Henrique do Nascimento Silva

Uso de Invariância de Base em Aplicações da Lei de Newcomb-Benford

Santo André, 2024

UNIVERSIDADE FEDERAL DO ABC

CENTRO DE MATEMÁTICA, COMPUTAÇÃO E COGNIÇÃO

HENRIQUE DO NASCIMENTO SILVA

USO DE INVARIÂNCIA DE BASE EM APLICAÇÕES DA LEI DE NEWCOMB-BENFORD

Orientador: Prof. Dr. Eduardo Gueron

Dissertação de mestrado apresentada ao
Centro de Matemática, Computação e Cognição
para obtenção do título de Mestre pelo Programa de
Mestrado Profissional em Matemática em Rede
Nacional (PROFMAT).

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO
DEFENDIDA PELO ALUNO HENRIQUE DO NASCIMENTO SILVA
E ORIENTADA PELO PROF. DR. EDUARDO GUERON.

SANTO ANDRÉ, 2024

Sistema de Bibliotecas da Universidade Federal do ABC
Elaborada pelo Sistema de Geração de Ficha Catalográfica da UFABC
com os dados fornecidos pelo(a) autor(a).

do Nascimento Silva, Henrique
Uso de invariância de base em aplicações da lei de Newcomb-Benford / Henrique do Nascimento Silva. — 2023.

81 fls. : il.

Orientador: Eduardo Gueron

Dissertação (Mestrado) — Universidade Federal do ABC, Mestrado Profissional em Matemática em Rede Nacional - PROFMAT, Santo André, 2023.

1. Lei de Newcomb-Benford, invariância de escala, Fraude Contábil. I. Gueron, Eduardo. II. Mestrado Profissional em Matemática em Rede Nacional - PROFMAT, 2023. III. Título.

Este exemplar foi revisado e alterado em relação à versão original, de acordo com as observações levantadas pela banca examinadora no dia da defesa, sob responsabilidade única da autora e com a anuência do orientador.

MINISTÉRIO DA EDUCAÇÃO
Fundação Universidade Federal do ABC
Avenida dos Estados, 5001 – Bairro Santa Terezinha – Santo André – SP
CEP 09210-580 · Fone: (11) 4996-0017

FOLHA DE ASSINATURAS

Assinaturas dos membros da Banca Examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato, HENRIQUE DO NASCIMENTO SILVA realizada em 28 de Novembro de 2023:

Documento assinado digitalmente
gov.br
DENILSON GOMES
Data: 29/11/2023 14:56:31-0300
Verifique em https://validar.iti.gov.br

_____
Prof.(a) DENILSON GOMES
UNIVERSIDADE FEDERAL DE SANTA MARIA

Documento assinado digitalmente
gov.br
SANDRA MARIA ZAPATA YEPES
Data: 29/11/2023 23:05:16-0300
Verifique em https://validar.iti.gov.br

_____
Prof.(a) SANDRA MARIA ZAPATA YEPES
UNIVERSIDADE FEDERAL DO ABC

_____
Prof.(a) LEONARDO PAULO MAIA
UNIVERSIDADE DE SÃO PAULO

_____
Prof.(a) RAFAEL DE MATTOS GRISI
UNIVERSIDADE FEDERAL DO ABC

Documento assinado digitalmente
gov.br
EDUARDO GUERON
Data: 30/11/2023 00:47:13-0300
Verifique em https://validar.iti.gov.br

_____
Prof.(a) EDUARDO GUERON
UNIVERSIDADE FEDERAL DO ABC - Presidente

* Por ausência do membro titular, foi substituído pelo membro suplente descrito acima: nome completo, instituição e assinatura

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoas de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

Dedico este trabalho aos meus pais, Jeremias e Maria.

AGRADECIMENTOS

Agradeço profundamente a Deus pela saúde e resiliência que Ele me concedeu, permitindo-me concluir este trabalho. Em momentos de dificuldade, pensei em desistir, mas a fé me deu forças para persistir e revelou o meu potencial.

Quero expressar minha gratidão ao meu orientador, Eduardo Gueron. Sua motivação, paciência e constante apoio foram fundamentais. Ele não apenas orientou, mas também estimulou novas ideias, encorajou experiências e me auxiliou a refinar e pesquisar, avaliando minhas fontes. Suas contribuições foram cruciais para o desenvolvimento deste trabalho.

À minha turma do PROFMAT, mesmo com a pandemia de 2020 limitando nosso contato, construímos boas amizades, trocando conhecimento e motivação. Agradeço por essa amizade que não esquecerei.

Não posso deixar de mencionar meus pais, Jeremias e Maria, e minha irmã, Lilian, pelo apoio incondicional que sempre me deram.

Por último, mas não menos importante, agradeço aos amigos e professores que foram essenciais em minha jornada acadêmica e na vida em geral. Suas palavras e incentivo foram como luzes no meu caminho.

A todos vocês, meu sincero agradecimento.

“Eu só posso lhe mostrar a porta. Você tem que atravessá-la.”

(Morpheus, *Personagem do filme Matrix*)

RESUMO

Neste trabalho, a Lei de Newcomb-Benford (LNB), uma abordagem probabilística que analisa a distribuição do primeiro dígito em números, independentemente de sua origem, é explorada. São apresentadas definições, aspectos probabilísticos e históricos que fundamentam a LNB, e a lei é utilizada para a análise de dados contábeis relacionados à execução orçamentária das despesas decorrentes da pandemia de COVID-19, empregando o método da proporção de dígitos. Os resultados indicaram que os dados não seguem a distribuição esperada pela LNB. Além disso, no estudo, a generalização da LNB e suas propriedades, incluindo a invariância na mudança de base e de escala, são abordadas, bem como é proposta uma abordagem pedagógica para a educação básica.

Palavras-chave: Lei de Newcomb-Benford, invariância de escala, Fraude Contábil

ABSTRACT

In this study, we explore the Newcomb-Benford Law (NBL), a probabilistic approach that examines the distribution of the first digit in numbers, regardless of their origin. Definitions, probabilistic aspects, and historical foundations supporting the NBL are presented, and the law is employed to analyze accounting data related to the budget execution of expenses resulting from the COVID-19 pandemic, using the digit ratio method. The results indicated that the data do not adhere to the expected distribution according to the NBL. Furthermore, in the study, the generalization of the NBL and its properties, including scale invariance, are addressed, along with a pedagogical approach proposed for elementary education.

Keywords: Newcomb-Benford Law, scale invariance, Accounting Fraud

CONTEÚDO

INTRODUÇÃO

A Lei de Newcomb-Benford (LNB), também conhecida como a Lei dos Primeiros Dígitos, é um conceito matemático intrigante que tem sido objeto de estudo e aplicação em diversas áreas. Neste capítulo introdutório, exploraremos em detalhes o contexto geral da LNB, suas raízes em observações empíricas, sua aplicação a conjuntos de dados e sua relevância no estudo de dados contábeis e no ensino médio. Além disso, discutiremos o uso do método da proporção dos dígitos para analisar dados contábeis relacionados aos gastos na epidemia de COVID-19 no Estado do Paraná, bem como a comparação com trabalhos previamente publicados. Através deste estudo, buscamos entender como os dados contábeis se comportam de acordo com a LNB, apresentando uma nova abordagem e sugestões para o ensino da LNB no ensino médio. Finalmente, examinaremos a frequência teórica da LNB em várias bases de dados e discutiremos nosso público-alvo, que inclui professores do ensino médio e indivíduos interessados na área.

A Lei de Newcomb-Benford, também referida como a Lei dos Primeiros Dígitos, é um fenômeno estatístico que ganhou destaque nas últimas décadas. Ela se origina de uma observação simples, porém profunda: em conjuntos de dados do mundo real, os primeiros dígitos dos números naturais não são igualmente distribuídos. Em vez disso, números começando com dígitos menores, como 1, 2 ou 3, ocorrem com mais frequência do que aqueles começando com dígitos maiores. Por exemplo, é mais provável encontrar um número como 134 do que um como 987.

As observações empíricas sobre a distribuição dos primeiros dígitos levaram a uma generalização importante: a frequência de ocorrência de dígitos iniciais segue um padrão previsível, conhecido como distribuição de Benford. Essa distribuição estabelece as expectativas de frequência para cada dígito inicial em uma escala logarítmica, e ela é universal, aplicando-se a conjuntos de dados em diversas áreas, desde contabilidade até fenômenos naturais.

A relevância da LNB transcende o campo acadêmico. Ela tem aplicações cruciais em conjunto de dados contábeis, onde a detecção de irregularidades nos números iniciais pode sinalizar fraudes ou erros. Além disso, a compreensão da LNB é fundamental para

o pensamento organizacional de dados, uma habilidade cada vez mais necessária em nossa era de informações abundantes.

Neste estudo, utilizaremos o método da proporção dos dígitos para comparar dados contábeis com os gastos relacionados à epidemia de COVID-19 no Estado do Paraná. Este método envolve calcular a frequência de cada dígito inicial em um conjunto de dados e compará-la com a distribuição teórica da LNB. Essa análise nos permitirá identificar desvios e padrões notáveis.

Ao longo deste trabalho, estabeleceremos uma conexão com pesquisas e estudos anteriores que abordaram a LNB. Isso nos permitirá situar nosso trabalho no contexto acadêmico e científico existente e destacar as contribuições únicas desta pesquisa.

Uma das metas centrais deste estudo é compreender como os dados contábeis se comportam em relação à LNB. Pretendemos lançar luz sobre se esses dados seguem ou desviam do padrão esperado da LNB, o que pode ter implicações significativas para a análise contábil.

Propomos uma nova abordagem para a aplicação da LNB a conjuntos de dados contábeis, visando simplificar conceitos matemáticos complexos e torná-los acessíveis a um público mais amplo.

Além disso, nosso estudo tem como objetivo fornecer sugestões para o ensino da LNB no ensino médio. Acreditamos que a compreensão dos princípios estatísticos subjacentes à LNB é valiosa e pode ser transmitida de maneira eficaz a estudantes nessa faixa etária.

Realizaremos análises detalhadas comparando a frequência teórica da LNB com as frequências observadas em várias bases de dados, incluindo os dados contábeis e os gastos relacionados à epidemia de COVID-19 no Estado do Paraná.

Com este trabalho, esperamos contribuir para uma nova forma de pensar sobre a LNB e sua aplicação prática. Além disso, almejamos oferecer uma abordagem acessível a professores do ensino médio e a indivíduos curiosos interessados na área.

É importante ressaltar que este trabalho não se aprofundará excessivamente em conceitos matemáticos complexos, pois nossa proposta é tornar a LNB mais acessível e prática. No entanto, sugerimos a consulta aos artigos de Hill, [13] e [15] e ao livro de Miller [18] para aqueles que desejam explorar os aspectos matemáticos com mais profundidade.

Nosso público-alvo principal consiste em professores do ensino médio fundamental, bem como em indivíduos curiosos sobre a área da LNB e sua aplicação em diversos contextos.

Este capítulo introdutório estabelece a base para a pesquisa que se segue, abordando o contexto histórico, aspectos matemáticos essenciais, aplicação prática do método da proporção dos dígitos e possibilidades de implementação em sala de aula. À medida que avançamos neste estudo, exploraremos em detalhes cada um desses aspectos, aprofundando nossa análise e apresentando resultados significativos.

A LNB tem suas raízes em uma série de estudos e descobertas que remontam ao século XIX, quando o astrônomo Simon Newcomb notou regularidades nos primeiros dígitos das constantes matemáticas. No entanto, foi apenas no século XX que o físico Frank Benford formalizou essa observação e a estendeu para um amplo espectro de dados. Desde então, a LNB tem sido objeto de curiosidade e pesquisa interdisciplinar, transcendo barreiras entre matemática, estatística, economia, contabilidade e muito mais.

Embora este trabalho não se aprofunde em detalhes matemáticos complexos, é essencial compreender alguns conceitos fundamentais subjacentes à LNB. A distribuição de Benford estabelece que a probabilidade de um dígito inicial ocorrer segue uma escala logarítmica, o que significa que dígitos menores têm uma probabilidade maior de aparecer. Esta distribuição é expressa por uma fórmula matemática que relaciona os dígitos iniciais e suas probabilidades.

O método da proporção dos dígitos é a ferramenta-chave para aplicar a LNB a conjuntos de dados. Consiste em contar a frequência de cada dígito inicial em um conjunto de dados, calcular as proporções e compará-las com as expectativas da distribuição de Benford. Desvios significativos podem indicar anomalias nos dados, levando a investigações adicionais.

Uma das dimensões mais emocionantes deste trabalho é a aplicação da LNB no ensino médio. Propomos uma abordagem didática que torna esse conceito complexo mais acessível aos estudantes, destacando sua relevância em várias disciplinas, desde matemática até ciências sociais. Acreditamos que essa abordagem prática pode enriquecer o currículo escolar e promover o pensamento crítico.

Ao longo deste estudo, esperamos fornecer uma nova perspectiva sobre a LNB, uma teoria que tem desempenhado um papel cada vez mais importante em nosso mundo de dados. Além disso, nossas sugestões para o ensino da LNB podem impactar positi-

vamente a educação de jovens estudantes, capacitando-os com ferramentas analíticas valiosas.

Este capítulo introdutório lançou as bases para nossa exploração detalhada da Lei de Newcomb-Benford. No próximo capítulo, mergulharemos no contexto histórico e nas observações empíricas que levaram à formulação desta lei. Examinaremos as aplicações práticas da LNB e como ela se relaciona com dados contábeis e o ensino médio. Além disso, abordaremos o uso do método da proporção dos dígitos e a análise comparativa com trabalhos anteriores. Com estas fundações estabelecidas, estamos prontos para investigar mais profundamente os intricados padrões que a LNB revela e as implicações que esses padrões podem ter em diversos domínios.

Este trabalho visa ampliar a compreensão da LNB e tornar seus conceitos mais acessíveis, proporcionando benefícios tanto no âmbito acadêmico quanto no educacional. Com um olhar crítico e inovador, esperamos contribuir significativamente para a exploração contínua deste fenômeno matemático intrigante.

1

PANORAMA HISTÓRICO

Neste capítulo, será abordada a história da descoberta da Lei de Benford, explorando seus principais colaboradores, Simon Newcomb e Frank Benford Jr. Além disso, será apresentado um resumo de suas carreiras e como suas pesquisas culminaram na descoberta dessa lei.

1.1 SIMON NEWCOMB

Simon Newcomb (Figura 1) foi um astrônomo americano-canadense que nasceu na Nova Escócia, no Canadá, em 1835 e morreu em Washington, nos Estados Unidos, em 1909. Suas principais contribuições foram nas áreas de cronometragem, matemática aplicada, economia e estatística. Mas uma de suas contribuições mais importantes foi a pesquisa sobre o movimento da Lua e suas influências. Em 1877, tornou-se diretor do Nautical Almanac Office e, em 1878, dirigiu um projeto para medir com mais precisão a velocidade da luz. Em 1874, recebeu a medalha de ouro da Royal Astronomical Society. Esse era o histórico de Newcomb como pesquisador, quando em 1881, publicou no *American Journal of Mathematics*, Vol. 4, No. 1 (pp. 39-40) [19], um artigo de duas páginas descrito como "Nota sobre a Frequência de uso dos Dígitos nos Números Naturais". (*Note on the Frequency of Use of the Different Digits in Natural Numbers*. (Ver Figura 4)). Nesse artigo, Newcomb escreve que: "*A lei da probabilidade de ocorrência dos números é tal que todas as mantissas de seus logaritmos são igualmente prováveis*". E para mostrar tal fato, ele apresenta a seguinte tabela 1:

O interessante é que no início de seu artigo, Newcomb afirma, o que é confirmado pela Tabela 1, que os dígitos significativos dos números em uma amostra de dados não

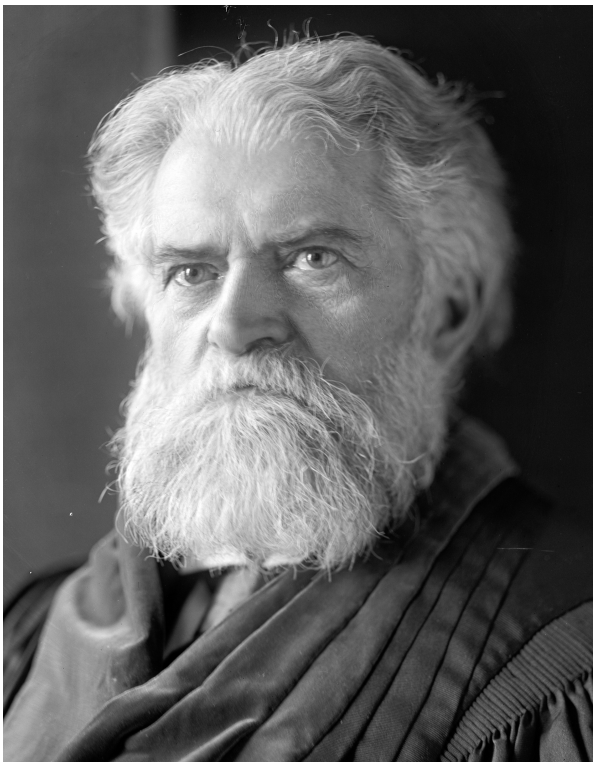


Figura 1: Simon Newcomb (1835-1909)

Dígito.	Primeiro Dígito.	Segundo Dígito.
0	...	0,1197
1	0,3010	0,1139
2	0,1761	0,1088
3	0,1249	0,1043
4	0,0969	0,1003
5	0,0792	0,0967
6	0,0669	0,0934
7	0,0580	0,0904
8	0,0512	0,0876
9	0,0458	0,0850

Tabela 1: Tabela de frequência dos primeiros e segundos digitos significativos

são igualmente distribuídos e que isso é facilmente percebido por quem usa tabelas de logaritmos, em que as folhas iniciais eram mais gastas do que as finais. Ou seja, a pesquisa de números começados com primeiro dígitos significativos de 1 a 9 não era igualmente consultada, sendo os primeiros dígitos com maior frequência de consulta.

