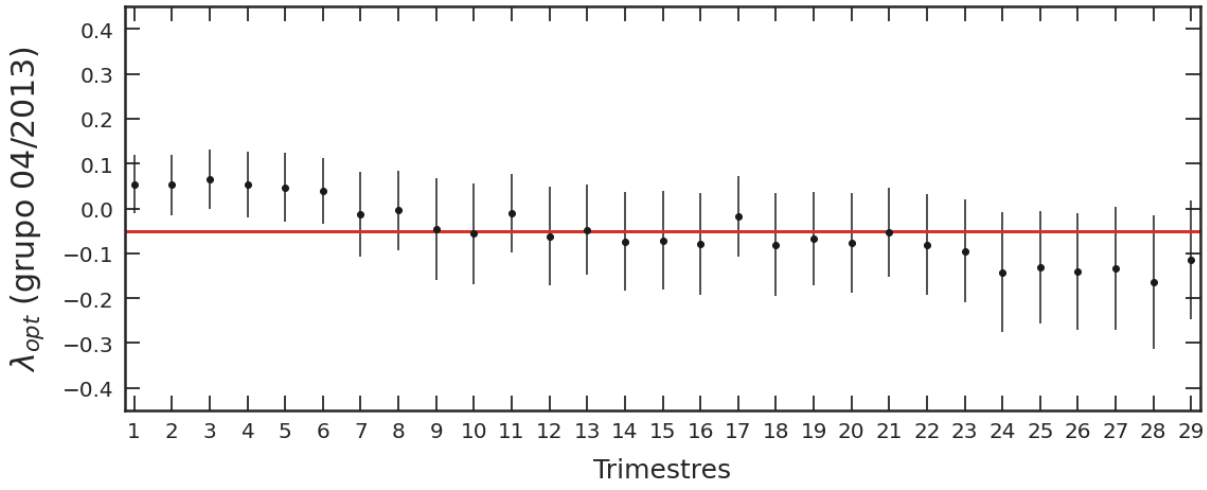


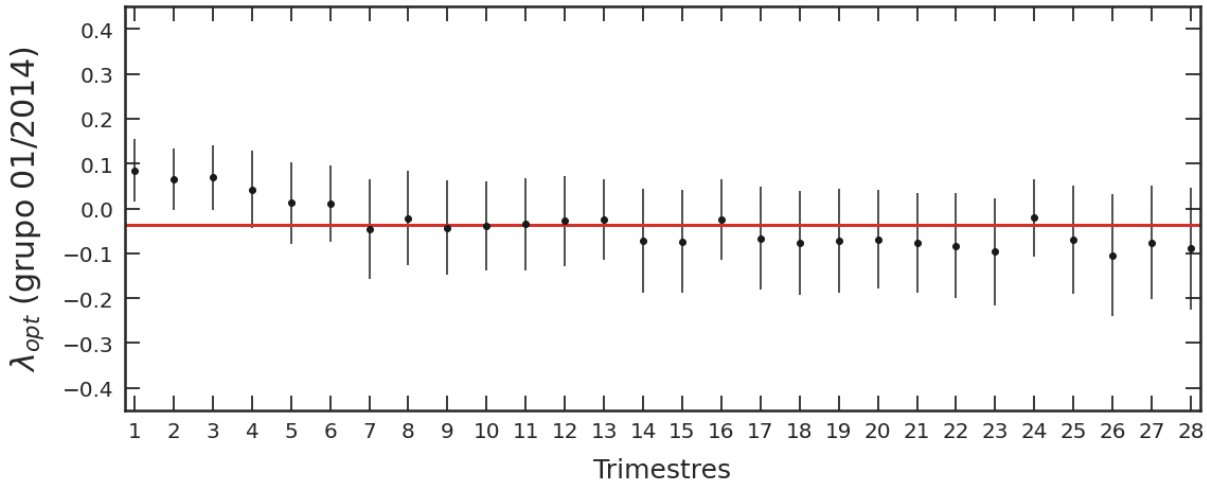


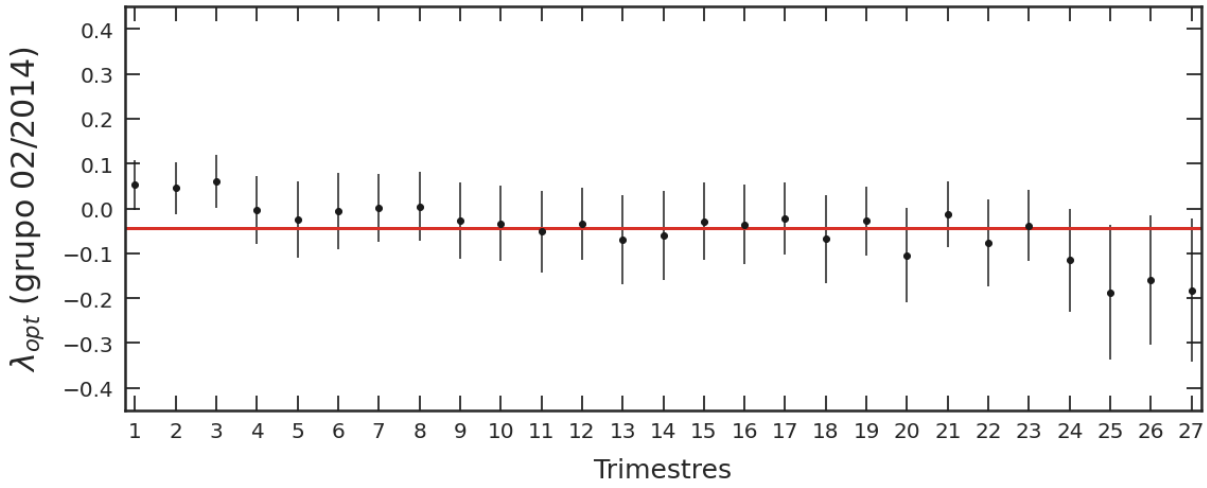

λ_{opt} (grupo 01/2013)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Trimestres

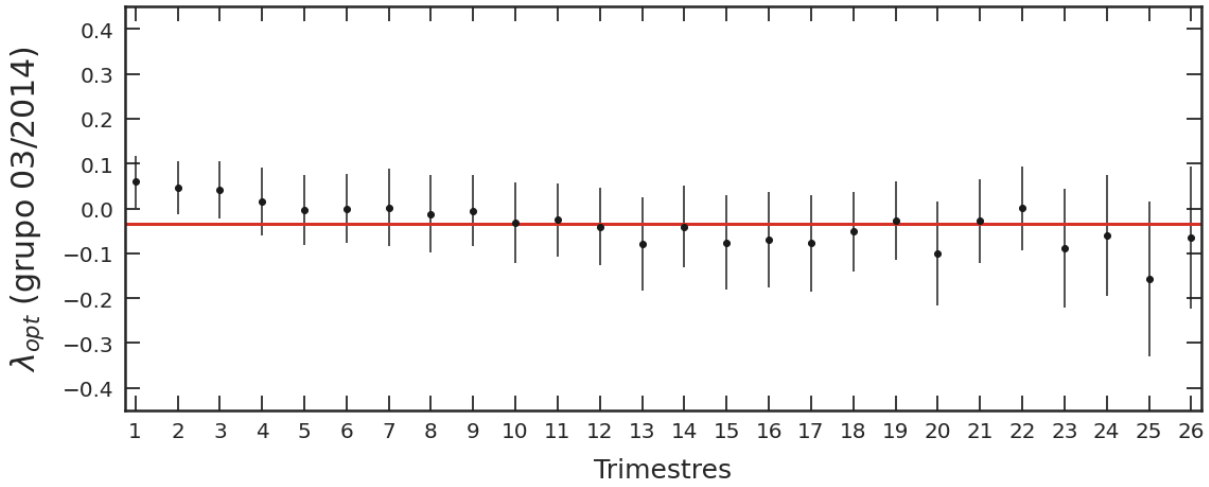

λ_{opt} (grupo 02/2013)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Trimestres

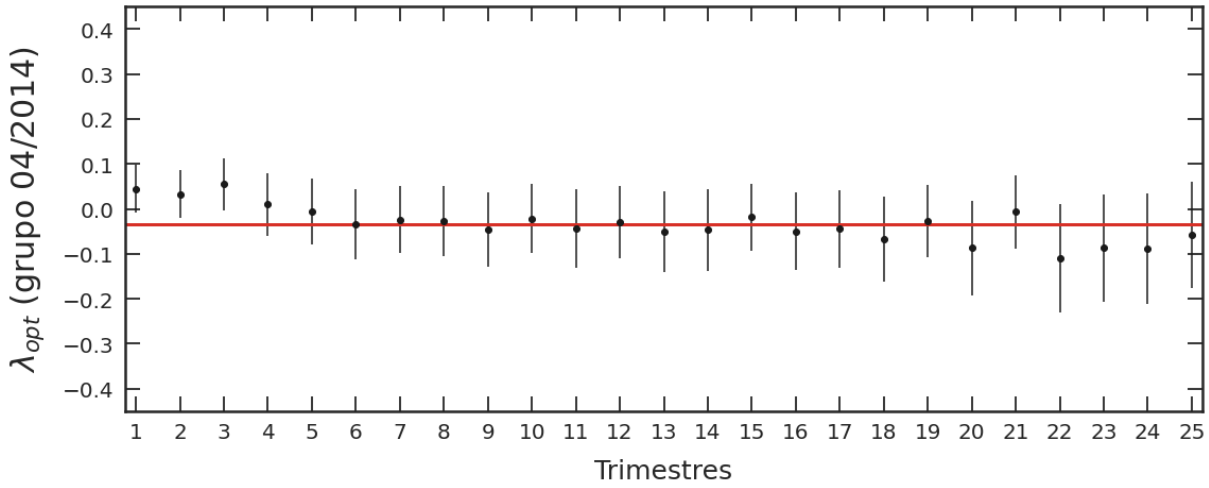

λ_{opt} (grupo 03/2013)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Trimestres

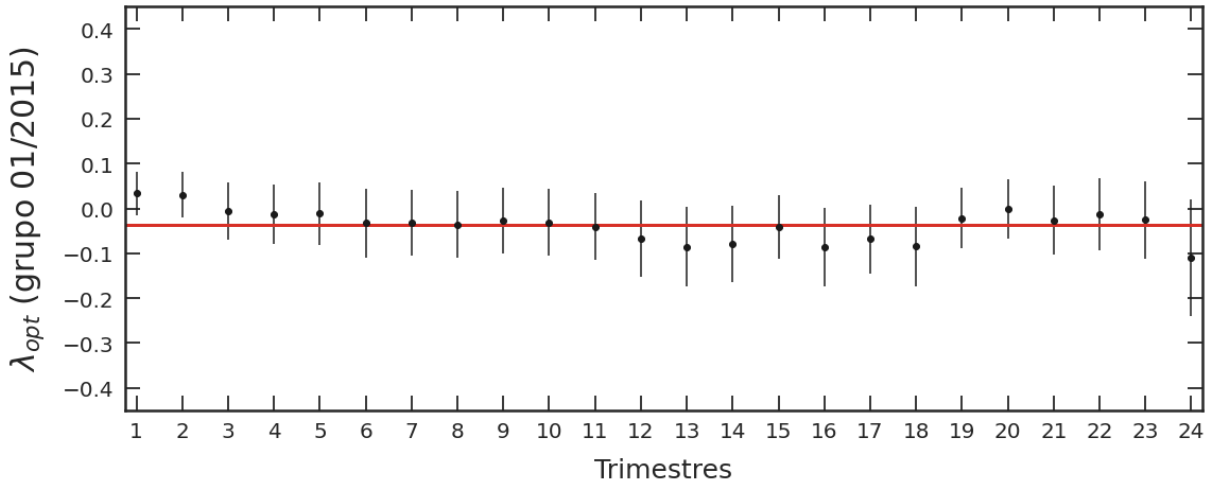

λ_{opt} (grupo 04/2013)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
Trimestres

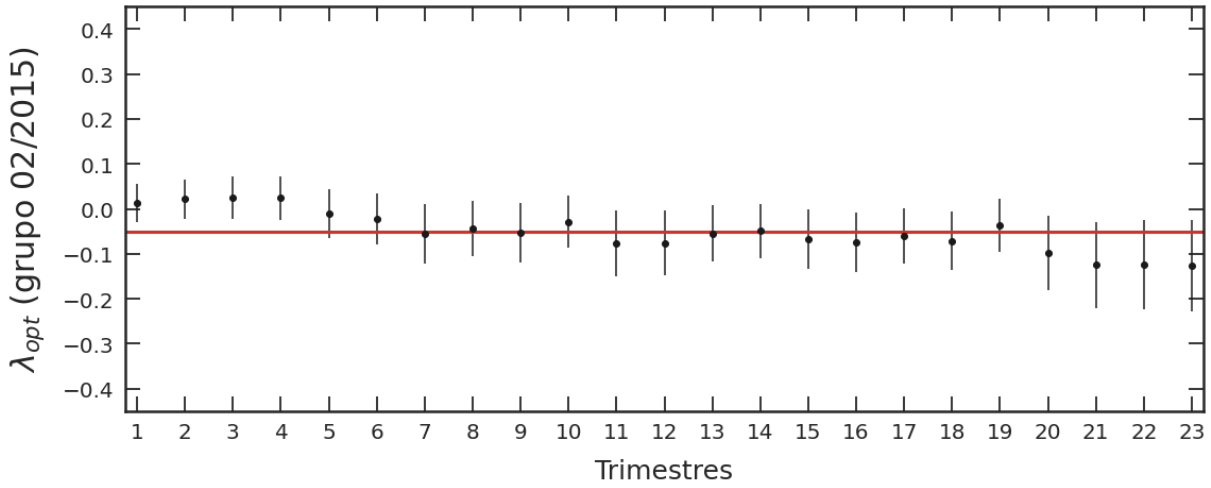

λ_{opt} (grupo 01/2014)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
Trimestres

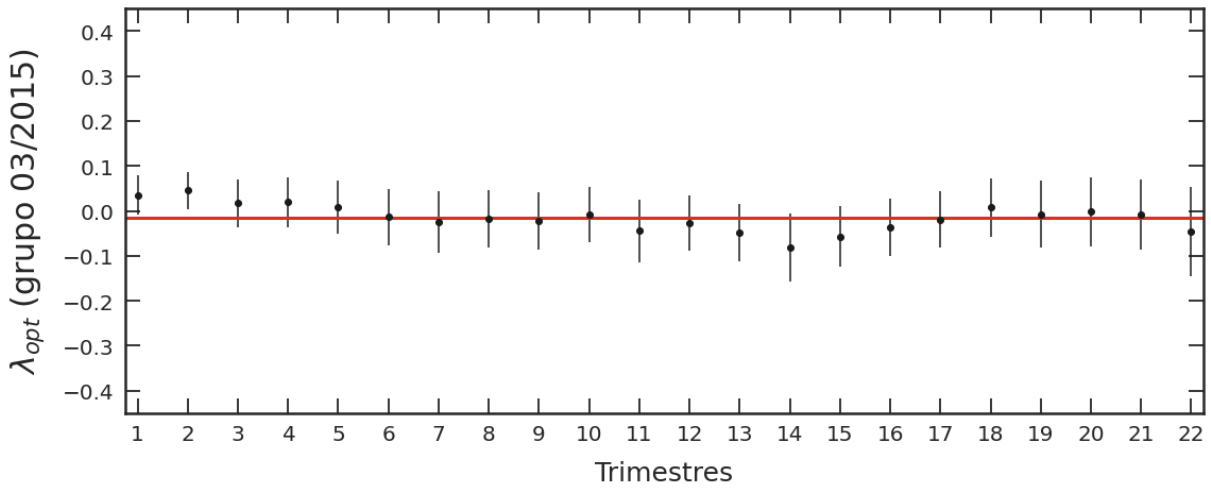

λ_{opt} (grupo 02/2014)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
Trimestres

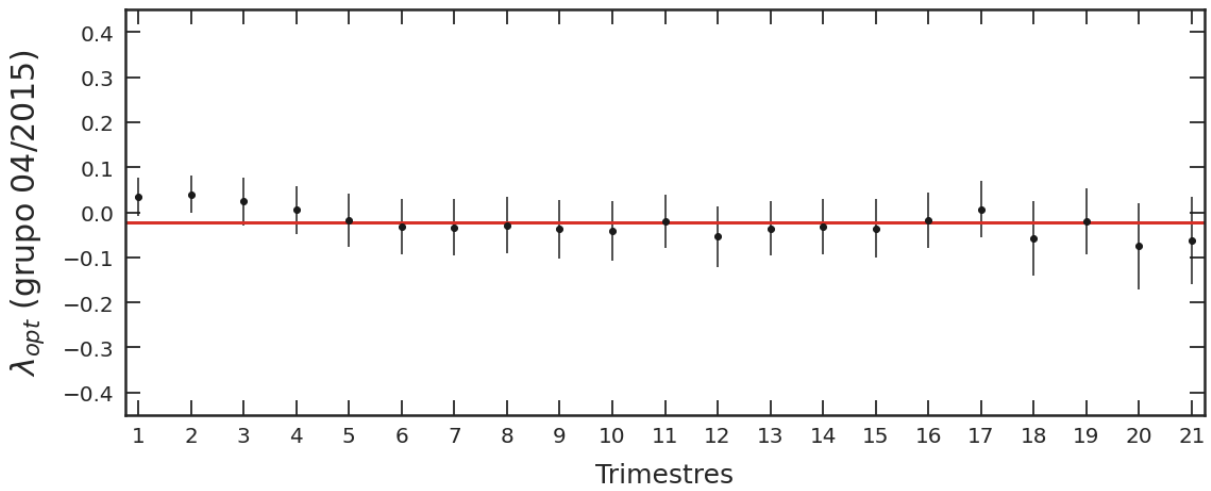

λ_{opt} (grupo 03/2014)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Trimestres

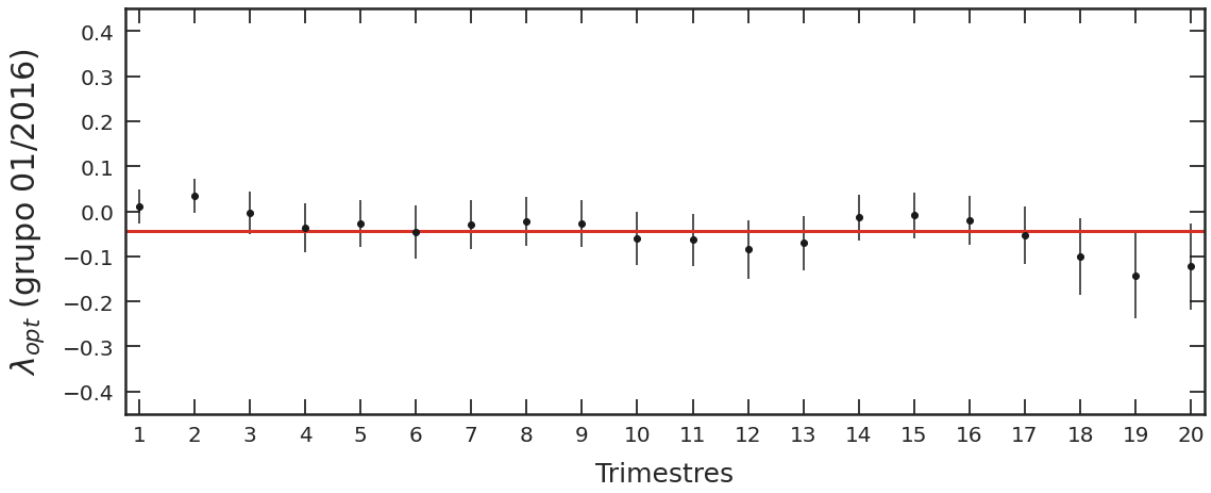

λ_{opt} (grupo 04/2014)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Trimestres

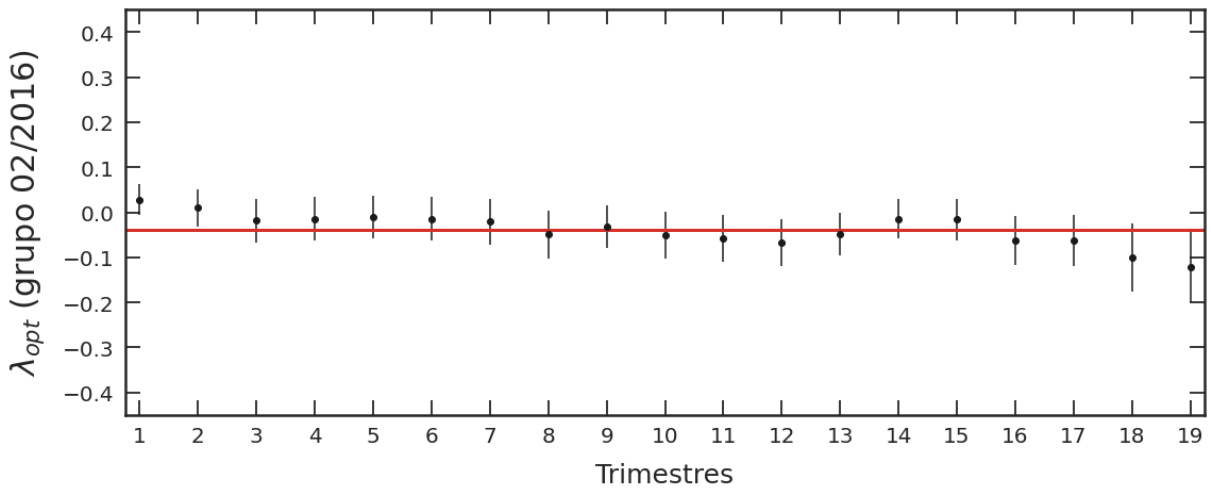

λ_{opt} (grupo 01/2015)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Trimestres

