DOI: 10.36558/rsc.v14i1.8675

A Lei de Benford e o Coeficiente de Gini Como Métodos Estatísticos na Análise de Dados Eleitoriais

Benford's Law and the Gini Coefficient as Statistical Methods in the Analysis of Electoral Data

Maciel de Lima Souza
Instituto Federal do Ceará
CE-187, s/n - Estádio - 62.320-000,
Tianguá – CE – Brasil
maciells0002@gmail.com

Rhyan Ximenes de Brito
Instituto Federal do Ceará
CE-187, s/n - Estádio - 62.320-000,
Tianguá – CE – Brasil
rxbrito@gmail.com

Janaide Nogueira de Sousa Ximenes
Instituto Federal do Ceará
Rua Luiz Cunha, 178 - Monte Castelo - 62.350-000,
Ubajara – CE – Brasil
nogueirajanaide@gmail.com

Paulo César de Almeida Júnior
Instituto Federal do Ceará
CE-187, s/n - Estádio - 62.320-000,
Tianguá – CE – Brasil
paulo.almeida@ifce.edu.br

ABSTRACT

The article addresses the application of statistical methods in the analysis of electoral data from the state of Ceará in 2022, utilizing Benford's Law as a tool to detect anomalies in the numerical distribution of votes. Additionally, the Chi-square test and Z statistical test are discussed as techniques for assessing the statistical significance of observations and results. The Gini coefficient and the Lorenz Curve are presented as tools for evaluating inequality in data distribution. Together, these methodologies provide a comprehensive approach to the statistical analysis of datasets, highlighting patterns, identifying discrepancies, and assessing inequality. This study is relevant for researchers, students, and professionals seeking a comprehensive understanding and interpretation of data.

Keywords

Statistical Methods; Analysis of Electoral Data; Electoral Data.

RESUMO

O artigo aborda a aplicação de métodos estatísticos na análise de dados eleitorais do Ceará em 2022, utilizando a lei de Benford para detectar anomalias na distribuição numérica dos votos. Além disso, são discutidos o teste Quiquadrado e o teste estatístico Z para avaliação da significância estatística. O coeficiente de Gini e a Curva de Lorenz são apresentados como ferramentas para avaliar a desigualdade na distribuição de dados. Em conjunto, essas metodologias proporcionam uma abordagem abrangente para identificar padrões, discrepâncias e avaliar a desigualdade em conjuntos de dados, sendo relevante para pesquisadores, alunos e profissionais.

Palavras-Chave

Métodos estatísticos; Análise de Dados Eleitorais; Dados Eleitorais.

1. INTRODUÇÃO

A análise de dados numéricos fundamental em várias áreas, abrangendo desde a auditoria financeira, compreendida como uma subdivisão da contabilidade que busca assegurar a conformidade dos procedimentos contábeis realizados em uma entidade específica [6], até a detecção de fraudes e a avaliação de fenômenos naturais. Nesse contexto, a lei de Benford, também conhecida como a lei dos primeiros dígitos, tem ganhado destaque como uma ferramenta poderosa para a detecção de irregularidades em conjuntos de dados. Esta lei postula que, em muitos conjuntos de dados do mundo real, os primeiros dígitos não ocorrem igualmente com frequência, mas sim de acordo com uma distribuição logarítmica específica. A hipótese de que dados fabricados ou falsificados são identificados mediante o desvio dos dígitos em relação à distribuição de Benford foi testada recentemente em diversos contextos [4].

Esta relação logarítmica intrigante foi primeiramente observada por Simon Newcomb em 1881, um astrônomo e matemático do século XIX. Ele observou que as primeiras

páginas das tábuas de logaritmos se apresentavam mais desgastadas do que as últimas, indicando que o valor usualmente mais acessado era o 1, e que a frequência diminuía até o 9. [4], no entanto foi Frank Benford, um físico, em 1938 que apresentou a esta observação um caráter mais popular. A observação de Newcomb foi demonstrada e difundida por Benford, que a analisou em conjuntos de números oriundos de diferentes contextos não relacionados, indicando-a como uma lei de probabilidade geral de ampla aplicação [17].

Portanto, a aplicação da lei de Benford estende-se além do campo da análise estatística, tendo implicações práticas em auditorias e investigações de fraudes. A discrepância entre a distribuição esperada pela lei de Benford e a distribuição real de dígitos iniciais pode sugerir manipulação de dados ou inconsistências nos registros financeiros. Além disso, a lei de Benford tem sido adotada em estudos de autenticidade de dados científicos e eleitorais, onde a detecção de anomalias pode indicar resultados suspeitos ou até mesmo falsificações [24].

Neste artigo será explorada a lei de Benford aplicando-a em uma página *web* que usa essa prerrogativa por meio da análise de conjuntos de dados reais extraídos da base de dados do TRE-CE (Tribunal Regional Eleitoral do Ceará) com o intuito de investigar a utilidade dessa lei na detecção de padrões anômalos nos dados eleitorais, além de discutir sobre seu potencial como ferramenta de análise de dados em votos eleitorais. Por parâmetro de dados será utilizado os votos dados a todos os candidatos elegíveis por município do estado do Ceará nas últimas eleições gerais do ano de 2022 para os cargos de Deputado Estadual, Deputado Federal, Senador, Governador e Presidente.

Em conjunto com a lei de Benford será usado o coeficiente de Gini e a curva de Lorenz como ferramentas de validação de distribuição dos dados. Gini e Lorenz são ferramentas analíticas amplamente utilizadas em diferentes contextos. O coeficiente de Gini é uma medida estatística que avalia a desigualdade de distribuição de renda em uma determinada população, variando de 0 (igualdade perfeita) a 1 (desigualdade total) [22]. Já a curva de Lorenz, por sua vez, visualiza graficamente essa desigualdade, representando a acumulação percentual da renda em relação à acumulação percentual da população. Apesar de aplicações em áreas diferentes, neste trabalho sua aplicação faz-se sentido em avaliar a distribuição de dados dentro de um padrão seguido pela lei de Benford, além disso, também será usado os testes Qui-quadrado e teste Z para medir a aderência desses dados a mesma lei.

O objetivo desse trabalho é aplicar a lei de Benford e o coeficiente de Gini na análise de dados eleitorais do estado do Ceará, especificamente nas últimas eleições gerais do país no ano de 2022. A intenção é avaliar a aderência dos dados eleitorais a essas ferramentas estatísticas, buscando identificar possíveis anomalias, fraudes ou desigualdades na distribuição dos votos. Além disso, o estudo visa contribuir para o aprimoramento de ferramentas de análise estatística e para a compreensão da integridade e confiabilidade do processo eleitoral.

