Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 10, n. 1, 2023.

Trabalho apresentado no XLII CNMAC, Universidade Federal de Mato Grosso do Sul - Bonito - MS, 2023

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Distribuição de Newcomb-Benford, Mudanças de Bases e Aplicações Eleitorais

Eduardo Gueron[1], Jerônimo Pellegrini[2]
CMCC-UFABC, Santo André-SP
Bruno Aristimunha[3]
CMCC-UFABC, Santo André-SP;
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique

Resumo.

Neste estudo, usamos a distribuição de Newcomb Benford como ferramenta para analisar os primeiros turnos das eleições presidenciais brasileiras desde 1994, utilizando a invariância de base presente nessa distribuição. Ao comparar os resultados das eleições com o número de votantes por Zona Eleitoral, verificamos que não há indícios de fraude nas eleições analisadas. Além disso, avaliamos a aderência à distribuição de Benford e observamos que algumas Zonas Eleitorais apresentaram menor aderência em relação às eleições em si. Esses resultados sugerem que a distribuição de Benford pode ser uma ferramenta útil para detectar anomalias nas eleições, embora não tenhamos encontrado evidências de fraude nas eleições presidenciais no Brasil.

Palavras-chave. Distribuição de Benford; Fraude Eleitoral; Invariância de Base.

1 Introdução

Simon Newcomb, um astrônomo canadense, e mais tarde Frank Benford, um físico americano, notaram uma curiosa tendência em alguns conjuntos de dados: a frequência com que o algarismo 1 aparecia como primeiro dígito era consideravelmente maior do que a de outros dígitos, como o 7. Essa observação intrigante foi ainda mais evidente em tábuas de logaritmo antigas, onde as páginas associadas a números com primeiro dígito 1 ou 2 apresentavam mais desgaste [8].

A partir de análises empíricas e resultados, Newcomb, e, posteriormente, Benford, derivaram a probabilidade com que o primeiro dígito d deveria aparecer em conjuntos que seguem a distribuição que estudaram.

$$P(d) = \log\left(1 + \frac{1}{d}\right). \tag{1}$$

Desde a publicação do trabalho original de Newcomb-Benford, diversos autores têm estudado conjuntos de dados que seguem essa distribuição, por mera curiosidade ou para fins de registro. Um bom compilado histórico desses estudos foi apresentado em [6] no aniversário de 125 anos do trabalho original. A questão que surge é até que ponto essa ubiquidade pode ser útil em um sistema de verificação. Para responder a essa pergunta, o pesquisador inglês Mark Nigrini propôs em sua tese de doutorado [10] o uso da distribuição de Newcomb-Benford como uma ferramenta de análise forense [9].

[1]eduardo.gueron@ufabc.edu.br
[2]jeronimo.pellegrini@ufabc.edu.br
[3]b.aristimunha@gmail.com

A detecção de fraudes eleitorais é uma área de grande interesse e a Lei de Newcomb-Benford é considerada uma ferramenta possível para auxiliar nesse processo. No entanto, sua eficácia é alvo de muita controvérsia, conforme discutido em [3]. É importante salientar que nem todas as fraudes eleitorais podem ser identificadas através desta abordagem, além disso, existe um desafio em definir claramente o nível de desvio da distribuição que caracteriza uma fraude.

Em uma investigação recente, identificamos um possível caso de fraude eleitoral nas eleições para o Senado no estado da Bahia em 1994. Para isso, foram comparadas dezenas de eleições semelhantes para o cargo de senador no Brasil, utilizando a invariância de base da distribuição como base para os testes estatísticos, como descrito em [4].

Vale a pena reforçar que, em uma base qualquer, b a frequência com que o dígito d deve aparecer em um conjunto que segue a Distribuição de Newcomb-Benford deve ser:

$$f(d) = \log_b \left(1 + \frac{1}{d}\right), \tag{2}$$

em que $\log_b$ é logaritmo na base b. A invariância de base, na realidade, implica (e é implicada) Benford [5].

Visando estabelecer parâmetros de comparação, este artigo revisita alguns resultados obtidos e analisa o comportamento de eleições presidenciais passadas no Brasil. O estudo pretende aprofundar a compreensão do cenário eleitoral brasileiro e contribuir para a identificação de possíveis padrões em eleições futuras.