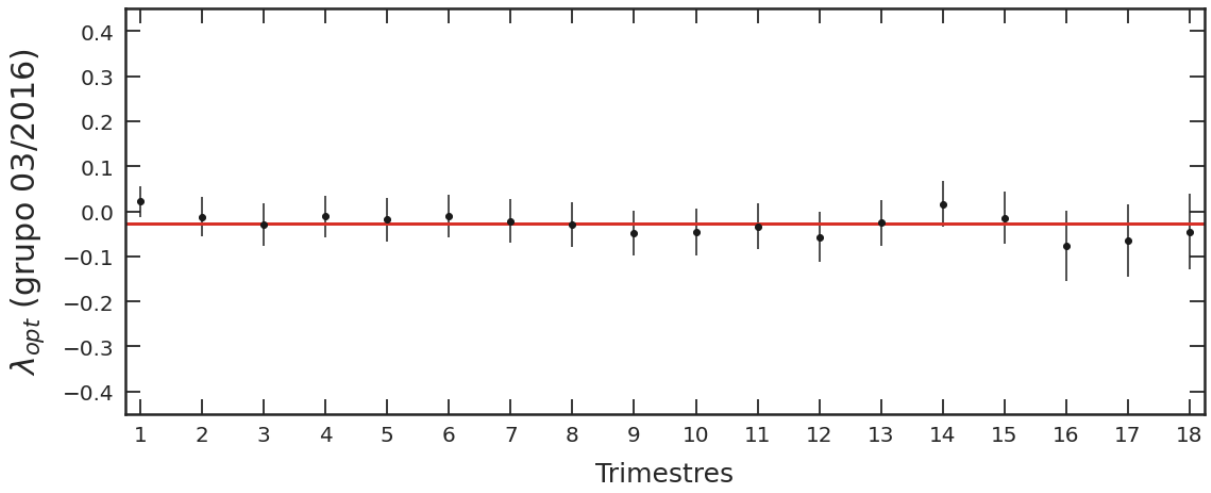

λ_{opt} (grupo 02/2015)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Trimestres

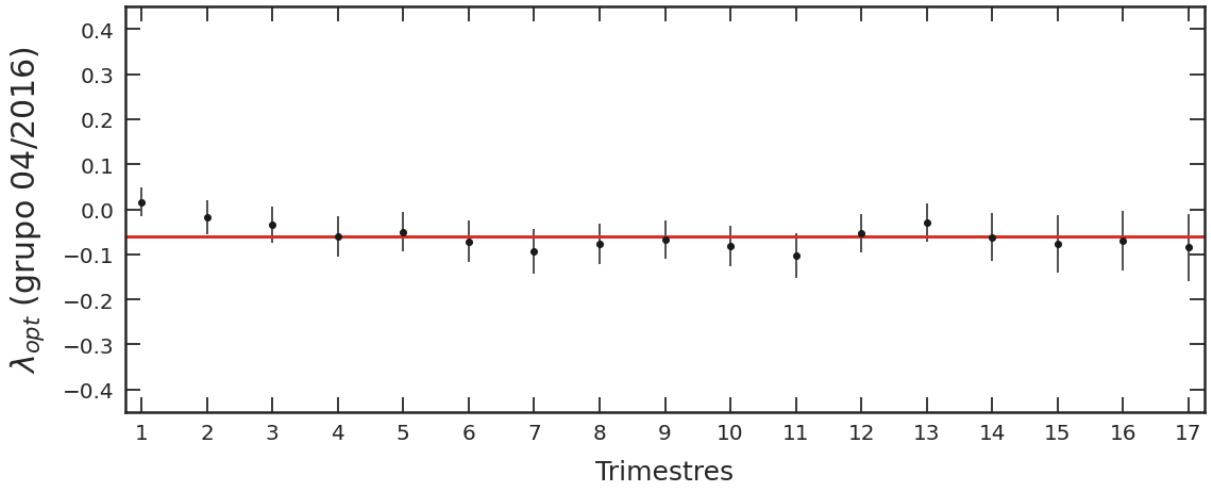

λ_{opt} (grupo 03/2015)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Trimestres

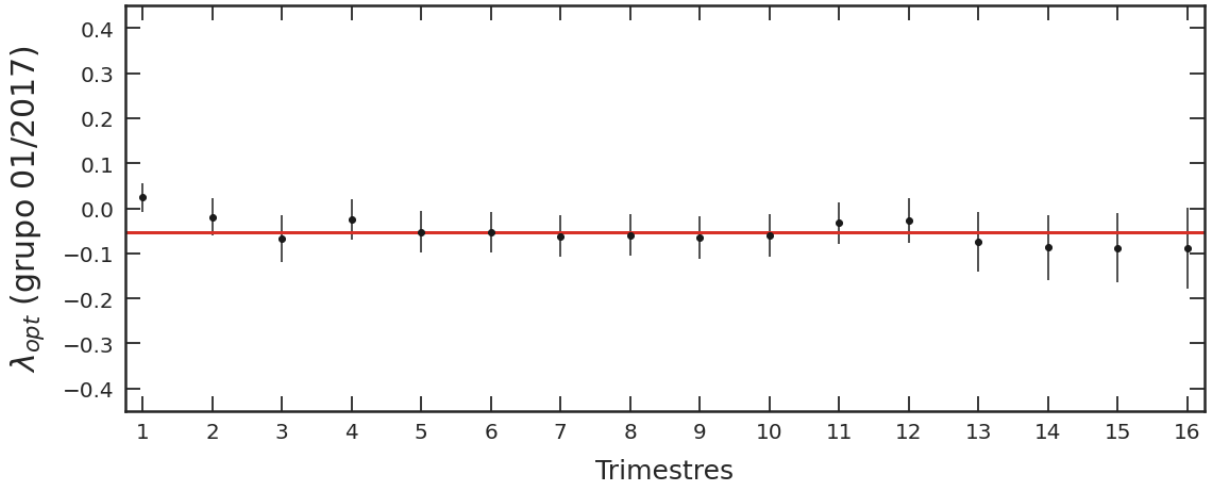

λ_{opt} (grupo 04/2015)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Trimestres

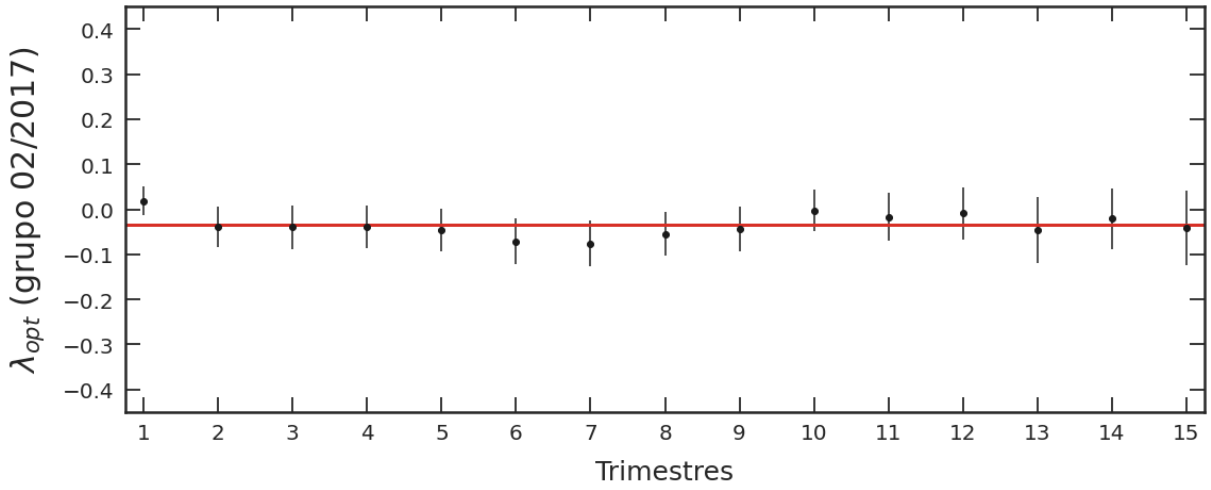

λ_{opt} (grupo 01/2016)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Trimestres

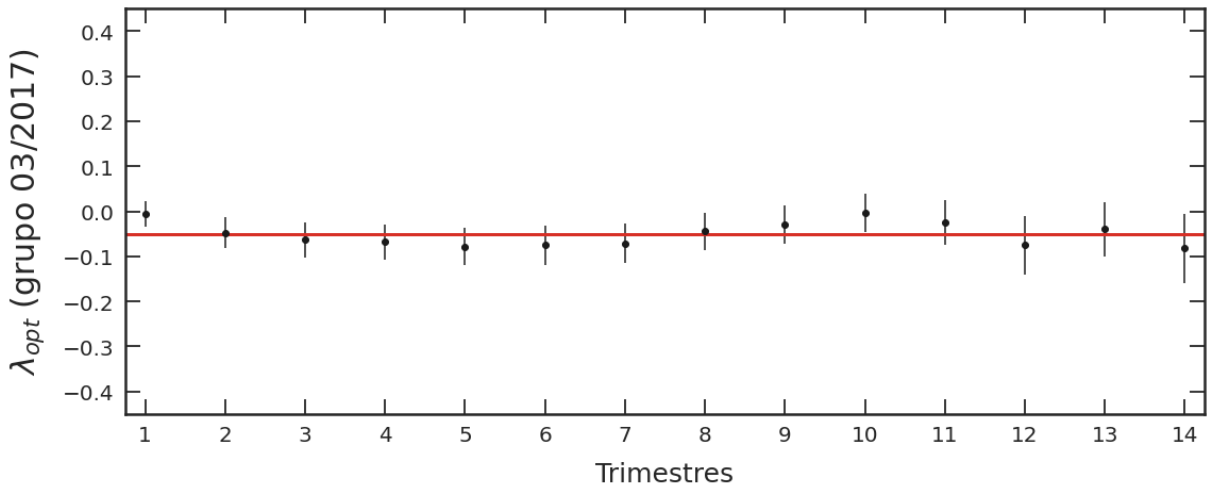

λ_{opt} (grupo 02/2016)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Trimestres

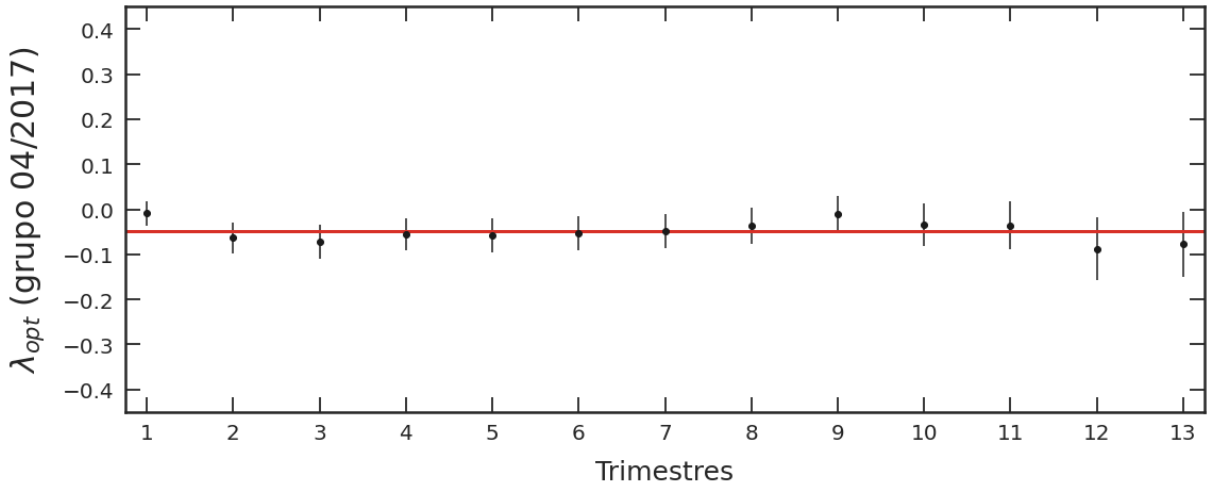

λ_{opt} (grupo 03/2016)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Trimestres

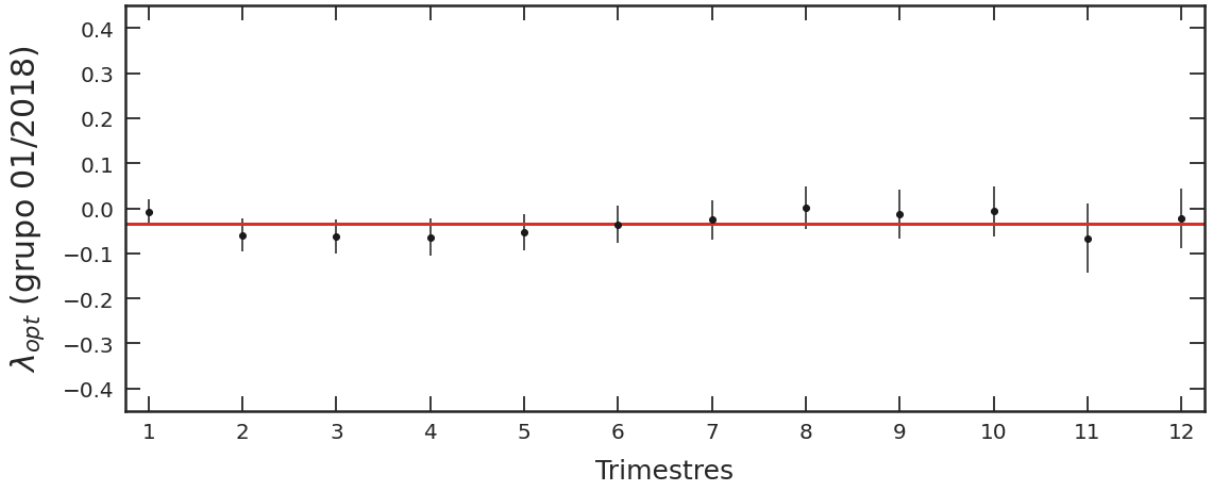

λ_{opt} (grupo 04/2016)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Trimestres

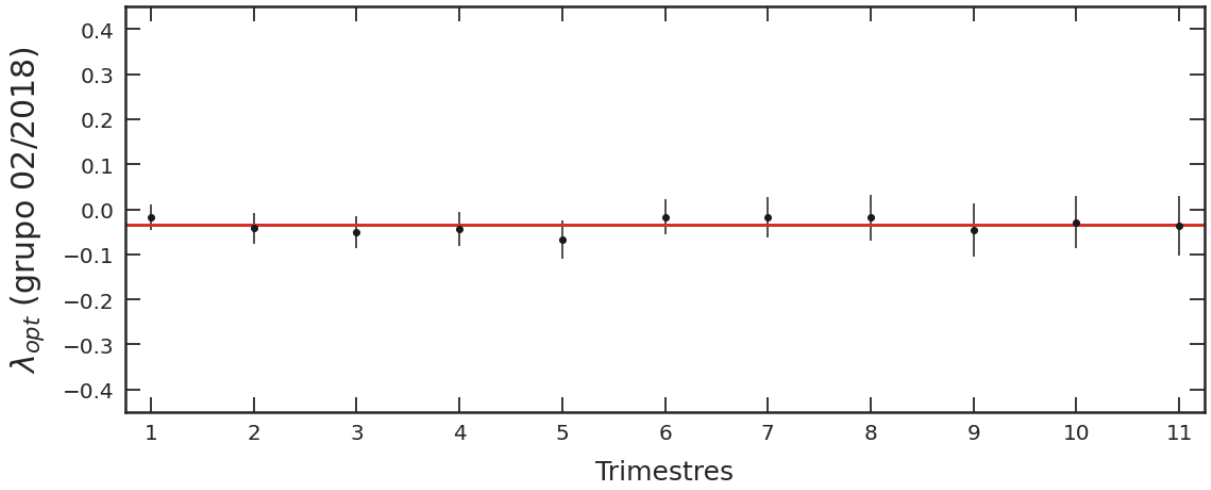

λ_{opt} (grupo 01/2017)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Trimestres

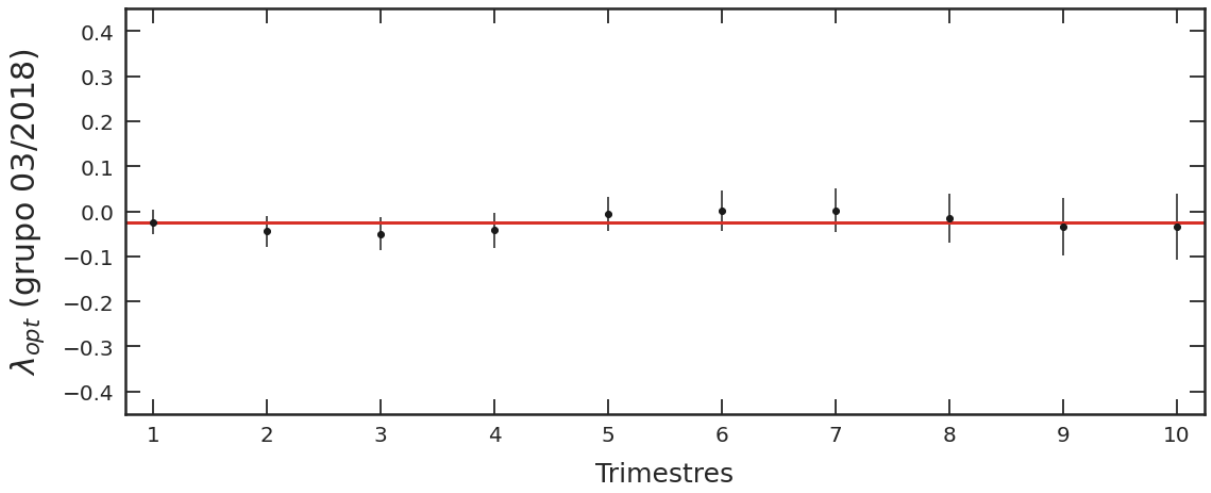

λ_{opt} (grupo 02/2017)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Trimestres

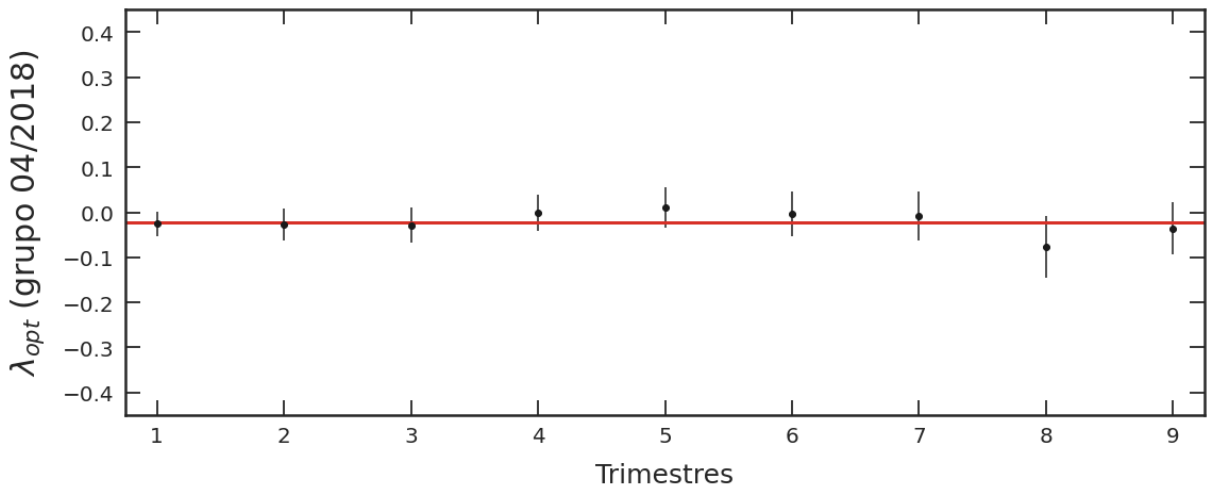

λ_{opt} (grupo 03/2017)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Trimestres

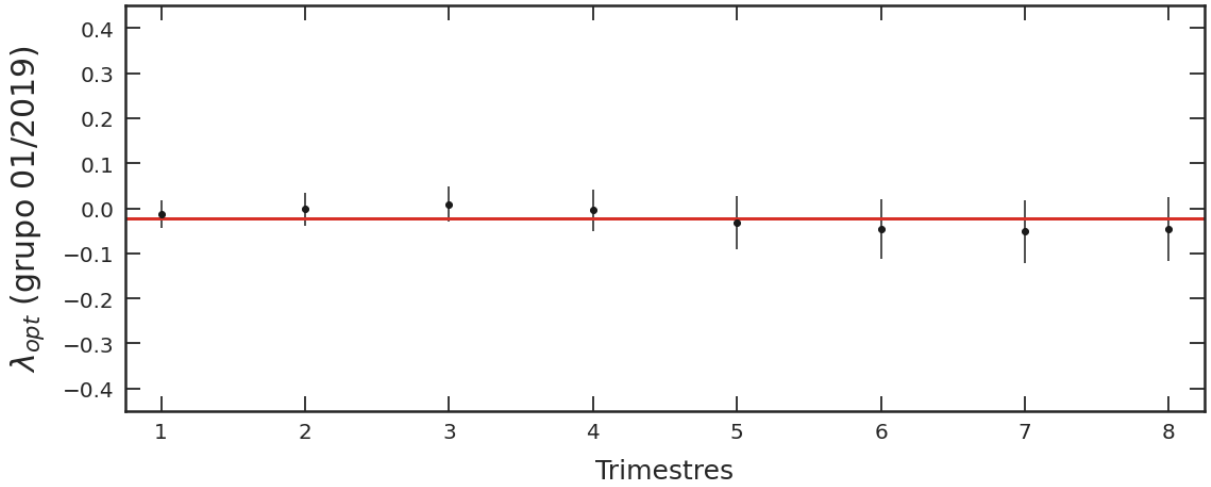

λ_{opt} (grupo 04/2017)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
4
5
6
7
8
9
10
11
12
13
Trimestres