Essa informação é interessante e contra intuitiva. Na verdade, ele percebe que as probabilidades desses dígitos significativos seguiam uma determinada distribuição.

Quando lançou seu artigo, mesmo sendo muito curioso, não trouxe muita atenção da comunidade científica. Afinal, não tinha uma rigorosa formalização matemática, pois Newcomb era astrônomo.

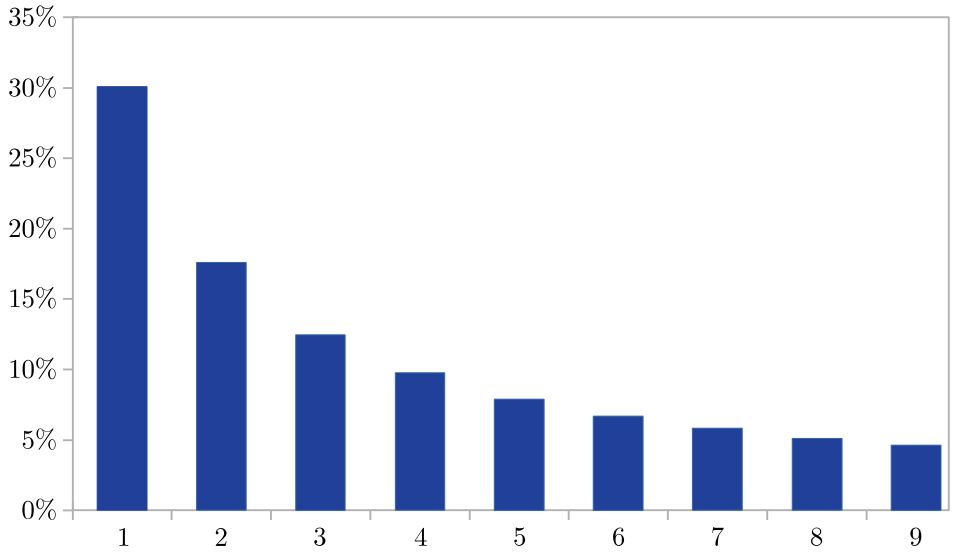


Figura 2: Gráfico de frequência dos primeiros digitos significativos

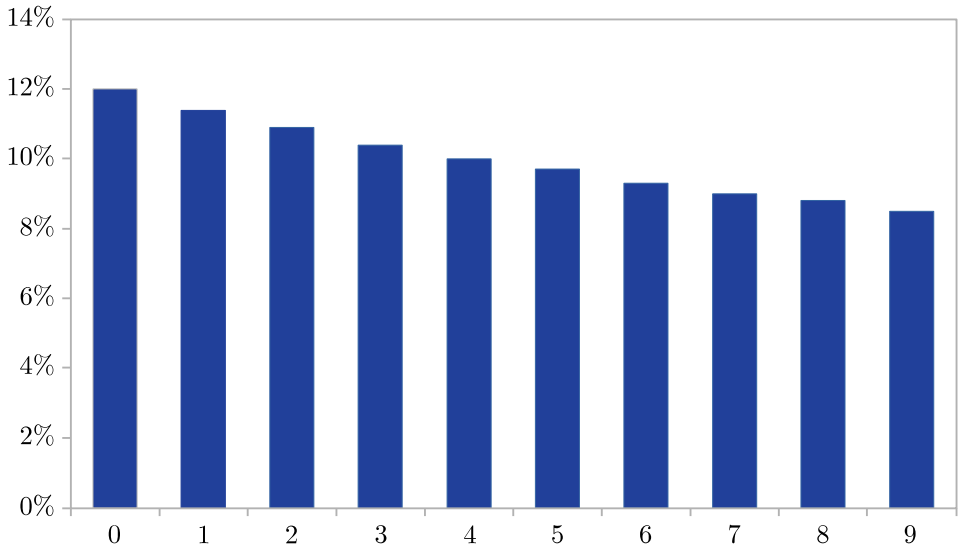
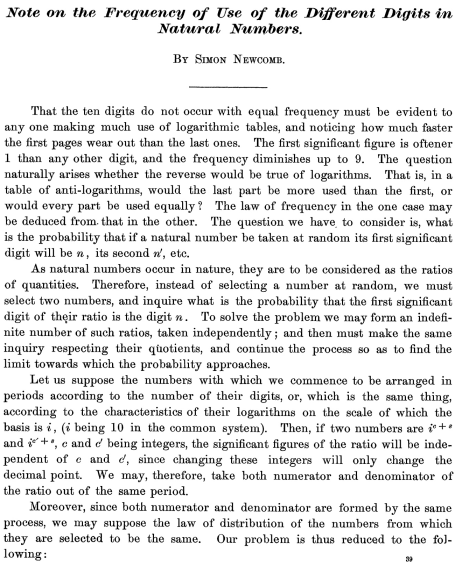


Figura 3: Gráfico de frequência dos segundos digitos significativos

Newcomb percebeu que as mantissas dos dígitos significativos tendiam a se tornar uniformes. Se observarmos o gráfico (Figura 3) de distribuição de probabilidade do segundo dígito significativo, já podemos notar uma assimetria menor em relação aos do primeiro dígito significativo. As distribuições de probabilidade dos dígitos significativos subsequentes, como a do terceiro dígito, tornam-se ainda menos assimétricas, sendo quase imperceptível a partir do quarto dígito.



Note on the Frequency of Use of the Different Digits in Natural Numbers.

BY SIMON NEWCOMB.

That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones. The first significant figure is oftener 1 than any other digit, and the frequency diminishes up to 9. The question naturally arises whether the reverse would be true of logarithms. That is, in a table of anti-logarithms, would the last part be more used than the first, or would every part be used equally? The law of frequency in the one case may be deduced from that in the other. The question we have to consider is, what is the probability that if a natural number be taken at random its first significant digit will be n, its second n', etc.

As natural numbers occur in nature, they are to be considered as the ratios of quantities. Therefore, instead of selecting a number at random, we must select two numbers, and inquire what is the probability that the first significant digit of their ratio is the digit n. To solve the problem we may form an indefinite number of such ratios, taken independently; and then must make the same inquiry respecting their quotients, and continue the process so as to find the limit towards which the probability approaches.

Let us suppose the numbers with which we commence to be arranged in periods according to the number of their digits, or, which is the same thing, according to the characteristics of their logarithms on the scale of which the basis is i, (i being 10 in the common system). Then, if two numbers are i^{c+s} and $i^{c'+s'}$, c and c' being integers, the significant figures of the ratio will be independent of c and c', since changing these integers will only change the decimal point. We may, therefore, take both numerator and denominator of the ratio out of the same period.

Moreover, since both numerator and denominator are formed by the same process, we may suppose the law of distribution of the numbers from which they are selected to be the same. Our problem is thus reduced to the following:

39

Figura 4: O artigo *Note on the Frequency of Use of the Different Digits in Natural Numbers* publicado no *American Journal of Mathematics*.

No ano de 1912, por meio do estudo de probabilidades chamado *"Calcul des Probabilités"*, Henri Poincaré (1854 - 1912) formalizou a Lei de Newcomb-Benford ao examinar uma lista de logaritmos. Nessa análise, ele observou que, em uma determinada posição de destaque, a ocorrência de números pares ou ímpares, dentre os dígitos de zero a nove, ocorre de maneira igualmente provável. Posteriormente, em 1917, Franel fez algumas correções no trabalho de Poincaré e verificou novamente os cálculos de probabilidade [3]. Apesar das contribuições de Poincaré e Franel, as pesquisas sobre o assunto sofreram uma pausa até a publicação de Frank A. Benford em 1938.

1.2 FRANK BENFORD

Figura 5: Frank Benford Jr. (1883-1948)

Frank A. Benford (Figura 5) desempenhou sua função como físico no Laboratório de Pesquisa da General Electric Company em Schenectady, NY, e sua área principal de trabalho estava relacionada à óptica na empresa. Demonstrando um compromisso exemplar, ele dedicou seu próprio tempo para conduzir todas as pesquisas e escrever "A Lei dos Números Anômalos". Em um breve trecho autobiográfico de três páginas que ele redigiu para Leonard Clark, do Union College, em 1939, apenas um pequeno parágrafo abordou sua descoberta da Lei dos Números Anômalos.

Frank Benford tinha orgulho de seu artigo e sentia satisfação com ele, mas não era alguém que se vangloriava. Provavelmente ele tenha se resignado à ideia de que sua Lei dos Números Anômalos eventualmente cairia no esquecimento. Ele certamente ficaria contente (e talvez até surpreso) em ver que o fenômeno dos primeiros dígitos, que ele redescobriu, permanece um assunto de interesse constante [18].

O próximo passo no estudo da distribuição dos dígitos iniciais dos números foi dado com a publicação do trabalho de Frank Benford intitulado "A Lei dos Números Anômalos"

(*The Law of Anomalous Numbers*) nos Proceedings of the American Philosophical Society em 1938 [4]. Além de fornecer explicações sobre a razão por trás da distribuição dos dígitos, o artigo também apresentava justificativas convincentes para a importância desse problema.

Assim como Newcomb, Benford observou que as páginas mais utilizadas em uma tabela de logaritmos comuns mostravam sinais de uso seletivo dos números naturais. As páginas contendo os logaritmos de números baixos, como 1 e 2, tendiam a ficar mais sujas e desgastadas devido ao uso, em comparação com aquelas contendo os logaritmos de números mais altos, como 8 e 9. Embora a condição de uma tabela de logaritmos não seja um assunto que desperte muito interesse, essa questão se torna digna de estudo quando consideramos que essa tabela é amplamente utilizada na construção de literatura científica, engenharia e informações gerais. A limpeza relativa das páginas de uma tabela de logaritmos pode fornecer insights sobre nossa forma de pensar e reagir ao lidar com coisas que podem ser descritas por meio de números.

Benford conduziu estudos sobre a distribuição dos dígitos iniciais em mais de 20 conjuntos de dados, abrangendo rios, áreas geográficas, populações, constantes físicas, sequências matemáticas (como $\sqrt{n}$, $n!$, n^2, etc.), esportes, uma edição da *Reader's Digest* e os primeiros 342 endereços de ruas listados no *American Men of Science* da época. Suas observações foram reproduzidas na Tabela 2 e na figura 6.

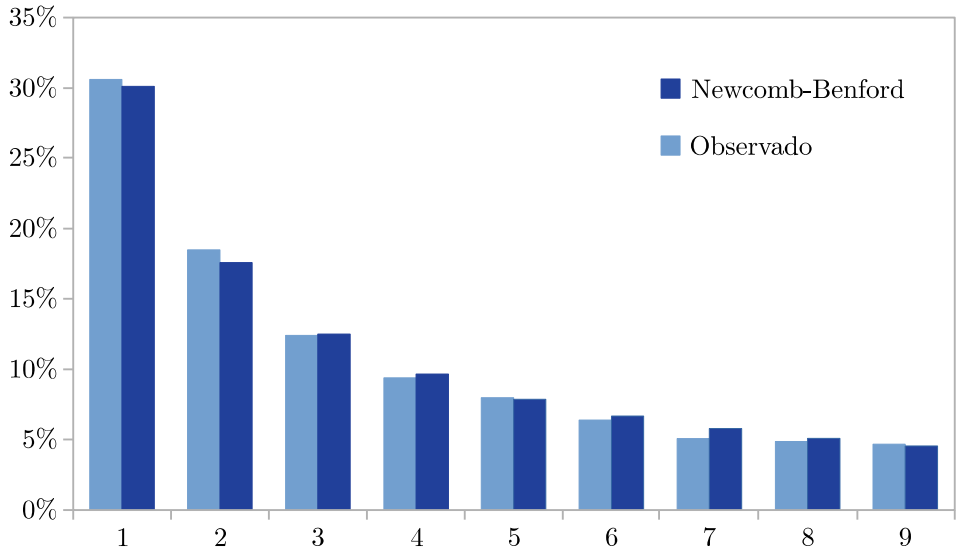


Figura 6: Comparação entre os dados observados por Benford e a distribuição dos números naturais.

Título	1	2	3	4	5	6	7	8	9	Amostras
Rios, Área	31,0	16,4	10,7	11,3	7,2	8,6	5,5	4,2	5,1	335
População	33,9	20,4	14,2	8,1	7,2	6,2	4,1	3,7	2,2	3259
Constantes	41,3	14,4	4,8	8,6	10,6	5,8	1,0	2,9	10,6	104
Itens de jornal	30,0	18,0	12,0	10,0	8,0	6,0	6,0	5,0	5,0	100
Calor específico	24,0	18,4	16,2	14,6	10,6	4,1	3,2	4,8	4,1	1389
Pressão	29,6	18,3	12,8	9,8	8,3	6,4	5,7	4,4	4,7	703
Potência perdida	30,0	18,4	11,9	10,8	8,1	7,0	5,1	5,1	3,6	690
Peso molecular	26,7	25,2	15,4	10,8	6,7	5,1	4,1	2,8	3,2	1800
Drenagem	27,1	23,9	13,8	12,6	8,2	5,0	5,0	2,5	1,9	159
Peso atômico	47,2	18,7	5,5	4,4	6,6	4,4	3,3	4,4	5,5	91
Design	25,7	20,3	9,7	6,8	6,6	6,8	7,2	8,0	8,9	5000
n^{-1}, $\sqrt{n}$	26,8	14,8	14,3	7,5	8,3	8,4	7,0	7,3	5,6	560
Digestão	33,4	18,5	12,4	7,5	7,1	6,5	5,5	4,9	4,2	308
Dados de custo	32,4	18,8	10,1	10,1	9,8	5,5	4,7	5,5	3,1	741
Volts de raios X	27,9	17,5	14,4	9,0	8,1	7,4	5,1	5,8	4,8	707
Liga Americana	32,7	17,6	12,6	9,8	7,4	6,4	4,9	5,6	3,0	1458
Corpo negro	31,0	17,3	14,1	8,7	6,6	7,0	5,2	4,7	5,4	1165
Endereços	28,9	19,2	12,6	8,8	8,5	6,4	5,6	5,0	5,0	342
n^1, n^2, ... , $n!$	25,3	16,0	12,0	10,0	8,5	8,8	6,8	7,1	5,5	900
Taxa de mortalidade	27,0	18,6	15,7	9,4	6,7	6,5	7,2	4,8	4,1	418
Média	30,6	18,5	12,4	9,4	8,0	6,4	5,1	4,9	4,7	1011
Erro provável (±)	0,8	0,4	0,4	0,3	0,2	0,2	0,2	0,2	0,3	

Tabela 2: Distribuição dos dígitos principais dos conjuntos de dados do artigo de Benford; a de todas as observações é indicada por “Média”. Observe que a concordância com a Lei de Benford é melhor para alguns exemplos do que para outros, e a amalgamação de todos os exemplos está bastante próxima da Lei de Benford.

O artigo de Benford trouxe à tona várias observações cruciais sobre o tema em questão. Entre essas, destaca-se uma das mais significativas: embora conjuntos de dados individuais possam não conformar-se estritamente à Lei de Newcomb-Benford, quando agregados em um conjunto maior, emerge uma nova sequência cujo comportamento tende a se aproximar mais dessa lei. Esse fenômeno pode ser observado em diversas escalas, como nas linhas correspondentes a n, n^2, e assim por diante, onde se pode demonstrar que cada uma delas não adere estritamente à Lei de Benford. Além disso, ao analisar a média de todos os conjuntos de dados, mesmo sem manipulações, é possível perceber um ajuste notavelmente preciso. Vale mencionar que Diaconis e Freedman [10] apresentaram evidências sólidas de que Benford manipulou erros de arredondamento para obter um ajuste ainda melhor, mas mesmo os dados não manipulados exibem uma

correspondência notável. Como resultado, a "lei" de Benford ganhou ampla aceitação, e, de acordo com Nigrini (2000) [21], a tabela de Benford foi a mais completa disponível até a década de 1990.

O destino do artigo de Benford foi muito mais promissor do que o de Newcomb, possivelmente em parte porque foi publicado imediatamente antes de um artigo de física de Bethe, Rose e Smith sobre o espalhamento múltiplo de elétrons. Enquanto levou décadas para surgir outro artigo baseado no trabalho de Newcomb, o próximo artigo após o de Benford foi publicado seis anos depois, por S. A. Goutsmit e W. H. Furry, intitulado "*Significant Figures of Numbers in Statistical Tables*", na revista Nature. A partir desse ponto, os artigos sobre o assunto começaram a ser publicados com mais frequência [18].

1.3 UM PANORAMA APÓS A PUBLICAÇÃO DE BENFORD

Desde a publicação pioneira de Benford em 1938, a Lei de Newcomb-Benford tem sido objeto de interesse em diversas áreas. Nesta seção histórica, destacamos as principais contribuições que moldaram a compreensão e a aplicação dessa lei ao longo das décadas.

No início dos anos 1960, Roger Pinkham [26] conduziu pesquisas cruciais para a Lei de Newcomb-Benford, investigando situações em que a lei poderia ser aplicável e fornecendo evidências empíricas adicionais para confirmar sua validade. Pinkham também explorou a invariância da lei quando multiplicada por uma constante, demonstrando sua robustez.

Na década de 1970, Hal Varian [27] sugeriu que a Lei de Newcomb-Benford poderia ser aplicada a dados socioeconômicos, ampliando suas possibilidades ao revelar uniformidades na fabricação de números. Isso abriu caminho para a detecção de erros deliberados ou fraudes.