2 Aplicação em Detecção de Fraude Eleitoral

Com o intuito de verificar a ocorrência de fraude em uma eleição para senador na Bahia em 1994, utilizamos métodos baseados na Distribuição de Newcomb-Benford. Para tanto, analisamos as eleições para o senado federal entre 1994 e 2018 nos sete estados brasileiros com maior número de municípios: Minas Gerais, São Paulo, Rio Grande do Sul, Bahia, Paraná, Santa Catarina e Goiás (sendo o estado da Bahia, o mediano). Nosso objetivo foi identificar casos atípicos de votação em cada zona eleitoral e compará-los com os demais.

A ideia era verificar se, no espectro de votação dos senadores em cada zona eleitoral, havia um caso atípico (outlier) entre semelhantes. À época, houve questionamento por parte do candidato preterido, negado pelo TRE por falta de evidências – tal questionamento corrobora a tese de que a eleição de fato tenha sido fraudada [2].

A verificação do desvio da Lei de Benford na eleição para senador pela Bahia foi cerificada por três métodos: a observação da razão entre os dois dígitos líderes; cálculo da divergência de Kullback-Leibler; e a observação do desvio padrão no primeiro método. (Na próxima seção, explicaremos os métodos quando aplicarmos no estudo desse artigo).

Em todos os casos, a eleição para senador na Bahia em 1994 se destaca muito claramente, aumentando os indícios de fraude naquele pleito. Acreditamos termos apresentado um dos poucos resultados onde a fraude eleitoral se caracterizou indubitavelmente por meio da distribuição de Newcomb-Benford e, pelo que sabemos, foi a primeira vez que se aplicou invariância por mudança de base como técnica forense.

Para mencionar alguns dos valores numéricos obtidos para a eleição da Bahia em 1994, em um histograma que escrevemos, Figura 3 de [4], pode se estimar que a chance de ter sido 'sorte' o desvio da distribuição de Benford é menor que $0,5\%$.

DOI: 10.5540/03.2023.010.01.0042

3 Na Direção de uma Caracterização de Quando (ou como) a Lei de Benford Vale em Eleições

Como mencionado anteriormente, uma questão fundamental a ser abordada é a proximidade entre a distribuição da votação para candidatos a cargos majoritários e a distribuição de Newcomb-Benford. Neste contexto, uma hipótese que decidimos investigar é se o número de eleitores nas Zonas Eleitorais (ZEs) em todo o Brasil segue a distribuição de Newcomb-Benford. Além disso, buscamos realizar uma comparação com as votações majoritárias nas mesmas ZEs.

Para tanto, os dados utilizados para comparação foram os primeiros turnos das últimas oito eleições presidenciais (1994 a 2022) além de 7 eleições para governador (1994 a 2018). Em todos os casos, foram considerados apenas os 3 candidatos mais votados.

Criamos, também, três eleições simuladas com dois candidatos nas Zonas Eleitorais brasileiras definidas no ano de 2022. A votação de um dos candidatos é obtida por sorteio e a do outro com os votos restantes (não consideramos brancos e nulos nesse exemplo). Nas simulações, Tabela 1, foram consideradas respectivamente 2, 3 e 4 distribuições (a fim de simular regiões onde o candidato é proporcionalmente mais forte ou fraco). Na simulação, sorteamos em cada zona eleitoral uma proporção de votos para o primeiro candidato usando uma distribuição normal escolhida de acordo com as probabilidades mostradas na tabela a seguir (contabilizamos para o primeiro candidato os votos obtidos pela distribuição Normal sorteada, e os outros para o segundo candidato).

Simulação	Distribuição			
1	$p = 0.1$	$p = 0.4$	$p = 0.15$	$p = 0.35$
	$\mathcal{N}(0.5, 0.2)$	$\mathcal{N}(0.2, 0.1)$	$\mathcal{N}(0.4, 0.2)$	$\mathcal{N}(0.1, 0.1)$
2	$p = 0.333$	$p = 0.333$	$p = 0.334$	
	$\mathcal{N}(0.4, 0.2)$	$\mathcal{N}(0.3, 0.1)$	$\mathcal{N}(0.1, 0.1)$	
3	$p = 0.5$	$p = 0.5$		
	$\mathcal{N}(0.4, 0.7)$	$\mathcal{N}(0.7, 0.2)$		

Tabela 1: Simulação de modelos de votações por zona eleitoral.