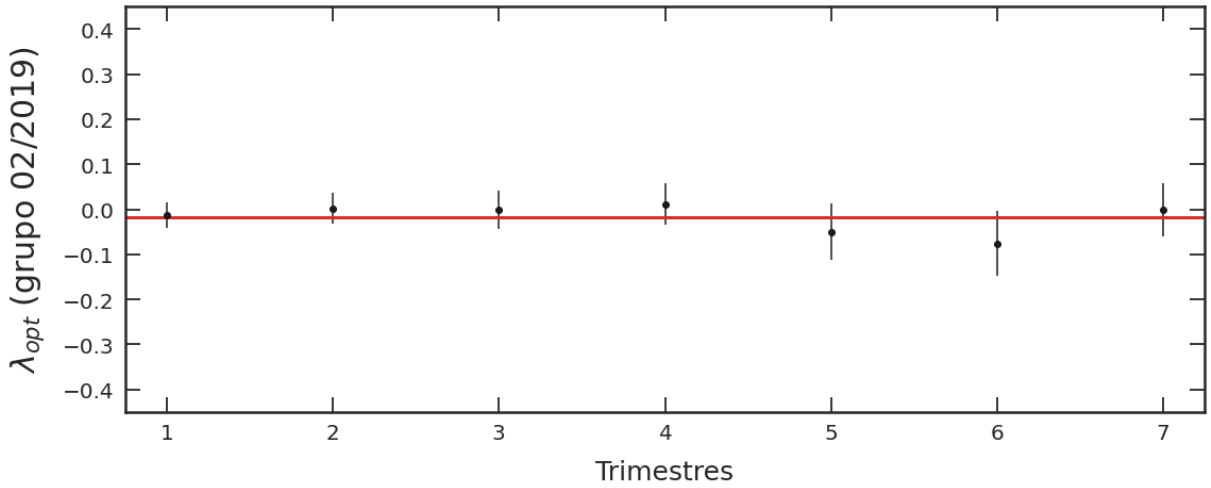

λ_{opt} (grupo 01/2018)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
4
5
6
7
8
9
10
11
12
Trimestres

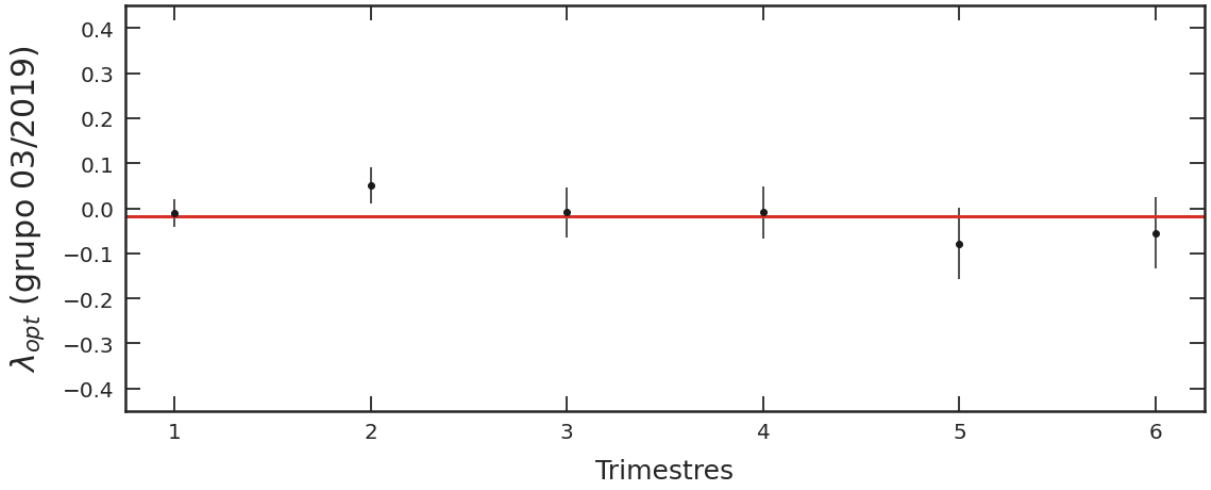

λ_{opt} (grupo 02/2018)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
4
5
6
7
8
9
10
11
Trimestres

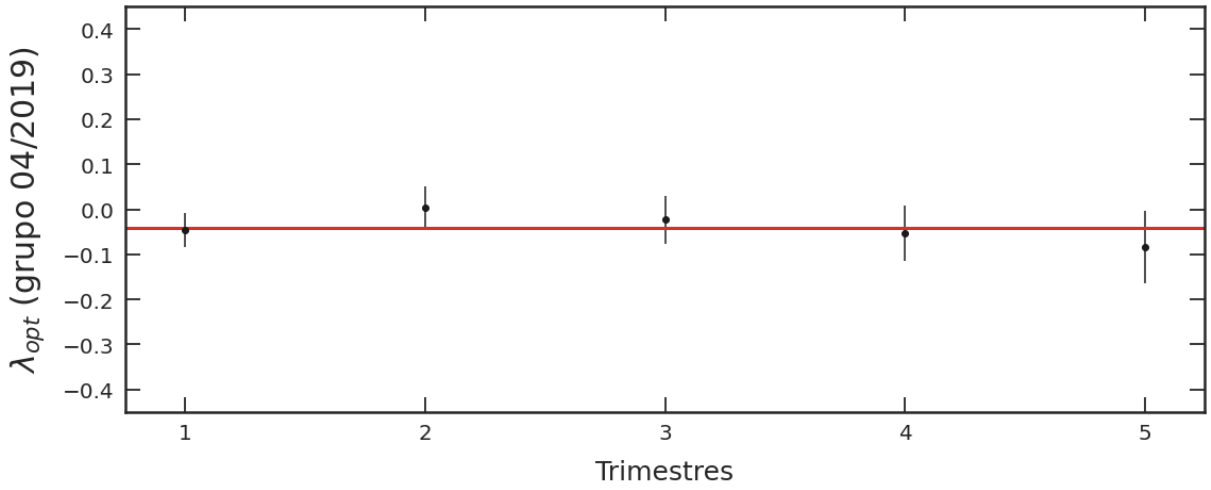

λ_{opt} (grupo 03/2018)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
4
5
6
7
8
9
10
Trimestres

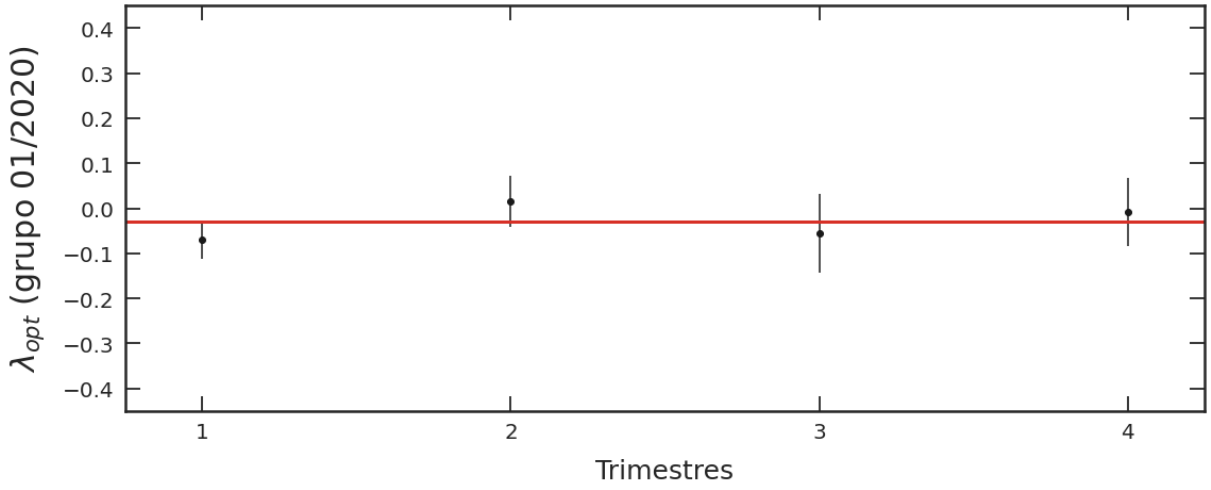

λ_{opt} (grupo 04/2018)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
4
5
6
7
8
9
Trimestres

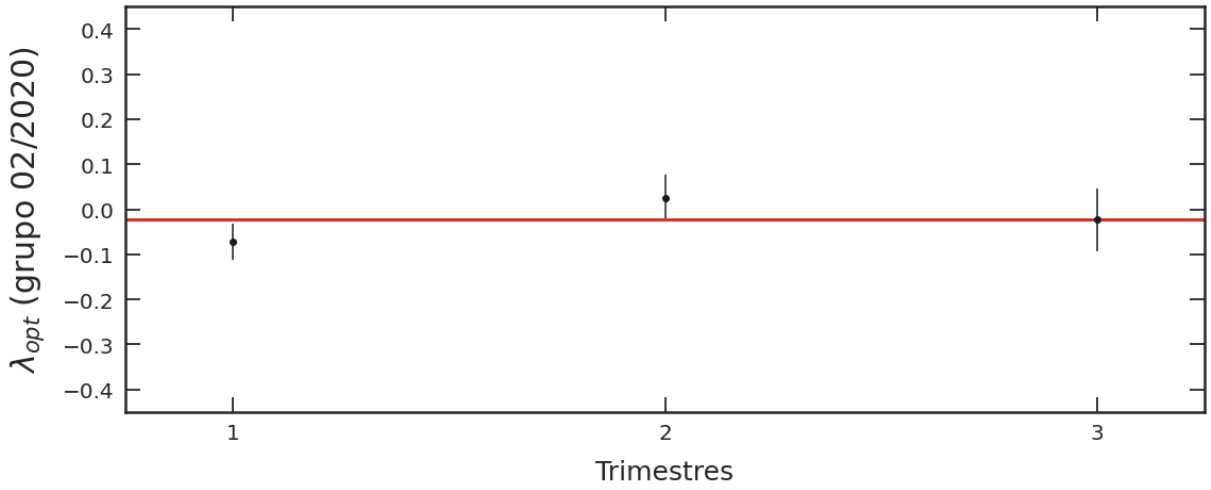

λ_{opt} (grupo 01/2019)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
4
5
6
7
8
Trimestres


λ_{opt} (grupo 02/2019)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
4
5
6
7
Trimestres


λ_{opt} (grupo 03/2019)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
4
5
6
Trimestres

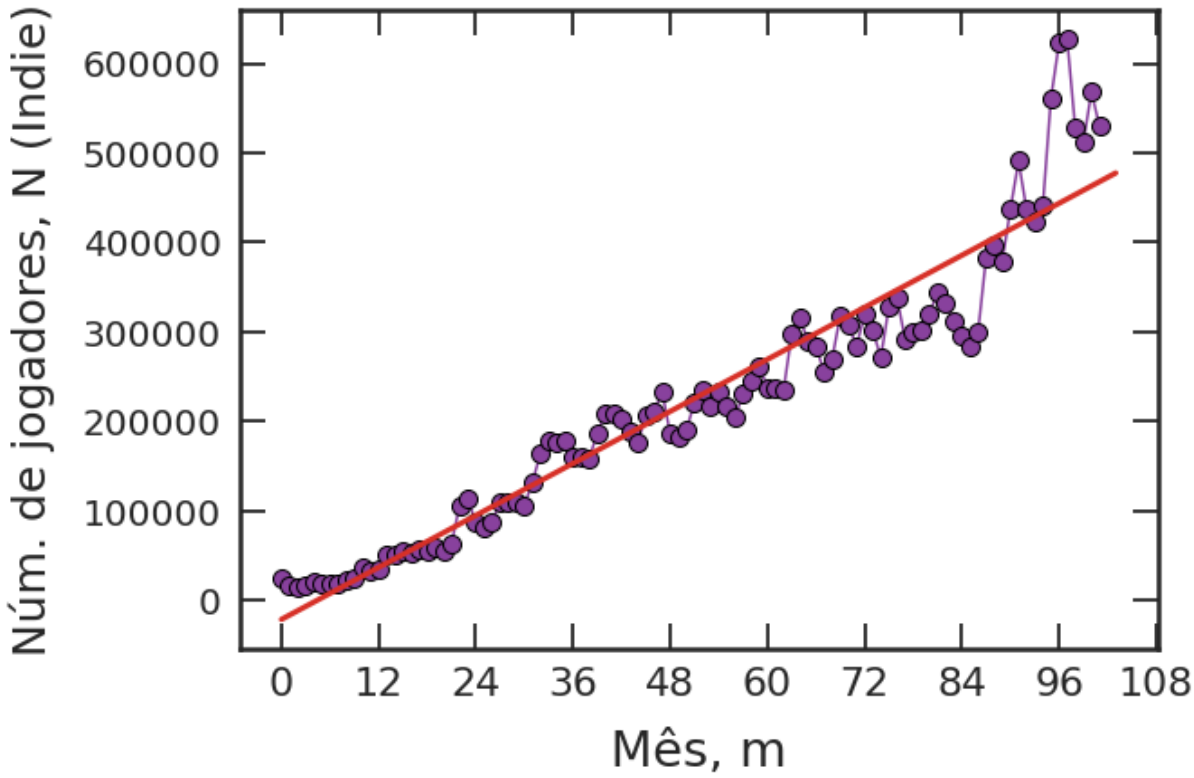

λ_{opt} (grupo 04/2019)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
4
5
Trimestres

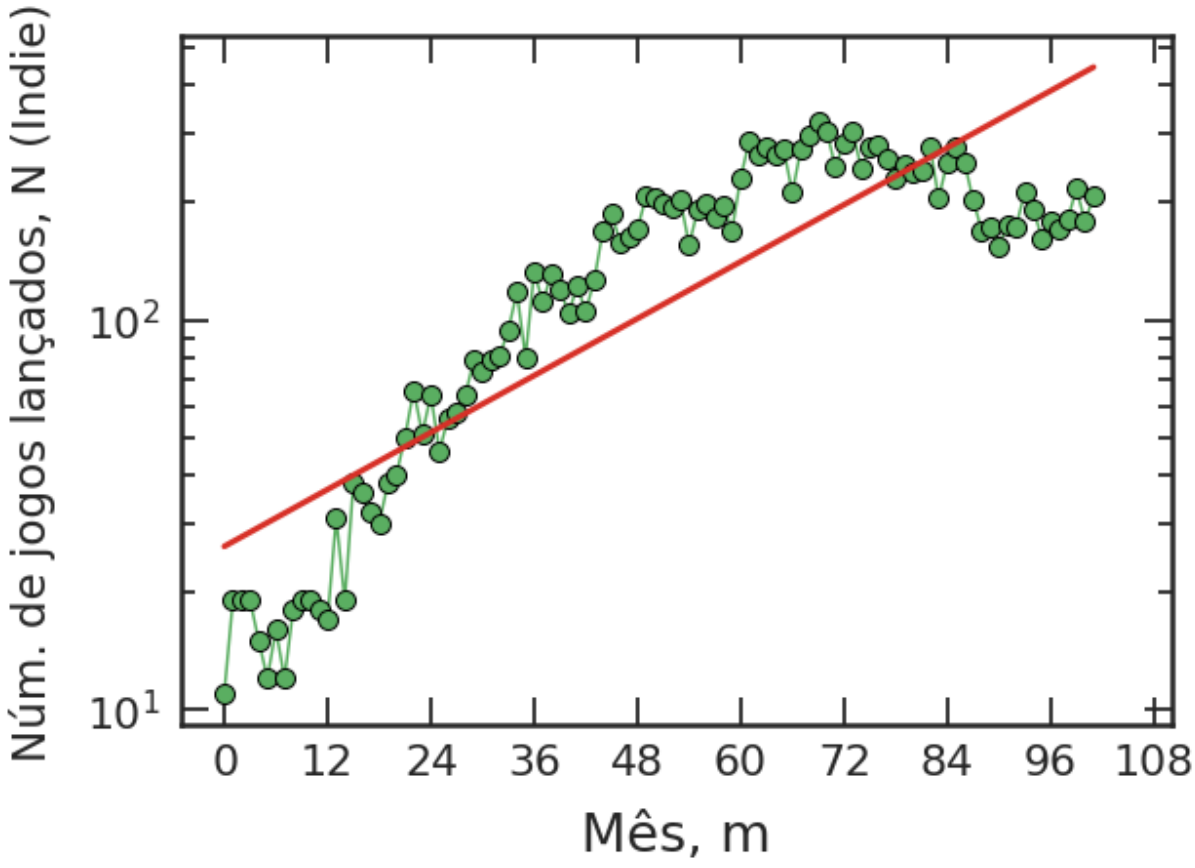

λ_{opt} (grupo 01/2020)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
4
Trimestres

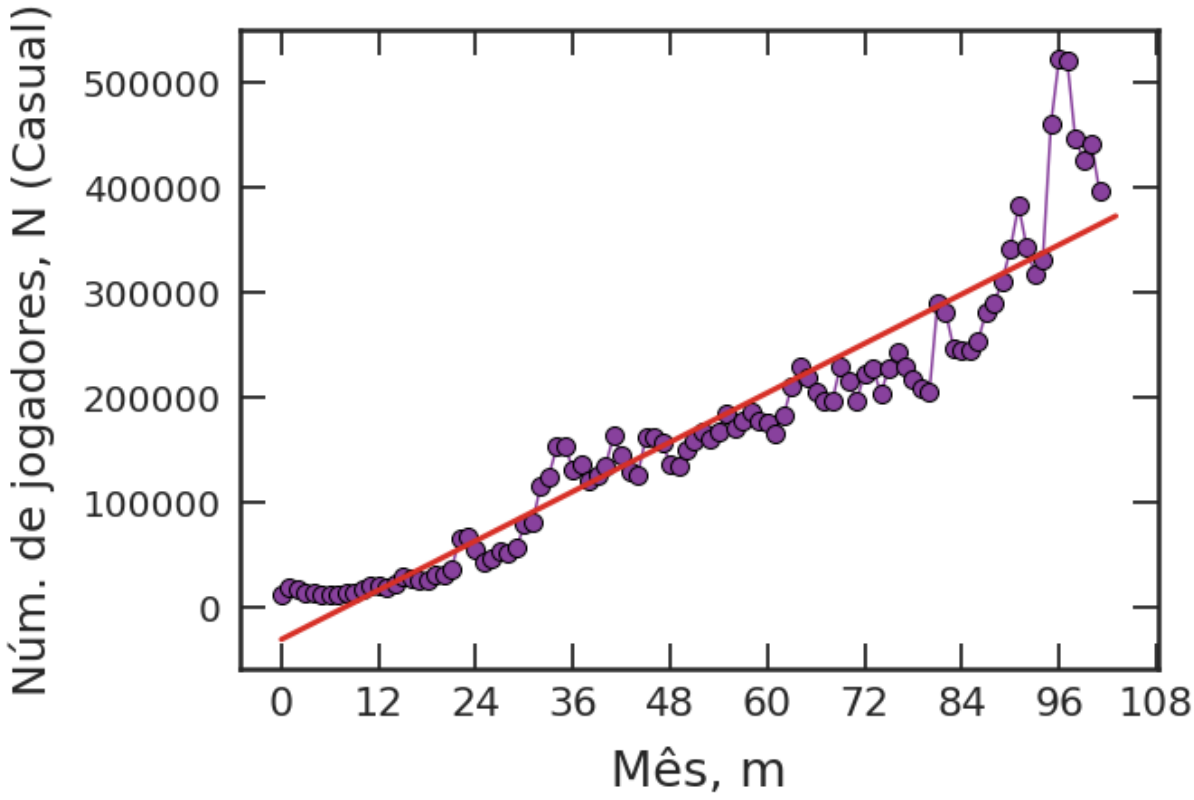

λ_{opt} (grupo 02/2020)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
Trimestres

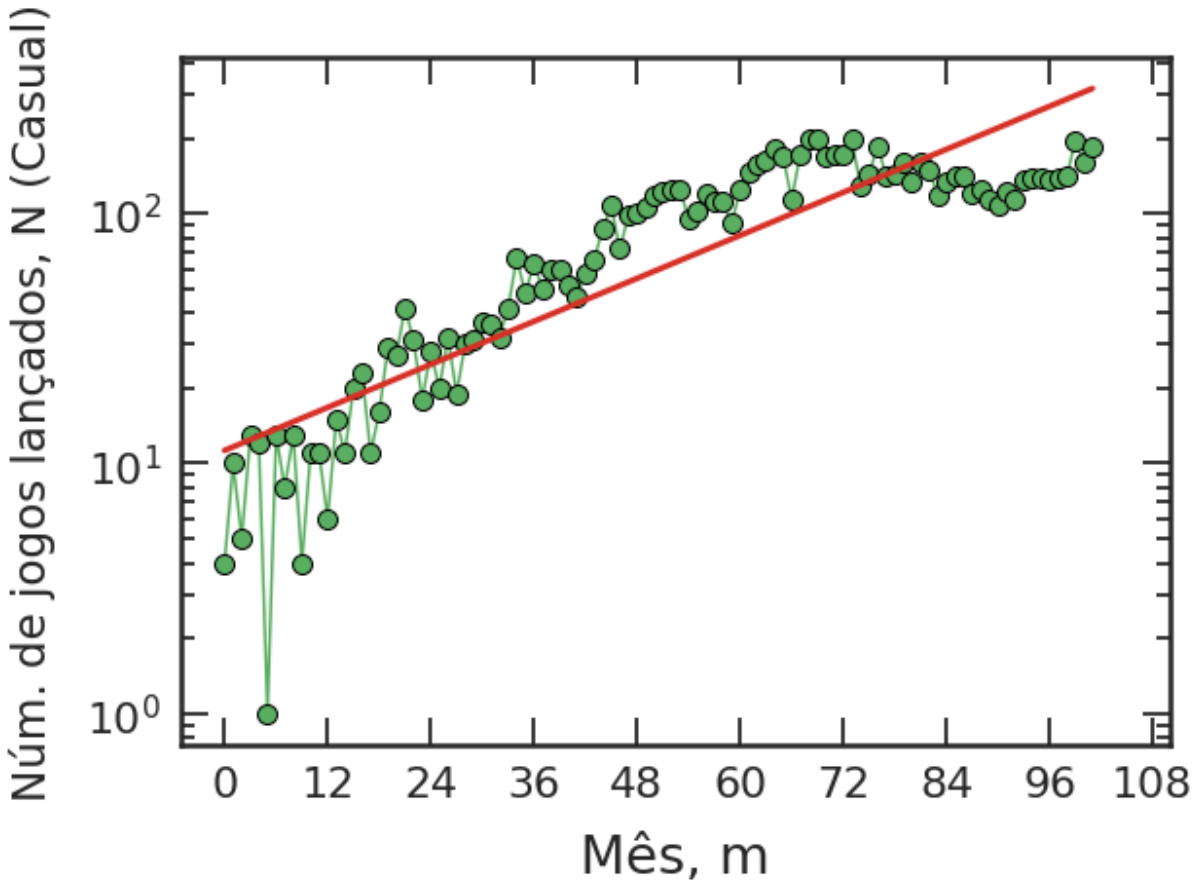
APÊNDICE D – Padrões de crescimento e funções de distribuição das categorias principais

