UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Lei de Benford e detecção de fraudes

Pedro Fernando da Silva Biajotto

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática em Rede Nacional (PROFMAT)

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: ______________________

Pedro Fernando da Silva Biajotto

Lei de Benford e detecção de fraudes

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Mestrado Profissional em Matemática em Rede Nacional. *VERSÃO REVISADA*

Área de Concentração: Mestrado Profissional em Matemática em Rede Nacional

Orientador: Prof. Dr. Miguel Vinícius Santini Frasson

USP – São Carlos
Janeiro de 2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Biajotto, Pedro Fernando da Silva
B576l Lei de Benford e detecção de fraudes / Pedro Fernando da Silva Biajotto; orientador Miguel Vinícius Santini Frasson. -- São Carlos, 2024.
39 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Mestrado Profissional em Matemática em Rede Nacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Lei de Benford. 2. detecção de fraude. 3. logaritmo. 4. estatística. I. Frasson, Miguel Vinícius Santini, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Pedro Fernando da Silva Biajotto

Benford’s law and fraud detection

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of Mathematics Professional Master’s Program. *FINAL VERSION*

Concentration Area: Professional Master Degree Program in Mathematics in National Network

Advisor: Prof. Dr. Miguel Vinícius Santini Frasson

USP – São Carlos
January 2025

À minha família,
que sempre me incentivou
a continuar meus estudos.

AGRADECIMENTOS

Agradeço primeiramente a Deus pela minha vida e por permitir que eu alcançasse e conquistasse feitos que jamais imaginei.

À minha família, pelo incentivo, pelo amor e pelo apoio nos momentos difíceis. À minha esposa Elaine Cristina Gomes, por ser imprescindível na minha vida, por ter participado, apoiado e compartilhado das alegrias e das angústias. Ao meu filho Lucas Gomes Reis, por ter tornado os meus dias muito mais felizes. Agradeço pelo amor, companheirismo e pelo apoio constante.

À minha sogra, Maria Auxiliadora Teodoro Gomes,ao meu sogro José David Gomes (*in memoriam*), agradeço pelo carinho que vocês sempre tiveram comigo, obrigado por sempre acreditarem em mim, tenho imensa gratidão por vocês.

Aos meus irmãos, Danilo e William, por tudo que representam para mim.

Especial gratidão aos meus pais, Antônio Carlos Biajotto (*in memoriam*) e Jacira Fátima Silva Biajotto, por incentivarem meus estudos e por me apoiarem em todos os momentos.

E por fim gostaria de expressar minha gratidão ao meu professor e orientador Prof. Dr. Miguel Vinícius Santini Frasson, que desempenhou um papel fundamental em meu enriquecimento acadêmico. Seu incentivo, orientações firmes quando necessárias e a dedicação durante nossas reuniões e conversas, demonstram seu grande comprometimento. Agradeço por nunca me abandonar e por ser uma fonte constante de inspiração ao longo desta jornada.

"Nossa exploração não deve nunca cessar
E o fim de toda exploração
Será voltar ao lugar de onde partimos
E o conhecer pela primeira vez"
T.S. Eliot.

RESUMO

BIAJOTTO, P. F. DA S. **Lei de Benford e detecção de fraudes**. 2025. 40 p. Dissertação (Mestrado em Ciências – Mestrado Profissional em Matemática em Rede Nacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

A lei de Benford foi cunhada pela observação do astrônomo Simon Newcomb de seu livro de tábuas de logaritmos tinha marcas de uso mais nas primeiras páginas (para as mantissas que começam com 1) do que nas finais (para mantissas que iniciam com 9). Isso significa que os números que iniciavam com 1 eram muito mais comuns de serem verificados. Décadas mais tarde, o físico Frank Benford redescobriu e a testou numa grande quantidade de dados, como comprimentos de rios, populações, constantes científicas, números usados por pessoas num livro de romances etc. Newcomb e Benford notaram que os dígitos significativos dos números têm um padrão logaritmico. Por exemplo, números que iniciam com o dígito 1 aparecem com frequência próxima de $30\% = \log 2$. A aplicação mais impressionante é a detecção de dados gerados por pessoas, que tendem a distribuir os primeiros dígitos de uma maneira uniforme, sendo uma importante ferramenta para a detecção de fraudes. A lei de Benford pode ser usada para estimular o estudo de estatística, logaritmos e de números em notação científica. Planilhas podem ser usadas para calcular a frequência de dados.