Os resultados são apresentados a seguir, com métodos específicos.

Razão de Frequências

A probabilidade de ser d o algarismo inicial de um número escrito na base b em um conjunto de dados que segue a distribuição de Benford tem a expressão dada na Equação 2.

Podemos, ainda, calcular a razão $R(d_1, d_2)$ da frequência com que os números escritos na base b se iniciam com os dígitos d_1 e d_2. Utilizando logaritmo natural, ficamos com

$$R(d_1, d_2) = \frac{\ln(1 + \frac{1}{d_1})}{\cancel{\ln(b)}} \frac{\cancel{\ln(b)}}{\ln(1 + \frac{1}{d_2})}, \tag{3}$$

ou seja, a razão independe da base. Baseados nesta ideia, pode-se definir a função:

$$F(d_1, d_2) = R(d_1, d_2)\frac{N(d_2)}{N(d_1)} - R(d_2, d_1)\frac{N(d_1)}{N(d_2)}, \tag{4}$$

em que $N(d)$ é o número de vezes em que d aparece como dígito inicial na amostra estudada. É fácil verificar que, no caso Benford idealizado, $F(d_1, d_2)$ deve se anular.

DOI: 10.5540/03.2023.010.01.0042

A seguir, apresentamos o resultado de $F(1,2)$ para as eleições para presidente e senador citadas anteriormente, além do total de votos por Zona Eleitoral e as simulações de candidatos fictícios, Figura 1.

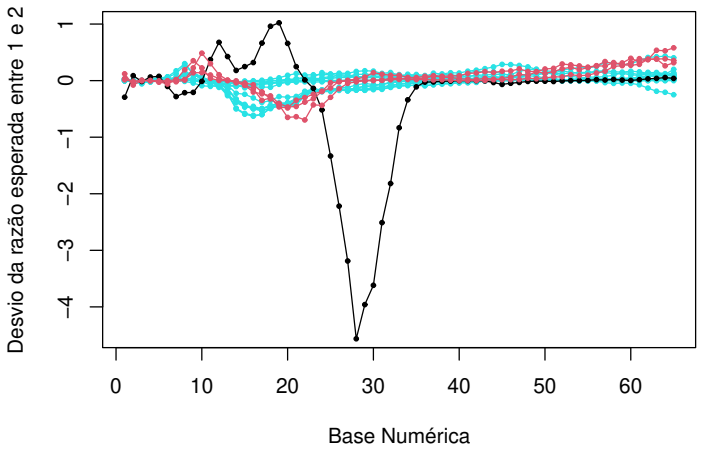

Figura 1: Gráfico da equação Equação 4, com valores $F(1,2)$, para as bases entre 6 a 65. As curvas acumuladas em azul representam as eleições reais. As curvas em vermelho são as três simulações e a curva preta representa os votos totais por ZE.

Distribuição de Desvios

Utilizando o mesmo conjunto de dados, conduzimos uma análise estatística para calcular o Desvio Padrão de $F(1,2)$ na base b para cada um dos conjuntos de dados. Para visualizar a distribuição desses desvios padrão, construímos um histograma, como mostrado na Figura abaixo:

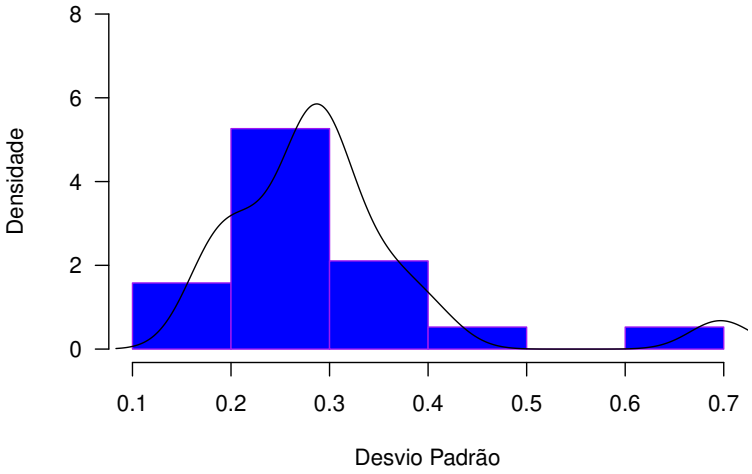

Figura 2: Histograma dos Desvios Padrão (DP) para cada uma das amostras de eleições presidenciais (1994-2022). À direita, está o desvio padrão calculado para o número de votantes por Zona Eleitoral.