O artigo está organizado em seis seções, a contar da introdução. Na seção 2 está posta uma análise em trabalhos relacionados da literatura sobre a ótica da lei de Benford e seu uso na identificação de fraudes além de sua utilização em outras áreas em conjunto com o coeficiente de Gini. Então, na seção 3, foi apresentada a fundamentação teórica do trabalho e de forma detalhada o funcionamento da lei de Benford, teste Qui-quadrado, Teste Z e coeficiente de Gini em conjunto com a Curva de Lorenz. Já na seção 4, foi exposto a base de dados e os testes estatísticos utilizados. Após isso, na seção 5, há a discussão dos resultados obtidos. Por fim, a conclusão na seção 6.

2. TRABALHOS RELACIONADOS

A revisão literária desempenha um papel crucial no desenvolvimento de um trabalho acadêmico além de proporcionar uma compreensão aprofundada do tópico de pesquisa em questão. Nesta seção, serão explorados estudos anteriores relacionados a lei de Benford e suas aplicações na detecção de fraudes, assim como também sua utilização em outras áreas, mas que corroboram para o estudo proposto neste artigo, além de apresentar o coeficiente de Gini como ferramenta de análise das distribuição de proporção sobre Benford.

É de conhecimento que muitos pesquisadores já se aventuraram no estudo sobre a aplicação desta lei, exemplo disso foi [17] que conduziu um estudo notável nesse campo. Seu estudo adotou uma abordagem investigativa sobre os casos de COVID-19 em diferentes países pelo mundo. Seu trabalho analisou o Brasil, a China, a Índia entre outros, utilizando a lei de Benford para validar a veracidade dos dados disponibilizados por esses países revelando descobertas importantes sobre a aderência desses dados a esta lei. Menezes[17] por vez entende seu artigo como sendo do tipo quantitativo, na medida em que busca verificar a lei de Benford nos números confirmados de COVID-19 em diferentes países.

Já [25] em seu trabalho analisou gastos públicos em emendas parlamentares a luz da lei de Benford, por fim este concluiu que a desconformidade com a lei de Benford é um indicador que sugere análises minuciosas daquele grupo de dados financeiros com o propósito de avaliar o que deu causa à discrepância. Isso permite o aperfeiçoamento dos controles a que os gastos públicos estão sujeitos em vários contextos, como por exemplo, durante a realização de licitações a partir da análise de planilhas de preços, ou por meio da despesa empenhada, liquidada ou paga.

Por outro [26] usou o índice de Gini como ferramenta para medir a desigualdade de renda em diferentes grupos e populações. Em seu trabalho, Silva[25] calculou o índice de Gini a partir da distribuição da população e da renda em diferentes grupos brasileiros. O resultado obtido foi de 0,464, indicando uma desigualdade moderada no país. Além disso, o índice de Gini foi utilizado em conjunto com outros índices, como o índice de Kakwani e o índice de Reynolds-Smolensky, para avaliar a progressividade e o efeito redistributivo do Imposto de Renda de Pessoa Física.

Já [15] trouxe a luz resultados que indicaram que a implementação de subsídios para alguns passageiros do Metrô-DF contribuiu para uma redução na desigualdade de renda dos passageiros que utilizam o serviço, mesmo que a variação do índice de Gini tenha sido baixa. Em seu trabalho o coeficiente de Gini foi utilizado como parâmetro para verificar a distribuição de renda entre os usuários do Metrô-DF. Além disso, o índice de Kakwani confirmou a progressividade da política de gratuidade do Metrô-DF, indicando que a concentração da renda bruta dos passageiros do Metrô-DF após o subsídio é menor do que a concentração do subsídio. A relevância dessa pesquisa se dá por gerar contribuições empíricas à literatura sobre o efeito da gratuidade dos trans-

portes públicos na redistribuição de renda e desigualdade social.

Na Tabela 1 apresentada a seguir, destacam-se outros estudos relacionados que empregam a lei de Benford como parâmetro para a avaliação de dados eleitorais ou correlatos.

Table 1: Síntese demonstrativa de outros trabalhos relacionados.

Autor	Título
[13]	Uma avaliação da qualidade dos dados reportados sobre financiamento de campanha com base na Lei de Benford.
[14]	As aplicabilidades da Lei de Benford na análise de um conjunto de dados eleitorais.
[18]	As últimas eleições e a Lei de Benford (ou Lei do Primeiro Dígito).
[27]	Em busca de transparência: a Lei de Benford aplicada às despesas eleitorais.
Nossa proposta	A utilização da Lei de Benford e do coeficientede Gini como métodos estatísticos na análise de um conjunto de dados eleitorais.

Fonte: Elaboração própria.

Em última análise, a revisão da literatura evidencia a presença de um conjunto substancial de pesquisas que delineiam o campo de estudo em questão. Trabalhos anteriores estabelecem uma base para a investigação proposta, ao mesmo tempo em que apontam áreas específicas em que nossa pesquisa pode contribuir para preencher lacunas no conhecimento existente. Similarmente a outros estudos mencionados, empregamos técnicas estatísticas para a avaliação dos dados como o teste Qui-quadrado e o teste Z. No entanto, o trabalho possui una analíse inovadora, por combinar a lei de Benford e o coeficiente de Gini como ferramentas complementares na análise de dados. Essa abordagem integrada representa um diferencial significativo em relação aos estudos prévios, proporcionando uma perspectiva única e aprofundada à nossa investigação.

3. FUNDAMENTAÇÃO TEÓRICA

Nesta seção, será apresentada uma revisão técnica e teórica dos temas que fundamentam este trabalho, organizando-se da seguinte maneira: A Subseção 3.1 introduz a definição da lei de Benford, fornecendo uma base conceitual para compreensão do fenômeno. As Subseções 3.1.1 e 3.1.2 exploram, respectivamente, o teste Qui-quadrado e o teste Z, expandindo o entendimento sobre as metodologias estatísticas aplicadas neste contexto. A Subseção 3.2 aprofunda as definições relacionadas ao coeficiente de Gini, um indicador crucial na avaliação da desigualdade em diferentes conjuntos de dados. Por fim, a Subseção 3.3 aborda o funcionamento dos pleitos eleitorais, detalhando o sistema eleitoral brasileiro e contextualizando as nuances desse processo fundamental para o contexto político.

3.1 Lei de Benford

O desenvolvimento da lei de Benford tem raízes históricas que remontam ao final do século XIX, quando o astrônomo Simon Newcomb observou uma tendência interessante nos primeiros dígitos de tabelas de logaritmos. Ele observou que, ao examinar conjuntos de dados numéricos do mundo real, os primeiros dígitos dos números não pareciam ser igualmente distribuídos. Em vez disso, dígitos menores, como 1, eram mais frequentes como primeiros dígitos, enquanto dígitos maiores, como 9, eram menos comuns, ou seja, se uma extensa coleção de dados numéricos for classificada conforme seu primeiro dígito significativo, então as nove classes possíveis resultantes não possuirão geralmente o mesmo tamanho [8].