Os anos 1990 marcaram um avanço significativo na pesquisa sobre a aplicação da Lei de Newcomb-Benford na detecção de fraudes. Mark Nigrini [23], em 1999, descreveu como a lei poderia ser usada para identificar irregularidades em dados financeiros e contábeis [20]. Ele também demonstrou a impossibilidade de criar dados artificiais que seguissem a lei de Newcomb-Benford, reforçando sua utilidade na detecção de anomalias [22], [24]. Theodore Hill [13], [14], em 1995, demonstrou a invariância

da Lei de Newcomb-Benford em relação à base, estendendo ainda mais seu escopo de aplicação.

No início dos anos 2000, a Lei de Newcomb-Benford continuou a ser estudada em diversas aplicações, incluindo fraudes contábeis, microeconomia e análise de dados eleitorais. Diversos pesquisadores, como Durtschi (2004) [11], Nye e Moul (2007) [25], destacaram sua relevância como ferramenta de detecção de fraudes.

À medida que a pesquisa avançou, a Lei de Newcomb-Benford expandiu suas fronteiras. Arno Berger, Theodore Hill [5], [6] e outros pesquisadores, em 2009, exploraram as propriedades matemáticas fundamentais da lei e propuseram explicações para sua ocorrência.

Em 2015, Mario Livio [17] destacou seu uso bem-sucedido na detecção de fraudes científicas e falsificação de dados, ampliando sua aplicação para o campo da pesquisa científica.

Finalmente, em 2015, Steven J. Miller [18] ofereceu uma introdução abrangente à Lei de Newcomb-Benford, discutindo suas aplicações, propriedades matemáticas e destacando as áreas de pesquisa em andamento, consolidando seu lugar como uma ferramenta valiosa em diversas disciplinas. A Lei de Benford continua a evoluir e encontrar novas aplicações em campos diversos, demonstrando sua relevância e versatilidade ao longo do tempo.

Em 2017, Michel Ausloos e Roy Cerqueti [2], [1]investigaram a aplicação da Lei de Newcomb-Benford em conjuntos de dados complexos, que possuem várias subpopulações, propondo uma extensão do princípio para esses casos.

2

ASPECTOS TEÓRICOS DA LEI DE NEWCOMB-BENFORD

Neste capítulo, introduziremos os conceitos fundamentais relacionados à Lei de Newcomb-Benford, abordando sua definição, conceitos-chave e a enunciação de alguns teoremas. Além disso, discutiremos aspectos importantes, como a σ-álgebra, bem como resultados relevantes como a invariância de escala e base, que desempenham um papel crucial em aplicações futuras, a serem exploradas no próximo capítulo.

2.1 A DISTRIBUIÇÃO DO PRIMEIRO DÍGITO

Para estabelecer uma base sólida, designaremos os dígitos significativos da ordem i como D_i, em que i pertence ao conjunto dos números naturais ($\mathbb{N}$). Por exemplo, consideremos o número e cuja representação decimal é dada por 2,7182818285... . Nesse contexto, o primeiro dígito significativo de e é $D_1(e) = 2$, e o segundo dígito significativo é $D_2(e) = 7$, e assim por diante.

O fundamento teórico da Lei de Newcomb-Benford é encapsulado na afirmação de Newcomb de que "a lei da probabilidade de ocorrência dos números é tal que todas as mantissas de seus logaritmos são igualmente prováveis," especialmente quando se trata do primeiro dígito. Essa afirmação encontra sua tradução matemática na seguinte definição:

Definição 2.1. A função de densidade logarítmica (discreta) para o primeiro dígito D_1 é definida por:

$$\Pr(D_1 = d) = \log\left(1 + \frac{1}{d}\right) \tag{2.1}$$

onde $d \in \{1, ..., 9\}$.

A probabilidade de ocorrência do primeiro dígito 1 em uma distribuição de dados é dada por $\Pr(D_1 = 1) = \log_{10}(1 + 1/1) \approx 0,301$. Similarmente, a probabilidade do primeiro dígito ser 7, por exemplo, é expressa como $\Pr(D_1 = 7) = \log_{10}(1 + 1/7) \approx 0,057$.

Em particular, exploraremos a probabilidade de o primeiro dígito ser 1, o segundo dígito ser 3 e o terceiro dígito ser 5 na Lei de Benford. A definição a seguir generalizará essa situação.

Definição 2.2. A distribuição conjunta logarítmica dos primeiros dígitos significativos D_1, D_2, ... , D_m (para cada $m \in \mathbb{N}^*$) é definida por:

$$\Pr(D_1 = d_1, D_2 = d_2, ..., D_m = d_m) = \log\left(1 + \left(\sum_{i=1}^{m} 10^{m-i} d_i\right)^{-1}\right) \tag{2.2}$$

em que $d_1 \in \{1, ..., 9\}$ e todos os outros $d_j \in \{0, ..., 9\}$.

A definição nos proporciona a probabilidade de o primeiro dígito na Lei de Benford ser 1 ($D_1 = 1$), o segundo dígito ser 3 ($D_2 = 3$) e o terceiro dígito ser 5 ($D_3 = 5$), e é expressa como:

$$\begin{aligned} \Pr(D_1 = 1, D_2 = 3, D_3 = 5) &= \log_{10}\left(1 + \frac{1}{135}\right) \\ &= \log_{10}\left(\frac{136}{135}\right) \\ &= 0,0032051399.... \end{aligned}$$

Às vezes, para simplificar nossa notação, optaremos por $\Pr(135)$.

O aspecto interessante aqui é que, para calcular a probabilidade do segundo dígito ser igual a 5, devemos realizar o seguinte cálculo:

$$\begin{aligned} \Pr(15) + \Pr(25) + ... + \Pr(95) &= \log_{10}\left(1 + \frac{1}{15}\right) + \log_{10}\left(1 + \frac{1}{25}\right) + ... + \log_{10}\left(1 + \frac{1}{95}\right) \\ &= \log_{10}\frac{16}{15} + \log_{10}\frac{26}{25} + ... + \log_{10}\frac{96}{95} \\ &= \log_{10}\left(\frac{16}{15} \cdot \frac{26}{25} \cdot ... \cdot \frac{96}{95}\right) \\ &= 0,09672258.... \end{aligned}$$

Isso nos permite estender nosso entendimento sobre as probabilidades associadas aos dígitos significativos além do primeiro dígito.

2.2 DÍGITOS SIGNIFICATIVOS

Definição 2.3. (Primeiro digito significativo) Para cada número real x diferente de zero, o primeiro dígito significativo de x (na forma decimal), denotado por $D_1(x)$, é o único inteiro $j \in \{1, 2, ..., 9\}$ satisfazendo

$$10^k j \leq |x| < 10^k(j+1) \tag{2.3}$$

para algum k, necessariamente único, em $\mathbb{Z}$.

Definição 2.4. (m-ésimo dígitos significativo) Da mesma forma, sendo $x \in \mathbb{R}$ e para cada $m \geq 2, m \in \mathbb{N}$, o m-ésimo dígito significativo de x (na forma decimal), denotado por $D_m(x)$, é definido indutivamente como o único inteiro $j \in \{0, 1, ..., 9\}$ tal que

$$10^k \left(\sum_{i=1}^{m-1} D_i(x) 10^{m-i} + j \right) \leq |x| < 10^k \left(\sum_{i=1}^{m-1} D_i(x) 10^{m-i} + j + 1 \right) \tag{2.4}$$

para alguns (necessariamente únicos) $k \in \mathbb{Z}$.

Definição 2.5. Definiremos por conveniência, $D_m(0) := 0$ para todo $m \in \mathbb{N}$.

2.3 SIGNIFICANDO

Definição 2.6. (Função significado) A função significado $S : \mathbb{R} \to [1, 10)$ é definida da seguinte forma:

- Se $x \neq 0$ então $S(x) = t$, em que t é o único número em $[1, 10)$ com $|x| = 10^k t$ para algum $k \in \mathbb{Z}$, k necessariamente único;
- Se $x = 0$ então, por conveniência, $S(0) := 0$.

(**Observação:** x está na forma decimal)

Podemos explicitar S da seguinte forma:

$$S(x) = 10^{\log|x| - \lfloor \log|x| \rfloor}$$

para todo $0 \neq x \in \mathbb{R}$ ($\lfloor \cdot \rfloor$ é a função menor inteiro);

Proposição 2.7. *Seja x um número real. Então:*

(i) $S(x) = \sum_{m\in\mathbb{N}} 10^{1-m} D_m(x)$;

(ii) $D_m(x) = \lfloor 10^{m-1} \cdot S(x) \rfloor - 10\lfloor 10^{m-2} \cdot S(x) \rfloor$ *para todo* $m \in \mathbb{N}$.

Definição 2.8. Seja S a função significado. Então:

$$\Pr(S \leq t) = \log t, \text{para todo} 1 \leq t < 10. \tag{2.5}$$

2.4 A σ-ÁLGEBRA DO SIGNIFICADO

Definição 2.9. A σ-álgebra do significado $\mathcal{S}$ é a σ-álgebra em $\mathbb{R}^+$ gerado pela função significado S, ou seja, $\mathcal{S} = \mathbb{R}^+ \cap \sigma(S)$.

A fim de iniciar as discussões, é imperativo estabelecer algumas definições. O símbolo $\mathcal{B}[1,10)$ será utilizado para denotar a σ-álgebra de Borel no intervalo $[1,10)$. Essa σ-álgebra é, portanto, obtida a partir de todos os subconjuntos gerados a partir do intervalo real $[1,10)$. É fundamental compreender essa construção, uma vez que ela constitui a base para a definição do lema que será apresentado a seguir.

Lema 2.10. *Para $f : \mathbb{R}^+ \to \mathbb{R}$ são equivalentes as seguintes afirmações:*

(i) f é determinado por S, isto é, existe uma função $\varphi : [1,10) \to \mathbb{R}$ com $\sigma(\varphi) \subset \mathcal{B}[1,10)$ tal que $f(x) = \varphi(S(x))$ para todo $x \in \mathbb{R}^+$;

(ii) $\sigma(f) \subset \mathcal{S}$.

A demonstração desse lema pode ser encontrada com detalhes no livro *"An Introduction to Benford's Law"* de Arno Berger e Theodore Hill [7], na página 16.

Teorema 2.11. *Para cada $A \in \mathcal{S}$,*

$$A = \bigcup_{k\in\mathbb{Z}} 10^k S(A), \tag{2.6}$$

em que $S(A) = \{S(x) : x \in A\} \subset [1,10)$. Além disso,

$$\mathcal{S} = \mathbb{R}^+ \cap \sigma(D_1, D_2, D_3, ...) = \left\{ \bigcup_{k\in\mathbb{Z}} 10^k B : B \in \mathcal{B} \right\}. \tag{2.7}$$

Demostração: Pela definição,

$$\begin{aligned}\mathcal{S} &= \mathbb{R}^+ \cap \sigma(S) \\ &= \mathbb{R}^+ \cap \{S^{-1}(B) : B \in \mathcal{B}\} \\ &= \mathbb{R}^+ \cap \{S^{-1}(B) : B \in \mathcal{B}[1,10)\}.\end{aligned}$$

Portanto, dado qualquer $A \in \mathcal{S}$, existe um conjunto $B \in \mathcal{B}[1,10)$ tal que

$$\begin{aligned}A &= \mathbb{R}^+ \cap S^{-1}(B) \\ &= \bigcup_{k\in\mathbb{Z}} 10^k B.\end{aligned}$$

Uma vez que $S(A) = B$, segue-se que (2.6) vale para todos $A \in \mathcal{S}$. Para provar (2.7), primeiro observe que, de acordo com a Proposição 2.7 (*i*), a função do significado S é completamente determinada pelos dígitos significativos D_1, D_2, D_3, ..., então $\sigma(S) \subset \sigma(D_1, D_2, D_3, ...)$ e, portanto, $\mathcal{S} \subset \mathbb{R}^+ \cap \sigma(D_1, D_2, D_3, ...)$. Por outro lado, de acordo com a Proposição 2.7 (*ii*), cada D_m é determinado por S, assim

$$\sigma(D_m) \subset \sigma(S) \text{ para todo } m \in \mathbb{N},$$

mostrando que $\sigma(D_1, D_2, D_3, ...) \subset \sigma(S)$ também. Para verificar a igualdade restante em (2.6), note que para todo $A \in \mathcal{S}$, $S(A) \in \mathcal{B}[1,10)$ e, portanto,

$$A = \bigcup_{k\in\mathbb{Z}} 10^k B$$

para $B = S(A)$, por (2.6). Por outro lado, todo conjunto da forma

$$\bigcup_{k\in\mathbb{Z}} 10^k B = \mathbb{R}^+ \cap S^{-1}(B)$$

com $B \in \mathcal{B}[1,10)$ obviamente pertence a S. □

Lema 2.12. *As seguintes propriedades são válidas para a σ-álgebra do significado S:*

(i) *S é autossimilar em relação à multiplicação por potências inteiras de 10, isto é, $10^k A = A$ para cada $A \in S$ e $k \in \mathbb{Z}$;*

(ii) *S está fechado para a multiplicação por escalares, ou seja, $aA \in S$ para cada $A \in S$ e $a > 0$;*

(iii) *S é fechado sob raízes integrais, ou seja, $A^{1/n} \in S$ para cada $A \in S$ e $n \in \mathbb{N}$*

Demonstração: (i) Isso é claro a partir de (2.6) porque $S(10^k A) = S(A)$ para todo $k \in \mathbb{Z}$.

(ii) Dado $A \in \mathcal{S}$, por (2.7) existe $B \in \mathcal{B}[1,10)$ tal que

$$A = \bigcup_{k\in\mathbb{Z}} 10^k B.$$

Em vista de (i), assume-se sem perda de generalidade que $1 < a < 10$. Então

$$\begin{aligned} aA &= \bigcup_{k\in\mathbb{Z}} 10^k aB \\ &= \bigcup_{k\in\mathbb{Z}} 10^k \left((aB \cap [a,10)) \cup \left(\frac{1}{10} aB \cap [1,a) \right) \right) \\ &= \bigcup_{k\in\mathbb{Z}} 10^k C, \end{aligned}$$

com

$$C = (aB \cap [a,10)) \cup \left(\frac{1}{10} aB \cap [1,a) \right) \in \mathcal{B}[1,10),$$

mostrando que $aA \in \mathcal{S}$.

(iii) Uma vez que intervalos da forma $[1,10^s]$ com $0 < s < 1$ geram $\mathcal{B}[1,10)$, ou seja, $\mathcal{B}[1,10) = \sigma(\{[1,10^s] : 0 < s < 1\})$, é suficiente verificar a afirmação para o caso especial

$$A = \bigcup_{k\in\mathbb{Z}} 10^k [1,10^s] \text{ para todo } 0 < s < 1.$$

Nesse caso,

$$\begin{aligned} A^{1/n} &= \bigcup_{k\in\mathbb{Z}} 10^{k/n} [1,10^{s/n}] \\ &= \bigcup_{k\in\mathbb{Z}} 10^k \bigcup_{j=0}^{n-1} [10^{j/n}, 10^{(j+s)/n}] \\ &= \bigcup_{k\in\mathbb{Z}} 10^k C, \end{aligned}$$

com

$$C = \bigcup_{j=0}^{n-1} [10^{j/n}, 10^{(j+s)/n}] \in \mathcal{B}[1,10).$$

Portanto, $A^{1/n} \in \mathcal{S}$. □

Lema 2.13. *A função $\ell : \mathbb{R}^+ \to [0,1)$ definida por $\ell(x) = \log S(x)$ estabelece uma correspondência bijetiva (isomorfismo de medida) entre medidas de probabilidade em $(\mathbb{R}^+, \mathcal{S})$ e em $([0,1), \mathcal{B}[0,1))$.*

2.5 INVARIÂNCIA DE ESCALA

Nesta seção abordaremos a importante questão da invariância de escala da Lei de Newcomb-Benford (LNB). Para compreender plenamente este conceito, é fundamental explorar o significado subjacente à invariância de escala.

A invariância de escala, em termos gerais, refere-se à capacidade de um fenômeno ou lei de se manter consistente, independentemente da escala na qual é observado. Por exemplo, quando analisamos um conjunto de dados relacionados à bacia hidrográfica de rios em uma determinada região e percebemos que ele segue a LNB, isso significa que a aplicação da LNB continua válida, independentemente da unidade de medida escolhida. Seja em quilômetros quadrados (km^2), metros quadrados (m^2), milhas quadradas (mi^2), jardas quadradas (yd^2) ou até mesmo em unidades mais incomuns, como comprimentos de palitos de fósforos ao quadrado, a LNB permanece aplicável.

É importante notar que existe uma constante multiplicativa que permite a conversão de uma medida para outra. Por exemplo, se uma bacia hidrográfica tem uma área de 40 km^2, ao multiplicarmos 40 por $0,38610192$, obtemos a mesma área da bacia hidrográfica expressa em milhas quadradas (mi^2). Isso demonstra que estamos operando em escalas diferentes, mas a essência da LNB permanece inalterada.

Em termos mais rigorosos, podemos afirmar que se A representa um conjunto de dados e a é uma constante positiva, então afirmar que a LNB é invariante de escala é equivalente a dizer que se um conjunto de dados segue a LNB, então o conjunto transformado $aA = \{ax : x \in A\}$ também segue a LNB.

Nesta seções, iremos aprofundar nossa discussão sobre a invariância de escala da LNB e explorar suas implicações em nosso contexto de pesquisa.

Definição 2.14. Seja $\mathcal{A} \supset \mathcal{S}$ uma σ-álgebra em $\mathbb{R}^+$. Uma medida de probabilidade P em $(\mathbb{R}^+, \mathcal{A})$ tem dígitos significativos invariantes de escala se

$$P(aA) = P(A) \text{ para todo } a > 0 \text{ e } A \in \mathcal{S},$$

ou, de forma equivalente, se para todo $m \in \mathbb{N}$, todo $d_1 \in \{1, 2, ..., 9\}$, todos $d_j \in \{0, 1, ..., 9\}$, $j \geq 2$, e todos $a > 0$,

$$P(\{x : D_j(ax) = d_j \text{ para } j = 1, 2, 3, ..., m\}) = P(\{x : D_j(x) = d_j \text{ para } j = 1, 2, 3, ..., m\}).$$

Antes de adentrar na discussão, faz-se necessário estabelecer algumas definições fundamentais:

Definição 2.15. Dados dois números reais x e y, diremos que x e y são equivalentes, e expressaremos isso como $x \sim y$, se, e somente se, existe um número inteiro $n \in \mathbb{Z}$ tal que $x = y \cdot 10^n$.