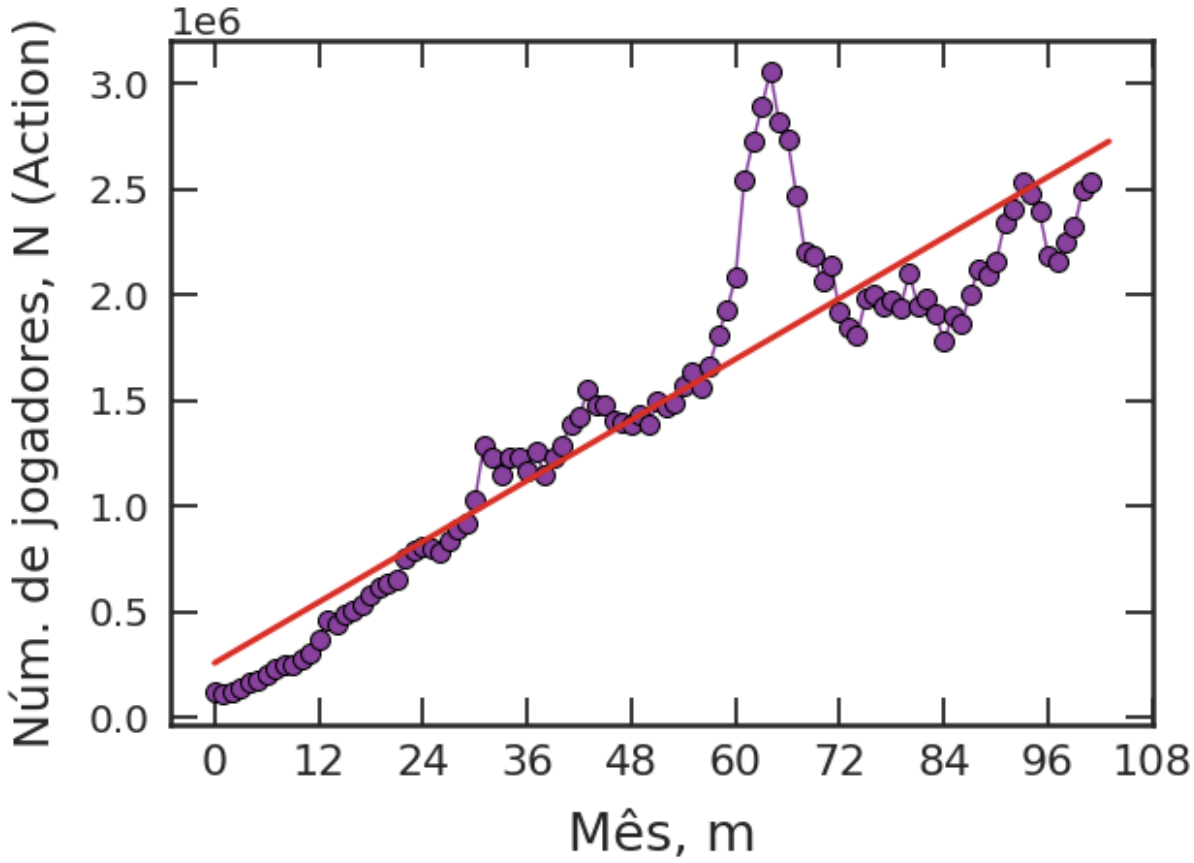
Aqui, apresentaremos os padrões de crescimento das doze categorias principais analisadas neste trabalho (vide capítulo 4), assim como suas funções de distribuição em conjunto tanto com o ajuste lognormal quanto com o Box-Cox.

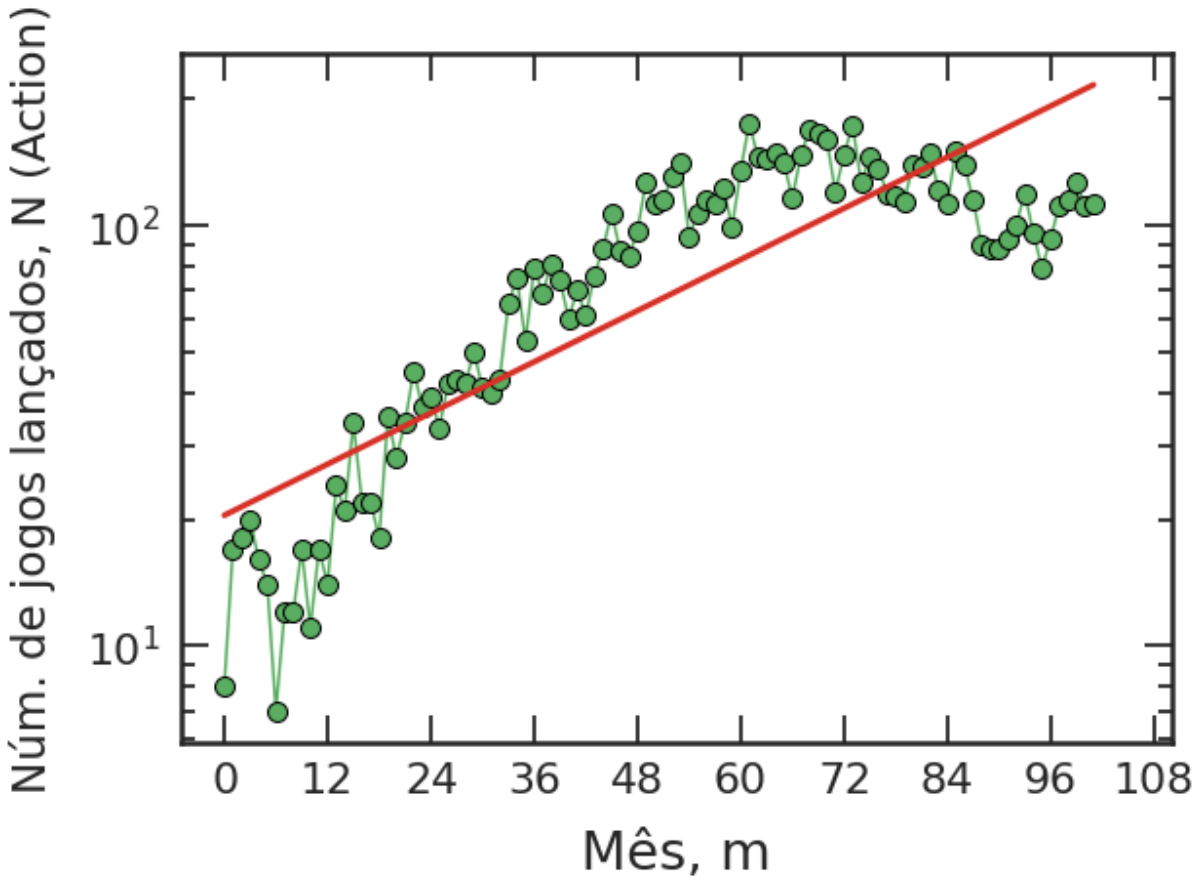

Núm. de jogadores, N (Indie)
600000
500000
400000
300000
200000
100000
0
0
12
24
36
48
60
72
84
96
108
Mês, m

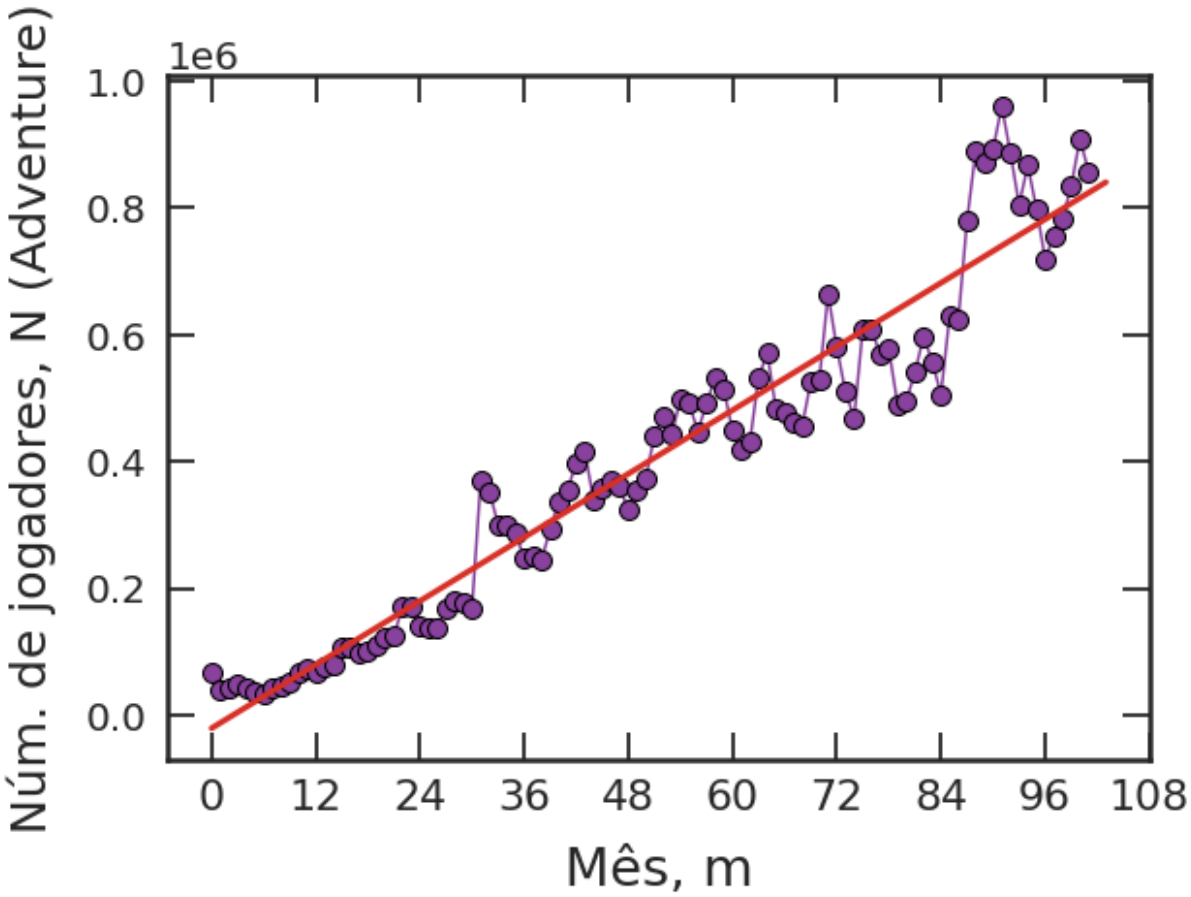

Núm. de jogos lançados, N (Indie)
10^2
10^1
0
12
24
36
48
60
72
84
96
108
Mês, m

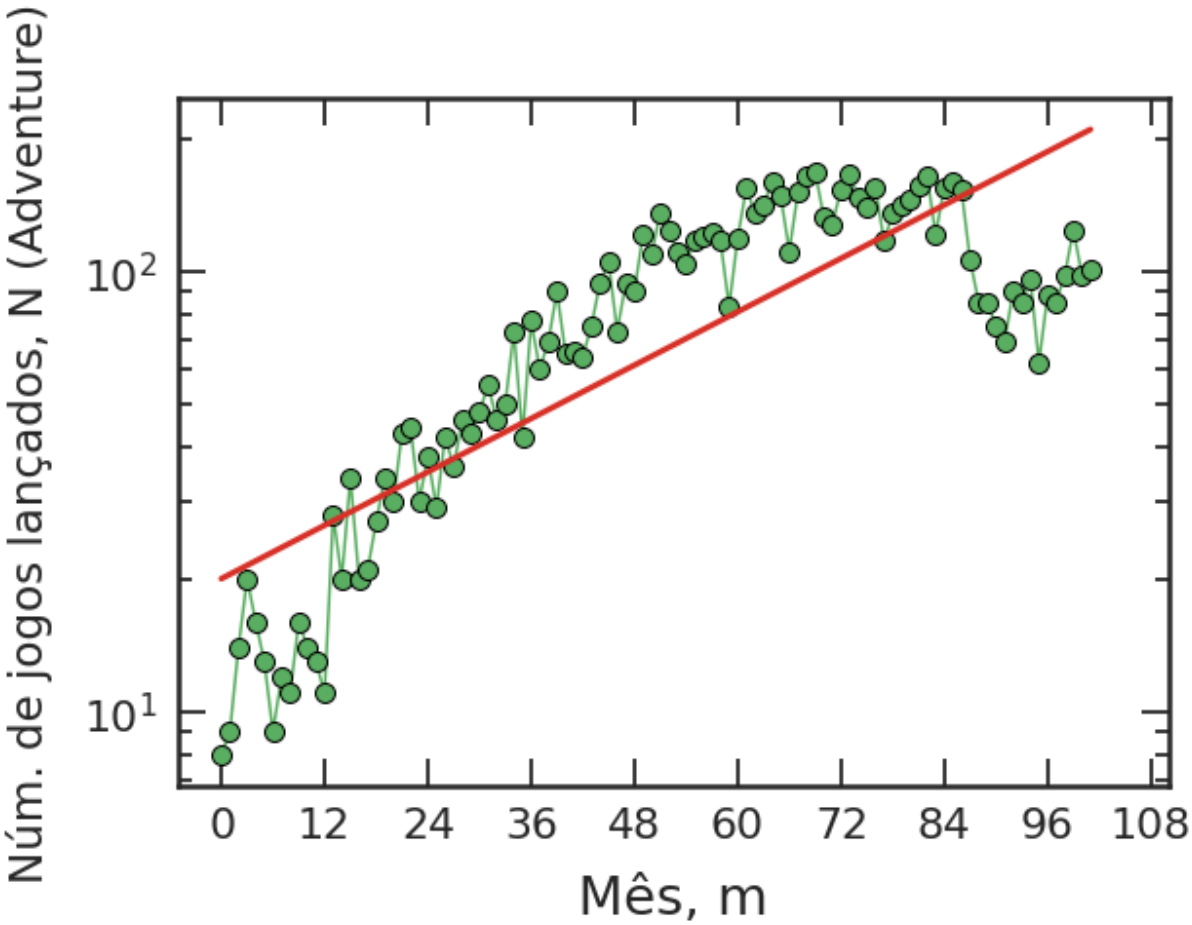

Núm. de jogadores, N (Casual)
500000
400000
300000
200000
100000
0
0
12
24
36
48
60
72
84
96
108
Mês, m

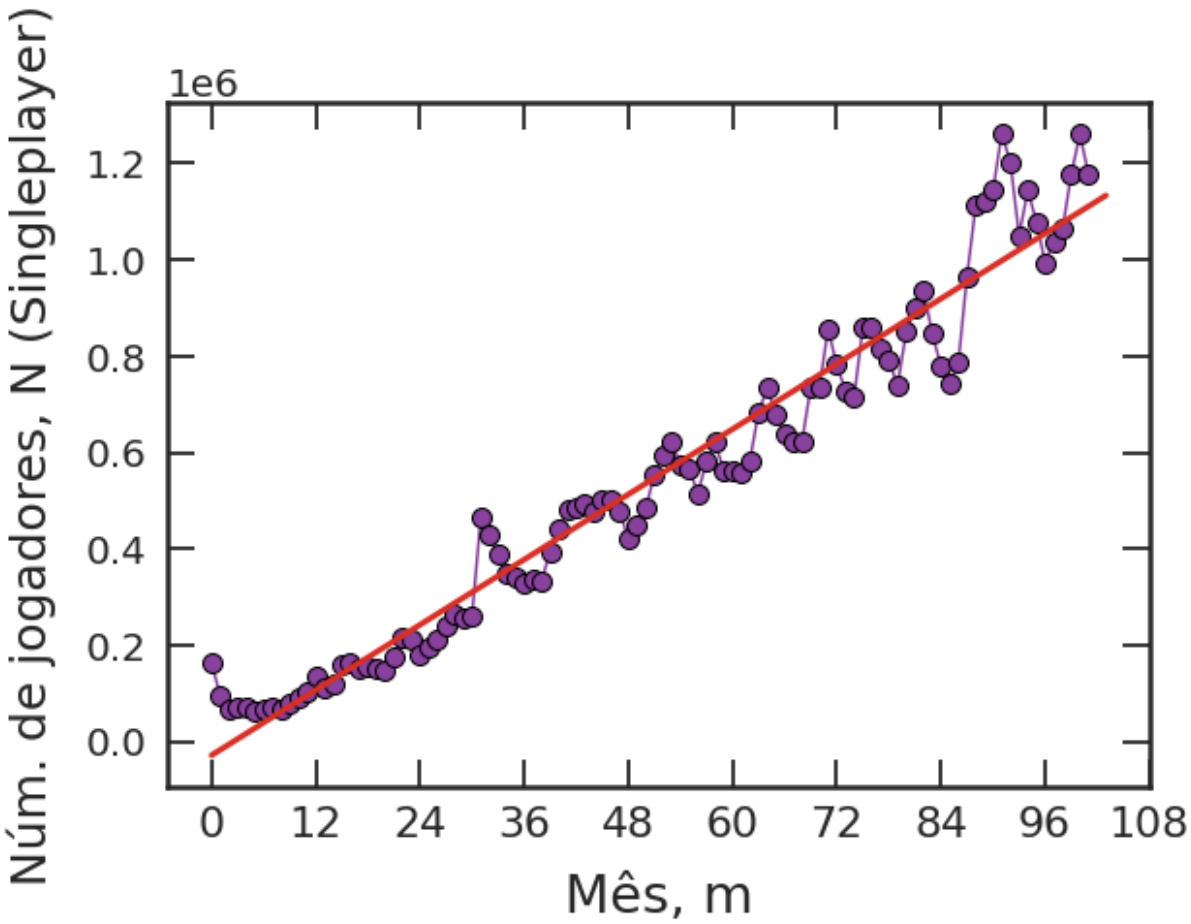

Núm. de jogos lançados, N (Casual)
10^2
10^1
10^0
0
12
24
36
48
60
72
84
96
108
Mês, m

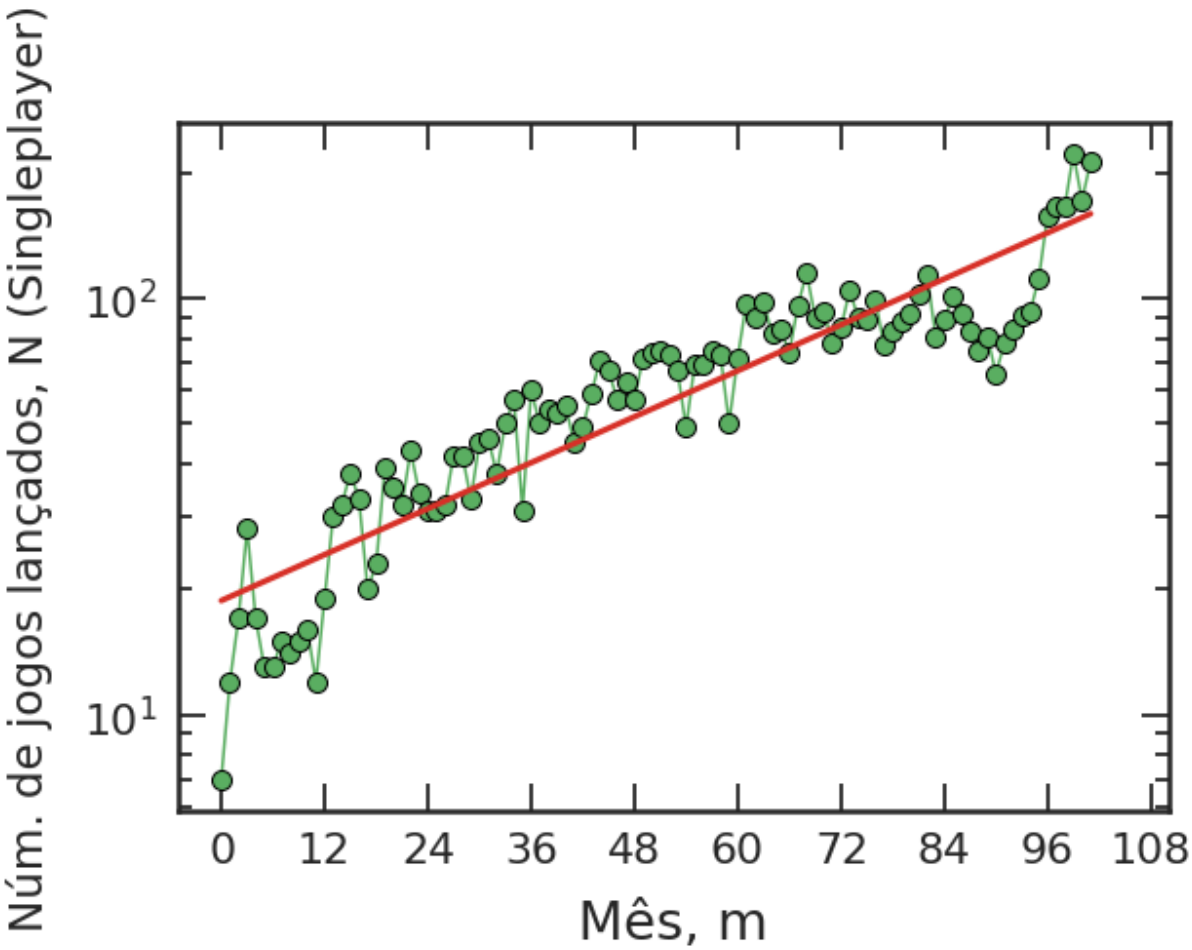

1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Núm. de jogadores, N (Action)
0
12
24
36
48
60
72
84
96
108
Mês, m