Embora o fenômeno da lei do primeiro dígito tenha sido inicialmente observado por Newcomb, a associação preeminente a esse conceito é frequentemente atribuída a Frank Benford, um físico e engenheiro que redescobriu e popularizou a lei quase cinquenta anos após a sua primeira notação. Apesar de sua descoberta remontar ao final do século XIX, a lei do primeiro dígito experimentou uma crescente atenção acadêmica, especialmente a partir da publicação acadêmica de Benford em 1938. Este trabalho é reconhecido como o estudo empírico mais abrangente da lei do primeiro dígito até a década de 1990, destacando-se pela inclusão da mais extensa tabela de frequências de dígitos disponível para investigação na época. A obra de Benford proporcionou uma base sólida para estudos subsequentes sobre a aplicação e interpretação da lei do primeiro dígito em diversas disciplinas acadêmicas [21].

Portanto é possível inferir que os números anômalos são um fenômeno observado em muitos conjuntos de dados do mundo real, onde os primeiros dígitos não ocorrem igualmente, mas seguem uma distribuição logarítmica específica. Essa distribuição contrasta com a expectativa intuitiva de que todos os dígitos tenham igual probabilidade de aparecer como o primeiro dígito [19].

A Tabela 2 abaixo evidencia a distribuição esperada dos primeiros dígitos conforme a lei de Belford:

Table 2: Probabilidade associada ao primeiro dígito.

Dígito(d)	P(d)
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%
Total	100%

Fonte: Elaboração propria.

Essas proporções segundo [7] são dadas pela fórmula: $P(d) = \log_{10}(1+1/d)$, $d \in \{1, 2, 3..., 9\}$.

Desta forma em muitos campos distintos de conhecimento Benford pode ser aplicado e na análise de dados eleitorais pode-o ser também. É imprescindível que, no atual ambiente de variações políticas, crises econômicas, escândalos e as incertezas que a administração pública brasileira vem enfrentando, ferramentas e estudos estatísticos possam oferecer uma metodologia científica indispensável no processo decisório e transparência fidedigna dos resultados gerais das eleições [20].

Quando se trata da análise de dados eleitorais sobre a luz de Benford, segundo [23] é preferível utilizar o segundo dígito mais significativo uma vez que esta lei, em seu julgamento, não funciona bem quando os números na base de dados são limitados. Isso ocorre com as urnas eletrônicas segundo o mesmo autor, pois há um limite no número de eleitores em cada seção eleitoral já que a alocação máxima é de 400 eleitores em cada seção nas cidades do interior e 500 eleitores nas capitais dos estados do país. Já [16] destaca que dados de contagem de votos é um processo que se assemelha fortemente à distribuição da lei de Benford quando estudados os segundos dígitos.

No entanto foi utilizado neste trabalho como dígito mais significativo o primeiro, pois o objetivo é verificar de forma indiscriminada os votos por seção sem fazer distinção entre candidatos e/ou cargos, assim será verificado a consonância entre todos os votos computados e ao final será feito um paralelo entre os trabalhos que utilizaram outras abordagens afim de julgar a assertividade do processo utilizado aqui.

3.1.1 Teste Qui-quadrado

O teste Qui-quadrado, também conhecido como teste X^2 (qui ao quadrado), é uma técnica estatística usada para determinar se existe uma associação significativa entre duas variáveis categóricas. Ele é amplamente utilizado em estatística e pesquisa em diversas áreas, como ciências sociais, medicina, biologia, entre outras [2].

Segundo [3], o teste é utilizado para verificar se a frequência com que um determinado acontecimento observado em uma amostra se desvia significativamente ou não da frequência com que ele é esperado. Assim como também observam que os graus de liberdade no teste são determinados pelo número de categorias nas duas variáveis. Geralmente, é calculado como (número de linhas - 1) * (número de colunas - 1). Com os graus de liberdade e o valor da estatística, podemos calcular o valor-p.

O valor-p é fundamental para a interpretação do teste. Se o valor-p for menor que um nível de significância pré-definido (geralmente 0,05), isso sugere que as duas variáveis não são independentes e que há uma associação estatisticamente significativa entre elas. Nesse caso, a hipótese nula é rejeitada em favor da hipótese alternativa, concluindo que existe uma relação entre as variáveis categóricas.

No entanto existem duas principais variações do teste Qui-quadrado: o teste Qui-quadrado de independência e o teste Qui-quadrado de ajuste. O teste de independência é usado quando se deseja determinar se duas variáveis categóricas são independentes uma da outra, enquanto o teste de ajuste é usado para verificar se uma amostra segue uma distribuição de probabilidade teórica, como por exemplo as estimativas da lei de Benford.

Em resumo, o teste Qui-quadrado é uma ferramenta versátil e fundamental na análise de dados. Ele ajuda a responder perguntas sobre a relação entre variáveis e é uma parte essencial da caixa de ferramentas estatísticas para pesquisadores e analistas de dados em diversas áreas do conhecimento.

3.1.2 Teste Z

O teste estatístico Z é uma técnica estatística que compara uma medida observada em seus dados com uma distribuição normal padrão. Ele é usado para avaliar se uma amostra é significativamente diferente da população da qual foi retirada. O valor Z indica quantos desvios padrão uma medida está em relação à média da distribuição normal padrão [5].

A fórmula para calcular o valor Z é:

$$Z = \frac{(X - \mu)}{\sigma}$$

onde:

- X é a medida observada na amostra,
- μ é a média da população,
- σ é o desvio padrão da população.

Para exemplificar, suponha-se que ao estudar a altura de uma amostra de estudantes de uma determinada escola. A altura média da população é conhecida por ser 170 cm, com um desvio padrão de 10 cm. Se um estudante específico tem 175 cm de altura, pode-se calcular o valor Z para verificar o quão longe essa medida está da média da população:

$$Z = \frac{(175 - 170)}{10} = 0.5$$

Isso significa que a altura do estudante está 0.5 desvios padrão acima da média da população. A partir dessa observação pode-se consultar tabelas estatísticas Z para determinar a probabilidade associada a esse valor Z específico. Se a probabilidade for suficientemente baixa geralmente abaixo de 0,05, a hipótese nula é rejeitada, pois a altura do estudante não é significativamente diferente da média da população.