É crucial observar que a relação "$\sim$" está devidamente definida. Para verificar consideremos $x, y, z \in \mathbb{R}$. A relação é reflexiva, uma vez que se $x \sim x$, então $x = x \cdot 10^0$, sendo $0 \in \mathbb{Z}$. Além disso, ela é simétrica, pois se $y \sim x$, então $x \sim y$. Isso pode ser evidenciado ao notar que se $y = x \cdot 10^n$, então $y \cdot 10^{-n} = (x \cdot 10^n) \cdot 10^{-n} \implies y \cdot 10^{-n} = x$, dado que $-n \in \mathbb{Z}$.

Ademais, a relação é transitiva. Se $x \sim y$ e $y \sim z$, então $x \sim z$. Essa propriedade torna-se evidente ao considerarmos as igualdades:

$$\begin{cases} x = y \cdot 10^n \\ y = z \cdot 10^m \end{cases}$$

Com $n, m \in \mathbb{Z}$, obtemos $x = y \cdot 10^n \implies x = (z \cdot 10^m) \cdot 10^n \implies x = z \cdot 10^{m+n}$, visto que $m + n \in \mathbb{Z}$.

Assim, a relação "$\sim$" é reflexiva, simétrica e transitiva, cumprindo os requisitos de uma relação de equivalência. Essa definição é essencial para a compreensão de conceitos relacionados à equivalência entre números reais no contexto da potência de 10.

Seguindo a definição estabelecida, podemos afirmar que a expressão 2,34 é numericamente equivalente a 234 ($2,34 \sim 234$) quando consideramos a base 10 elevada à potência de -2, isto é, $2,34 = 234 \cdot 10^{-2}$.

A invariância de escala da LNB é um conceito fundamental que não depende do dígito considerado. Para ilustrar esse princípio, consideremos um conjunto $A = [a, b) \subset [1, 10)$. Sem perda de generalização, podemos calcular a probabilidade $P(A)$ da seguinte forma:

$$\begin{aligned} P(A) &= P([a, b)) \\ &= \log b - \log a. \end{aligned}$$

Assim, para valores de $a > 0$, podemos estabelecer:

$$\begin{aligned} P(aA) &= P(tA) \\ &= P([ta, tb)), \end{aligned}$$

onde $a \sim t \in [1, 10)$ é uma constante positiva. Existem dois casos possíveis a serem considerados:

1º Caso: Quando $tb < 10$, temos:

$$\begin{aligned} P(aA) &= P([ta, tb)) \\ &= \log(tb) - \log(ta) \\ &= \log t + \log b - \log t - \log a \\ &= \log b - \log a \\ &= P(A). \end{aligned}$$

2º Caso: Quando $tb \geq 10$ e $ta < 10$, podemos dividir o intervalo em duas partes:

$$\begin{aligned} P(aA) &= P([ta, tb)) \\ &= P([ta, 10) \cup [10, tb)) \\ &= P([ta, 10)) + P([10, tb)) \\ &= P([ta, 10)) + P\left(\left[1, \frac{tb}{10}\right)\right) \\ &= \log 10 - \log(ta) + \log(tb) - \log 10 - \log 1 \\ &= \log t + \log b - \log t - \log a \\ &= \log b - \log a \\ &= P(A). \end{aligned}$$

Portanto, em ambos os casos, obtemos que $P(aA) = P(A)$, o que demonstra que a LNB é invariante de base.

Esta análise evidencia a robustez da LNB em relação à escala adotada, independentemente dos dígitos considerados. No contexto da nossa pesquisa, essa invariância de escala desempenha um papel crucial e será explorada em detalhes nas próximas seções.

Teorema 2.16. *(Caracterização de invariância de escala). Uma medida de probabilidade P em $(\mathbb{R}^+, \mathcal{A})$ é invariante à escala se e somente se P satisfaz a Lei de Benford.*

Demonstração: Seja P uma medida de probabilidade em $(\mathbb{R}^+, \mathcal{A})$ e

$$A_d = \bigcup_{n=-\infty}^{\infty} [1, 10^d) \cdot 10^n$$

para qualquer $d \in [0,1)$ (uma medida de probabilidade em $\mathcal{A}$ é inteiramente definida por seus valores em conjuntos desse tipo).

Sejam $\bar{P}$ e $\hat{P}$ as medidas de probabilidade definidas, respectivamente, nos espaços mensuráveis $([0,1), \mathcal{B}[0,1]))$ e $((1,10), \mathcal{B}[1,10))$ por:

$$\forall d \in [0,1), \bar{P}[0,d) = \hat{P}[1,10^d) = P(A_d) \tag{2.8}$$

Essa relação, na verdade, define uma correspondência útil e biunivoca entre os espaços mensuráveis de P, $\hat{P}$ e $\bar{P}$.

Agora, para a prova em si. No espaço mensurável $(\mathbb{R}^+, \mathcal{A})$, P satisfaz a Lei de Newcomb-Benford se e somente se $P(A_d) = d$ para todos $d \in [0,1)$ (então, de acordo com (2.8), se e somente se $\bar{P}$ for a distribuição uniforme em $[0,1]$ e se e somente se $\hat{P}[1,10^d) = d, \forall d \in [0,1)$).

Agora, suponha que P satisfaz a LNB e devemos provar que para todos $a \in \mathbb{R}$:

$$P(aA_d) = P\left(\bigcup_{n=-\infty}^{\infty} [a, a \cdot 10^d) \cdot 10^n\right) = P(A_d)$$

Sem perda de generalidade, a pode ser restrito a $[1,10)$ (caso contrário, faça $a \bmod 10$). Dois casos agora são distinguíveis:

- Se $a \cdot 10^d \leq 10$:

$$\begin{aligned} P(aA_d) &= \hat{P}[a, a \cdot 10^d) \\ &= \hat{P}[1, a \cdot 10^d) - \hat{P}[1,a) \\ &= \log(a \cdot 10^d) - \log(a) \\ &= d \\ &= P(A_d) \end{aligned}$$

- Se $a \cdot 10^d > 10$: (já que $a \leq 10$, $a \cdot 10^d$ está em $[10, 100)$)

$$\begin{aligned} P(aA_d) &= \hat{P}[a,10) + \hat{P}([10, a \cdot 10^d) \bmod 10) \\ &= (1 - \log(a)) + \hat{P}\left[1, \frac{a \cdot 10^d}{10}\right) \\ &= 1 - \log(a) + \log\left(\frac{a \cdot 10^d}{10}\right) \\ &= 1 - \log(a) + \log(a) + d - 1 \\ &= d \\ &= P(A_d) \end{aligned}$$

Por outro lado, suponha que P é invariante à escala, ou seja, $P(A_d) = P(aA_d)$ para todos $a \in \mathbb{R}$ e $d \in [0,1)$, agora mostraremos que P satisfaz a LNB, ou seja, $P(A_d) = d$.

Seja α um irracional arbitrário em $\mathbb{R}$. Então $P(A_d) = P(10^\alpha A_d)$, para todos d em $[0,1)$. Sem perda de generalidade, pode-se supor que $10^\alpha \in [1,10)$ (caso contrário, faça $10^\alpha \bmod 10$).

O isomorfismo definido em (2.8) implica então que:

$$\hat{P}[1,10^d) = \hat{P}([10^\alpha, 10^{\alpha+d}) \bmod 10), \forall d \in [0,1)$$

[Aqui, a notação $[a,b) \bmod 10$ significa $[a,b)$ se $b \leq 10$, ou $[a,10) \cup \left[1, \frac{b}{10}\right)$ se $b \in [10,100)$]

E, como consequência:

$$\bar{P}[0,d) = \bar{P}([\alpha, d+\alpha) \bmod 10), \forall d \in [0,1)$$

[Com a notação equivalente.]

Essa última igualdade significa que $\bar{P}$ é invariante por uma rotação irracional no círculo unitário. É conhecido há muito tempo que a única distribuição em $[0,1)$ que possui tal propriedade é a uniforme (veja, por exemplo, [28]), e consequentemente, P satisfaz a Lei de Benford. □

Teorema 2.17. *Para cada variável aleatória X, sendo $P(X = 0) = 0$, as seguintes afirmações são equivalentes:*

(i) *X é Benford;*

(ii) *Existe $d \in \{1,2,...,9\}$, tal que*

$$P(D_1(aX) = d) = P(D_1(X) = d), \text{ para todo } a > 0;$$

(iii) *Existe $d \in \{1,2,...,9\}$, tal que*

$$P(D_1(aX) = d) = \log\left(1 + \frac{1}{d}\right), \text{ para todo } a > 0.$$

A demonstração desse teorema pode ser encontrada com detalhes no livro *"An Introduction to Benford's Law"* de Arno Berger e Theodore Hill [7], na página 72.

2.6 INVARIÂNCIA DE BASE

A invariância de base, uma hipótese sutil, desempenha um papel crucial na fundamentação da Lei de Benford. Para ilustrar essa invariância, voltamo-nos para a reflexão sobre as leis físicas universais, as quais proporcionam uma compreensão aprofundada desse princípio subjacente. Tomemos como exemplo a Lei da Gravitação Universal de Newton, expressa pela equação $F = G \cdot m_1 \cdot m_2/r^2$, onde F representa a força gravitacional, m_1 e m_2 são as massas das entidades em questão, e r é a distância entre os centros de massa.

É notável que, independentemente da base numérica escolhida para representação, a Lei da Gravitação Universal descreverá a mesma força gravitacional. A invariância de base nesta lei implica que sua formulação permanece consistente, quer os cálculos sejam realizados na base "10", como comumente feito por humanos, ou na base "2", como é típico em ambientes computacionais. Essa constância revela a robustez da lei diante da variação na escolha da base numérica, ressaltando sua invariância fundamental.

A analogia entre a Lei da Gravitação Universal e a invariância de base oferece uma compreensão clara de como leis naturais tendem a manter-se invariantes em diferentes bases numéricas. Esta analogia serve como uma transição natural para abordar a Lei de Newcomb-Benford, que, ao descrever a distribuição de dígitos significativos em conjuntos de dados, também exibe invariância de base.

A hipótese subjacente é que, se a determinada lei é observada em conjuntos de dados "naturais" na base 10, essa mesma lei deve manter-se ao utilizar outras bases numéricas. A Lei de Newcomb-Benford, ao ser transposta para uma nova base numérica, mantém a mesma distribuição de dígitos significativos, embora possa apresentar um fator de normalização diferente. Essa variação é atribuída à diferente quantidade de elementos presentes em cada base, evidenciando a invariância de base como um princípio consistente na análise estatística de conjuntos de dados.

Definição 2.18. Seja $\mathcal{A} \supset \mathcal{S}$ uma σ-álgebra em $\mathbb{R}^+$. Uma medida de probabilidade P em $(\mathbb{R}^+, \mathcal{A})$ tem dígitos significativos invariante de base se, $P(A) = P(A^{1/n})$, para todos $A \in \mathcal{S}$ e todos $n \in \mathbb{N}$.

Procederemos à definição da Lei de Newcomb-Benford para uma base genérica.

Definição 2.19. A função de densidade logarítmica (discreta) para o primeiro dígito D_1 na base $b > 1$ é definida por:

$$\Pr_{(b)}(D_1 = d) = \log_b \left(1 + \frac{1}{d}\right) \tag{2.9}$$

onde $d \in \{1, ..., b-1\}$.

A probabilidade de ocorrência do primeiro dígito 1, na base 3, em uma distribuição de dados é dada por $\Pr_{(3)}(D_1 = 1) = \log_3(1 + 1/1) \approx 0,631$. Similarmente, a probabilidade do primeiro dígito ser 2 na base 3, por exemplo, é expressa como $\Pr_{(3)}(D_1 = 2) = \log_3(1 + 1/2) \approx 0,369$.

A definição a seguir generalizará essa situação.

Definição 2.20. A distribuição conjunta logarítmica dos primeiros dígitos significativos D_1, D_2, ... , D_m (para cada $m \in \mathbb{N}^*$) na base (b > 1) é definida por:

$$\Pr_{(b)}(D_1 = d_1, D_2 = d_2, ..., D_m = d_m) = \log_b \left(1 + \left(\sum_{i=1}^{m} 10^{m-i} d_i\right)^{-1}\right) \tag{2.10}$$

em que $d_1 \in \{1, ..., b-1\}$ e todos os outros $d_j \in \{0, ..., b-1\}$.

No trabalho de Adrien Jamain [16], as demonstrações destas propriedades são apresentadas de maneira sistemática e organizada.

3

LEI DE NEWCOMB-BENFORD, UMA APLICAÇÃO

Este capítulo aborda o contexto que motivou esta pesquisa, fornecendo um resumo dos principais estudos e conceitos que deram origem às investigações realizadas. Em particular, exploramos a problemática inicial e a metodologia aplicada para abordá-la.

3.1 PROBLEMA MOTIVADOR

O estudo de Vitória Eduardo Bello & Anselmo Chaves Neto lançou luz sobre a execução orçamentária de despesas relacionadas à pandemia de COVID-19 no Estado do Paraná no período de 2020 a 2022 [3] [8]. A pesquisa adotou a metodologia desenvolvida por Nigrini para identificar possíveis fraudes e irregularidades em conjuntos de dados contábeis. Um resultado surpreendente dessa investigação foi a observação de que os dígitos 20 e 80 não obedeciam estritamente à distribuição de Newcomb-Benford. Os resultados mencionados motivaram a análise desses dados sob a perspectiva de uma abordagem metodológica alternativa.

Por outro lado, o estudo de Gueron & Pellegrini concentrou-se no desvio dos dados observados em relação à distribuição de Newcomb-Benford. Gueron & Pellegrini analisou as eleições para o Senado no Estado da Bahia no ano de 1994 [12], um cenário já amplamente considerado como fraudulento. Os resultados revelaram que os dados da eleição fraudulenta divergiam consideravelmente da distribuição esperada pela Lei de Newcomb-Benford, enquanto as eleições em outros estados apresentavam um comportamento distinto. O estudo serviu de incentivo para empregar o método utilizado por Gueron na análise dos registros de execução orçamentária de despesas durante a pandemia de COVID-19 no Estado do Paraná. Bello & Chaves Neto conduziu

essa análise, aplicando o método de Nigrini, o que nos proporcionou um resultado que desejamos comparar com os obtidos através da abordagem metodológica descrita no trabalho de Gueron.

3.2 DESENVOLVIMENTO

Para abordar o problema de desvio em dados contábeis e eleitorais em relação à Lei de Newcomb-Benford, primeiramente, demonstramos que a razão entre a distribuição dos primeiros dígitos significativos, d_1 e d_2, da distribuição teórica da LNB é independente da base numérica. Sendo b ser maior que zero e diferente de um, essa razão pode ser expressa da seguinte maneira:

$$\begin{aligned}
\frac{\Pr_{(b)}(D_1 = d_1)}{\Pr_{(b)}(D_2 = d_2)} &= \frac{\log_b\left(1 + \frac{1}{d_1}\right)}{\log_b\left(1 + \frac{1}{d_2}\right)} \\
&= \frac{\log\left(1 + \frac{1}{d_1}\right)}{\cancel{\log b}} \cdot \frac{\cancel{\log b}}{\log\left(1 + \frac{1}{d_2}\right)} \\
&= \frac{\log\left(1 + \frac{1}{d_1}\right)}{\log\left(1 + \frac{1}{d_2}\right)}
\end{aligned}$$

Com base nesse resultado, definimos $Q(d_1, d_2)$ como a razão entre as distribuições do primeiros dígitos teóricos da Lei de Newcomb-Benford:

$$Q(d_1, d_2) = \frac{\log\left(1 + \frac{1}{d_1}\right)}{\log\left(1 + \frac{1}{d_2}\right)}$$

Neste contexto, evidencia-se a vantagem intrínseca do método, destacando sua independência em relação à base do logaritmo. Adicionalmente, apresentamos a notação $n(d_1, d_2)$ para representar a razão entre os primeiros dígitos significativos na distribuição dos dados reais. Essa abordagem oferece uma perspectiva analítica que

contribui para a compreensão mais aprofundada dos padrões subjacentes na distribuição dos dados.

Um exemplo prático para ilustrar essas razões é apresentado, destacando a relevância dessa abordagem para a análise do comportamento de conjuntos de dados. A razão entre a LNB e os dados observados, denotada por $f(d_1, d_2)$, é definida como:

$$f(d_1, d_2) = \frac{Q(d_1, d_2)}{n(d_1, d_2)} \tag{3.1}$$

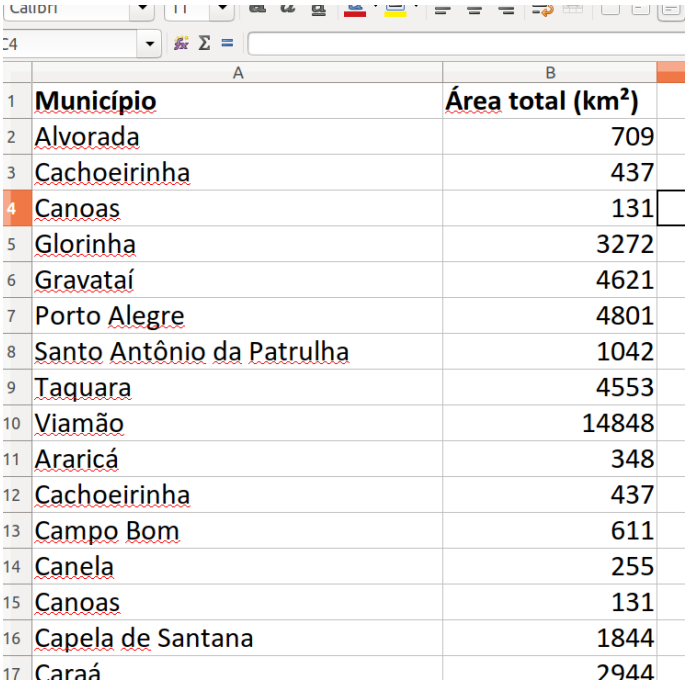
Esta razão desempenha um papel fundamental na análise de dados sob a perspectiva da LNB. Quando os dados seguem a distribuição de Newcomb-Benford, o desvio da razão esperada entre os primeiros dígitos, isto é, $f(d_1, d_2) - f(d_2, d_1)$, tende a ser nulo. Isso sugere que, ao plotar essa razão em um gráfico de base numérica versus desvio da razão esperada entre os dígitos, espera-se observar uma reta horizontal com ordenada zero.