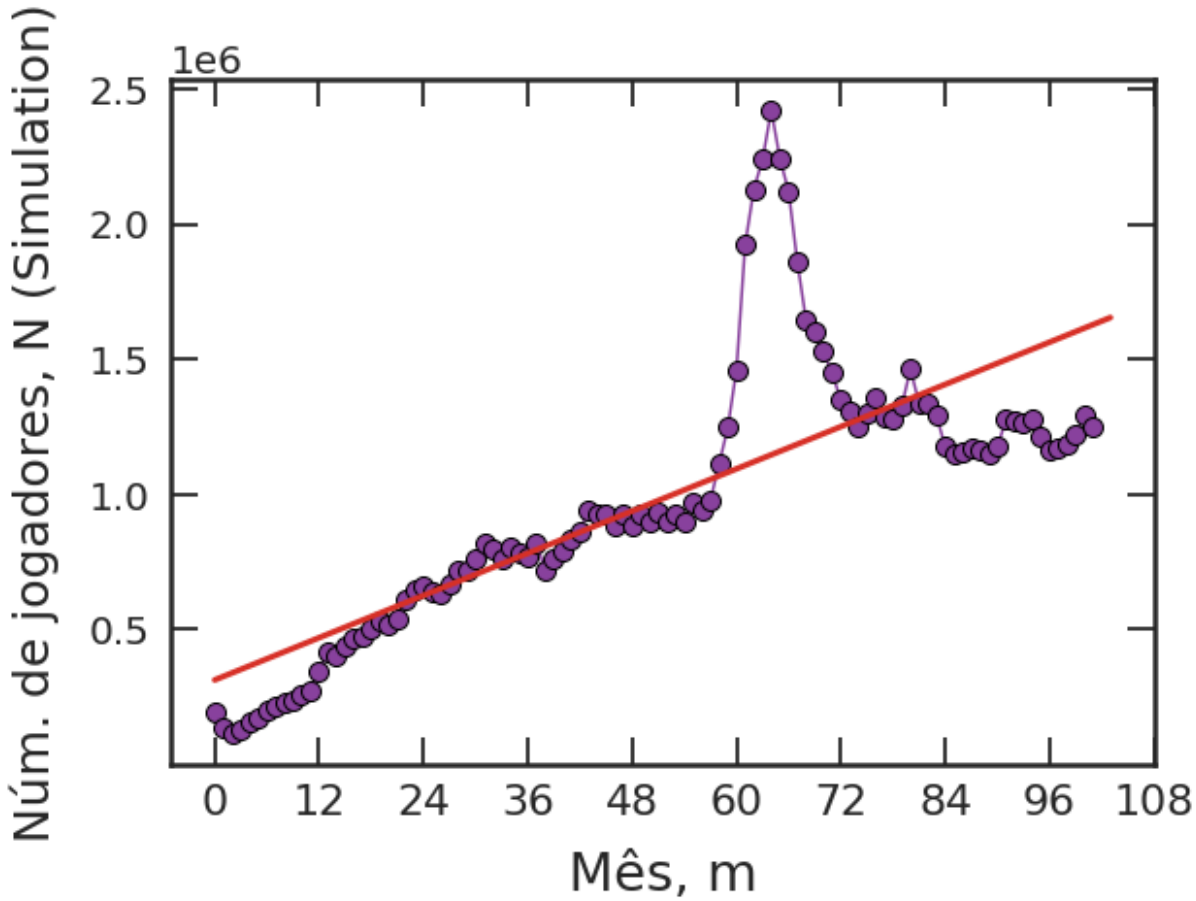

10^2
10^1
Núm. de jogos lançados, N (Action)
0
12
24
36
48
60
72
84
96
108
Mês, m

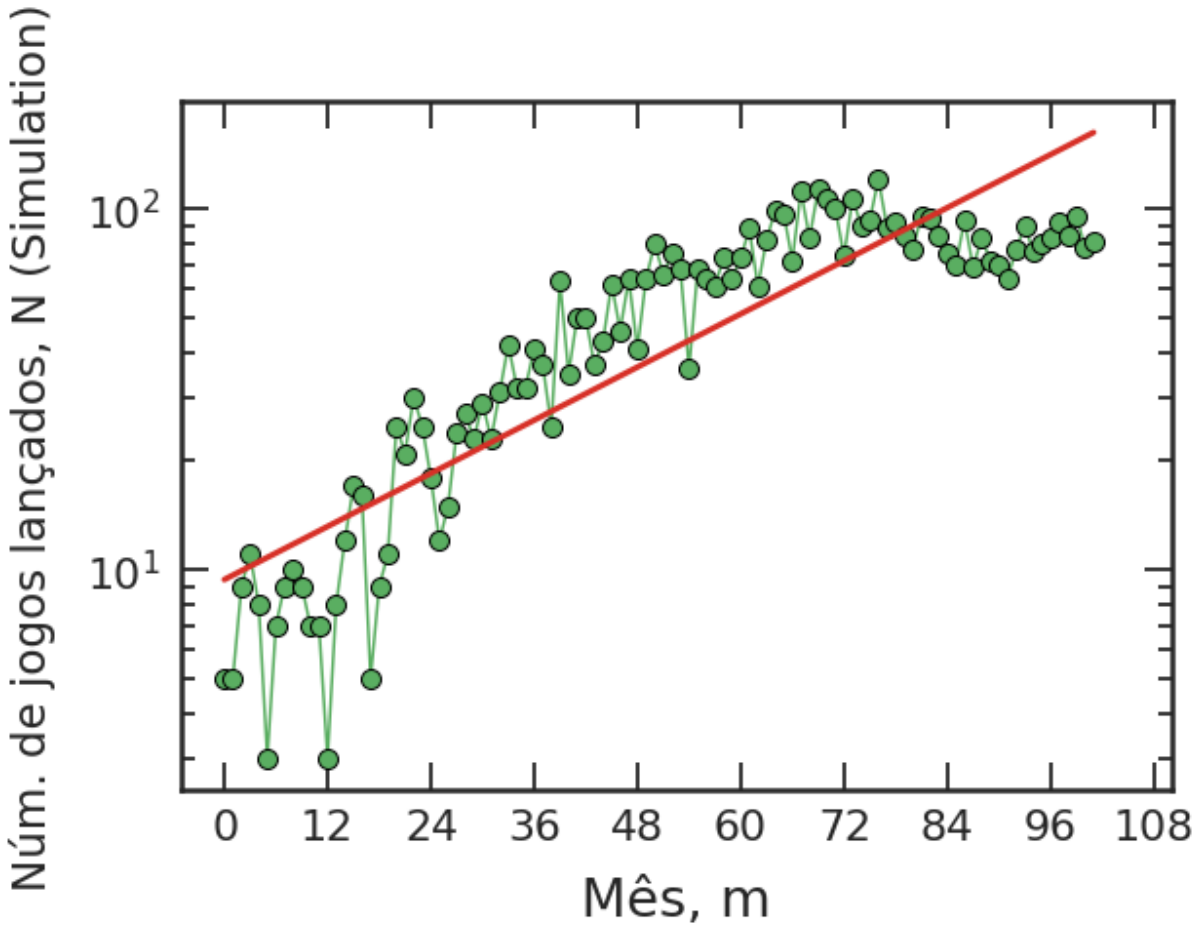

1e6
1.0
0.8
0.6
0.4
0.2
0.0
Núm. de jogadores, N (Adventure)
0
12
24
36
48
60
72
84
96
108
Mês, m

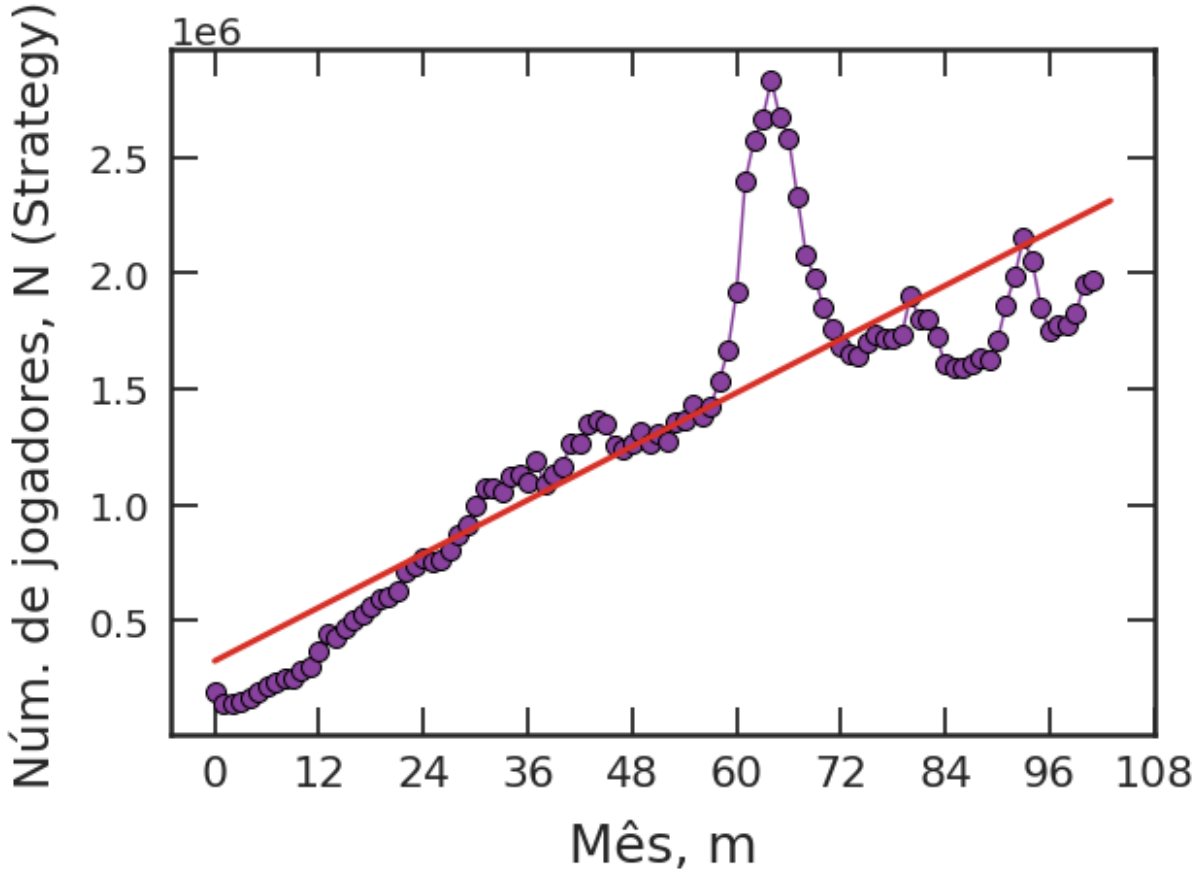

10^2
10^1
Núm. de jogos lançados, N (Adventure)
0
12
24
36
48
60
72
84
96
108
Mês, m

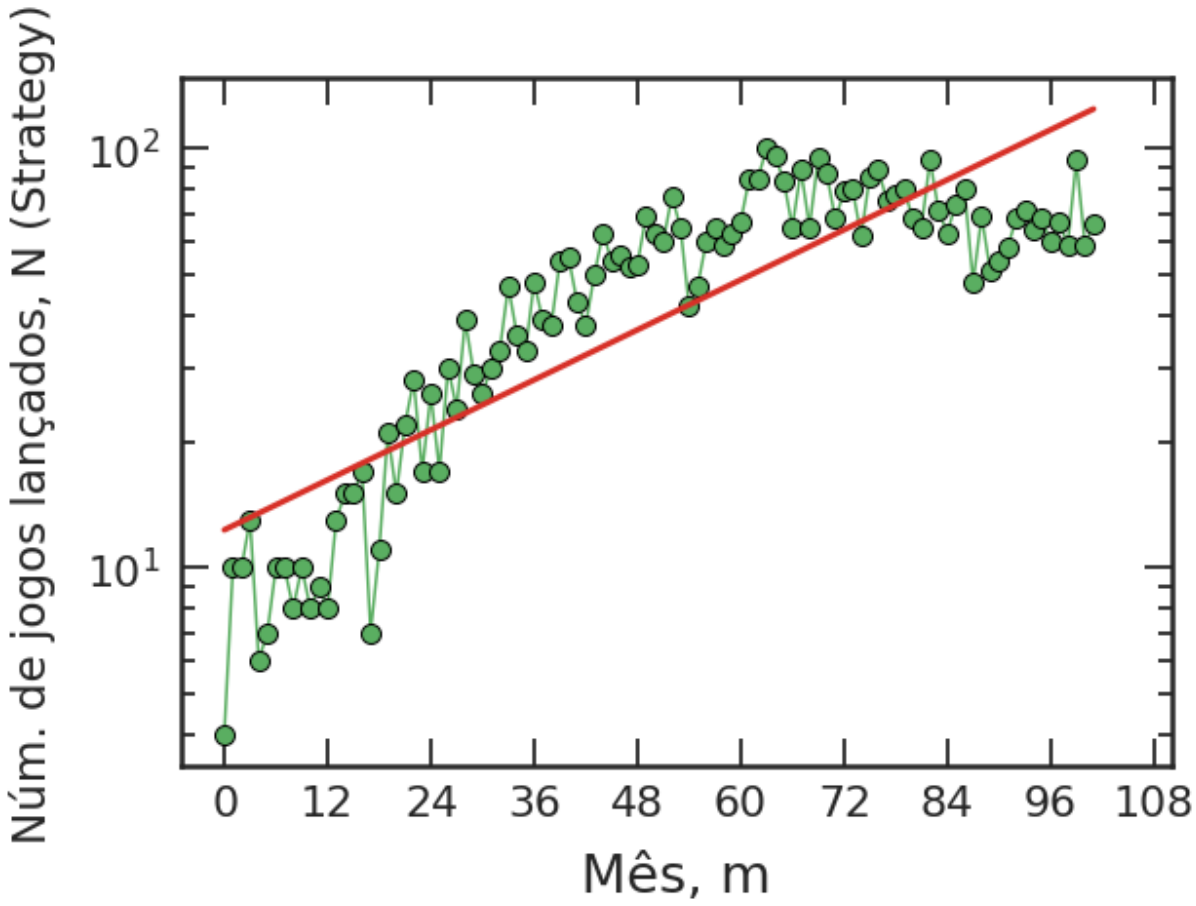

1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Núm. de jogadores, N (Singleplayer)
0
12
24
36
48
60
72
84
96
108
Mês, m

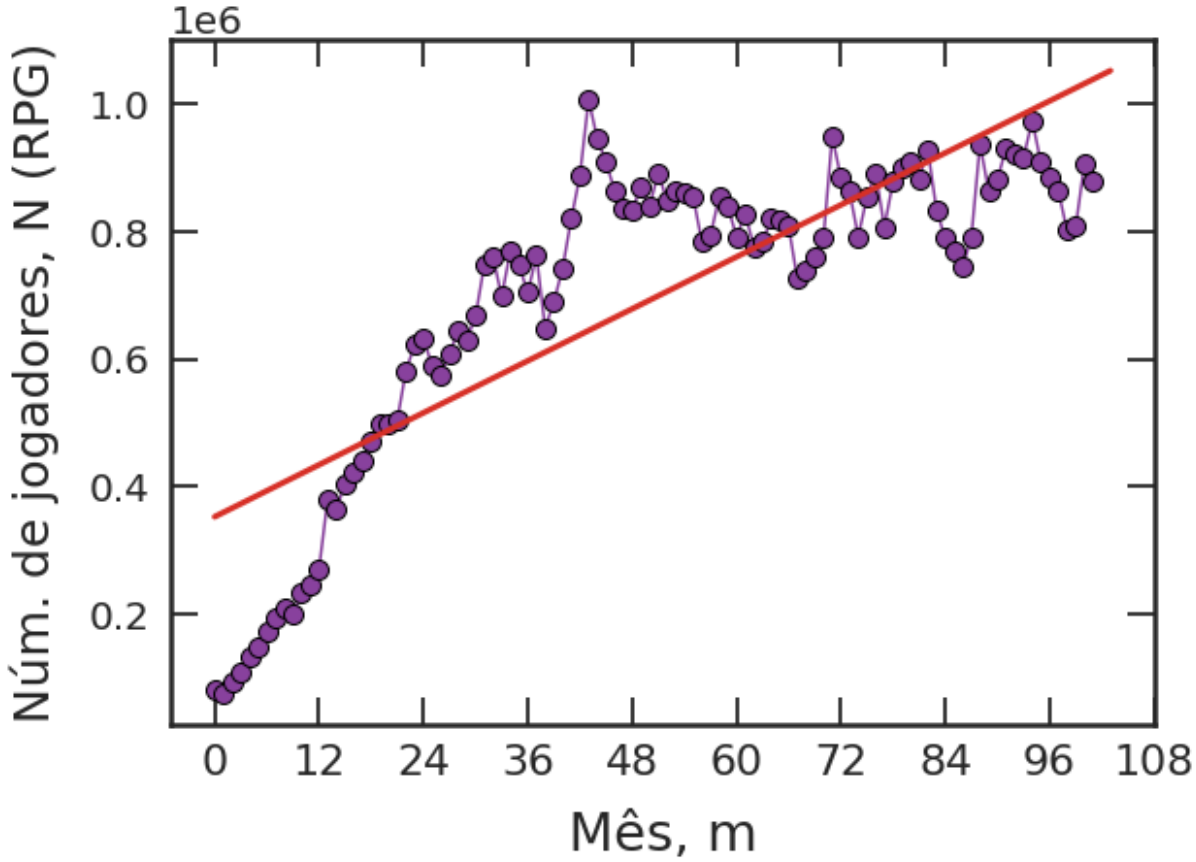

10^2
10^1
Núm. de jogos lançados, N (Singleplayer)
0
12
24
36
48
60
72
84
96
108
Mês, m

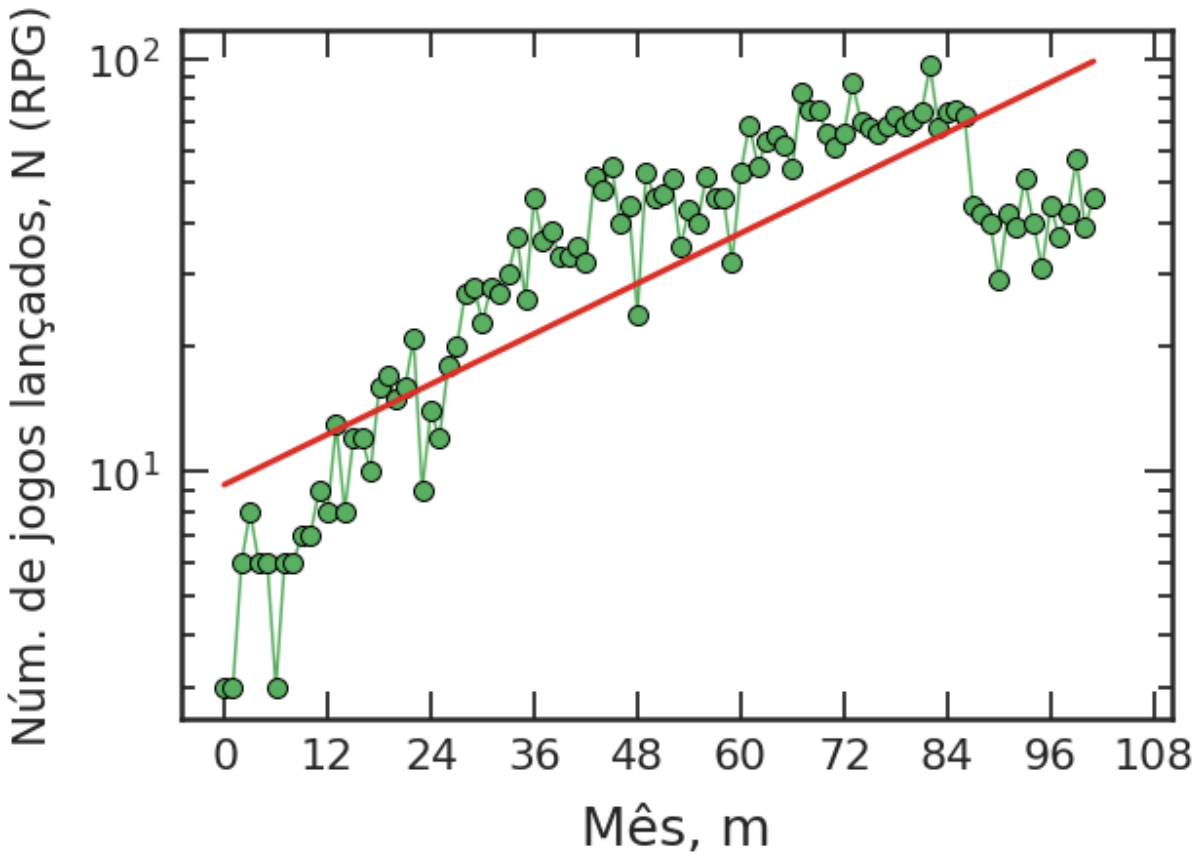

1e6
2.5
2.0
1.5
1.0
0.5
Núm. de jogadores, N (Simulation)
0
12
24
36
48
60
72
84
96
108
Mês, m