O teste estatístico Z é comumente usado em situações em que a distribuição das amostras é conhecida e assume uma distribuição normal. É uma ferramenta valiosa na inferência estatística, especialmente em testes de hipóteses. Para a proposta deste trabalho o valor de *score* adotado foi o intervalo entre -1,96 a 1.96 com um grau de significância de 0.05%.

3.2 Coeficiente de Gini

Desenvolvido pelo estatístico italiano Corrado Gini, o coeficiente de Gini é um índice que proporciona uma medida numérica variando de 0 a 1. O valor 0 representa uma distribuição perfeitamente igualitária, enquanto o valor 1 indica uma distribuição totalmente desigual. Este índice vai além do âmbito econômico, pois, ao oferecer uma perspectiva quantitativa da desigualdade, torna-se uma ferramenta valiosa para pesquisadores, formuladores de políticas e analistas. Sua utilidade reside na capacidade de proporcionar uma compreensão mais precisa das disparidades socioeconômicas, orientando a formulação de estratégias eficazes para promover a equidade. O coeficiente de Gini, aliado à curva de Lorenz, são medidas empregadas para avaliar a desigualdade na distribuição de uma variável, sendo comumente aplicadas em contextos econômicos para analisar

a distribuição de renda ou riqueza entre os membros de uma população. Ambas as medidas fornecem uma abordagem quantitativa para avaliar e expressar a disparidade na distribuição de uma característica específica [28].

Sendo a curva de Lorenz um gráfico que ilustra a distribuição cumulativa percentual de uma variável em relação à distribuição cumulativa percentual da população. Ela é construída plotando a proporção acumulativa da variável. Um exemplo pode ser a renda contra a proporção acumulativa da população. Se a distribuição fosse completamente igual, a curva de Lorenz seria uma linha diagonal, indicando que cada segmento igual de população detém uma proporção igual da variável em questão [28].

Já o coeficiente de Gini é um número entre 0 e 1 que resume a informação contida na curva de Lorenz em uma única medida. Ele é calculado como a razão entre a área, entre a curva de Lorenz e a linha de igualdade, área de desigualdade e a área total abaixo da linha de igualdade. Um coeficiente de Gini de 0 representa uma distribuição perfeitamente igual, enquanto um coeficiente de 1 indica extrema desigualdade, onde uma única unidade detém toda a variável em questão.

A fórmula do coeficiente de Gini é:

$$G = \frac{A}{A + B}$$

Onde:

- G é o coeficiente de Gini,
- A é a área entre a curva de Lorenz e a linha de igualdade,
- B é a área abaixo da linha de igualdade.

Em resumo, a curva de Lorenz e o coeficiente de Gini são ferramentas importantes para analisar a desigualdade em distribuições de variáveis econômicas, sendo amplamente utilizados para entender a disparidade na distribuição de renda ou riqueza em uma sociedade.

3.3 Sistema Eleitoral Brasileiro

O Sistema Eleitoral Brasileiro é baseado em princípios democráticos e visa a representatividade política da população. Utiliza o voto como meio de escolha dos representantes em diversos níveis de governo, incluindo a Presidência da República, o Congresso Nacional, as Assembleias Legislativas Estaduais e as Câmaras Municipais. O Brasil adota o sistema de voto proporcional para eleições legislativas, onde os partidos recebem cadeiras de acordo com a proporção de votos obtidos.

As eleições presidenciais ocorrem a cada quatro anos, enquanto as eleições legislativas e municipais têm intervalos menores. O voto é obrigatório para cidadãos entre 18 e 70 anos, sendo facultativo para jovens de 16 e 17 anos, idosos acima de 70 anos e analfabetos. O eleitor vota em candidatos específicos para cargos proporcionais e majoritários.

No sistema proporcional, os votos dados aos partidos são contabilizados para a distribuição de cadeiras, levando em consideração a votação total e as coligações formadas entre partidos. Já no sistema majoritário, como nas eleições presidenciais e para alguns cargos no legislativo, o candidato que obtiver a maioria dos votos é eleito [9].

O Brasil também utiliza o sistema de dois turnos nas eleições presidenciais e municipais, exigindo que um candidato obtenha mais de 50% dos votos válidos para ser eleito no primeiro turno. Caso nenhum candidato alcance esse percentual, os dois mais votados disputam um segundo turno.

Além disso, o país adota medidas para promover a participação política das mulheres e de minorias, como a reserva de vagas para candidaturas femininas e a distribuição equitativa do Fundo Partidário. A Justiça Eleitoral é responsável por organizar e fiscalizar o processo eleitoral, garantindo a lisura e a transparência das eleições [12].

4. MATERIAIS E MÉTODOS

Nesta seção será apresentada as ferramentas utilizadas e o modo como foram utilizadas neste trabalho para a análise dos dados e posteriormente apresentado os resultados colhidos no tópico seguinte.

4.1 Descrição da Base de Dados

Os dados utilizados neste artigo foram extraídos do site oficial do Tribunal Regional Eleitoral do Ceará (TRE-CE), disponível em http://www.tre-ce.jus.br/www.tre-ce.jus.br, que serve como um portal de dados e consulta pública. O TRE-CE desempenha um papel essencial como órgão da Justiça Eleitoral brasileira, responsável por supervisionar e organizar as eleições no estado do Ceará desde sua criação, contribuindo significativamente para a promoção da democracia e facilitação do exercício do direito de voto. Além disso, o tribunal atua na garantia da lisura das eleições e transparência no processo eleitoral, reforçando seu compromisso com a integridade do sistema democrático.

O acervo do TRE-CE é completo e robusto possui em sua base dados eleitorais desde o ano de 1930 até as ultimas eleições, estes dados estão separados por seções muito bem catalogadas possuindo arquivos eletrônicos do tipo: *.csv*, *JSON*, *.pdf* e *.jpg*. Essa organização é de profunda importância pois auxilia os pesquisadores que a procuram com o intuído de estudar algo relacionado a estes dados. São exemplos de estudo [1] e [11] que usaram essa base como apoio de suas pesquisas. [1] utilizou os dados desta base para avaliar os financiamentos políticos e verificar seus impactos e possíveis lacunas de fiscalização. Já [11] utilizou essa base para examinar a Justiça Eleitoral no Brasil, abordando a organização, composição, funções e competências conforme os parâmetros estabelecidos na Constituição.

A base de sustentação deste artigo se volta para as eleições gerais brasileiras, com especial ênfase no estado do Ceará durante o pleito de 2022, cujo propósito era eleger candidatos para os cargos de Deputado Estadual, Deputados Federal, Senador, Governador e Presidente. O estado do Ceará abrange 184 municípios, distribuídos por 109 zonas eleitorais em áreas urbanas e rurais, totalizando 25.064 seções eleitorais. O colégio eleitoral do estado alcança um total de 6.819.976 eleitores, sendo a capital, com 1.869.135 votantes, o município com o maior contingente eleitoral. O conjunto de dados compreende 185 arquivos no formato .csv, dos quais 184 correspondem aos municípios do estado, e 1 abrange a votação geral do estado. Esses arquivos apresentam 11 colunas com informações diversas, como cargo, número do candidato e nome do candidato, sendo que, para os propósitos deste trabalho, será focalizada exclusivamente a coluna referente aos votos computados por município.