Para ilustrar, consideremos um conjunto de dados de bacias hidrográficas de rios. A partir desses dados, obtemos uma tabela que registra as frequências dos primeiros dígitos 1 e 2, juntamente com as frequências teóricas esperadas para esses primeiros dígitos.

1º. Digito na Base 10	Frequência dos dados	Resultado teórico da LNB
1	32,0%	30,1%
2	21,2%	17,6%
1º. Digito na Base 9	Frequência dos dados	Resultado teórico da LNB
1	32,8%	33,3%
2	19,2%	19,5%
1º. Digito na Base 8	Frequência dos dados	Resultado teórico da LNB
1	33,0%	33,3%
2	20,6%	19,5%
1º. Digito na Base 7	Frequência dos dados	Resultado teórico da LNB
1	33,3%	33,3%
2	22,4%	20,8%

Tabela 3: tabela elaborada pelo autor.

Ao aplicarmos a diferença $f(1, 2) - f(2, 1)$, obtemos a seguinte tabela:



	A	B
1	**Município**	**Área total (km²)**
2	Alvorada	709
3	Cachoeirinha	437
4	Canoas	131
5	Glorinha	3272
6	Gravataí	4621
7	Porto Alegre	4801
8	Santo Antônio da Patrulha	1042
9	Taquara	4553
10	Viamão	14848
11	Araricá	348
12	Cachoeirinha	437
13	Campo Bom	611
14	Canela	255
15	Canoas	131
16	Capela de Santana	1844
17	Caraá	2944

Figura 7: Planilha com as áreas de algumas bacias hidrográfica do estado do estado do Rio Grande do Sul [9].

Base	$f(1,2) - f(2,1)$
7	$0,1483516484$
8	$0,1279399897$
9	$-0,0007506099$
10	$0,2504329524$

Tabela 4: tabela elaborada pelo autor.

Construindo um gráfico a partir desses dados, temos:

3.3 MÉTODOS E RESULTADOS

O conjunto de dados utilizado neste estudo é o mesmo utilizado por Bello & Chaves Neto para fins de comparação. Realizamos um estudo de caso com dados provenientes do Portal da Transparência do Governo do Estado do Paraná, referentes à execução orçamentária destinada à contingência da pandemia de COVID-19. Estes dados abrangem o período de junho a dezembro de 2020, fevereiro a dezembro de 2021 e fevereiro a dezembro de 2022. Para uma análise completa, organizamos os dados em planilhas mensais.

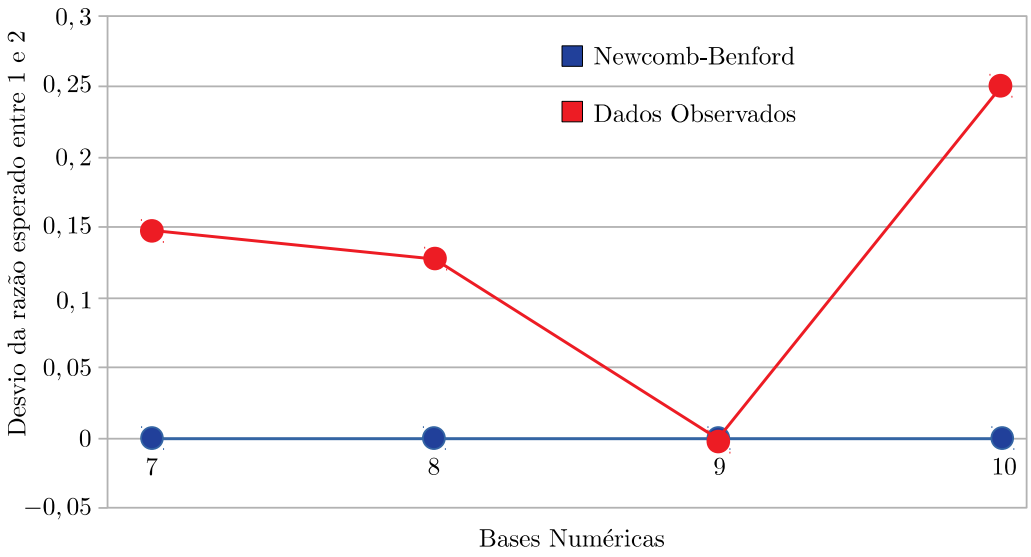


Figura 8: Comparação entre os dados observados por Benford e a distribuição dos números naturais.

Quanto à coleta de dados, utilizamos as planilhas de empenho, liquidação e pagamento de despesas disponíveis nos dados abertos. Além disso, aplicamos um procedimento de "tratamento de dados", no qual filtramos apenas os valores de orçamento que não estavam repetidos, uma vez que a duplicação desses valores poderia prejudicar a análise. Multiplicamos todos os números por 100 para remover as vírgulas e os convertemos em valores absolutos.

Portanto, após a filtragem e organização dos dados, trabalhamos com um total de 195.419 observações. É importante destacar que o número total de observações antes do processo de filtragem era inicialmente o mesmo, ou seja, 195.419 observações.

A Tabela 5 a seguir apresenta a quantidade de dados brutos antes do tratamento dos dados e o número de observações coletadas.

Tabela 5 – Número de observações por ano antes do processo de tratamento dos dados

	Ano	Número de Observações
1	2020	57.807
2	2021	100.798
3	2022	36.814

Tabela 5: Elaborada pelo autor com dados extraídos do Portal da Transparência do Estado do Paraná (2022)

3.3.1 *Análise dos Dados*

Neste estudo, os dados fornecidos foram analisados por meio do software R (R CORE TEAM, 2019), com a utilização de pacotes adicionais que permitem a aplicação de uma variedade de técnicas de análise gráfica e a criação de tabelas de resumo estatístico.

O método de análise discutido anteriormente, conhecido como o método da proporção de dígitos, oferece uma ferramenta valiosa para avaliar conjuntos de dados. Para melhor compreensão, introduzimos a função $R(d_1, d_2)$, definida como

$$R(d_1, d_2) = f(d_1, d_2) - f(d_2, d_1), \tag{3.2}$$

conforme 3.1.

O processo começa com a representação gráfica das Bases $\times$ $R(d_1, d_2)$, com relação aos dados de cada mês na coluna empenhado. Uma premissa fundamental é que, se o conjunto de dados seguisse perfeitamente a Lei de Newcomb-Benford (LNB), os valores de $R(d_1, d_2)$ seriam consistentemente iguais a zero. Isso resultaria em uma função constante no valor zero e, consequentemente, em um gráfico que se assemelharia a uma reta horizontal.

Portanto, qualquer desvio notável dessa linha horizontal pode nos dar uma indicação de irregularidade nos dados. No entanto, como enfatizado anteriormente, para uma análise completa, é necessário comparar todas as distâncias calculadas para cada mês dos dados, a fim de detectar eventuais valores discrepantes.

O presente estudo se propôs a analisar minuciosamente as colunas referentes a empenho, liquidação e pagamento no conjunto de dados [8]. Cada uma dessas categorias desempenha um papel crucial na gestão e compreensão das finanças públicas, contribuindo para uma visão abrangente do fluxo de recursos e gastos planejados.

- **Empenhado:** A variável “Empenhado” diz respeito ao comprometimento de recursos para uma despesa previamente planejada. Quando um valor é empenhado, isso implica que tal quantia está reservada para ser alocada em uma atividade específica, porém, crucialmente, ainda não foi despendida efetivamente. Em outras palavras, o empenho representa a reserva de fundos para garantir a disponibilidade financeira necessária antes da efetivação da despesa.
- **Liquidado:** Por sua vez, a categoria “Liquidado” indica que a despesa empenhada foi efetivamente realizada ou concluída. O valor liquidado corresponde à porção do empenho que foi utilizada ou consumida durante o processo. Assim, a

liquidação representa o estágio em que a despesa planejada foi concretizada, evidenciando a efetividade na utilização dos recursos em relação ao planejamento inicial.

- **Pago:** A terceira variável, "Pago", refere-se ao montante efetivamente desembolsado, ou seja, o valor que foi verdadeiramente pago. É importante ressaltar que o valor pago nem sempre coincide com o montante empenhado, uma vez que podem ocorrer variações ao longo do processo. Essas discrepâncias entre os valores empenhados e pagos podem resultar de ajustes, negociações ou outras circunstâncias que impactam a execução financeira, conferindo uma dinâmica complexa ao processo de gastos públicos.

Ao considerar essas três dimensões - empenhado, liquidado e pago - na análise dos dados, é possível obter insights valiosos sobre o ciclo de gastos, planejamento e execução financeira. Essa abordagem proporciona uma compreensão mais abrangente e aprofundada das práticas orçamentárias, permitindo uma gestão mais eficiente e transparente dos recursos públicos.

A partir dos dados da coluna empenhado mencionados anteriormente, calculamos a função $R(1,2)$ de acordo com a Eq.3.2, considerando diferentes bases (variando de 6 a 70). Os resultados são apresentados na Figura 9.

Na Figura 9, destacamos os dados da coluna empenhado de 2020 em verde, de 2021 em azul e de 2022 em preto. No entanto, o aspecto mais notável é a linha vermelha, que corresponde aos dados contábeis do mês de fevereiro de 2022, claramente identificados como um outlier.

Essa observação nos fornece uma indicação evidente de irregularidade nesse conjunto de dados contábeis. Além disso, a Figura destaca a importância de considerar várias bases, uma vez que, em algumas situações, a distribuição de dados pode parecer estar dentro do esperado, como exemplificado na base dez. No entanto, ao explorar várias bases, torna-se mais evidente a presença de anomalias.

Em resumo, esse método e a análise visual da Figura 9 nos capacitam a identificar e compreender melhor os outliers, contribuindo significativamente para a análise de dados contábeis e a detecção de irregularidades em diferentes contextos. Portanto, é fundamental explorar todas as bases disponíveis para uma avaliação abrangente e precisa dos conjuntos de dados.

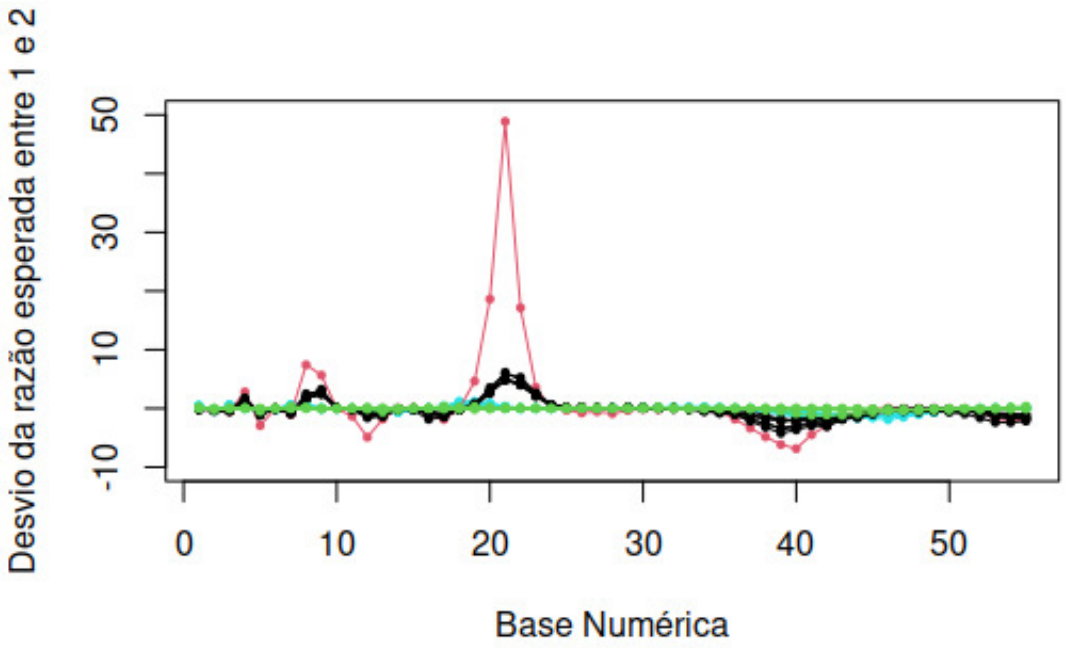


Figura 9: Função erro para relação entre o primeiro dígitos 1 e 2 da coluna empenhado do conjunto de dados.

A Figura 10 apresenta a aplicação da fórmula $R(1,2)$ ao conjunto de dados relacionados às despesas liquidadas. Nessa abordagem, as cores adotadas mantêm consistência com aquelas utilizadas na análise do conjunto de dados referente aos empenhos. A observação revela que o ponto de dados que exibe o maior desvio pertence ao ano de 2021, identificado pela coloração azul. Esse destaque está associado ao mês de fevereiro desse ano, indicando a presença de outliers que evidenciam um comportamento atípico no contexto deste conjunto de dados.

A análise aprofundada da Figura 10 permite identificar que o mês de fevereiro de 2021 emerge como um ponto de destaque, caracterizado pelo desvio da norma estabelecida. A adoção da fórmula (3.2) facilita a detecção de padrões anômalos, fornecendo uma base sólida para a identificação de comportamentos atípicos no âmbito das despesas liquidadas.

Por sua vez, a Figura 11 concentra-se na análise do conjunto de dados relacionados aos pagamentos efetuados. Intrigantemente, a identificação de outliers coincide com o mesmo mês de fevereiro, desta vez destacado na figura por meio da coloração vermelha.

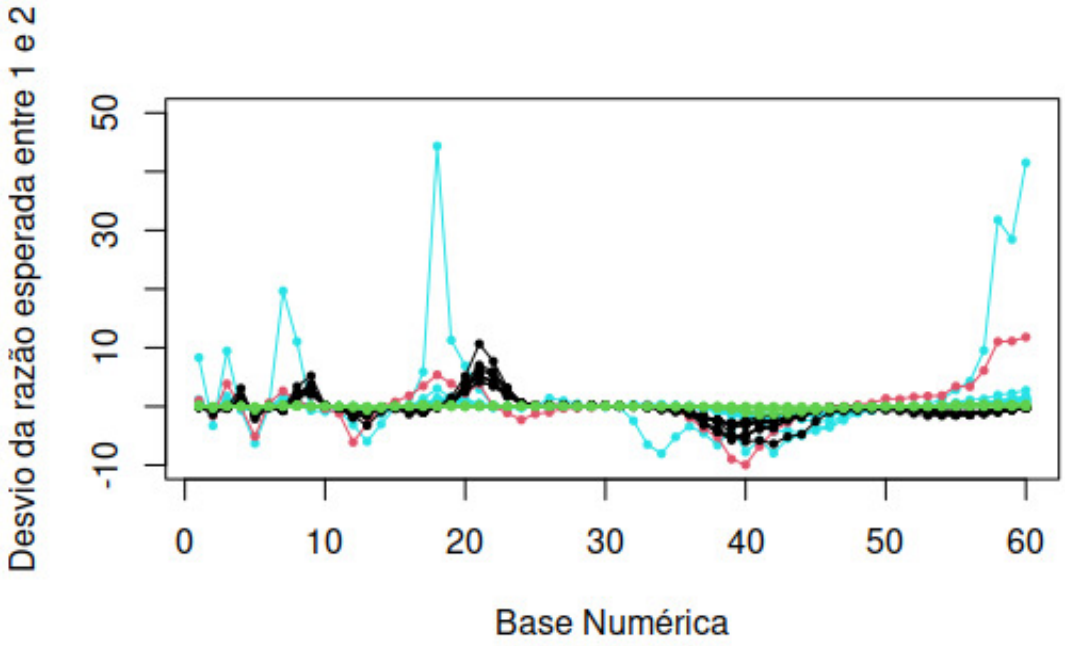


Figura 10: Função erro para relação entre o primeiro dígitos 1 e 2 da coluna liquidado do conjunto de dados.

Esse alinhamento sugere uma sincronia nos comportamentos atípicos observados nos conjuntos de dados de despesas empenhadas e pagas.

A convergência desses resultados destaca a relevância do mês de fevereiro de 2022 como um período de interesse para uma análise mais aprofundada. A presença consistente de outliers nesse intervalo temporal sugere a necessidade de uma investigação mais detalhada sobre as circunstâncias que levaram a tais comportamentos atípicos nos processos de empenho, liquidação e pagamento. Essa análise visual, fundamentada na fórmula (3.2), revela-se instrumental na identificação e compreensão de padrões irregulares nos dados, contribuindo para uma interpretação mais robusta e informada das dinâmicas financeiras em questão.