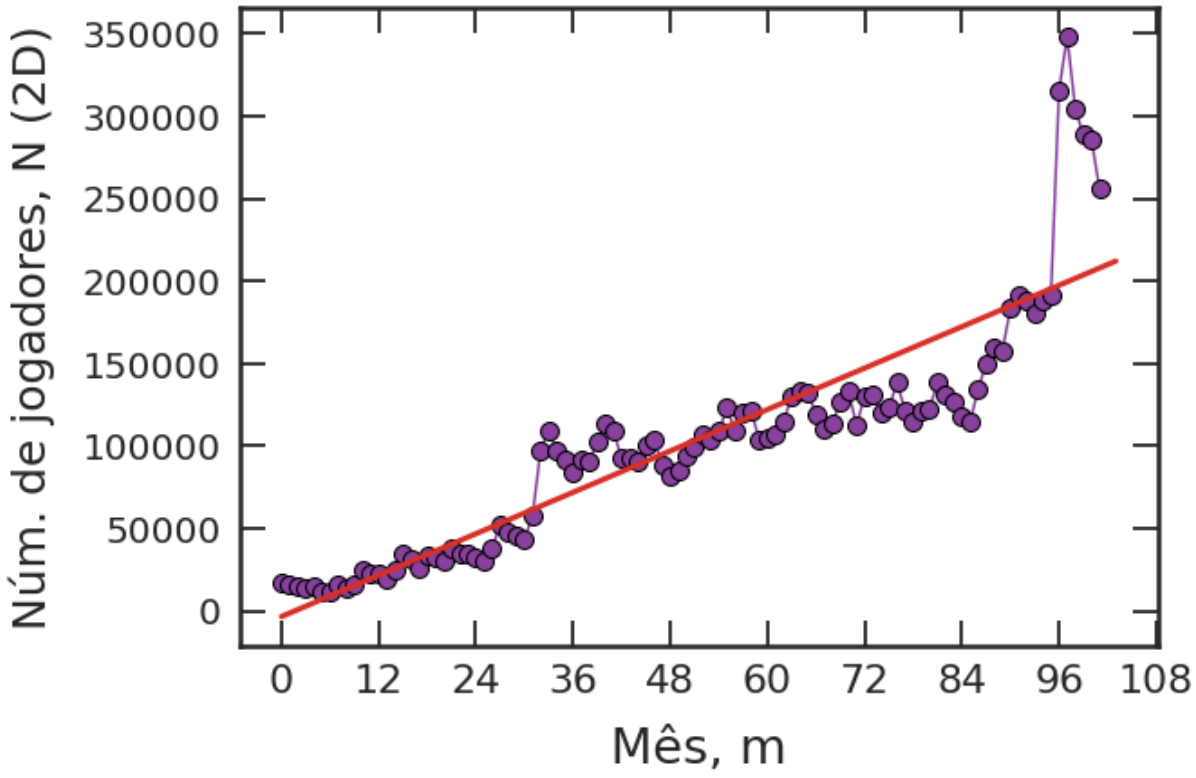

10^2
10^1
Núm. de jogos lançados, N (Simulation)
0
12
24
36
48
60
72
84
96
108
Mês, m

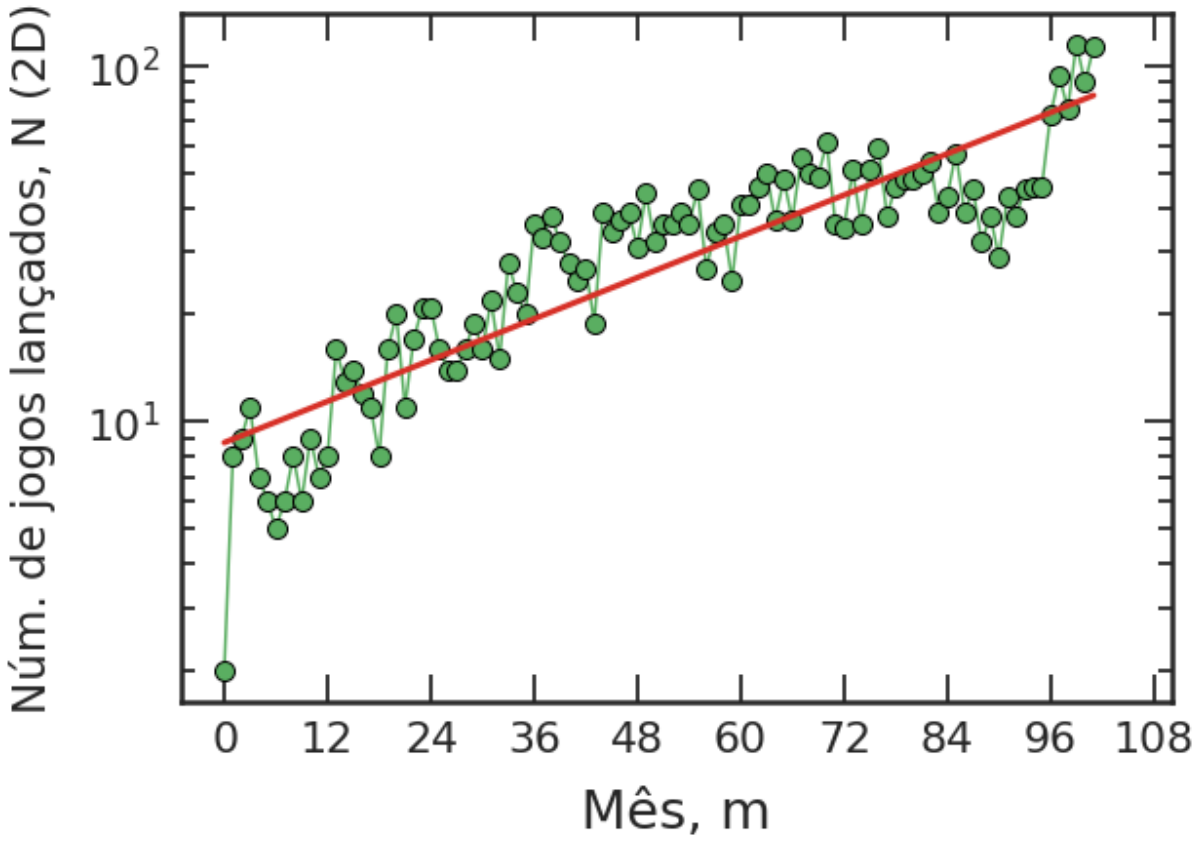

1e6
2.5
2.0
1.5
1.0
0.5
Núm. de jogadores, N (Strategy)
0
12
24
36
48
60
72
84
96
108
Mês, m

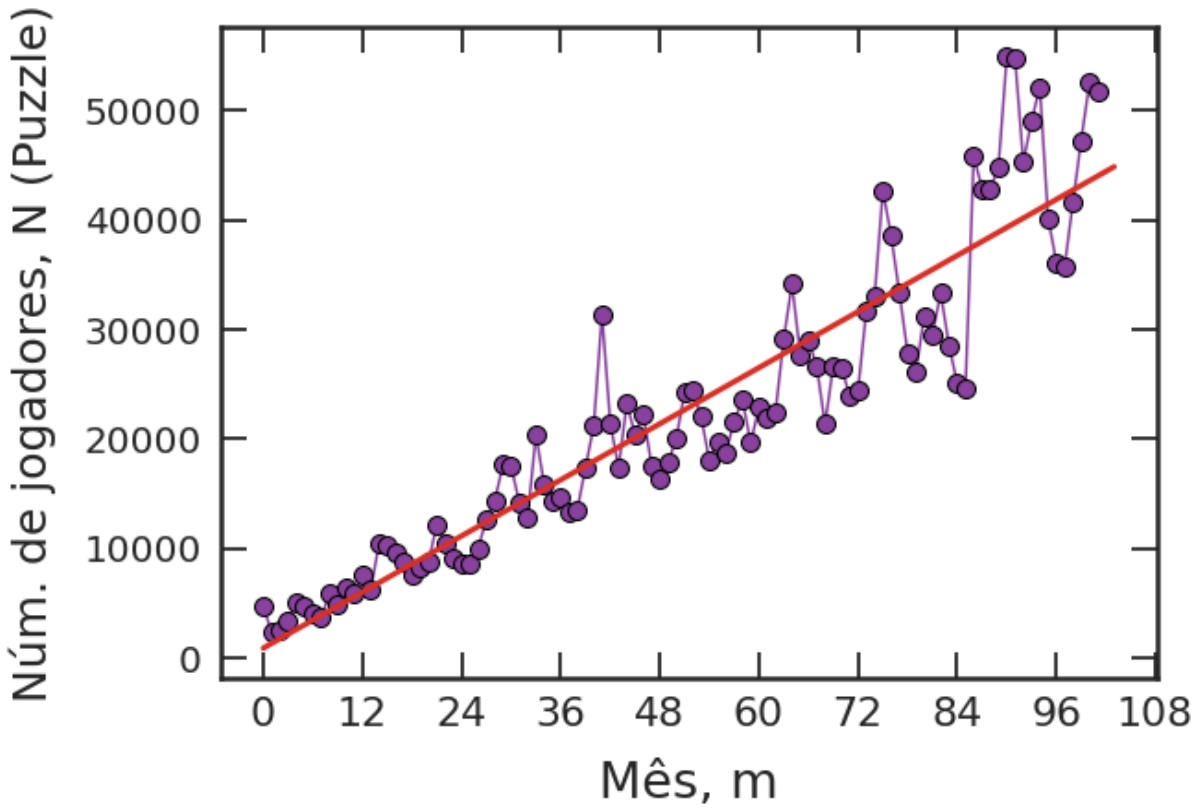

10^2
10^1
Núm. de jogos lançados, N (Strategy)
0
12
24
36
48
60
72
84
96
108
Mês, m

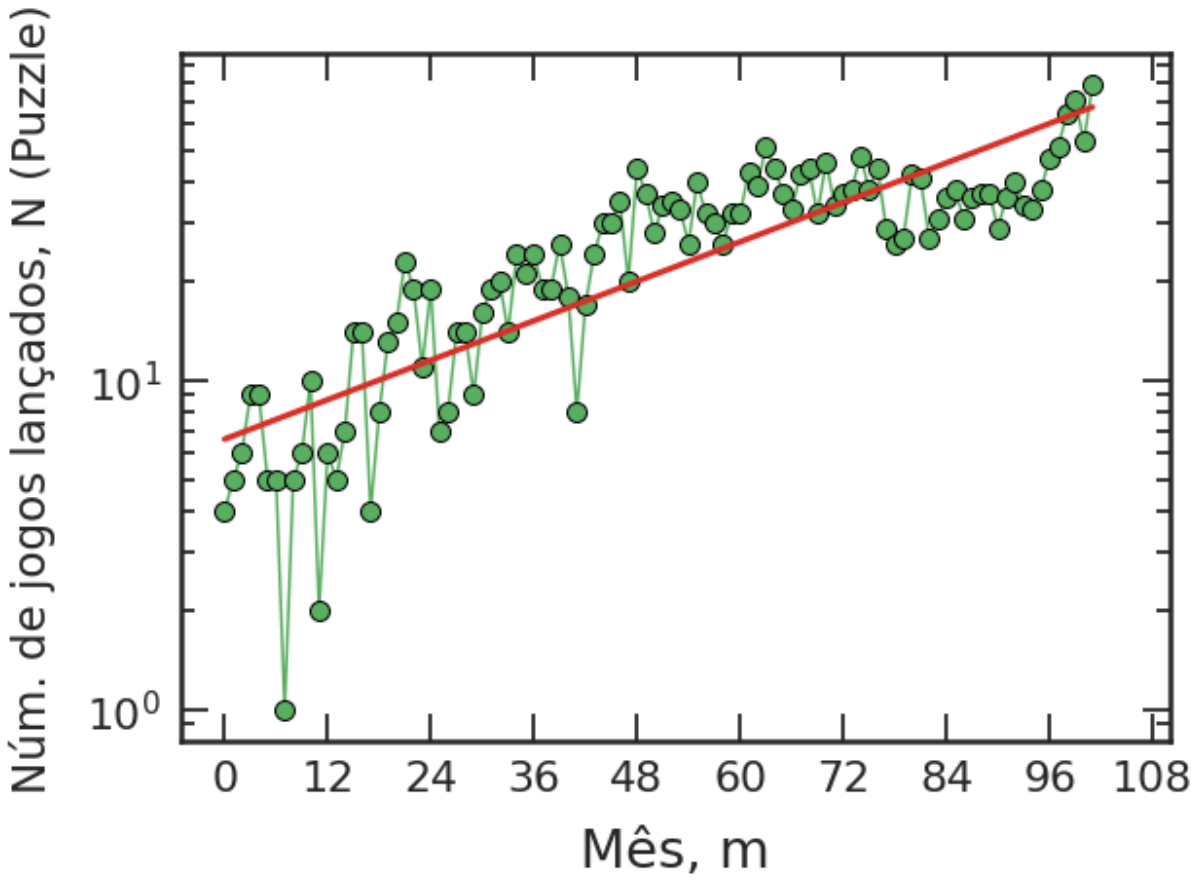

1e6
1.0
0.8
0.6
0.4
0.2
Núm. de jogadores, N (RPG)
0
12
24
36
48
60
72
84
96
108
Mês, m

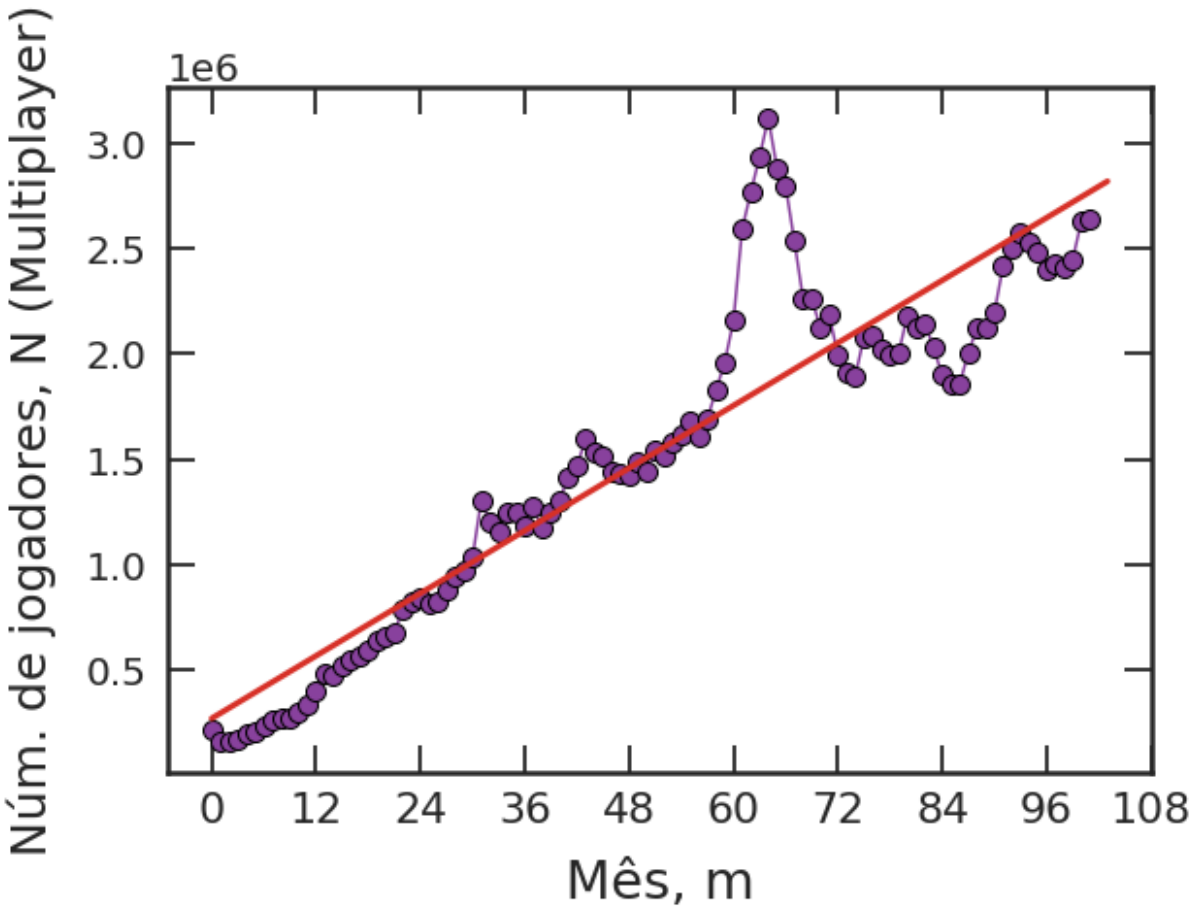

10^2
10^1
Núm. de jogos lançados, N (RPG)
0
12
24
36
48
60
72
84
96
108
Mês, m

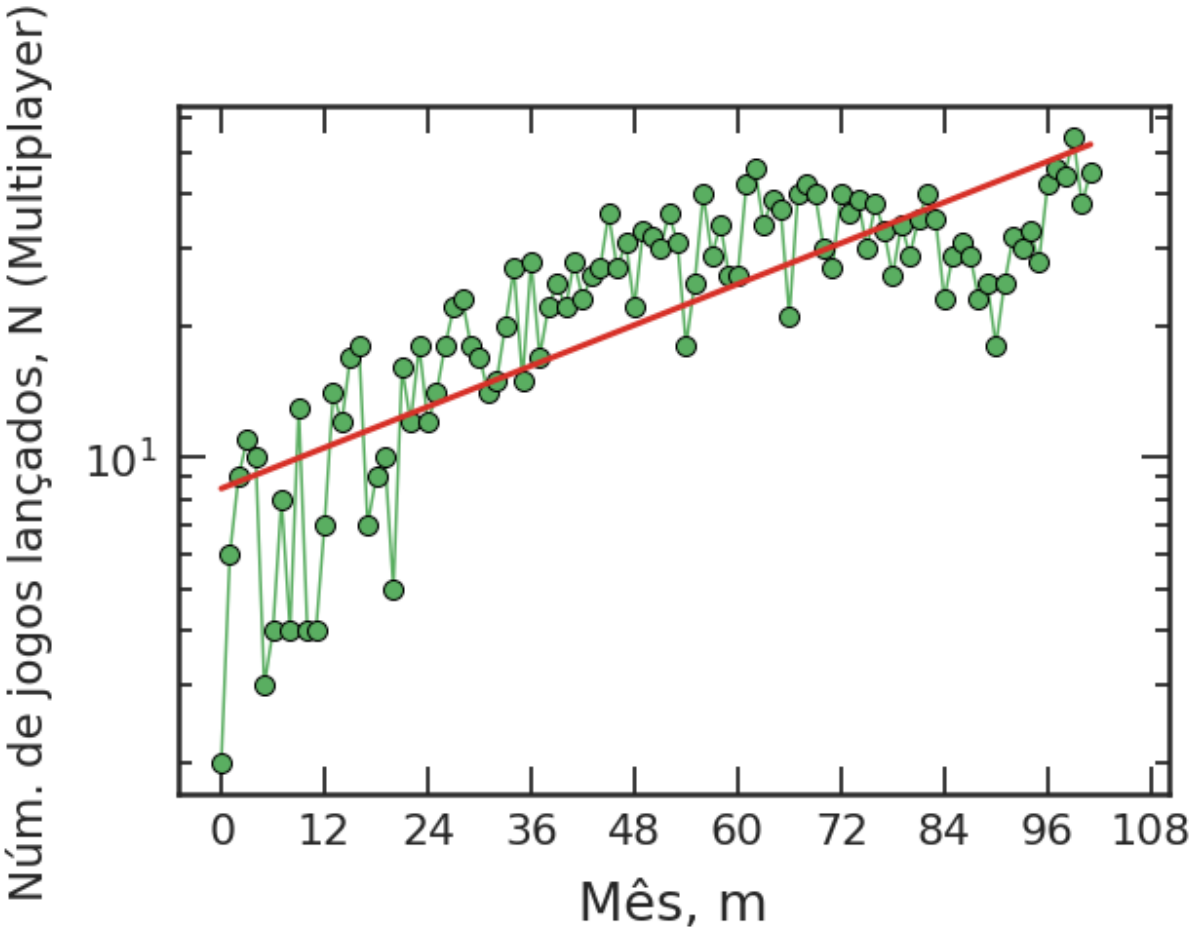

Núm. de jogadores, N (2D)
350000
300000
250000
200000
150000
100000
50000
0
0
12
24
36
48
60
72
84
96
108
Mês, m