A escolha desse conjunto de dados permitiu uma avaliação abrangente da aplicabilidade da lei de Benford, pois é possível organiza-los da melhor forma possível. As eleições de 2022 contaram com número total de candidatos passíveis de

votos igual a 991, onde existiam seis chapas de candidatos a governador, também seis candidatos ao senado, 407 candidatos a deputado federal e 554 candidatos a deputado estadual, todos esses concorrendo a 73 vagas, sendo uma para governador, uma para senador, 22 para deputados federais e 46 para deputados estaduais.

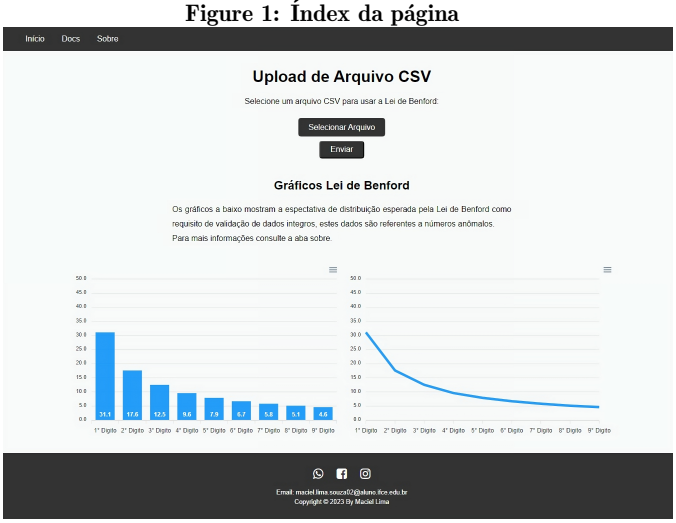
A análise dos dados foi realizada de acordo com a metodologia estabelecida por [10] e [20], para cada conjunto de dados os primeiros dígitos significativos de cada número dos votos por seção foram registrados pela aplicação e analisados, após isso a frequência observada de cada dígito foi calculada e comparada com a distribuição esperada e já conhecida da lei de Benford, após isso a página retorna o resultado da análise sobre os dados para o usuário.

4.2 Metodologia de Implementação da Ferramenta Utilizada

A página *web* utilizada como método de avaliação da lei de Benford sobre os dados eleitorais foi construída com as tecnologias *HTML*, *CSS* e *Javascript*, tem por nome *Received data* encontrada através do *link*: https://receiveddata.netlify.app. Esta página tem por objetivo carregar dados através de um *input* aceitando arquivos *.csv* disponibilizados pelo usuário. Vale salientar que a página em questão, por padrão, só valida a primeira coluna do arquivo por isso há a necessidade de tratar os dados antes de utilizar a página, no mais, esse arquivo pode possuir cabeçalho ou não pois ele será desconsiderado pela aplicação.

A Figura 1 abaixo exibe a visão inicial da página *web* desenvolvida, nela é possível observar um campo de entrada *(input)* destinado ao carregamento de dados, bem como dois gráficos que ilustram a distribuição desses dados de acordo com a lei de Benford.

Figure 1: Índex da página



Elaboração própria.

Após o carregamento dos dados a página faz o processamento destes avaliando somente o primeiro dígito de cada número, esse número avaliado corresponde a quantidade de votos computados por cada seção. Posteriormente a página trabalha os dados para após esse processamento *plotar* gráficos através da biblioteca *apexcharts* mostrando uma projeção desses em relação a expectativa esperada com lei de Benford através de um gráfico do tipo barras e outro do tipo linha, a partir desses gráficos é possível avaliar a integridade desses dados de forma visual basta verificar os valores apresentados nos gráficos com a expectativa esperada pela lei de Berford e avaliar essa condição.

Portanto o teste de adesão envolve a comparação das frequências observadas com as frequências esperadas para os dígitos de 1 a 9. No entanto havendo discrepâncias entre os valores obtidos e a expectativa não se pode concluir que houve fraude ou algo do tipo, mas apenas que há alguma anomalia nos dados divergentes e que há a necessidade de uma avaliação minuciosa nestes para confirmar ou negar as hipóteses levantadas.

4.3 Implementação do Método de Avaliação Qui-quadrado

Neste estudo, optou-se pelo uso do teste Qui-quadrado de ajuste, uma escolha justificada pela melhor adequação desse método ao objeto de análise dos dados. Vale ressaltar que, o valor-p no teste Qui-quadrado é fundamental para a interpretação do teste, determinando se os dados fornecem evidências estatísticas suficientes para rejeitar a hipótese nula de igualdade entre a distribuição observada e a distribuição esperada. Entretanto, é crucial comparar os dados obtidos com essa frequência esperada, para o qual foi adotado um limiar de tolerância de 0.05% para um valor crítico de 15.51 dentro de um universo de grau de liberdade igual a 8. Este limiar representa a margem de erro convencional associada à técnica do teste Qui-quadrado. A seguir é apresentada a função responsável por realizar esse processo e fornecer o valor do limiar encontrado para cada dígito:

Figure 2: Função que valida o teste Qui-quadrado

```
/*------------ Test Qui-quadrado ----------------*/

const testQui = (observedFrequencies, sampleSize, benfordDistribution) => {

    const expectedFrequencies = benfordDistribution.map(digit => digit * sampleSize)

    let chiSquare = 0
    for (let i = 0; i < 9; i++) {
        chiSquare += Math.pow(expectedFrequencies[i] - observedFrequencies[i], 2) / expectedFrequencies[i]
    }

    if (sampleSize < 1000) {
        return (chiSquare / 100).toFixed(2)
    } else if (sampleSize < 10000) {
        return (chiSquare / 1000).toFixed(2)
    } else {
        return (chiSquare / 10000).toFixed(2)
    }
}
```

Elaboração própria.

A Figura 2 acima apresenta a função que, ao receber a frequência observada e a frequência esperada para cada dígito como parâmetros, calcula o valor do teste Qui-quadrado para os dados. No entanto, o retorno da função exibe o valor obtido em relação aos dados testados, deixando ao usuário a responsabilidade de interpretar a adequação do mesmo. A página destaca que o limiar de erro estabelecido é de 0.05%, informando que se o valor calculado for superior a esse limiar, a probabilidade de os dados estarem em desacordo é mais elevada, indicando uma disparidade significativa. Em outras palavras, quanto maior for a diferença entre o valor obtido e 0.05%, maior é a possibilidade de existir uma divergência substancial entre os dados.