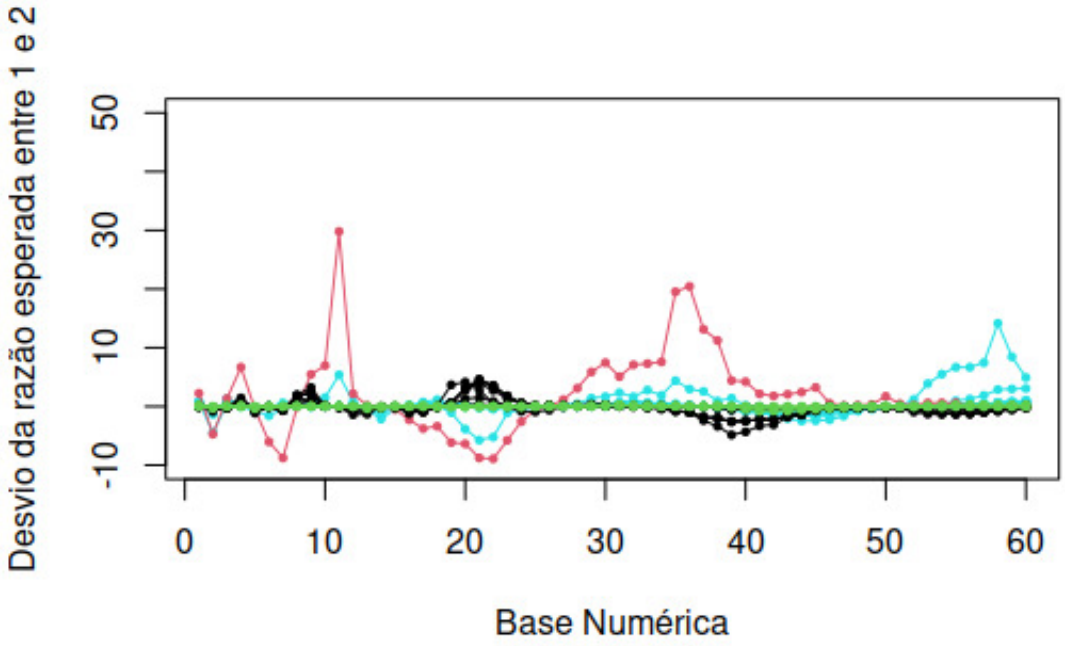


Figura 11: Função erro para relação entre o primeiro dígitos 1 e 2 da coluna pago do conjunto de dados.

3.3.2 *Conclusão da Análise dos Dados*

A conclusão da análise empreendida revela uma congruência com as observações de Bello & Chaves Neto [3], indicando que os dados orçamentários destinados ao enfrentamento da pandemia de COVID-19 no estado do Paraná não aderem à lei de Newcomb-Benford. O exame detalhado realizado neste estudo demonstrou que tanto em fevereiro de 2021 quanto em fevereiro de 2022, ocorreu uma dispersão significativa nos desvios dos dados. A presença de outliers sugere a possibilidade de irregularidades, indicando a necessidade de uma investigação mais aprofundada.

É digno de nota que, nos mencionados meses, embora tenha sido observada uma quantidade menor de dados, esta não foi suficiente para explicar a ocorrência dos outliers. Tal discrepância levanta a hipótese de potenciais problemas orçamentários nesses períodos. Contudo, devido à falta de acesso a informações detalhadas, não é possível fazer afirmações conclusivas sobre a natureza desses desvios. O que se torna evidente, no entanto, é que os dados desses meses não seguem o comportamento padrão observado nos demais meses.

A constatação de tais divergências nos dados orçamentários, particularmente em meses críticos como fevereiro, sugere a pertinência de uma investigação mais aprofundada. A possibilidade de irregularidades nos recursos destinados ao combate à pandemia destaca a importância de uma análise mais minuciosa desses períodos específicos. Em virtude disso, uma investigação mais detalhada é recomendada para elucidar as razões por trás dessas disparidades, visando assegurar a transparência e integridade na gestão dos recursos públicos voltados para o enfrentamento de crises de saúde pública.

4

NEWCOMB-BENFORD NA SALA DE AULA

Nesta capítulo, apresentaremos uma proposta didática para a aplicação da Lei de Newcomb-Benford (LNB) no ensino médio. A LNB aborda diversos tópicos presentes no currículo de matemática da educação básica, incluindo aspectos de estatística e probabilidade.

Inicialmente, acreditávamos que a aplicação da Lei de Newcomb-Benford, por meio de métodos pedagógicos, poderia atribuir significados práticos aos conteúdos tradicionais do currículo. Portanto, essa abordagem tinha duas bases sólidas e interligadas: a revisão e prática de diversos conteúdos relacionados e o estímulo à curiosidade dos alunos em relação ao conhecimento matemático.

O estímulo à curiosidade é uma conquista significativa, pois quando os alunos desenvolvem interesse por um tópico, tendem a fazer perguntas e realizar pesquisas independentes, adquirindo conhecimento de forma autônoma. Em outras palavras, o sucesso de uma aula que visa instigar a curiosidade pode resultar em maior atenção e motivação por parte dos alunos nas aulas subsequentes.

A turma escolhida para a aplicação desses conceitos foi o terceiro ano do ensino médio.

4.1 A PROPOSTA DIDÁTICA

4.1.1 *Objetivos gerais*

A abordagem da Lei de Newcomb-Benford (LNB) baseou-se em conceitos relacionados à Estatística e Probabilidade do ensino médio. O objetivo geral deste estudo foi tríplice:

1. Apresentar, definir e generalizar a Lei de Newcomb-Benford (LNB).

2. Demonstrar o teste qui-quadrado e sua aplicação no contexto da LNB.

3. Proporcionar aos alunos, tanto teoricamente quanto por meio de experimentos práticos, uma aplicação direta dos conteúdos estudados em sala de aula, ao mesmo tempo em que verificavam a aplicação da LNB.

Através desses objetivos, buscamos não apenas ensinar conceitos matemáticos, mas também aproximar os alunos dos procedimentos matemáticos do mundo real. O intuito era fazer com que percebessem que os conteúdos estudados têm um significado prático e real, estimulando-os a buscar essa compreensão de forma mais autônoma, dentro ou fora do ambiente escolar.

4.1.2 *Objetivos específicos.*

Em conformidade com a Base Nacional Comum Curricular (BNCC) atualmente em vigor, estabelecemos os seguintes objetivos específicos:

1. Analisar tabelas, gráficos e amostras de pesquisas estatísticas apresentados em relatórios divulgados por diversos meios de comunicação, com o propósito de identificar padrões e tendências.

2. Planejar e conduzir pesquisas amostrais abordando questões relevantes, fazendo uso de dados obtidos tanto por meio de coleta direta quanto de fontes variadas.

3. Elaborar e interpretar tabelas e gráficos de frequência, com base em dados adquiridos por meio de pesquisas amostrais, incluindo a aplicação de softwares como EXCEL, LibreOffice Calc ou ferramentas similares.

Estes objetivos foram delineados de acordo com as diretrizes da BNCC, buscando promover a capacidade dos alunos de analisar criticamente informações estatísticas, conduzir pesquisas com rigor e competência, e aplicar ferramentas computacionais para a representação e interpretação de dados.

A matemática é uma disciplina que permeia todas as esferas de nossas vidas, e o ensino médio é o momento ideal para não apenas transmitir conhecimentos matemáticos, mas também para revelar a aplicabilidade prática desses conceitos. Neste contexto, apresentamos uma proposta didática que introduzirá os alunos à Lei de Newcomb-

Benford (LNB), uma ferramenta poderosa de análise de dados, enquanto estimula o pensamento crítico e o engajamento prático.

O ensino médio é uma fase crucial na formação dos estudantes, onde consolidam sua compreensão da matemática. No entanto, muitos alunos veem a matemática como uma disciplina teórica e desafiadora, sem perceber sua relevância no mundo real. Nossa proposta visa abordar essa lacuna, tornando o ensino da probabilidade mais envolvente e demonstrando como a LNB, baseada em logaritmos, é aplicada na análise de dados reais.

4.1.3 *Fundamentando a Probabilidade*

Nossa jornada começa com uma base sólida nos princípios da probabilidade. Apresentamos a medida de probabilidade (P) como uma ferramenta essencial para lidar com a incerteza. Definimos P como uma função que associa a cada subconjunto de um espaço amostral um número real no intervalo de $[0,1]$. Esse conceito fundamental é alicerce para compreender eventos aleatórios e sua probabilidade de ocorrência.

O professor deve abordar as seguintes definições para garantir clareza ao definir a probabilidade em LNB:

• Experimentos aleatórios são aqueles que, quando repetidos em condições idênticas, geram resultados distintos. Embora o resultado específico de um experimento seja desconhecido, geralmente podemos descrever o conjunto completo de resultados possíveis que poderiam ocorrer. As variações nos resultados entre experimentos sucessivos são atribuídas a diversas causas não controláveis, comumente denominadas como aleatoriedade.

• O termo “espaço amostral”, representado por Ω, se refere a um conjunto que engloba todos os resultados possíveis de um experimento aleatório.

• Em um contexto de um experimento aleatório com espaço amostral Ω, qualquer subconjunto de Ω é denominado de evento.

Definição 4.1. Sendo Ω um espaço amostral e $\mathcal{P}(\Omega)$ o conjunto de todos os subconjuntos de Ω, probabilidade é uma função

$$P : \mathcal{P}(\Omega) \to [0,1]$$

satisfazendo as condições:

(i) Para todo $A \subset \Omega$, $0 \leq P(A) \leq 1$;

(ii) $P(\Omega) = 1$;

(iii) Para todo $A, B \subset \Omega$, com $A \cap B = \varnothing$, tem-se $P(A \cup B) = P(A) + P(B)$.

Essa abordagem é adotada em vez da abordagem baseada em espaço de medida, devido ao público-alvo, que consiste no ensino médio. É fundamental incluir demonstrações de conceitos, especialmente aquelas que os alunos podem compreender com as ferramentas à disposição, pois isso facilita a compreensão do assunto abordado.

4.1.4 *Propriedades Fundamentais da Probabilidade*

Aprofundamos nossa exploração, destacando propriedades cruciais da medida de probabilidade. Mostramos como a probabilidade se comporta diante de eventos disjuntos e independentes. Essas propriedades são fundamentais para calcular a probabilidade de eventos complexos e compreender seu impacto na análise de dados.

Proposição 4.2. *Sendo A e B eventos de Ω ($A, B \subset \Omega$), então*

(i) $\Omega - A = A^c$ *então,* $P(A^c) = 1 - P(A)$*;*

(ii) $P(\varnothing) = 0$*;*

(iii) $P(A - B) = P(A) - P(A \cap B)$*;*

(iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$*;*

(v) $A \supset B$ *então* $P(A) \geq P(B)$.

Demonstração:

(i) Temos que $A^c \cap A = \varnothing$ como $A^c \cup A = \Omega$, então

$$\begin{aligned} P(A^c \cup A) = P(\Omega) &\iff P(A^c) + P(A) = 1 \\ &\iff P(A^c) = 1 - P(A). \end{aligned}$$

(ii) Como $\Omega \cap \varnothing = \varnothing$ e $\Omega \cup \varnothing = \Omega$, então

$$\begin{aligned} P(\Omega \cup \varnothing) = P(\Omega) &\iff P(\Omega) + P(\varnothing) = P(\Omega) \\ &\iff P(\varnothing) = P(\Omega) - P(\Omega) \\ &\iff P(\varnothing) = 0. \end{aligned}$$

(iii) Segue que $(A - B) \cap (A \cap B) = \varnothing$ e $(A - B) \cup (A \cap B) = A$, então

$$\begin{aligned} P[(A - B) \cup (A \cap B)] = P(A) &\iff P(A - B) + P(A \cap B) = P(A) \\ &\iff P(A - B) = P(A) - P(A \cap B). \end{aligned}$$

(iv) Perceba que $A \cap (B - A) = \varnothing$ e $A \cup B = A \cup (B - A)$, sendo assim

$$P(A \cup B) = P[A \cup (B - A)] \iff P(A \cup B) = P(A) + P(B - A)$$

Portanto por (iii),

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

(v) $B \cap (A - B) = \varnothing$ e $A = B \cup (A - B)$, assim

$$P(A) = P[B \cup (A - B)] \iff P(A) = P(B) + P(A - B)$$

como $P(A - B) \geq 0$, portanto

$$P(A) \geq P(B).$$

□

Proposição 4.3. *Se $A_1, \ldots, A_n \subset \Omega$, com $A_i \cap A_j = \varnothing$, sendo $1 \leq i, j \leq n$ e $i \neq j$, então*

$$P(A_1 \cup \ldots \cup A_n) = P(A_1) + \ldots + P(A_n).$$

Demonstração: Sendo $A_1 \cap A_2 = \varnothing$, então, por definição de probabilidade,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

Supondo válido para n com $A_i \cap A_j = \varnothing$, $1 \leq i, j \leq n$ e $i \neq j$. Se A_{n+1} é tal que $(A_1 \cup \ldots \cup A_n) \cap A_{n+1} = \varnothing$, então

$$\begin{aligned} P(A_1 \cup \ldots \cup A_n \cup A_{n+1}) &= P[(A_1 \cup \ldots \cup A_n) \cup A_{n+1}] \\ &= P(A_1 \cup \ldots \cup A_n) + P(A_{n-1}) \\ &= P(A_1) + \ldots + P(A_n) + P(A_{n+1}) \end{aligned}$$

Portanto para todo $n \in \mathbb{N}$, temos que

$$P(A_1 \cup \ldots \cup A_n) = P(A_1) + \ldots + P(A_n).$$

□

Então pela Proposição 4.3, anteriormente citada, se $A_1, \ldots, A_n \subset \Omega$, com $A_i \cap A_j = \varnothing$, sendo $1 \leq i, j \leq n$ e $i \neq j$, se $A_1 \cup \ldots \cup A_n = \Omega$ e P uma probabilidade, consequentemente

$$\begin{aligned} P(A_1) + \ldots + P(A_n) &= P(A_1 \cup \ldots \cup A_n) \\ &= P(\Omega) \\ &= 1. \end{aligned}$$

Esse resultado é importante para posteriormente determinar se um conjunto de dados numéricos segue a LNB.

4.1.5 *Conexão com Logaritmos*

Aqui, introduzimos uma conexão intrigante entre a probabilidade e os logaritmos. Apresentamos a Lei de Newcomb-Benford (LNB), que se baseia em logaritmos para descrever a distribuição dos primeiros dígitos em conjuntos de dados. Essa lei, inicialmente aplicada em análises financeiras, revela-se uma ferramenta versátil e poderosa que transcende o campo das finanças.

Antes de iniciar a definição da Lei de Newcomb-Benford, é recomendável que o professor faça uma revisão sobre logaritmos, com suas principais propriedades. Aqui fica uma sugestão:

Definição 4.4. O logaritmo de um número em relação a uma base específica é o expoente ao qual a base deve ser elevada para obter esse número. Em outras palavras, se $b^x = a$, então o logaritmo de a (onde $a > 0$) na base b (onde $b > 0$ e $b \neq 1$) é x, e é denotado como $\log_b(a)$.

Observação: a é chamado de logaritmando.

Primeiramente, as propriedades que decorrem imediatamente das definições.

Proposição 4.5. *Se os números b ($b > 0$ e $b \neq 1$) e $a > 0$, então:*

(i) $b^{\log_b(a)} = a$

(ii) $\log_b(b) = 1$

(iii) $\log_b(1) = 0$

Demonstração:

i. Segue da definição que $b^x = a$ e $x = \log_b(a)$. Substituindo, $b^{\log_b(a)} = a$.

ii. Segue que $b^{\log_b(b)} = b = b^1$, então igualando os expoentes, $\log_b(b) = 1$.

iii. Segue que $b^{\log_b(1)} = 1 = b^0$, então igualando os expoentes, $\log_b(1) = 0$.

□

Agora, as propriedades que envolvem conhecimentos prévios de potenciação.

Proposição 4.6. *Se os números b ($b > 0$ e $b \neq 1$) e $a, c > 0$, então:*

(i) $\log_b(a \cdot c) = \log_b(a) + \log_b(c)$

(ii) $\log_b\left(\frac{a}{c}\right) = \log_b(a) - \log_b(c)$

(iii) $\log_b(a^c) = c \cdot \log_b(a)$

(iv) $\log_b(a) = \frac{\log_d(a)}{\log_d(b)}$ *(onde* $1 \neq d > 0$*)*

Demonstração:

i. Sendo os números $a, c > 0$ e $b \neq 1$ e $b > 0$, segue que

$$\begin{cases} ac = b^{\log_b(ac)} \\ ac = b^{\log_b(a)} \cdot b^{\log_b(c)} \end{cases} \iff \begin{cases} ac = b^{\log_b(ac)} \\ ac = b^{\log_b(a)+\log_b(c)} \end{cases}$$
$$\iff b^{\log_b(ac)} = b^{\log_b(a)+\log_b(c)}$$

Portanto,

$$\log_b(ac) = \log_b(a) + \log_b(c).$$

ii. Sendo os números $a, c > 0$ e $b \neq 1$ e $b > 0$, segue que

$$\begin{cases} \frac{a}{c} = b^{\log_b\left(\frac{a}{c}\right)} \\ \frac{a}{c} = \frac{b^{\log_b(a)}}{b^{\log_b(c)}} \end{cases} \iff \begin{cases} \frac{a}{c} = b^{\log_b\left(\frac{a}{c}\right)} \\ \frac{a}{c} = b^{\log_b(a)-\log_b(c)} \end{cases}$$
$$\iff b^{\log_b\left(\frac{a}{c}\right)} = b^{\log_b(a)-\log_b(c)}$$

Portanto,

$$\log_b\left(\frac{a}{c}\right) = \log_b(a) - \log_b(c).$$

iii. Sendo os números $a, c > 0$ e $b \neq 1$ e $b > 0$, segue que

$$\begin{cases} a^c = b^{\log_b(a^c)} \\ a^c = (b^{\log_b(a)})^c \end{cases} \iff \begin{cases} a^c = b^{\log_b(a^c)} \\ a^c = b^{c\cdot\log_b(a)} \end{cases}$$
$$\iff \quad b^{\log_b(a^c)} = b^{c\cdot\log_b(a)}$$

Portanto,

$$\log_b(a^c) = c \cdot \log_b(a).$$

iv. Sendo os números $a > 0$, $1 \neq b > 0$ e $1 \neq d > 0$, segue que

$$\begin{cases} a = b^{\log_b(a)} \\ a = d^{\log_d(a)} \end{cases} \iff \begin{cases} a = (d^{\log_d(b)})^{\log_b(a)} \\ a = d^{\log_d(a)} \end{cases}$$
$$\iff \quad d^{\log_d(b)\cdot\log_b(a)} = d^{\log_d(a)}$$
$$\iff \quad \log_d(b) \cdot \log_b(a) = \log_d(a)$$

Portanto,

$$\log_b(a) = \frac{\log_d(a)}{\log_d(b)}.$$

□

O professor deve definir a Lei de Newcomb-Benford (LNB) no contexto da probabilidade. A formulação proposta é a seguinte:

Definição 4.7. (Probabilidade do primeiro dígito) Dado um conjunto de dados D no qual todo o primeiro dígito $d \in \{1, 2, ..., 9\}$, a probabilidade P de D ter o dígito d é expressa como:

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

Essa definição teórica representa a expectativa para a distribuição dos primeiros dígitos. Por exemplo, considerando um conjunto de dados que segue a Lei de Newcomb-Benford, a frequência esperada de um determinado dígito, como o número 2, é aproximadamente o valor calculado por $\log\left(1 + \frac{1}{2}\right) = 0,176$.