Na Figura 2, vale observar que o histograma representa o desvio padrão de cada um dos conjuntos de dados. Esses DP constituem um novo conjunto de dados cuja média e desvio padrão foram calculados. O ponto correspondente ao número de eleitores por zona eleitoral está distante da média a quase 3.5 em unidades de desvio padrão.

Divergência Kullback-Liebler

A fim de quantificar a aderência do conjunto de dados eleitoras a uma distribuição teórica, uma medida comumente utilizada é a Divergência de Kullback-Liebler [1], uma métrica simétrica, calculada da seguinte maneira:

$$KLD_b = \min\left\{\sum_{d=1}^{b-1} P_B(d)\log\left(\frac{Q_S(d)}{P_B(d)}\right), \sum_{d=1}^{b-1} Q_S(d)\log\left(\frac{P_B(d)}{Q_S(d)}\right)\right\}, \tag{5}$$

em que $Q_S(d)$ é a frequência com que números iniciados com o dígito d na base b aparecem na amostra e $P_B(d)$ é a probabilidade se o conjunto segue exatamente a distribuição de Newcomb-Benford, dada pela Equação 2. O comportamento da divergência de Kullback-Liebler para o intervalo de base numéricas é apresentado na Figura 3.

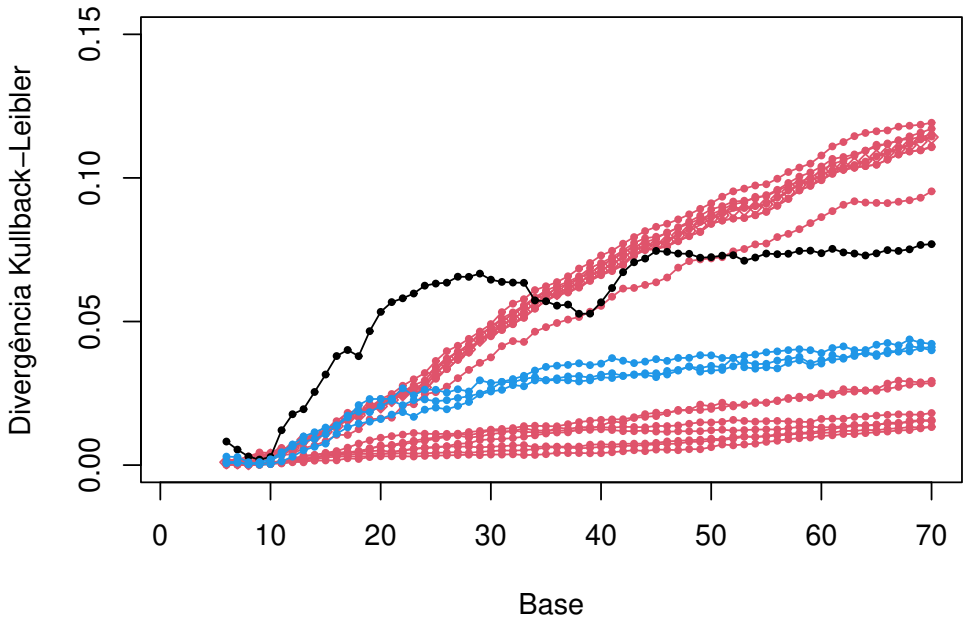

Figura 3: Valor da divergência KullBack-Leibler (KL) para cada base. As curvas vermelhas representam as eleições de 1994 a 2022, as azuis as simulações de votos e a preta os votantes por Zona Eleitoral (ZE).

Na Figura 3, vale destacar que as curvas vermelhas com menores valores da divergência são as eleições para governador, as curvas vermelhas com maiores valores são as para presidente, além da simulação (azul) e a de número de votos por zona eleitoral (preta).