Palavras-chave: Lei de Benford, detecção de fraudes, logaritmo, estatística.

ABSTRACT

BIAJOTTO, P. F. DA S. **Benford's law and fraud detection**. 2025. 40 p. Dissertação (Mestrado em Ciências – Mestrado Profissional em Matemática em Rede Nacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

Benford's law was coined by astronomer Simon Newcomb when he observed that his book of tables of logarithms had usage marks more on the first pages (for mantissas beginning with 1) than on the final ones (for mantissas beginning with 9). This means that numbers starting with 1 were much more common to be checked. Decades later, physicist Frank Benford rediscovered it and tested it on a large amount of data, such as river lengths, populations, scientific constants, numbers used by people in a book of novels, etc. Newcomb and Benford noticed that the significant digits of numbers have a logarithmic pattern. For example, numbers starting with the digit 1 appear with a frequency close to $30\% = \log 2$. The most impressive application is the detection of human-generated data, which tends to distribute the first digits evenly, making it an important tool for detecting fraud. Benford's law can be used to encourage the study of statistics, logarithms and numbers in scientific notation. Spreadsheets can be used to calculate data frequency.

Keywords: Benford's law, fraud detection, logarithm, statistics.

LISTA DE ILUSTRAÇÕES

Figura 1 – Frequências esperadas de acordo com a Lei de Benford. 22
Figura 2 – Captura de tela da planilha com os dados da população dos municípios do Brasil. 30
Figura 3 – Captura de tela da planilha com os dados do 2º dígito para a população dos municípios do Brasil. 33

SUMÁRIO

1 INTRODUÇÃO 19

2 LEI DE BENFORD 21

2.1 O que é a Lei de Benford 21

2.2 Como foi descoberta a Lei de Benford 21

2.3 A Lei de Benford para o primeiro dígito 22

2.4 A Lei de Benford para o segundo dígito 23

2.5 Detecção de fraudes 24

3 APLICANDO A LEI DE BENFORD A CONJUNTOS DE DADOS . 27

3.1 Introdução 27

3.2 Obtendo o primeiro dígito de um número 28

3.3 Exemplo 1 – população dos municípios no Brasil 29

3.4 Exemplo 2 – casos de COVID no Brasil 32

4 UM PLANO DE AULA 37

4.1 Plano de Aula 37

REFERÊNCIAS 39

CAPÍTULO

1

INTRODUÇÃO

Este trabalho visa explorar a Lei de Benford em profundidade, examinando suas origens históricas, princípios fundamentais e aplicações práticas. Ao longo do trabalho, serão abordados tópicos como a justificativa teórica por trás da lei, métodos de aplicação em diferentes contextos e estudos de casos emblemáticos que destacam seu impacto e utilidade. Além disso, serão discutidas as limitações e desafios associados à aplicação da Lei de Benford, promovendo uma compreensão abrangente de suas implicações e potenciais restrições.

No Capítulo 2, veremos o que é e e como foi descoberta a Lei de Benford, que se trata da distribuição do primeiro dígito de números em muitos conjuntos de dados do mundo real. A Lei de Benford é frequentemente utilizada em auditorias financeiras, contabilidade forense e detecção de fraudes devido à sua capacidade de identificar irregularidades nos dados. A extensão da Lei de Benford para o segundo dígito também é abordada. O capítulo também explora a aplicação da Lei de Benford em conjuntos de dados reais, destacando exemplos emblemáticos de detecção de fraudes.

O Capítulo 3 mergulha na aplicação prática da Lei de Benford em conjuntos de dados do mundo real, explorando sua relevância na detecção de padrões estatísticos e na identificação de possíveis irregularidades. Este capítulo aborda a metodologia detalhada para a aplicação da Lei de Benford, desde a coleta e organização dos dados em planilhas até a extração e análise dos primeiros dígitos significativos. Dois exemplos notáveis serão estudados: a distribuiçao da população das cidades brasileiras e dados relacionados à pandemia de COVID-19 no Brasil. O intuito é chamar a atenção dos alunos para aplicações interdisciplinares de ferramentas matemáticas, como a notação científica na extração dos primeiros dígitos será discutida em detalhes, destacando a importância dos logaritmos na obtenção da ordem de grandeza e da mantissa de um número, além de planilhas como uma ferramenta para organização, manipulação e contagem de dados.