4.4 Implementação dos Métodos de Avaliação Qui-quadrado e o Teste Z

A implementação do teste estatístico Z neste trabalho foi elaborado com o objetivo de investigar a possível discrepância entre a média de uma amostra de dados e a média de dados já conhecidos, estabelecendo uma conexão relevante com a lei de Benford. A escolha do teste Z de uma amostra foi motivada pela preexistência do conhecimento da variân-cia dos dados. Inicialmente, uma amostra representativa foi coletada, e as médias amostral e populacional foram calculadas. A hipótese nula foi formulada considerando a inexistência de diferença significativa, enquanto a hipótese alternativa refletiu a presença de uma disparidade estatisticamente relevante. Este contexto é particularmente relevante em análises que buscam entender se as distribuições de dados seguem padrões naturais, como os propostos pela lei de Benford

A Figura 3 abaixo apresenta uma função em *JavaScript* que implementa o teste Z. Essa função é projetada para calcular o valor Z com base em uma amostra e uma média populacional conhecida, permitindo assim a avaliação da significância estatística da diferença entre a média amostral e a média populacional. Essa implementação é particularmente u ´til em contextos de desenvolvimento *web* e análise de dados, proporcionando uma abordagem eficaz para a realização de testes estatísticos e apoio à tomada de decisões embasadas em dados quantitativos.

```
/*-----------------Teste Z--------------------------*/

function testZ(observedFrequencies, sampleSize, benfordDistribution) {

    const expectedFrequencies = benfordDistribution.map(digit => digit * sampleSize)

    let numerator = 0
    let denominator = 0

    for (let i = 0; i < 9; i++) {
        numerator += Math.pow(observedFrequencies[i] - expectedFrequencies[i], 2)
        denominator += expectedFrequencies[i]
    }

    const standardDeviation = Math.sqrt(denominator / 9)
    const zScore = numerator / Math.pow(standardDeviation, 2)

    if (sampleSize < 1000) {
        return (zScore / 100).toFixed(2)
    } else if (sampleSize < 10000) {
        return (zScore / 1000).toFixed(2)
    } else {
        return (zScore / 10000).toFixed(2)
    }
}
```

Elaboração própria.

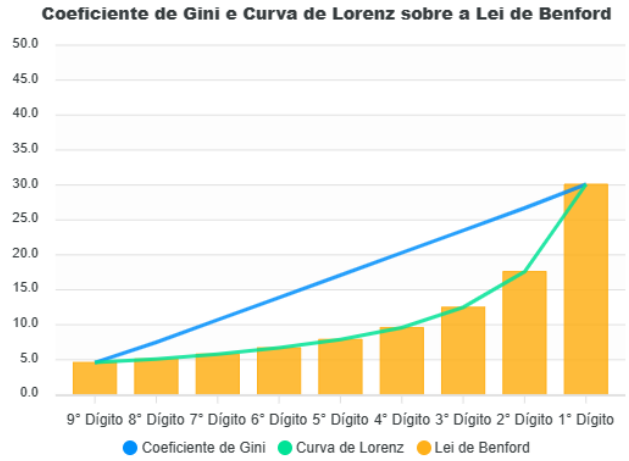
Durante o desenvolvimento do estudo, a aplicação do teste estatístico Z também permitiu uma comparação sistemática entre os resultados obtidos e as expectativas derivadas da lei de Benford. A interpretação dos resultados foi conduzida por meio da comparação do valor calculado do teste Z com o valor crítico associado a um nível de significância predeterminado, comumente estabelecido em 0.05% dentro de uma margem de significância de *score* entre -1.96 a 1.96. Isso proporcionou uma avaliação não apenas da significância estatística da diferença de médias, mas também da coerência dos dados com os padrões esperados pela lei de Benford.

4.5 Método de Avaliação Coeficiente de Gini e Curva de Lorenz

Nesta implementação o coeficiente de Gini e a curva de Lorenz foram utilizadas como ferramentas analíticas para avaliar a desigualdade no conjuntos de dados, estabelecendo uma conexão intrínseca com a lei de Benford. O coeficiente de Gini, comumente utilizado para mensurar a desigualdade na distribuição de riqueza, foi aplicado aqui para quantificar a disparidade entre os valores observados e os ideais propostos pela lei de Benford. A curva de Lorenz, por sua vez, proporciona uma representação visual da distribuição acumulada dos dados, permitindo a identificação de desvios significativos em relação às expectativas benfordianas.

Na Figura 4, o comportamento do coeficiente de Gini e da curva de Lorenz é vividamente ilustrado de forma gráfica, proporcionando *insights* valiosos sobre a interação dessas duas métricas quando são aplicadas em conjunto com a lei de Benford. A representação visual não apenas facilita uma análise clara da relação entre essas métricas, mas também destaca padrões e variações que emergem ao se contemplar a distribuição dos dados. Essa apresentação visual enriquece significativamente a compreensão da dinâmica entre o coeficiente de Gini, a curva de Lorenz e a aplicação da lei de Benford, proporcionando uma abordagem mais completa para compreender a distribuição e as nuances subjacentes nos dados em consideração.

Figure 4: Gráfico do coeficiente de Gini, Curva de Lorenz e lei de Benford



Elaboração própria.

Na página *web* desenvolvida o gráfico implementado para a lei de Benford mostra um desvio de 0,347 e portanto essa medida foi adotada como padrão para o trabalho, ou seja, ao verificar os dados a página faz uma distribuição desses dígitos de forma crescente plotando o gráfico e verificando quanto que a curva de Lorenz se distância do coeficiente de Gini. Como margem de erro foi adotando um percentual de 0.05% assim como também para o teste do Qui-quadrado e o teste estatístico Z.

5. RESULTADOS E DISCUSSÕES

Foram avaliados os votos computados de todas as cidades do estado do Ceará para todos os cargos eletivos possíveis (Deputado Estadual, Deputado Federal, Senador, Governador e Presidente) segundo a lei de Benford e dentro dos parâmetros incluídos na construção da página que foi pro-

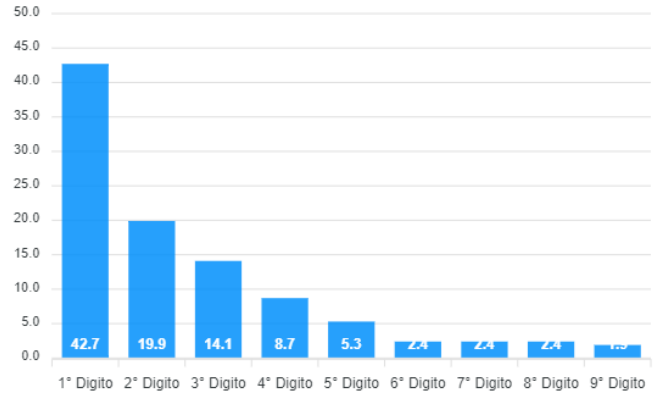
posta para isso, depois de aplicar em todos os municípios, foram plotados os gráficos e a classificação que a página retornou para que a partir desses gráficos e classificação fosse possível observar os resultados obtidos e avaliar o quando a aplicação é eficiente e ao mesmo tempo verificar a confiabilidade dos dados.