Se P é uma probabilidade, então $P(\Omega) = 1$, em que Ω é um espaço amostral que segue a LNB. Observa-se que Ω pode ser expresso como a união dos conjuntos $D_1, D_2, D_3, ..., D_9$, onde D_i é o conjunto de todos os números que começam com o dígito $i \in \{1, 2, ..., 9\}$. Como $D_i \cap D_j = \varnothing$ com $i \neq j$, temos:

$$\begin{aligned} P(\Omega) &= P(D_1 \cup D_2 \cup D_3 \cup \ldots \cup D_9) \\ &= P(D_1) + P(D_2) + \ldots + P(D_9) \\ &= \log\left(1+\frac{1}{1}\right) + \ldots + \log\left(1+\frac{1}{9}\right) \\ &= \log 2 + \log\frac{3}{2} + \ldots + \log\frac{10}{9} \\ &= \log\left(2\cdot\frac{3}{2}\cdot\frac{4}{3}\cdot\frac{5}{4}\cdot\ldots\cdot\frac{10}{9}\right) \\ &= \log 10 = 1 \end{aligned}$$

Portanto, $P(\Omega) = 1$.

4.1.6 *Exercícios Práticos*

Para reforçar o entendimento, propomos exercícios práticos que desafiam os alunos a aplicar as mudanças de base na LNB. Isso não apenas consolida o conhecimento teórico, mas também promove a compreensão prática desses conceitos.

Segue abaixo uma proposta de exercícios para os alunos:

Exercício 01: Supondo que o conjunto de dados sobre a área hidrográfica de rios de uma determinada região segue a Lei de Newcomb-Benford. Se um pesquisador estiver nessa região e pretender medir a área hidrográfica de um determinado rio, qual é a probabilidade da medida obtida por esse pesquisador começar com o número 3? (Dados: $\log 2 = 0,301$ e $\log 3 = 0,470$)

Solução:

$$\begin{aligned} \log\left(1+\frac{1}{3}\right) &= \log\left(\frac{4}{3}\right) \\ &= \log 4 - \log 3 \\ &= 2\cdot\log 2 - \log 3 \\ &= 2\cdot 0,301 - 0,470 \\ &= 0,602 - 0,470 \\ &= 0,132 = 13,2\% \end{aligned}$$

Exercício 02: Seja um conjunto numérico com 10 mil dados, todos na base 10. Se esse conjunto segue perfeitamente a Lei Newcomb-Benford, determine a quantidade

esperada desses dados que começam com o dígito 8? (Adote: $\log 2 = 0,301$ e $\log 3 = 0,470$)

Solução:

$$\begin{aligned}
\log\left(1+\frac{1}{8}\right) &= \log\left(\frac{9}{8}\right) \\
&= \log 9 - \log 8 \\
&= \log 3^2 - \log 2^3 \\
&= 2\cdot\log 3 - 3\cdot\log 2 \\
&= 2\cdot 0,470 - 3\cdot 0,301 \\
&= 0,940 - 0,903 = 0,037
\end{aligned}$$

Então,

$$0,037 \cdot 10000 = 370$$

Exercício 03: Quando não há distinção, denotamos $P(D = d)$ por $P(d)$, isto é, $P(D = 2) = P(2)$. Definimos a Lei de Newcomb-Benford para os dois primeiros dígitos significativos como sendo $P(d_1d_2) = \log\left(1+\frac{1}{d_1d_2}\right)$, em que d_1d_2 é o número de dois dígitos em que d_1 é o dígito da dezena e d_2 o da unidade. A partir disso, prove que $P(20) + P(21) + \ldots + P(29)$ é igual a $P(2)$.

Solução:

$$\begin{aligned}
P(21) + \ldots + P(29) &= \log\left(1+\frac{1}{20}\right) + \log\left(1+\frac{1}{20}\right) + \ldots + \log\left(1+\frac{1}{29}\right) \\
&= \log\left(\frac{21}{20}\cdot\frac{22}{21}\cdot\ldots\cdot\frac{30}{29}\right) \\
&= \log\frac{30}{20} \\
&= \log\frac{3}{2} \\
&= \log\left(1+\frac{1}{2}\right) \\
&= P(2)
\end{aligned}$$

Exercício 04: Sendo $b \geq 2$ ($b \in \mathbb{N}$) e $d \in \{1, ..., b-1\}$, mostre que

$$\log_b 10 \cdot P(d) = P_b(d)$$

em que $d \in \{1, ..., b-1\}$.

Solução:

$$\log_b 10 \cdot P(d) = \frac{\log 10}{\log b} \cdot \log\left(1 + \frac{1}{d}\right) = \frac{\log\left(1 + \frac{1}{d}\right)}{\log b} = \log_b\left(1 + \frac{1}{d}\right) = P_b(d)$$

Exercício 05: Na base 3, segundo a Lei de Newcomb-Benford, qual é a probabilidade do número começar com os dígitos $(100)_3$? (Adote: $\log 3 = 0,47$)

Solução:

$$100_3 = 1 \cdot 3^2 + 0 \cdot 3^1 + 0 \cdot 3^0 = (9)_{10}$$

Então,

$$\begin{aligned} P_3((100)_3) &= P_3(9) \\ &= \log_3\left(1 + \frac{1}{9}\right) \\ &= \frac{\log(10/9)}{\log 3} \\ &= \frac{\log 10 - \log 9}{\log 3} \\ &= \frac{1 - 2\log 3}{\log 3} \\ &= \frac{1 - 0,94}{0,47} \\ &= \frac{0,06}{0,47} \\ &= \frac{6}{47} \end{aligned}$$

Portanto,

$$P_3((100)_3) = 12,76\%$$

Exercício 06: Sendo $2 \leq b \in \mathbb{N}$ e $d \in \{1, \ldots, b-1\}$, mostre que

$$P_b(d \cdot b) + P_b(d \cdot b + 1) + \ldots + P_b(d \cdot b + (b-1)) = P_b(d)$$

em que $d \in \{1, \ldots, b-1\}$.

Solução:

$$\begin{aligned}
P_b(d\cdot b)+\ldots+P_b(d\cdot b+(b-1)) &= \log_b\left(1+\frac{1}{db}\right)+\ldots+\log_b\left(1+\frac{1}{db+(b-1)}\right)\\
&= \log_b\left[\frac{db+1}{db}\cdot\frac{db+2}{db+1}\cdot\ldots\cdot\frac{db+b}{db+(b-1)}\right]\\
&= \log_b\left(\frac{db+b}{db}\right)\\
&= \log_b\left(\frac{b(d+1)}{db}\right)\\
&= \log_b\left(\frac{d+1}{d}\right)\\
&= \log_b\left(1+\frac{1}{d}\right)\\
&= P_b(d)
\end{aligned}$$

Esses exercícios visam aprofundar a compreensão dos alunos sobre a Lei de Newcomb-Benford e suas propriedades, bem como desenvolver habilidades na manipulação de probabilidades e logaritmos em contextos específicos.

4.1.7 *Aplicabilidade da LNB*

Aprofundamos a aplicação da LNB em contextos do mundo real. Explicamos quando e por que normalmente se aplica a LNB, destacando suas implicações em áreas tão diversas quanto auditoria financeira e detecção de fraudes. Mostramos como a matemática da probabilidade e a LNB são ferramentas essenciais para tomar decisões informadas em situações complexas.

4.1.8 *Explorando o Mundo dos Dados*

Nossa jornada se estende à pesquisa de conjuntos de dados disponíveis na internet que seguem os princípios da LNB. Os alunos compreendem que a LNB tem aplicações em uma variedade de campos, desde informações financeiras até estatísticas populacionais, dados científicos e registros históricos.

4.1.9 *Ferramentas de Análise de Dados*

Para equipar os alunos com habilidades práticas, introduzimos o uso de ferramentas de planilha, como Excel, LibreOffice Calc ou Google Planilhas. Essas ferramentas, comuns no ambiente profissional, capacitam os alunos a organizar e analisar dados complexos de maneira eficaz.

4.1.10 *Conversão de Bases e Tabelas de Frequência*

Uma das partes mais envolventes da proposta é a conversão dos dados da base 10 para uma nova base escolhida pelos próprios alunos. Com os dados na nova base, eles criam tabelas de frequência para os primeiros dígitos, explorando a flexibilidade da LNB em diferentes contextos.

Função BASE: A função BASE(“Celula”; “Base numérica”) desempenha um papel fundamental na conversão de um número para uma representação textual em uma base específica. Sua estrutura básica consiste em dois argumentos: o número a ser convertido e a base para a qual realizar a conversão.

Argumentos:

- **Celula:** Este é o número que o aluno deseja converter. Ele deve inserir o valor desejado na célula B2.
- **Base numérica:** Este é o argumento que especifica a base para a qual o número será convertido. No caso de “=BASE(B2; 9)”, o número na célula B2 será convertido para sua representação na base 9.

Exemplo: Se, por exemplo, a célula B2 contiver o número 10, a função “=BASE(B2; 9)” retornará o valor “11” na célula onde a fórmula foi aplicada. Isso significa que 10 em base 10 é equivalente a 11 em base 9, representando $1 \cdot 9^1 + 1 \cdot 9^0$.

Aplicação em Outras Células: Uma vez que a fórmula foi inserida em uma célula, como C2, por exemplo, o aluno pode arrastar a alça de preenchimento para baixo, replicando a fórmula para outras células. Isso automatiza o processo de conversão para todos os números desejados na coluna B.

Portanto, a utilização da função “=BASE(B2; 9)” permite ao aluno converter números para sua representação na base 9 de maneira eficiente no LibreOffice Calc.

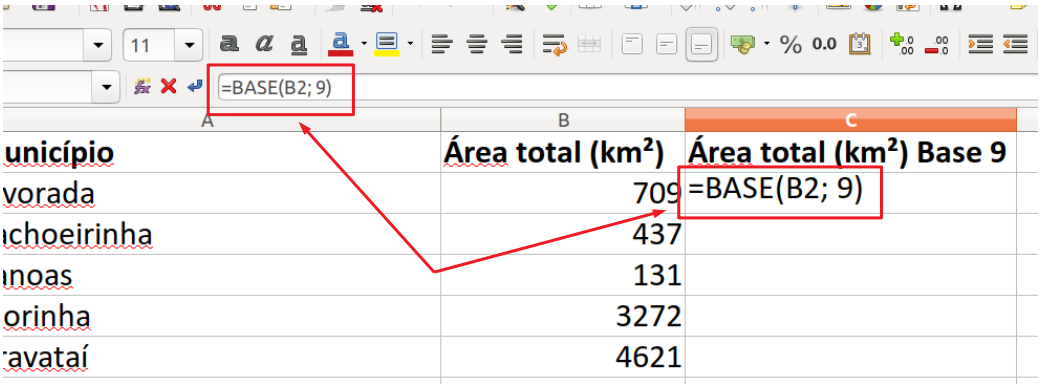


Figura 12: Convertendo o valor da célula B2 para a base 9.

4.1.11 *Explorando os Primeiros Dígitos*

Enfatizamos a importância de isolar os primeiros dígitos dos números nas análises. Essa etapa é essencial para a aplicação bem-sucedida da LNB, uma vez que a lei se concentra precisamente nesses dígitos.

O procedimento instrucional para extrair e converter o primeiro caractere de uma célula usando funções específicas no LibreOffice Calc é delineado a seguir:

O aluno iniciará a operação utilizando a função **ESQUERDA** disponível no LibreOffice Calc. Esta função é projetada para extrair um número determinado de caracteres do início de uma string. A sintaxe básica para a extração do primeiro caractere da célula B2 é exemplificada da seguinte maneira: “=ESQUERDA(B2; 1)”.

Em seguida, será necessário empregar a função **VALOR** para converter o texto extraído em um valor numérico. A função VALOR tenta interpretar o conteúdo da célula como um número.

Portanto, a expressão a ser inserida pelo aluno é “=VALOR(ESQUERDA(B2; 1))”, que essencialmente extrai o primeiro caractere da célula B2 e o converte em um valor numérico.

Para ilustrar, se a célula B2 contiver o texto “123”, então a função “ESQUERDA(B2; 1)” extrairá “1”, e a expressão “=VALOR(ESQUERDA(B2; 1))” converterá “1” em 1 (um número).

É fundamental salientar que o resultado pode variar de acordo com o conteúdo específico da célula B2. Caso o primeiro caractere não seja um dígito numérico, a

função “VALOR” poderá retornar um erro ou 0. Portanto, é crucial assegurar que os dados estejam alinhados conforme as expectativas dessas funções.

Posteriormente, os alunos deverão aplicar esse procedimento às demais células conforme necessário.

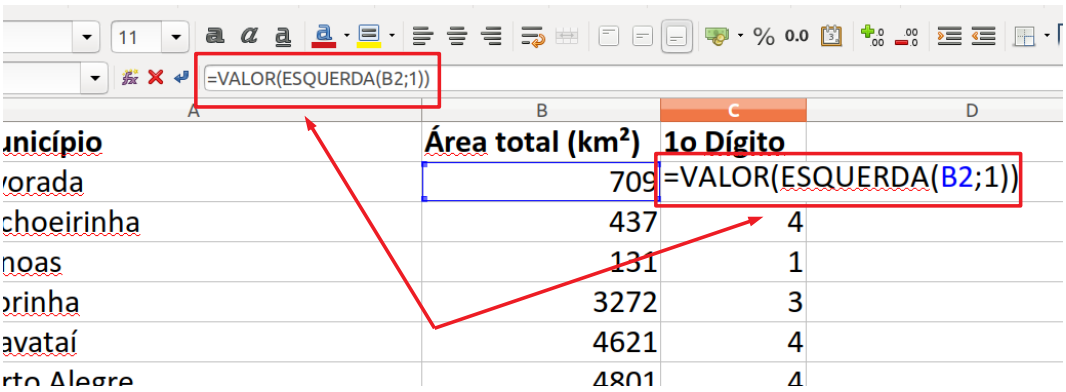


Figura 13: Como inserir somente o primeiro dígito?

4.1.12 *Registro da Contagem de Primeiros Dígitos na Base Selecionada*

Os alunos terão a tarefa de elaborar uma tabela de frequência para os primeiros dígitos significativos da base escolhida, pois essa frequência desempenha um papel crucial na compreensão da distribuição da Lei de Newcomb-Benford (LNB).

Primeiramente, eles realizarão a construção de uma coluna destinada a abrigar os primeiros dígitos correspondentes à base selecionada.

Em seguida, utilizarão a função “=CONT.SE(“Intervalo”; “Critério”)”.

Função CONT.SE: Essa função desempenha a contagem do número de células em um intervalo que atende a um critério específico.

Argumentos:

- **Intervalo:** Este é o conjunto de células a ser considerado para a contagem, por exemplo, $E$2:$E$694, abrangendo da célula E2 até a E694.
- **Critério:** Este é o critério adotado pela função para contabilizar as células no intervalo especificado. Por exemplo, a função visa contar quantas células no intervalo $E$2:$E$694 possuem o mesmo valor que a célula E2.

Exemplo: A função “=CONT.SE($D$2:$D$694;E2)” realiza a contagem de células no intervalo E2:E694 que compartilham o mesmo valor presente na célula E2. Essa abordagem é valiosa para determinar a ocorrência de um valor específico dentro de um intervalo.