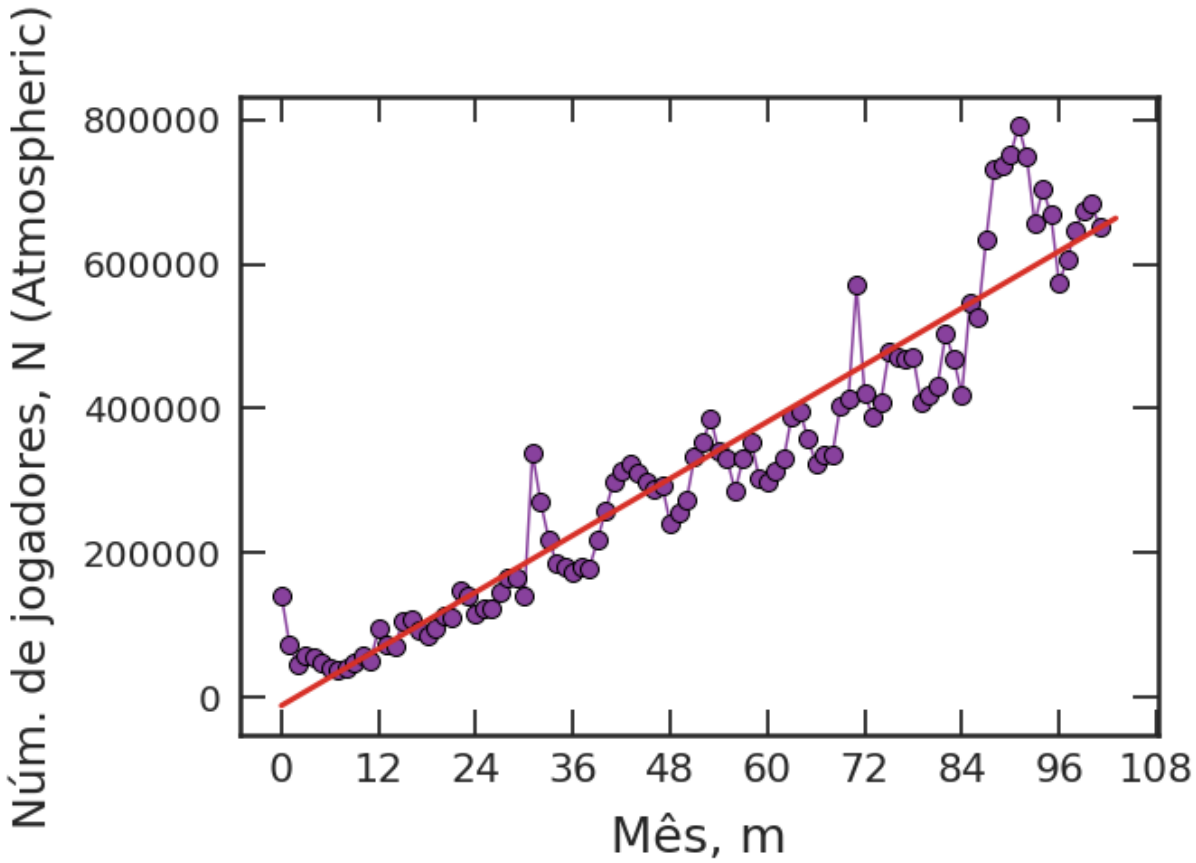

Núm. de jogos lançados, N (2D)
10^2
10^1
0
12
24
36
48
60
72
84
96
108
Mês, m

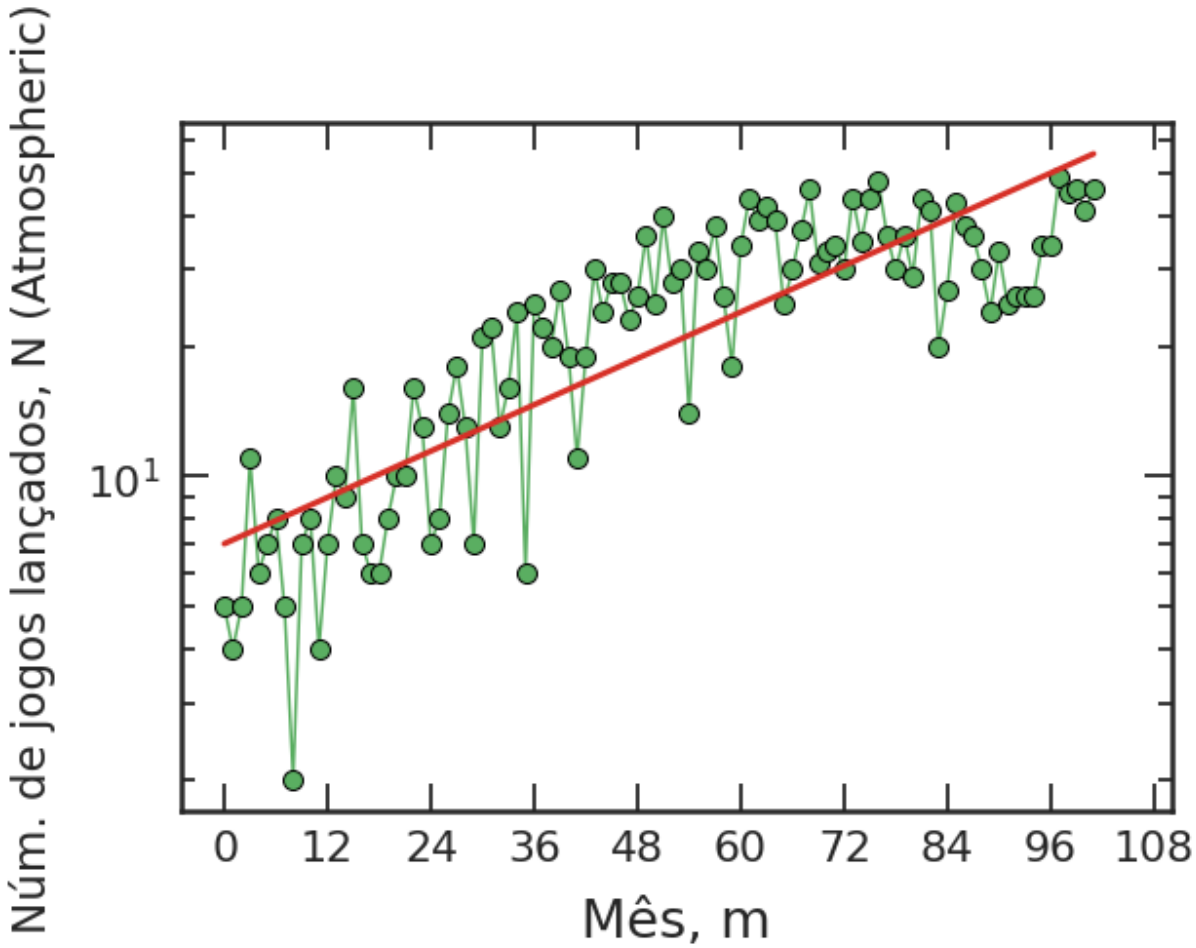

Núm. de jogadores, N (Puzzle)
50000
40000
30000
20000
10000
0
0
12
24
36
48
60
72
84
96
108
Mês, m

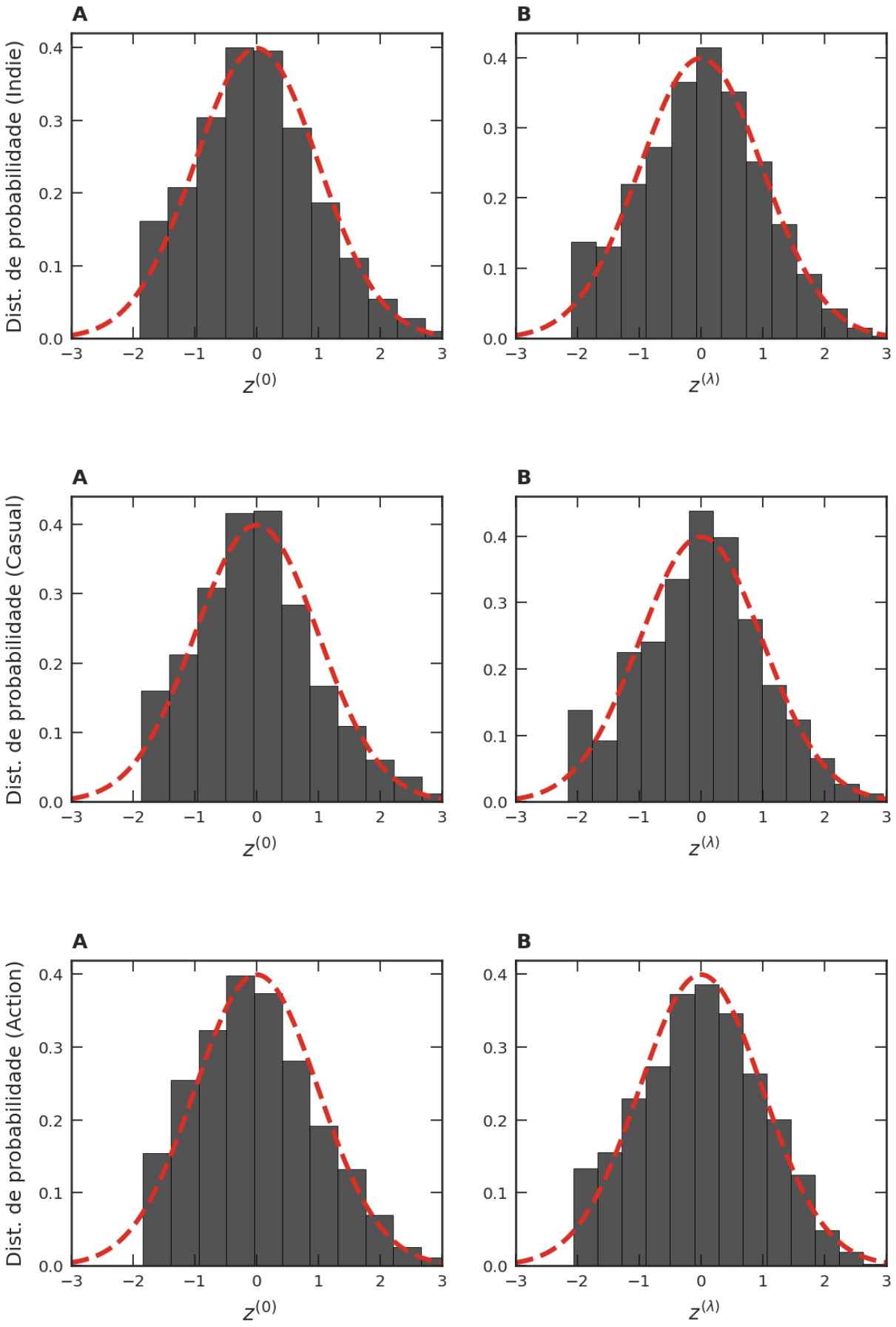

Núm. de jogos lançados, N (Puzzle)
10^1
10^0
0
12
24
36
48
60
72
84
96
108
Mês, m

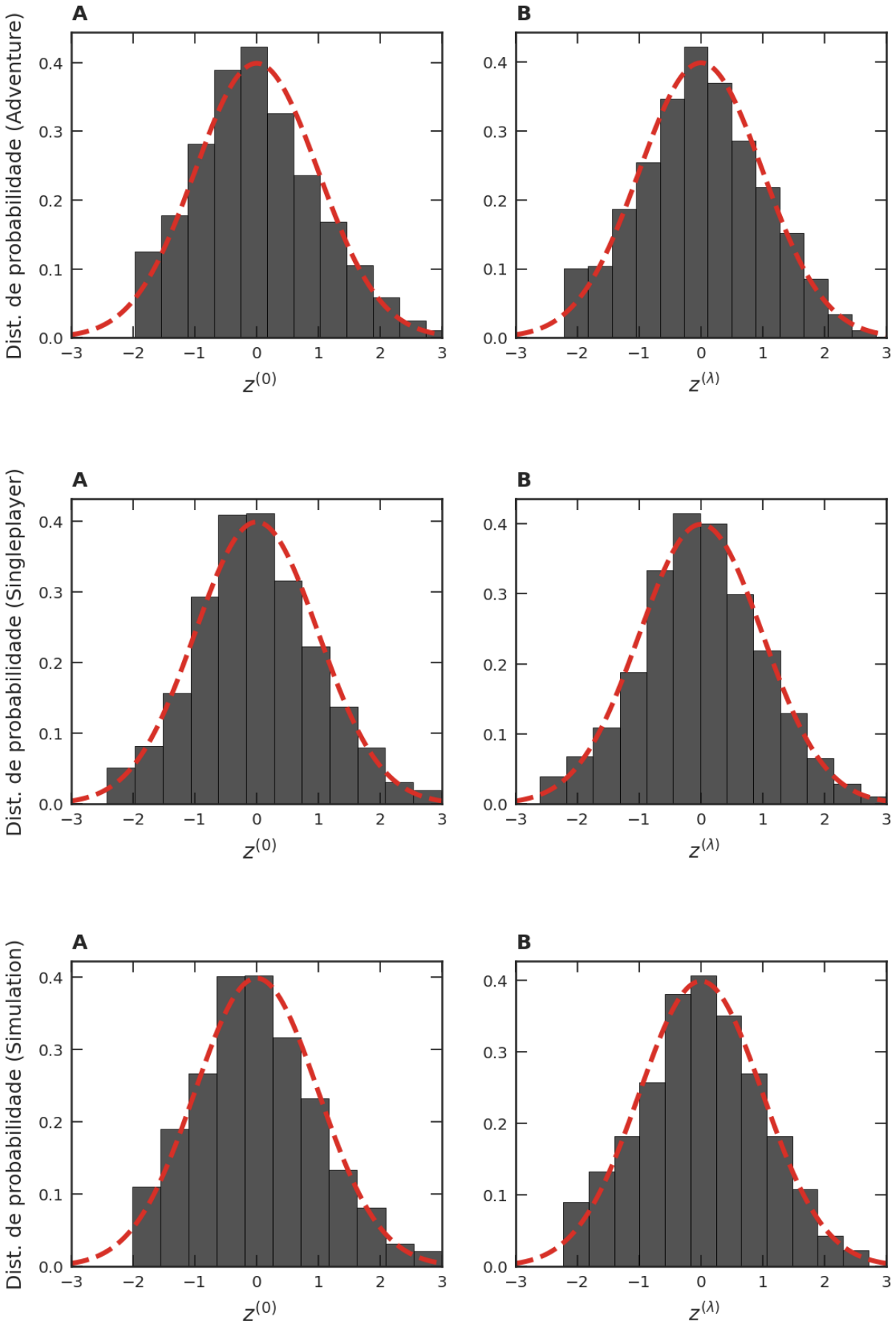

1e6
3.0
2.5
2.0
1.5
1.0
0.5
Núm. de jogadores, N (Multiplayer)
0
12
24
36
48
60
72
84
96
108
Mês, m

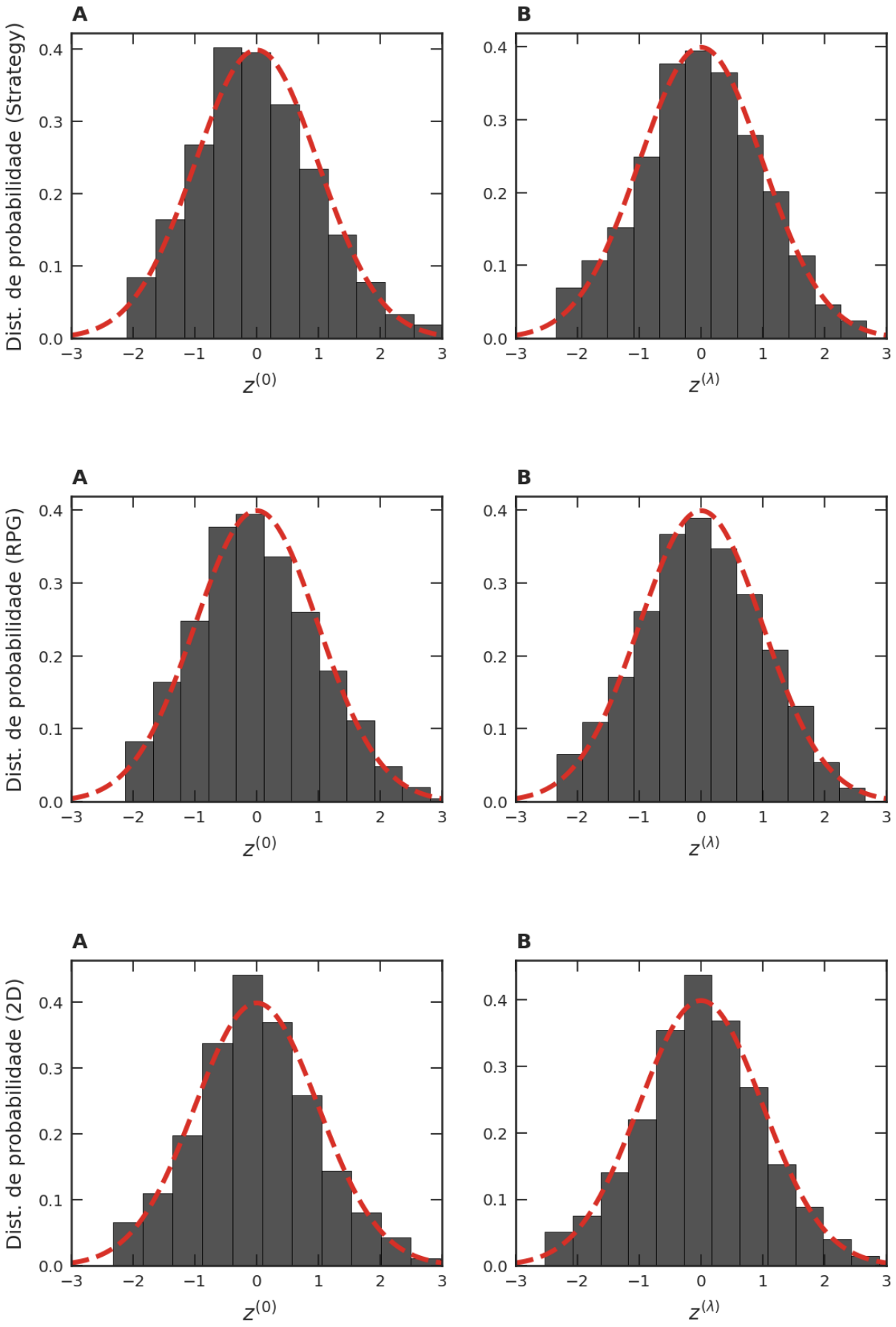

10^1
Núm. de jogos lançados, N (Multiplayer)
0
12
24
36
48
60
72
84
96
108
Mês, m

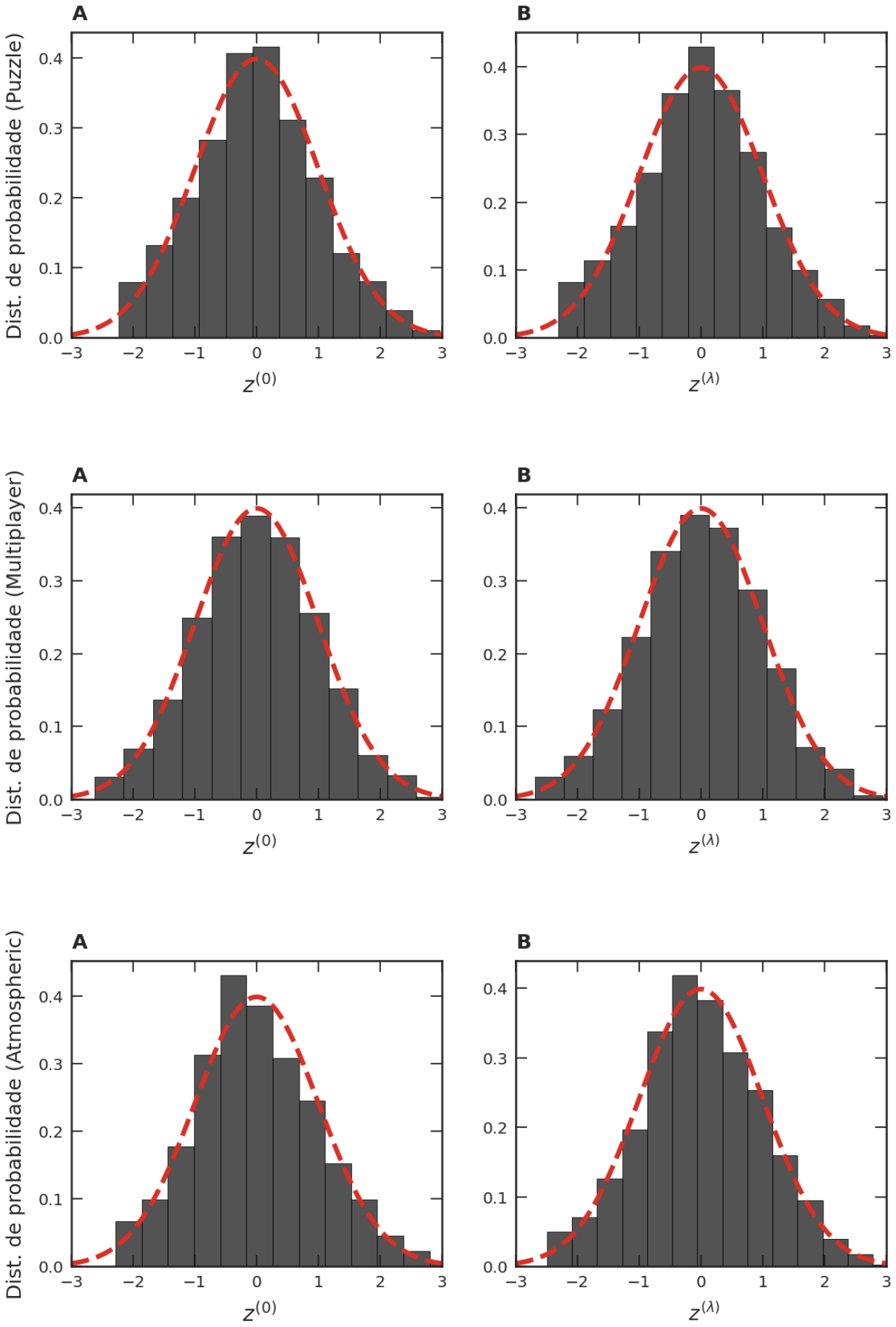

Núm. de jogadores, N (Atmospheric)
800000
600000
400000
200000
0
0
12
24
36
48
60
72
84
96
108
Mês, m


Núm. de jogos lançados, N (Atmospheric)
10^1
0
12
24
36
48
60
72
84
96
108
Mês, m


A
B
Dist. de probabilidade (Indie)
Dist. de probabilidade (Casual)
Dist. de probabilidade (Action)
$z^{(0)}$
$z^{(\lambda)}$

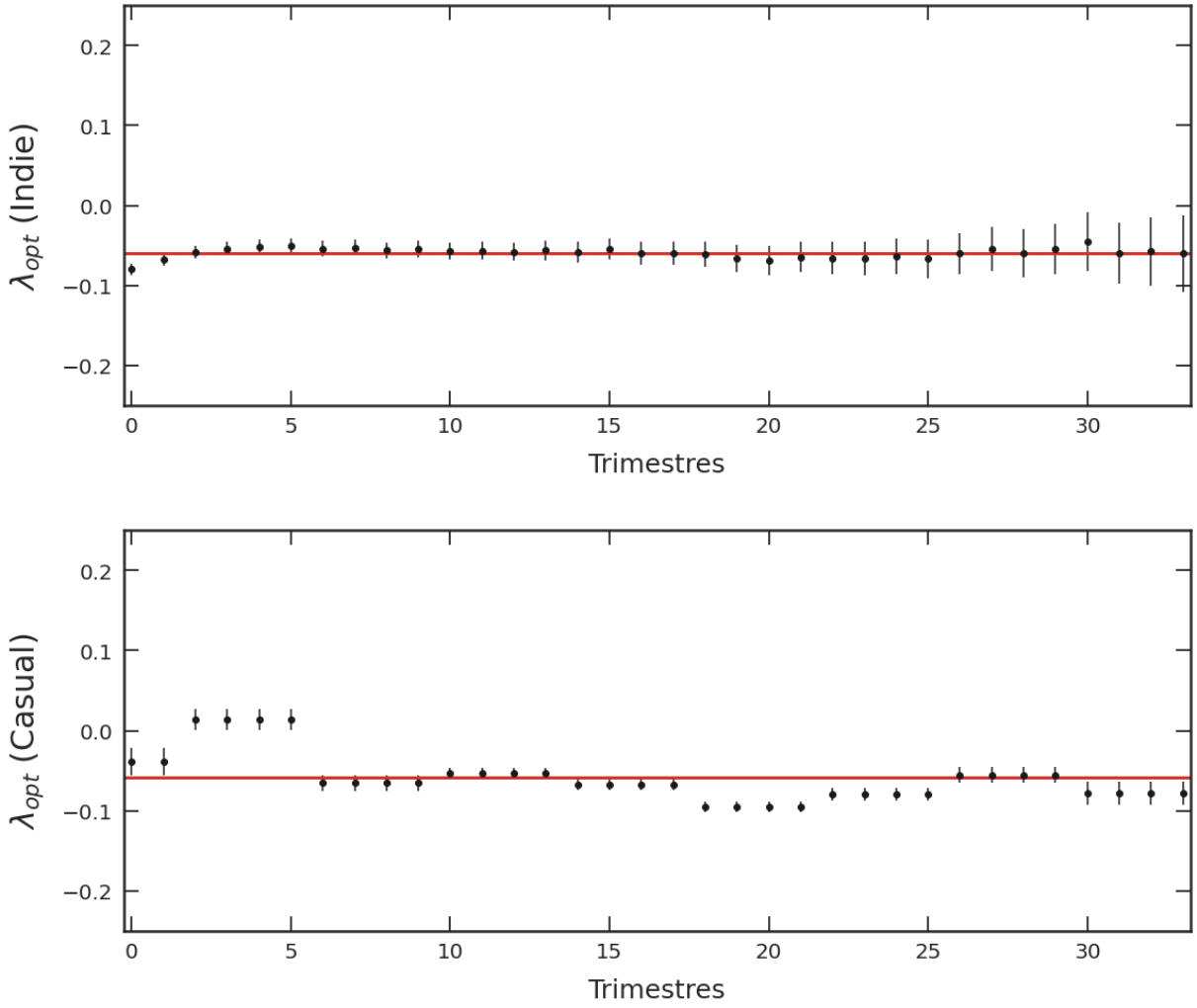

A
B
Dist. de probabilidade (Adventure)
Dist. de probabilidade (Singleplayer)
Dist. de probabilidade (Simulation)
$z^{(0)}$
$z^{(\lambda)}$
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3