DOI: 10.5540/03.2023.010.01.0042

4 Resultados e Considerações Finais

Observamos, principalmente em Figura 1 e Figura 2 que, enquanto o conjunto com as quantidades de votantes por zona eleitoral apresenta aderência não muito justa à distribuição de benford, a contabilização por candidato segue a distribuição de maneira mais próxima, bem como as simulações.

A fim de interpretar o resultado obtido, citemos o artigo de Miller e Nigrini [7], que afirmam:

> "muitos autores observaram que o produto (e de forma mais geral, qualquer operação aritmética bem comportada) de duas variáveis aleatórias fica mais próxima de satisfazer a Lei de Benford do que as variáveis de entrada. Mais ainda, à medida que o número de termos aumenta, a expressão resultante parece se aproximar da Lei de Benford"[4].

Podemos imaginar que diversos fatores constituem as votações de candidatos majoritários por zona eleitoral, um deles é o total de votantes por ZE. Tanto nas eleições que analisamos quanto nas simulações, outras variáveis aleatórias foram incluídas na composição do conjunto de dados, portanto, é esperado que se aproximem da distribuição de Benford.

Sobre os resultados com Divergência de Kullback-Liebler, Figura 3, há dois fatores destacados. O primeiro é que as eleições de presidente se comportaram de maneira muito semelhante (se acumulando na parte de cima do gráfico) bem como as eleições para governadores de estado (na parte de baixo) mostrando que não há qualquer indício de fraudes explicitado por este método já que não existem os chamados outliers. Diferente do que aconteceu nas eleições para senador da Bahia em 1994.

O segundo é que a divergência para o número de votantes se destaca bastante até aproximadamente base 40, mas depois se mantém razoavelmente constante, ficando menor do que a das eleições para presidente. Esse comportamento pode representar algumas limitações no método aplicado.

Futuramente, uma análise mais robusta de eleições em diversos países e contextos pode nos permitir definir parâmetros claros para identificar possíveis fraudes eleitorais. Certamente, a maneira com que os votos totais são separados em análogos às nossas Zonas Eleitorais são fundamentais para a definição destes parâmetros. Simulações similares às apresentadas na Tabela Tabela 1 devem ser utilizadas já que, pelos nossos resultados, comportaram-se de maneira muito similar às eleições reais em relação à distribuição de Newcomb-Benford.

Agradecimentos

E.G. foi Pesquisador Associado de projeto financiado pela FAPESP, recém finalizado (Processo: 2019/06174-7). O trabalho de B. A. é financiado pelo Intitut DATAIA.

Referências

[1] J.M. Bernardo e M. A. Juarez. "Intrinsic Estimation". Em: **Bayesian Statistics 7**. Ed. por J. M. Bernardo et al. Clarendon Press, Oxford, 2003, pp. 465–476.

[2] CONJUR. **Pedido Recontagem de Votos Senado Bahia**. Online. 1999, `https://www.conjur.com.br/1999-set-13/stf_decidir_pedido_recontagem_votos`.

[3] J. Deckert, M. Myagkov e P. C. Ordeshook. "Benford's Law and the Detection of Election Fraud". Em: **Political Analysis** 19 (2011), pp. 245–268.

[4] Tradução Nossa

[4] E. Gueron e J. Pellegrini. "Application of Benford–Newcomb law with base change to electoral fraud detection". Em: **Physica A: Statistical Mechanics and its Applications** (2022). DOI: 10.1016/j.physa.2022.128208.

[5] T. P. Hill. "A Statistical Derivation of the Significant-Digit Law". Em: **Statist. Sci.** (1995), pp. 354–363. DOI: 10.1214/ss/1177009869.

[6] W. Hurlimann. "Benford's Law from 1881 to 2006: a bibliography". Em: ().

[7] Steven J. Miller e Mark J. Nigrini. "The Modulo 1 Central Limit Theorem and Benford's Law for Products". Em: **International Journal of Algebra** 2.3 (2008), pp. 119–130.

[8] Simon Newcomb. "Note on the Frequency of Use of the Different Digits in Natural Numbers". Em: **American Journal of Mathematics** 4.1 (1881), pp. 39–40.

[9] M Nigrini e J.T. Wells. **Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection**. John Wiley & Sons, 2012. ISBN: 978-1-118-15285-0.

[10] M. J. Nigrini. "The detection of income tax evasion through an analysis of digital distributions". Tese de doutorado. University of Cincinatti, 1992.