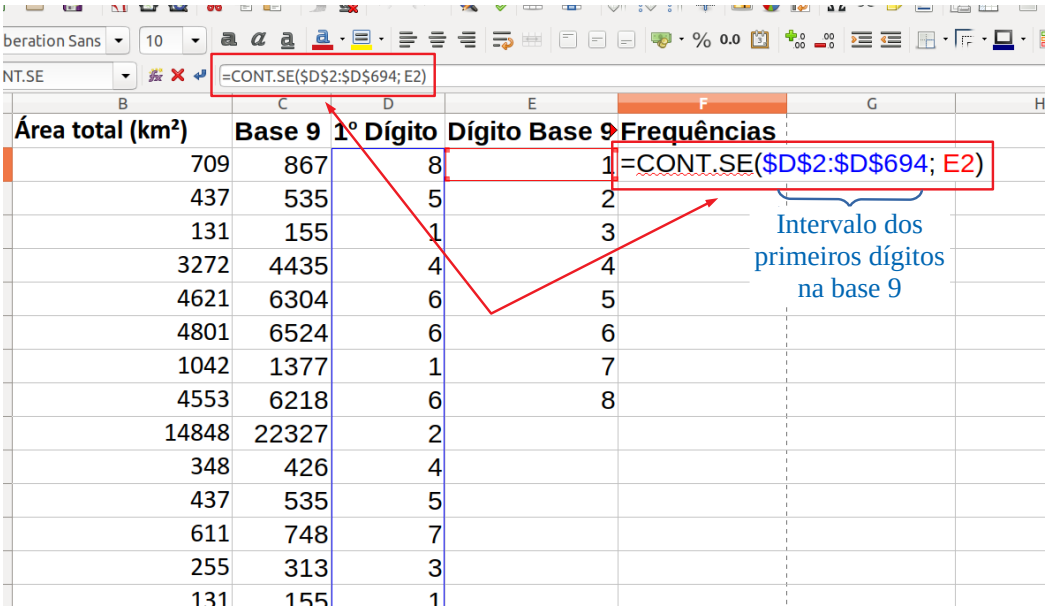


Figura 14: Fazendo o Frequencia do primeiro digito

Posteriormente, os estudantes estenderão essa função para as demais células, ampliando, assim, a análise para todo o conjunto de dados.

Adicionalmente, criarão uma coluna de porcentagem para os dígitos em relação à quantidade total de dados, utilizando a expressão “=Célula com quantidade do Dígito/Quantidade dos dados” no LibreOffice Calc.

Portanto, utilizando a fórmula “=K2/693”, por exemplo, realiza-se a divisão do valor na célula K2 por 693. O resultado da divisão seria 2, se K2 contiver, por exemplo, o valor 1386, pois 1386 dividido por 693 é igual a 2. Essa informação, quando expressa em porcentagem no LibreOffice Calc, seria arredondada para 200

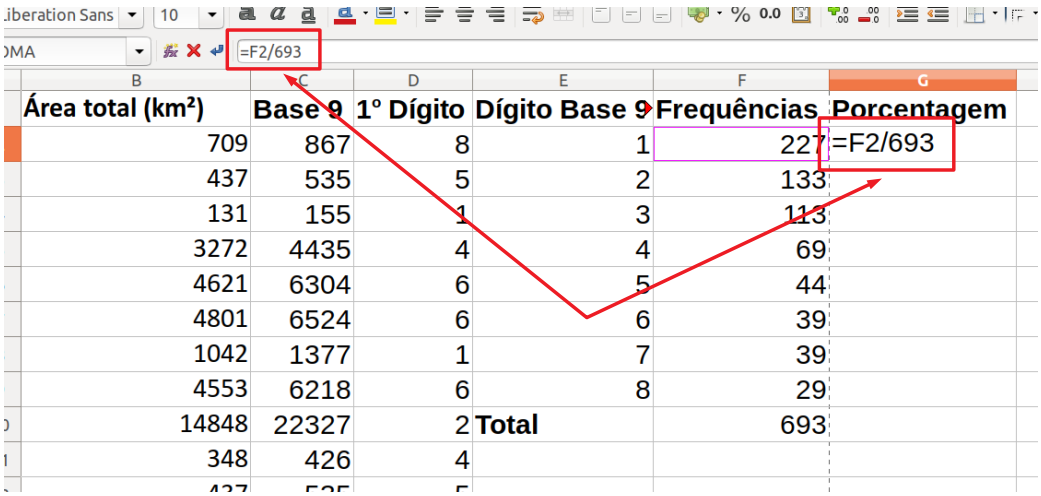


Figura 15: Fazendo o Porcentagem

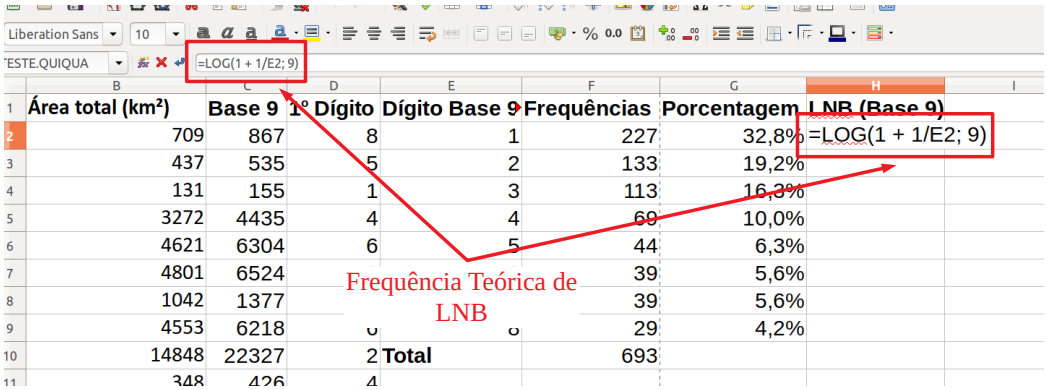
4.1.13 *Comparação com Valores Teóricos*

Os alunos estudam as sequências esperadas dos dados, tanto na base 10 quanto na nova base, de acordo com os valores teóricos da LNB. Isso ilustra a ligação entre a teoria e a aplicação prática da LNB.

O aluno, ao criar uma tabela teórica com os valores da Lei de Newcomb-Benford, utiliza a fórmula "=LOG(1+1/"Dígito", "Base")" em uma planilha eletrônica do LibreOffice. Essa fórmula calcula o logaritmo em uma base específica do valor resultante de "1 + 1/Dígito", onde "Dígito" representa um valor da coluna "Dígito" para aquela base. Essa expressão teórica, $P(d) = \log_b(1+1/d)$, representa o valor teórico esperado.

Em seguida, o aluno aplica essa fórmula para as demais células arrastando. Por exemplo, "=LOG(1+1/E2, 9)". Nesta fórmula, é calculado o logaritmo na base 9 do valor resultante de "1 + 1/E2". O termo "E2" provavelmente se refere a uma célula específica na planilha, e o uso de "1/E2" indica a inversa do valor contido na célula E2.

Com a criação da tabela contendo a porcentagem que indica a frequência observada dos dados e a tabela da Lei de Newcomb-Benford (Base " "), os alunos podem utilizar o teste qui-quadrado para avaliar quão próxima a distribuição dos dados está da Lei de Newcomb-Benford. Quanto mais próximo de um, mais próxima a distribuição dos dados está da lei.



	B	C	D	E	F	G	H	I
1	Área total (km²)	Base 9	1º Dígito	Dígito Base 9	Frequências	Porcentagem	LNB (Base 9)	
2	709	867	8	1	227	32,8%	=LOG(1 + 1/E2; 9)	
3	437	535	5	2	133	19,2%		
4	131	155	1	3	113	16,3%		
5	3272	4435	4	4	69	10,0%		
6	4621	6304	6	5	44	6,3%		
7	4801	6524			39	5,6%		
8	1042	1377			39	5,6%		
9	4553	6218			29	4,2%		
10	14848	22327	2	Total	693			
11	348	426	4					

Figura 16: Fazendo os valores teórico de LNB

A função **TESTE.QUIQUA** em uma planilha do LibreOffice realiza um teste qui-quadrado de independência entre duas variáveis categóricas, e a aplicação é feita da seguinte forma:

“=TESTE.QUIQUA(“Intervalo1”; “Intervalo2”)”

Os argumentos da função são:

- **Intervalo1:** Este é o intervalo de células que contém os dados da primeira variável categórica.
- **Intervalo2:** Este é o intervalo de células que contém os dados da segunda variável categórica.

A função compara as frequências observadas nas categorias cruzadas dessas duas variáveis com as frequências que seriam esperadas se as variáveis fossem independentes uma da outra.

O resultado do teste qui-quadrado é uma estatística qui-quadrado, que é comparada a uma distribuição qui-quadrado para determinar se as variáveis são independentes ou se há uma associação significativa entre elas.

É importante lembrar que a interpretação do teste qui-quadrado depende do contexto dos dados e do estudo específico em que está sendo aplicado. Por exemplo, o teste pode ser aplicado entre as variáveis “G2:G9” e “H2:H9” da seguinte forma: “=TESTE.QUIQUA(G2:G9; H2:H9)”.

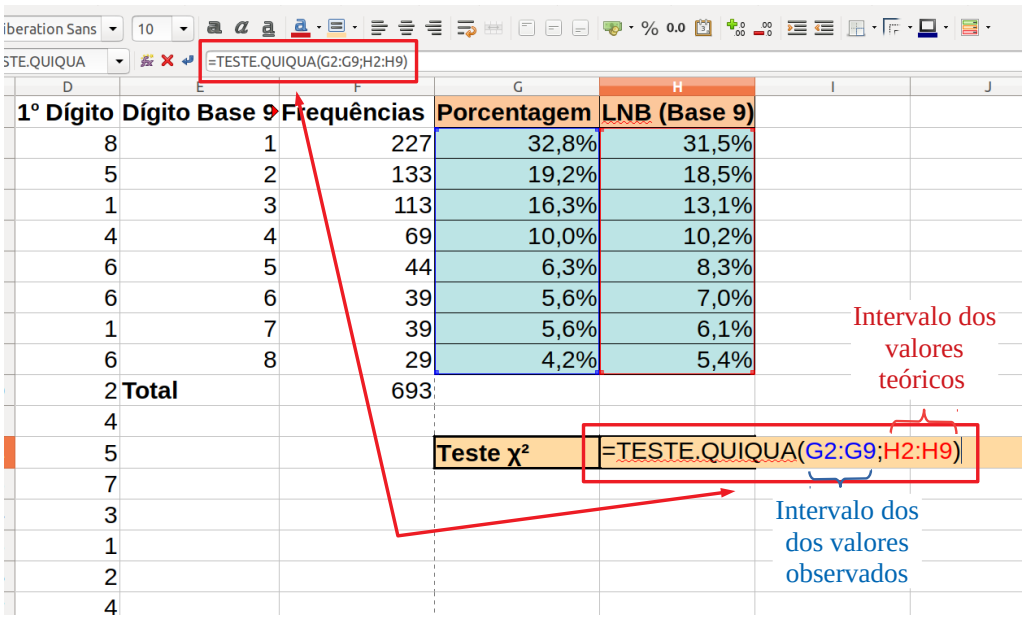


Figura 17: Fazendo o teste Qui

4.1.14 *Visualização e Análise dos Resultados*

Na etapa de visualização e análise dos resultados, os alunos entram em uma fase empolgante. Eles criam gráficos de barras que representam os primeiros dígitos nas tabelas da nova base. Essa visualização permite uma análise mais aprofundada das diferenças entre as frequências observadas e as esperadas.

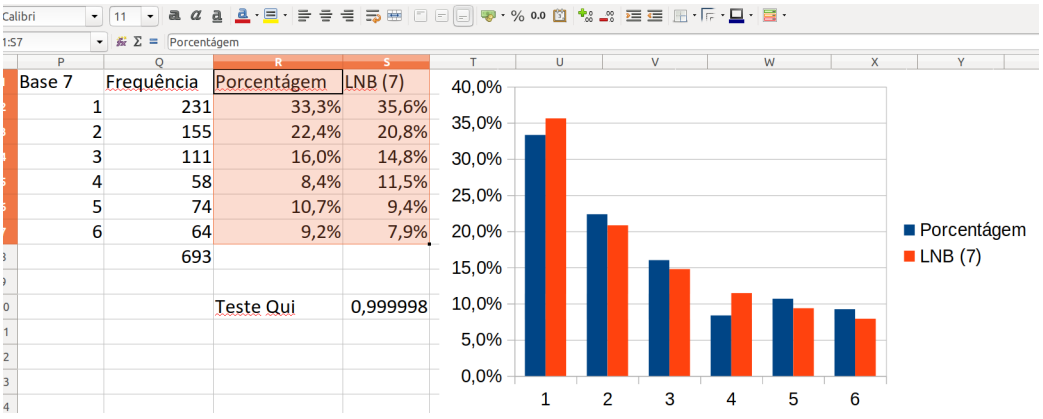


Figura 18: Comparando os Gráficos

4.1.15 *Discussão e Reflexão*

Na última etapa, os alunos se reúnem para uma discussão e reflexão sobre os resultados obtidos. Espera-se que, ao aplicar a LNB, eles não percebam uma diferença significativa entre as frequências dos primeiros dígitos nas duas bases. Isso demonstra a eficácia da LNB na análise de dados, independentemente da base original.

4.1.16 *Conclusão: Preparando para o Mundo Real*

Em resumo, esta proposta didática oferece aos alunos uma oportunidade única de explorar o mundo da análise de dados de maneira envolvente e prática. Mais do que apenas adquirir habilidades valiosas em matemática, os alunos aprendem a aplicar esses conhecimentos no mundo real. A LNB, uma ferramenta matemática poderosa, torna-se acessível e relevante, preparando os alunos para serem pensadores analíticos e informados.

Ao investir no potencial das próximas gerações, estamos construindo um futuro onde a matemática não é apenas uma disciplina abstrata, mas uma ferramenta essencial para a compreensão e transformação do mundo ao nosso redor. Esta proposta pedagógica visa iluminar esse caminho, capacitando os alunos a enfrentar desafios complexos com confiança, habilidades práticas e um profundo entendimento da matemática aplicada.

BIBLIOGRAFIA

[1] Marcel Ausloos, Roy Cerqueti e Claudio Lupi, *Long-range properties and data validity for hydrogeological time series: The case of the Paglia river*, Physica A: Statistical Mechanics and its Applications **470** (2017), 39–50.

[2] Marcel Ausloos, Roy Cerqueti e Tariq A Mir, *Data science for assessing possible tax income manipulation: The case of Italy*, Chaos, Solitons & Fractals **104** (2017), 238–256.

[3] Vitória Eduarda Bello e Anselmo Chaves Neto, *Lei de Newcomb-Benford: uma aplicação com dados de execução orçamentária de despesas com a pandemia de COVID-19 no estado do paraná*, Revista de Engenharia e Tecnologia **15** (2023), nº 1.

[4] Frank Benford, *The law of anomalous numbers*, Proceedings of the American philosophical society (1938), 551–572.

[5] Arno Berger, *Multi-dimensional dynamical systems and Benford's law*, Discrete Contin. Dyn. Syst **13** (2005), nº 1, 219–237.

[6] Arno Berger, Leonid Bunimovich e Theodore Hill, *One-dimensional dynamical systems and Benford's Law*, Transactions of the American Mathematical Society **357** (2005), nº 1, 197–219.

[7] Arno Berger e Theodore P Hill, *An introduction to Benford's law*, Princeton University Press, 2015.

[8] BRASIL, *Governo do Estado do Paraná. Portal da Transparência: Execução Orçamentária - Dados Abertos*, Disponível em: https://www.coronavirus.pr.gov.br/execucao-orcamentaria-dados-abertos . Acesso em: fevereiro, 2023.

[9] _____, *Governo do Estado do Rio Grande do Sul. Secretaria do Meio Ambiente e Infraestrutura. Anexo_1 - Sema RS*, Disponível em: https://www.sema.rs.gov.br/upload/arquivos/202008/07161358-nt-dipla-2020-002-municipios-e-bacias-anexos.xlsx . Acesso em: fevereiro, 2023.

[10] Persi Diaconis e David Freedman, *On rounding percentages*, Journal of the American Statistical Association **74** (1979), nº 366a, 359–364.

[11] Cindy Durtschi, William Hillison, Carl Pacini et al., *The effective use of Benford's law to assist in detecting fraud in accounting data*, Journal of forensic accounting **5** (2004), nº 1, 17–34.

[12] Eduardo Gueron e Jerônimo Pellegrini, *Application of Benford–Newcomb law with base change to electoral fraud detection*, Physica A: Statistical Mechanics and its Applications **607** (2022), 128208.

[13] Theodore P Hill, *Base-invariance implies Benford's law*, Proceedings of the American Mathematical Society **123** (1995), nº 3, 887–895.

[14] _____, *The significant-digit phenomenon*, The American Mathematical Monthly **102** (1995), nº 4, 322–327.

[15] _____, *The first digit phenomenon: A century-old observation about an unexpected pattern in many numerical tables applies to the stock market, census statistics and accounting data*, American Scientist **86** (1998), nº 4, 358–363.

[16] Adrien Jamain, *Benford's law*, Unpublished Dissertation Report, Department of Mathematics, Imperial College, London (2001).

[17] Mario Livio, *Why math works*, Scientific American **305** (2011), nº 2, 80–83.

[18] Steven J Miller, *Benford's law*, Princeton University Press, 2015.

[19] Simon Newcomb, *Note on the frequency of use of the different digits in natural numbers*, American Journal of mathematics **4** (1881), nº 1, 39–40.

[20] Mark J Nigrini, *I've got your number*, Journal of accountancy **187** (1999), nº 5, 79–83.

[21] _____, *Digital Analysis Using Benford's Law: Test and Statistics for Auditors*, Global Audit Publications, 2000.

[22] Mark J Nigrini e Linda J Mittermaier, *The use of Benford's Law as an Aid in Analytical Procedures.*, Auditing: A journal of practice & theory **16** (1997), nº 2.

[23] M Nigrini, *Using digital frequencies to detect fraud*, The white paper **8** (1994), nº 2, 3–6.

[24] M Nigrini e LJ Mittermaier, *Numerology for accountants*, Journal of Accountancy **186** (1998), nº 5, 15–16.

[25] John Nye e Charles Moul, *The political economy of numbers: on the application of Benford's law to international macroeconomic statistics*, The BE Journal of Macroeconomics **7** (2007), nº 1.

[26] Roger S Pinkham, *On the distribution of first significant digits*, The Annals of Mathematical Statistics **32** (1961), nº 4, 1223–1230.

[27] HR Varian, *Benford's Law (Letters to the Editor)(1972)*, The American Statistician **26**, nº 3, 62–65.

[28] Hermann Weyl, *Über die gleichverteilung von zahlen mod. eins*, Mathematische Annalen **77** (1916), nº 3, 313–352.