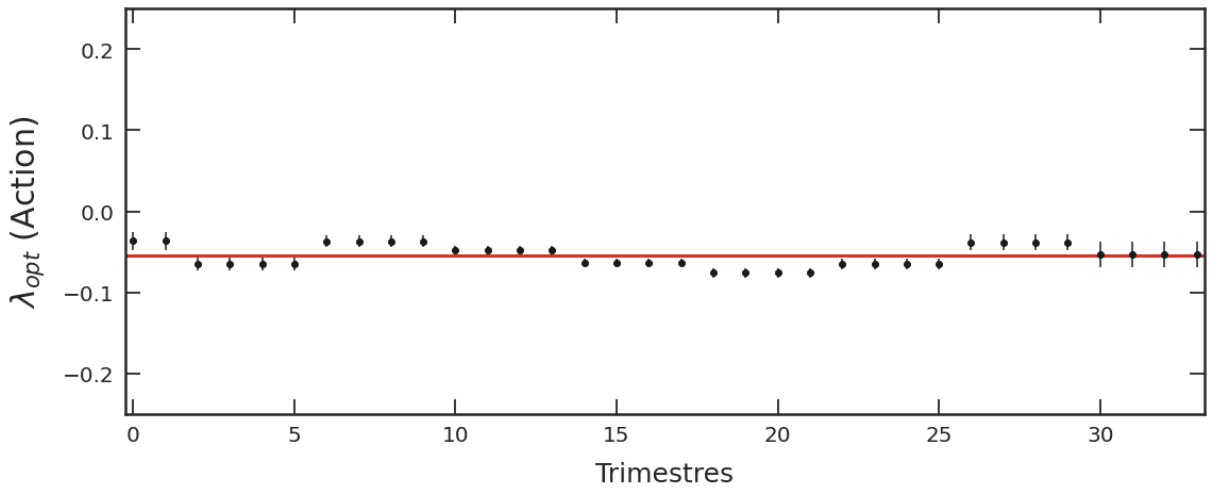

A
B
Dist. de probabilidade (Strategy)
Dist. de probabilidade (RPG)
Dist. de probabilidade (2D)
$z^{(0)}$
$z^{(\lambda)}$

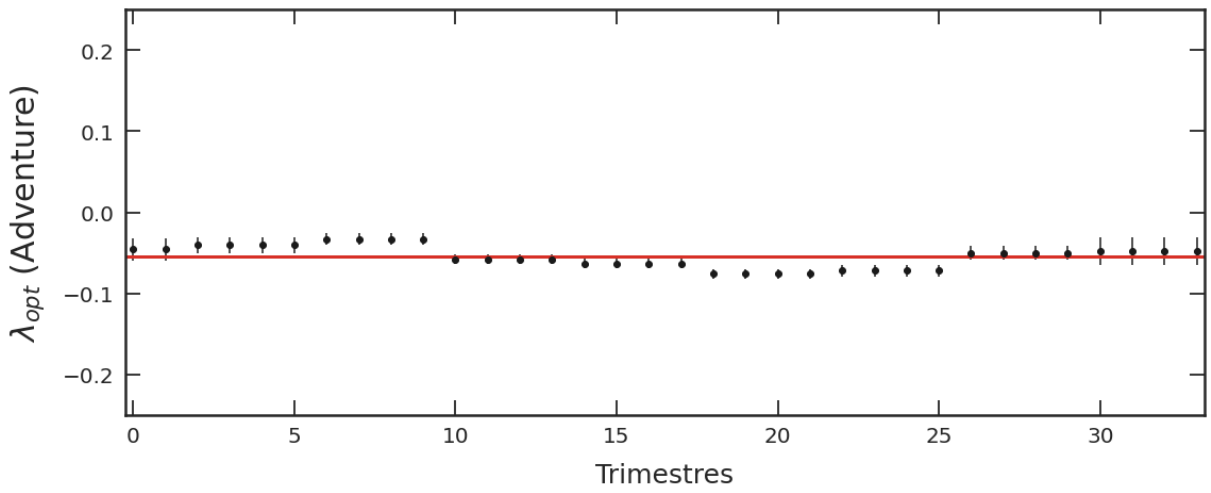

A
B
Dist. de probabilidade (Puzzle)
Dist. de probabilidade (Multiplayer)
Dist. de probabilidade (Atmospheric)
$z^{(0)}$
$z^{(\lambda)}$
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3

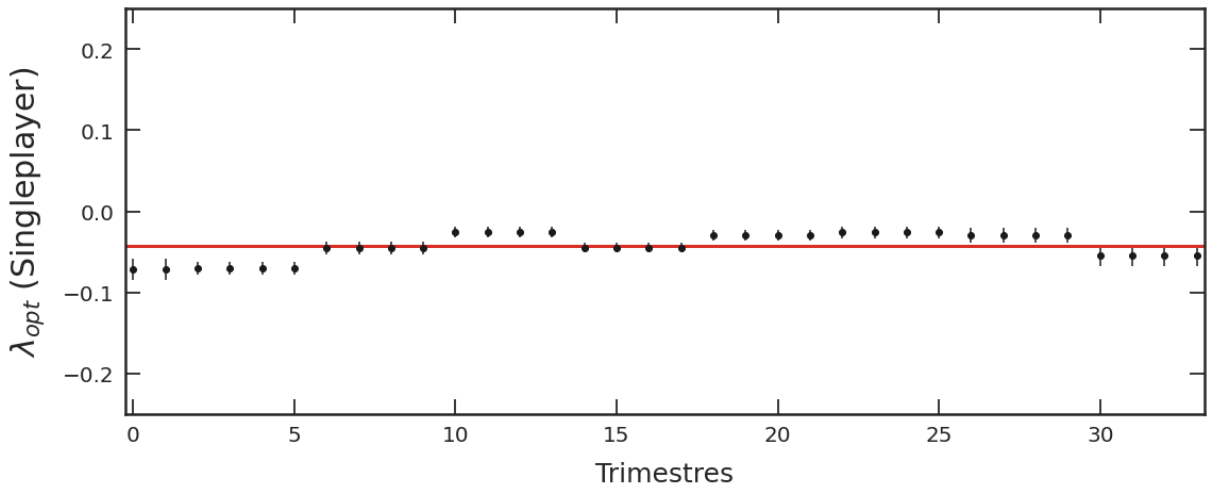
APÊNDICE E – Evolução temporal dos parâmetros de Box-Cox para as categorias principais

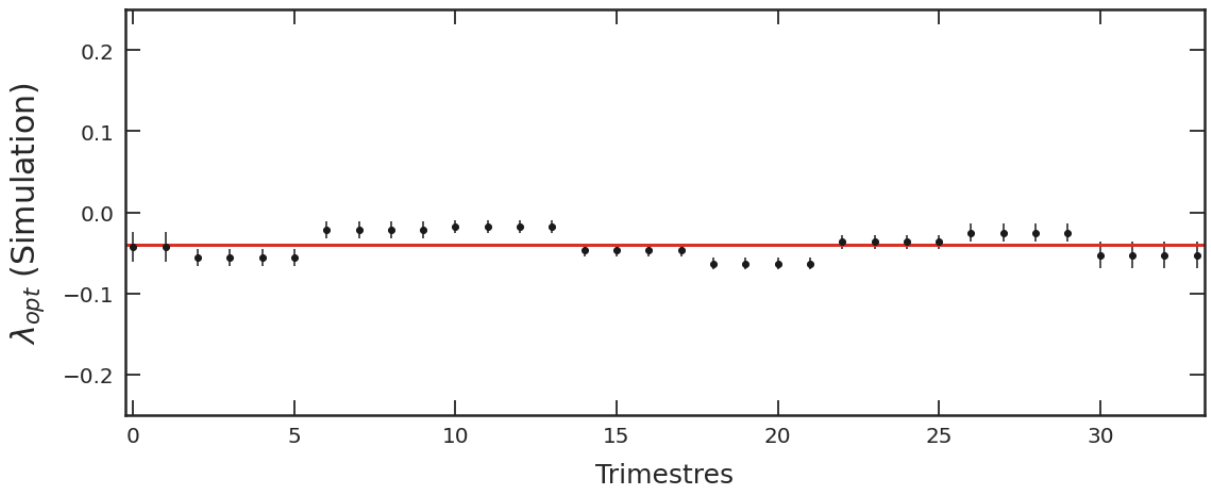
Fechando a parte dos apêndices, apresentamos aqui a evolução temporal do parâmetro de Box-Cox ótimo para cada uma das categorias principais. Como no texto, a reta vermelha representa o valor médio.

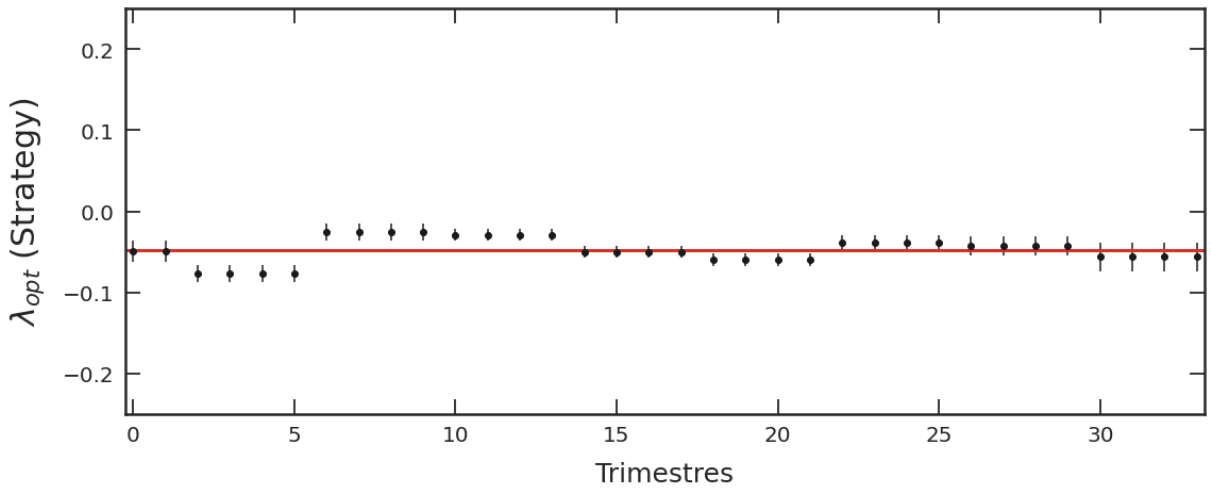


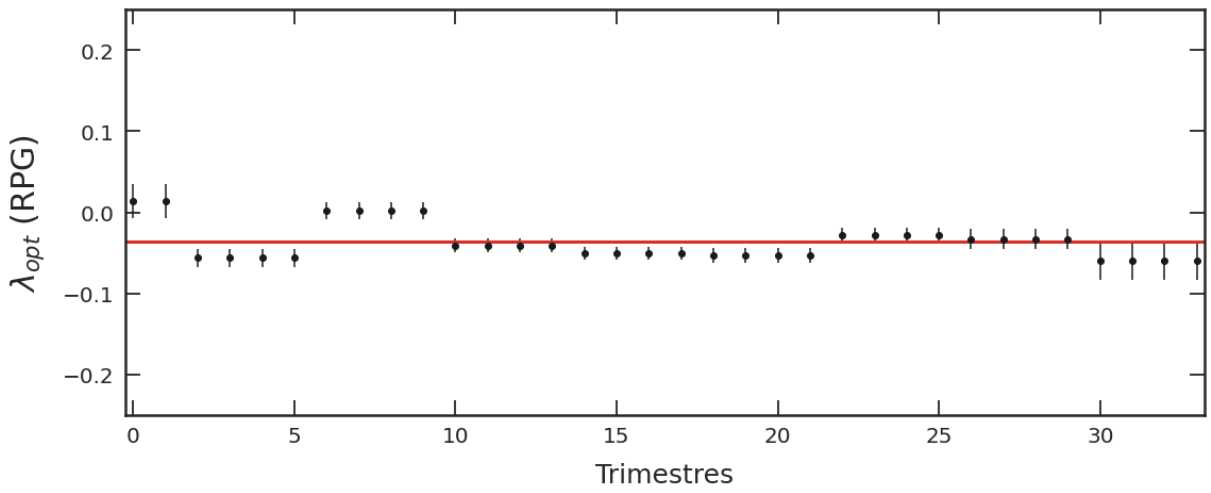

λ_{opt} (Action)
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20
25
30
Trimestres

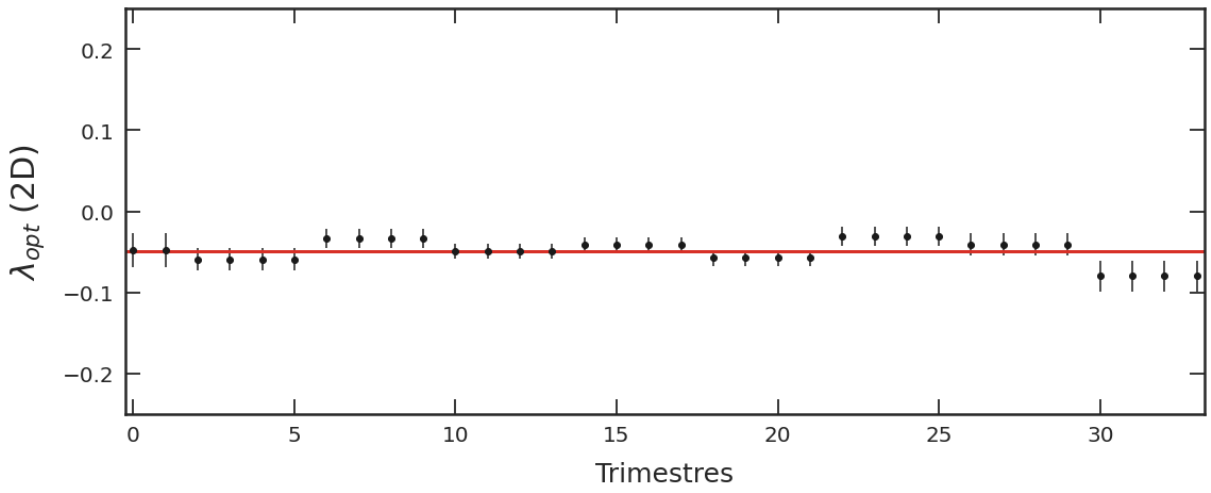

λ_{opt} (Adventure)
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20
25
30
Trimestres

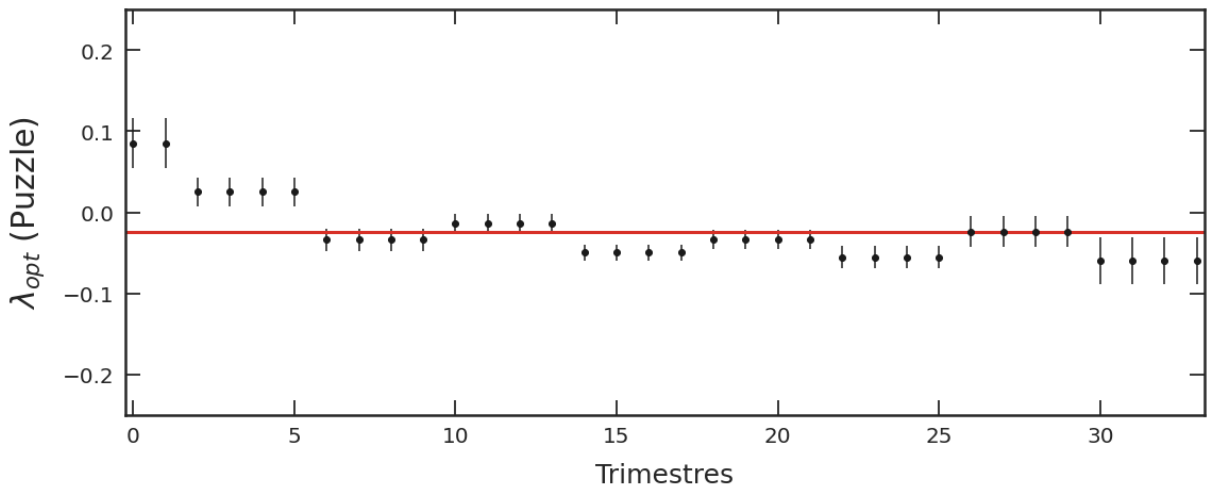

λ_{opt} (Singleplayer)
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20
25
30
Trimestres

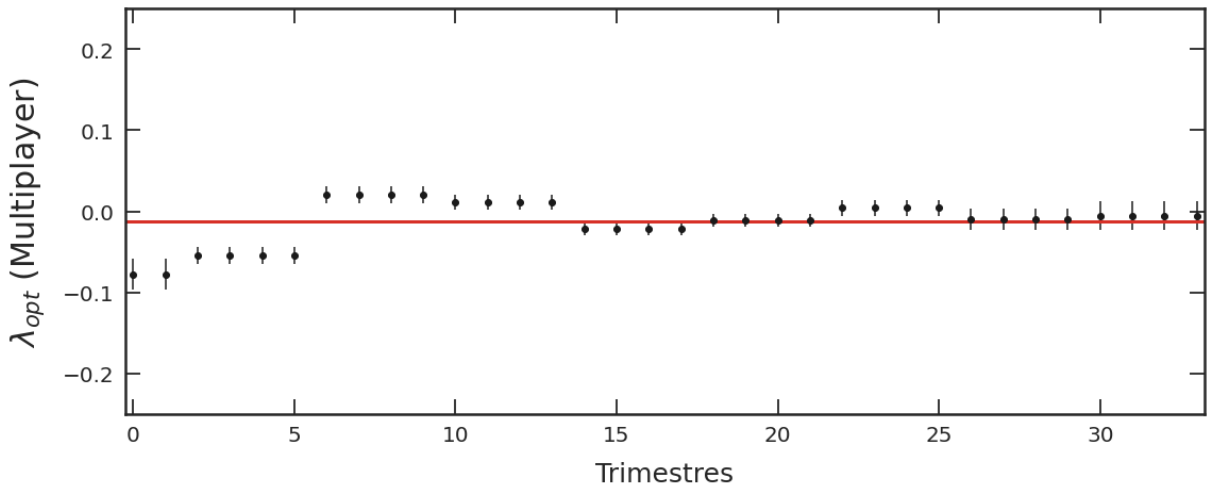

λ_{opt} (Simulation)
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20
25
30
Trimestres

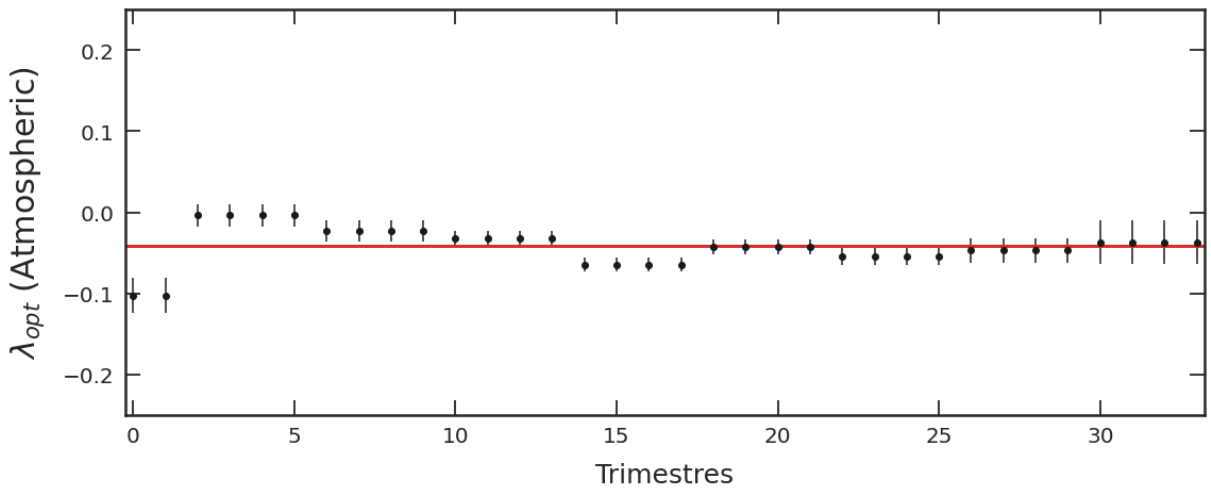

λ_{opt} (Strategy)
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20
25
30
Trimestres


λ_{opt} (RPG)
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20
25
30
Trimestres


λ_{opt} (2D)
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20
25
30
Trimestres


λ_{opt} (Puzzle)
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20
25
30
Trimestres


λ_{opt} (Multiplayer)
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20
25
30
Trimestres


λ_{opt} (Atmospheric)
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20
25
30
Trimestres