Dito isso, ao observar os gráficos é possível perceber que todos mostram uma inclinação a lei de Benford, até mesmo os que são classificados pela aplicação como não estando dentro do limiar de variação de 0,01 da lei de Benford que foi usado como margem de erro, e do teste Qui-quadrado que usa o valor crítico de 15,51 com grau de liberdade igual a 8, dentro da margem da significância de 0.05%, além do teste estatístico Z que tem como *score* a variância de -1,96 a 1.96 dentro da margem da significância de 0.05%.

O município de Abaixara serve como um exemplo ilustrativo de carregamento de dados que, ao ser submetido à avaliação na página, revela visualmente perturbações nas distribuições dos dígitos conforme os princípios da lei de Benford. Este cenário é fortalecido pela análise conjunta dos testes Qui-quadrado e Z, juntamente com a consideração do coeficiente de Gini e da curva de Lorenz. Essas avaliações combinadas fornecem uma visão sobre a presença de anomalias nos dados, destacando a importância de múltiplos indicadores para uma compreensão mais profunda e confiável das características estatísticas do conjunto de dados.

A Figura 5 apresenta a porcentagem da distribuição dos dígitos encontrados fazendo uma comparação com a lei de Benford.

Figure 5: Distribuição de dígitos segundo a Lei de Benford Abaixara.



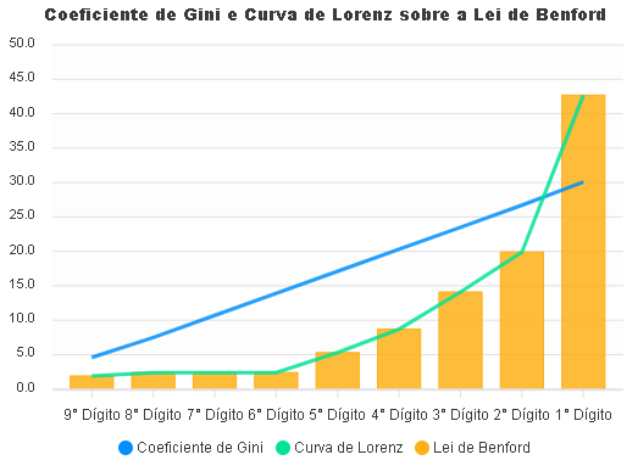
Elaboração própria.

A Figura 6 apresenta a porcentagem da distribuição dos dígitos encontrados traçando o coeficiente de Gini com a curva de Lorenz em conjunto com a distribuição sobre a lei de Benford.

Já a Figura 7 apresenta a avaliação que a página encontrou para os dados apresentados, mostra também em quais dígitos há a disparidade com a lei de Benford, assim com os valores calculados para a curva de Lorenz, teste Qui-quadrado e teste Z.

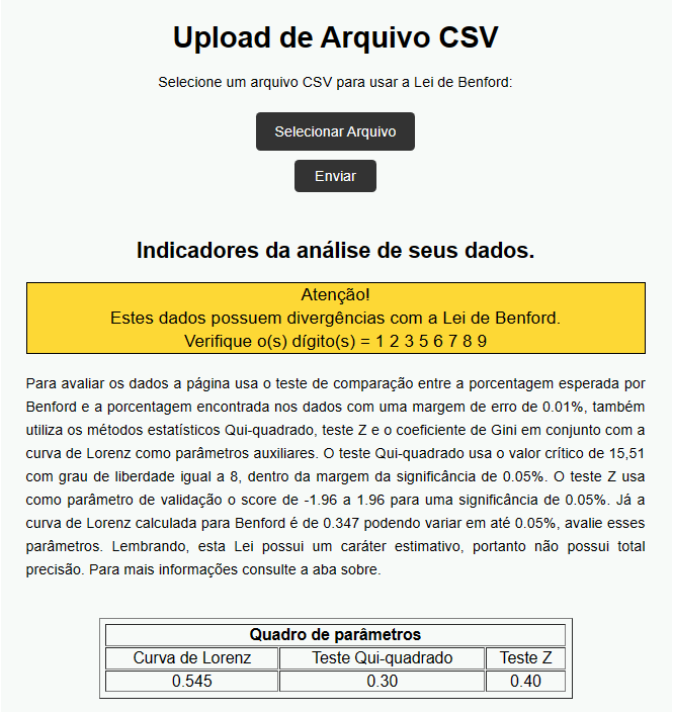
Dos dados analisados a aplicação retornou 184 municípios como não possuindo conformidade com Benford, ou seja, todos os municípios avaliados mostraram pelo menos um dígito em não conformidade, no entanto é possível considerá-

Figure 6: Distribuição de dígitos segundo coeficiente de Gini e Curva de Lorenz Abaixara



Elaboração própria.

Figure 7: Retorno da avaliação da página sobre os dados de Abaixara



Elaboração própria.

los passíveis de avaliações minuciosa para determinar algo sobre sua veracidade.

Ponto a ser considerado é a quantidades de candidatos que receberam votos nos municípios, pois em um universo de 991 candidatos possíveis (Deputado Estadual, Deputado Federal, Senador, Governador e Presidente), em muitos municípios vários não tiveram votos computados e devido a isso causando impacto na avaliação da porcentagem de cada dígito ao ser avaliado, pois os candidatos que não obtiveram votos são desconsiderados para efeito de avaliação já que so-

mente o primeiro dígito é considerado como significativo, ou seja, o dígito mais significativo do número encontra-se entre 1º e 9º, portanto para que a aderência fosse melhor a quantidade de candidatos votados a mesma se mostrou relevante neste estudo, pois nos municípios onde um números maior de candidatos foram votados a aderência se mostrou mais eficaz, no entanto somente nos votos consolidados o retorno foi positivo para os testes.