Finalmente, no Capítulo 4, apresentamos um plano de aula para o estudo da Lei de Benford, como um chamariz da utilidade da matemática, como também um auxílio na apresentação de ferramentas matemáticas como a notação científica e os logaritmos.

CAPÍTULO

2

LEI DE BENFORD

2.1 O que é a Lei de Benford

A Lei de Benford é um princípio estatístico que afirma que, em conjuntos de dados do mundo real, o primeiro dígito significativo (ou principal) não é distribuído uniformemente, como poderíamos esperar ingenuamente, mas segue uma distribuição logarítmica específica. Em termos simples, o dígito 1 tende a aparecer como o primeiro dígito mais frequentemente do que os outros dígitos em muitos conjuntos de dados.

A Lei de Benford é frequentemente utilizada em auditorias financeiras, contabilidade forense e detecção de fraudes porque dados fraudulentos muitas vezes não seguem essa distribuição esperada. Se um conjunto de dados não segue a Lei de Benford, pode ser um indicativo de manipulação ou erro nos dados.

2.2 Como foi descoberta a Lei de Benford

O fenômeno foi observado pela primeira vez pelo astrônomo Simon Newcomb por volta de 1880, que notou que as primeiras páginas de seu livro de tábua de logaritmos, que correspondiam aos primeiros algarismos significativos pequenos, como 1 ou 2, eram muito mais utilizados do que as páginas seguintes, que correspondiam aos primeiros algarismos significativos grandes, como 8 ou 9. Isso se constatava pelas bordas mais escurecidas das primeiras partes dos volumes de tábuas de logaritmos, ou seja, as primeiras páginas eram muito mais utilizadas do que as últimas. Newcomb sugere que a probabilidade de um algarismo d ser o primeiro dígito de um número era igual a $\log(d+1) - \log d$, onde log denota os logaritmos na base 10. Porém sua descoberta foi esquecida, sendo redescoberta por Frank Benford (1938), que a havia chamado de Lei dos Números Anômalos. Benford estava analisando dezena de milhares de números de 20 domínios diferentes, dentre eles estavam áreas de superfície de 335 rios, tamanho de populações de 3259 locais dos EUA, 104 constantes físicas, 1800 pesos moleculares, 5000 entradas de um livro matemático, 308 números contidos em uma edição da Reader's Digest, os 342 primeiros endereços listados na *American Men of Science* e 418 taxas de mortalidade, quando notou uma distribuição não uniforme dos primeiros dígitos. Ele observou que o dígito 1 aparecia com mais frequência como o primeiro dígito do que os dígitos subsequentes, seguido pelo dígito 2, 3 e assim por diante, em uma progressão decrescente.

2.3 A Lei de Benford para o primeiro dígito

A Lei de Benford afirma que, em muitos conjuntos de dados do mundo real, os dígitos iniciais de números não são distribuídos de maneira uniforme, como seria esperado intuitivamente. Em vez disso, a probabilidade de um dígito ser o primeiro parece ser inversamente proporcional ao seu valor. Em outras palavras, os números que começam com dígitos menores são mais comuns do que aqueles que começam com dígitos maiores.

Benford deduziu que a probabilidade $P(d)$ de um dígito d ocorrer como o primeiro dígito sigfinicativo de um número é

$$P(d) = \log(d+1) - \log d = \log\left(\frac{d+1}{d}\right) = \log\left(1+\frac{1}{d}\right). \tag{2.1}$$

Podemos organizar esses valores na tabela abaixo, podendo ser visualizados na Figura 1:

dígito	1	2	3	4	5	6	7	8	9
freq. lei de Beford	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046

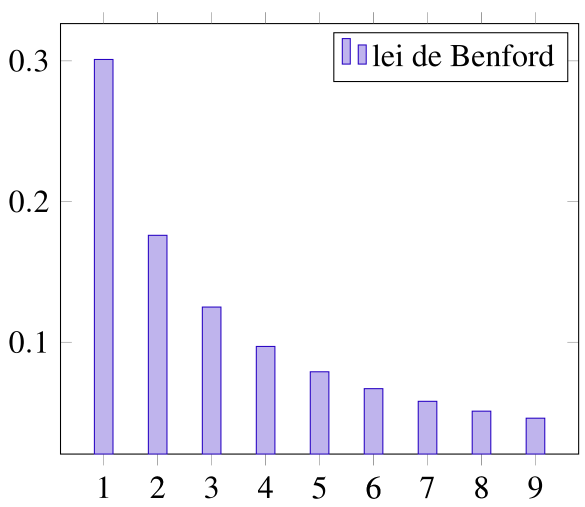


Figura 1 – Frequências esperadas para um dígito d, de acordo com a Lei de Benford.