Outra observação pertinente diz respeito ao testes estatísticos, pois para estes exemplos o percentual de 0.05 é mais eficaz quando há uma quantidade maior de dados, ou seja, quando a base de avaliação é pequena ele deixa a desejar em relação a avaliação, no entanto quando a base de avaliação é maior sua aderência é excelente, como o objetivo do trabalho é verificar se a lei de Benford cumpre seu papel na avaliação de dados eleitorais essa classificação tem peso secundário, pois Benfort na plotagem do gráfico por si só já é uma avaliação dos dados observados.

Por fim, os dados analisados neste artigo mostram-se pertinentes ao universo da análise de votos por Benford, pois é possível avaliá-los em contraponto a outros estudos, pois a aderência é uma marca essencial na detecção de possíveis fraudes uma vez que quando há alguma deformação nessa métrica essa deformação se mostra muito evidente. Sobre os dados aqui estudados era esperado que essas deformações não ocorressem pois como é sabido, ou não há provas cabais contrárias, as eleições no Brasil são seguras e confiáveis, portanto a ferramenta cumpre bem seu papel podendo assim auxiliar a outros estudos futuros que por ventura a utilizem.

6. CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho buscou aplicar a lei de Benford para avaliar dados eleitorais do estado do Ceará nas ultimas eleições gerais do país no ano de 2022 através de uma página *web* desenvolvida para esse propósito, a página se mostrou eficiente ao estudo da lei que é o propósito principal desse trabalho, no entanto para os dados propostos aparentemente essa abordagem não é a ideal pois a quantidade de retornos negativos se mostrou elevado. Para [13] a melhor forma de avaliar dados eleitorais sobre a visão de Benford é utilizando o segundo dígito, portanto é passível considerar que realmente a utilização do primeiro para esse tipo de estudo não se mostrou a melhor opção já que a margem de erro foi alta, mesmo assim, a aplicação é eficaz em sua proposta.

Os próximos passos serão avaliar melhorias na página transformando essa em uma aplicação *web* incrementando mais funcionalidades para que o usuário possa escolher qual dígito ele irá utilizar para validar sua base e assim minimizar o impacto dos testes estatísticos no início e posteriormente utilizar outras ferramentas para avaliar os resultados.

7. REFERENCES

[1] L. C. Akl. *Financiamento político subnacional, impacto e lacunas na fiscalização*. PhD thesis, 2023.

[2] J. A. Carneiro Filho and J. A. Falk. Impacto do controle sobre probabilidade de fraudes em dois municípios de pernambuco: Lei newcomb-benford. *Revista do TCU*, (149):200–216, 2022.

[3] A. CORREA, E. QUEIROZ, and N. TREVISAN. Teste do qui-quadrado, 2020, 2020.

[4] J. I. d. F. Costa, W. Silva, S. Travassos, and J. Santos. Análise de conformidade da lei de newcomb-benford no ambiente de auditoria contínua: uma proposta de identificação de desvios no tempo. *Anais do 37º Encontro Nacional da Associação Nacional de Pós-Graduação e Pesquisa em Administração*, 2013.

[5] F. C. R. d. Cunha. Aplicações da lei newcomb-benford à auditoria de obras públicas. 2014.

[6] V. H. da Fonseca, D. M. S. Nunes, and C. M. Santana. Amostragem: Conhecimento e uso em empresas de auditoria. *RAGC*, 4(16), 2016.

[7] H. C. da Silva. Equidistribuição e lei de benford em progressões geométricas. *Revista de Matemática*, 3(03):75–86, 2022.

[8] M. A. C. d. Freitas. A aplicação da lei newcomb-benford na auditoria governamental das despesas liquidadas do metrô-df. 2013.

[9] E. D. Gomes and B. B. Lechenakoski. Direito eleitoral e democracia: a problemática em torno do sistema eleitoral brasileiro. *Academia de Direito*, 5:191–217, 2023.

[10] T. P. Hill. Base-invariance implies benford's law. *Proceedings of the American Mathematical Society*, 123(3):887–895, 1995.

[11] A. T. N. Júnior. A justiça eleitoral no brasil e a garantia da democracia. *Revista Controle-Doutrina e Artigos*, 21(2):95–111, 2023.

[12] H. d. C. I. Júnior and F. F. M. Cunha. Crise de representação? uma análise contemporânea do contexto político-eleitoral brasileiro. *Revista Contemporânea*, 3(1):67–86, 2023.

[13] P. C. Lemos. Uma avaliação da qualidade dos dados reportados sobre financiamento de campanha com base na lei de benford. 2011.

[14] N. L. Madureira et al. Aplicabilidade da lei de benford na análise de um conjunto de dados eleitorais. 2012.

[15] L. M. Mangueira. Avaliação da redistribuição de renda pelas gratuidades do transporte público metroviário do distrito federal. 2023.

[16] W. R. Mebane Jr et al. Detecting attempted election theft: vote counts, voting machines and benford's law. In *Annual Meeting of the Midwest Political Science Association, Chicago, IL, April*, pages 20–23, 2006.

[17] R. O. Menezes. Aplicação da lei de benford nos números de casos confirmados de covid-19 em diferentes países. *REMAT: Revista Eletrônica da Matemática*, 7(1):e3005–e3005, 2021.

[18] S. S. Mizrahi. As últimas eleições e a lei de benford (ou lei do primeiro dígito). *American Journal of Mathematics*, 4(1):39–40, 1881.

[19] E. Muchine, E. Mabjaia, A. Vilanculos, and F. Mahaluça. Aplicabilidade da teoria das probabilidades na auditoria: Lei de newcomb-benford como procedimento para análise de conformidade de dados. *Revista da UI_IPSantarém*, 11(2):224–234, 2023.

[20] M. J. Nigrini. *Benford's Law: Applications for forensic accounting, auditing, and fraud detection*, volume 586. John Wiley & Sons, 2012.

[21] M. J. Nigrini and S. J. Miller. Data diagnostics using second-order tests of benford's law. *Auditing: A Journal of Practice & Theory*, 28(2):305–324, 2009.

[22] W. D. Nordhaus. Paul samuelson and global public

goods. *Samuelsonian Economics*, pages 88–98, 2006.
[23] É. d. S. G. Rabelo. A lei de benford e fraudes eleitorais: o caso das eleições presidenciais brasileiras de 2014. 2016.
[24] B. M. Schaefer. Autofinanciamento eleitoral no brasil: regulação, causas e consequências. 2022.
[25] J. A. B. d. Silva. Gastos públicos federais com emendas parlamentares: uma análise à luz da lei de benford. 2022.
[26] M. R. Silva. Imposto de renda de pessoa física brasileiro: uma análise de progressividade e efeito redistributivo. 2022.
[27] R. R. O. d. Silva. Em busca de transparência: a lei de benford aplicada às despesas eleitorais. 2015.
[28] S. Vietri and S. Del Duca. Curva de lorenz1.