No Capítulo 3 daremos exemplos da aplicação da Lei de Benford em alguns contextos.

2.4 A Lei de Benford para o segundo dígito

A Lei de Benford pode ser estendida para o segundo dígito. As aplicações da Lei de Benford para o segundo dígito são um complemento à lei do primeiro dígito, mais uma ferramenta de detecção de fraudes. Na auditoria financeira, pode ser utilizado para detectar irregularidades nas demonstrações financeiras, comparando as frequências observadas do segundo dígito com as frequências esperadas previstas na lei. Se for observado um desvio significativo, uma investigação mais aprofundada pode ser necessária.

Além disso, a Lei de Benford para o segundo dígito também encontrou aplicações na análise baseada em dígitos, especificamente na compreensão do comportamento de dados gerados por humanos versus dados gerados artificialmente. Observou-se que a lei é válida para conjuntos de dados que ocorrem naturalmente, enquanto desvios da lei indicam dados artificiais ou manipulados.

Por exemplo, na análise de redes sociais, a lei pode ser usada para identificar seguidores gerados artificialmente ou métricas de engajamento. Se a distribuição de seguidores ou curtidas de segundo dígito se desviar significativamente das frequências esperadas, isso sugere a presença de bots ou contas falsas.

Para uma dedução da Lei de Benford para o segundo dígito, vamos emprestar a explicação dada por Benford (BENFORD, 1938, p. 555):

> *Frequência dos dígitos na q-ésima posição*
>
> Os dígitos da segunda posição são dez, pois aqui devemos levar 0 em consideração. Além disso, ao considerar a frequência F_b de um dígito b da segunda posição, devemos levar em consideração o dígito a que o precedeu. O intervalo logarítmico entre dois dígitos deve agora ser dividido em dez partes correspondentes aos dez dígitos 0, 1, 2, ... 9. Seja a o primeiro dígito de um número e b o segundo dígito; então, usando o significado habitual de posição e ordem em nosso sistema decimal, um número de dois dígitos é escrito[1] ab, e o próximo número maior é escrito $ab+1$.
>
> O intervalo logarítmico entre ab e $ab+1$ é $\log(ab+1)-\log ab$, enquanto o intervalo coberto pelos dez dígitos possíveis do segundo lugar é $\log(a+1)-\log a$. Portanto, a frequência F_b de um dígito da segunda posição b após um dígito da primeira posição a é
>
> $$F_b = \log\frac{ab+1}{ab}\bigg/\log\frac{a+1}{a} \tag{2.2}$$

Benford explicou que a distribuição do segundo dígito é influenciada pelo primeiro dos números no conjunto de dados. Por exemplo, se o primeiro dígito for 1, é mais provável que o segundo dígito seja menor em comparação com quando o primeiro dígito é 9.

[1] Ou seja, Benford denota por ab o número $10a+b$

Por exemplo, de acordo com (2.2), a probabilidade de um 0 após um 1 e um 2 em primeiro lugar em um número aleatório são, respectivamente,

$$\frac{\log\left(1+\frac{1}{10}\right)}{\log\left(1+\frac{1}{1}\right)} \approx 0{,}1375, \qquad \frac{\log\left(1+\frac{1}{20}\right)}{\log\left(1+\frac{1}{2}\right)} \approx 0{,}1203.$$

Podemos escrever as ideias de Benford em linguagem estatística moderna, denotando o evento "primeiro dígito" por A e o "segundo dígito" por B, de forma que a Lei de Benford para o primeiro dígito, em (2.1), fica

$$P(A=a) = \log\frac{a+1}{a} = \log\left(1+\frac{1}{a}\right),$$

e a fórmula em (2.2) expressa a probabilidade condicional do segundo dígito ser b dado que o primeiro dígito é a, ou seja,

$$P(B=b|A=a) = \log\left(1+\frac{1}{10a+b}\right) \Big/ \log\left(1+\frac{1}{a}\right).$$

Para calcular a probabilidade de um segundo dígito b, independente do primeiro dígito a, que assume valores de 1 a 9, podemos utilizar a fórmula de Bayes para partições (BUSSAB; MORETTIN, 2010, sec. 5.4):

$$P(B=b) = \sum_{a=1}^{9} P(B=b|A=a)\cdot P(A=a) = \sum_{a=1}^{9} \log\left(1+\frac{1}{10a+b}\right)$$

Com isso, podemos montar a tabela abaixo, já dada em (BENFORD, 1938, p.556):

dígito	0	1	2	3	4	5	6	7	8	9
prob. 1º dígito		0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046
prob. 2º dígito	0,120	0,114	0,109	0,104	0,100	0,097	0,093	0,090	0,088	0,085

Podemos ver que as probabilidades do segundo dígito já são um tanto próximas da distribuição uniforme, onde cada dígito teria probabilidade de $10\% = 0{,}1$ de aparecer. Benford observou que a mesma técnica pode ser usada para deduzir as probabilidades para a q-ésima posição, mas que estas tendem à uniforme para posições além do segundo dígito.

2.5 Detecção de fraudes

A Lei de Benford tem sido utilizada para detectar irregularidades em conjuntos de dados financeiros e contábeis. Abaixo estão alguns exemplos de casos em que a Lei de Benford foi aplicada para detectar fraudes:

Caso da Enron (2001): Um dos casos mais famosos de fraude contábil. A análise dos relatórios financeiros da Enron, usando a Lei de Benford, revelou anomalias nos padrões de distribuição dos primeiros dígitos, o que levou a uma investigação mais aprofundada e à descoberta de uma fraude financeira significativa (TILDEN; JANES, 2012).

Escândalo da WorldCom (2002): A WorldCom, uma gigante das telecomunicações, foi envolvida em um escândalo de contabilidade. A Lei de Benford foi aplicada aos relatórios financeiros, e as discrepâncias nas distribuições dos primeiros dígitos ajudaram a identificar fraudes nas demonstrações financeiras (SILVA, 2013).

Fraude em Municípios no Brasil (2012): No Brasil, a Lei de Benford foi utilizada para analisar dados financeiros de municípios paranaenses e identificar possíveis fraudes. Casos foram relatados em que a distribuição dos primeiros dígitos não estava de acordo com o esperado, levando a investigações e descobertas de irregularidades (RODRIGUES; MIRANDA; MUSIAL, 2023).

Caso da Olympus (2011): A empresa japonesa Olympus enfrentou um escândalo de fraude contábil. A análise dos dados financeiros usando a Lei de Benford foi uma das ferramentas utilizadas para identificar padrões inconsistentes nos registros contábeis (GEPP; KUMAR; BHATTACHARYA, 2023).

Fraude em Eleições (Vários Casos): Em algumas situações, a Lei de Benford foi aplicada para analisar resultados de eleições. Variações significativas nos padrões de distribuição dos primeiros dígitos podem levantar suspeitas de manipulação nos resultados eleitorais (BREUNIG; GOERRES, 2011).

É importante notar que a Lei de Benford não é uma ferramenta definitiva para detectar fraudes, mas pode ser uma indicação inicial de irregularidades que justifiquem uma investigação mais detalhada. Outros fatores e métodos também são geralmente utilizados em conjunto para uma avaliação mais abrangente.

CAPÍTULO

3

APLICANDO A LEI DE BENFORD A CONJUNTOS DE DADOS

3.1 Introdução

Vamos explicar, através de exemplos, como aplicar a lei de Benford a um conjunto de dados que você deseja analisar. Pode ser um conjunto de números em uma planilha, registros contábeis, resultados eleitorais, dados financeiros, entre outros.

Todos os passos abaixo serão feitos explicitamente numa planilha, para possibilitar uma compreensão clara. Isso é especialmete necessário se for feito numa sala de aula, explorando a lei de Benford como uma aplicação interdisciplicar da matemática em que são trabalhados logaritmos, estatística, uso de planilhas (computação) e busca de dados estatísticos na geografia, biologia ou outras áreas.

1. **Colecione os números numa planilha:** É ideal que os dados coletados sejam postos em uma coluna de uma planilha.

2. **Extraia os primeiros dígitos:** Na próxima coluna, a planilha calculará o primeiro dígito, ou por uma fórmula matemática que calcule o primeiro dígito, como mostraremos na Seção 3.2, o que é ideal no ensino, ou por uma fórmula pronta de planilha, como `=ESQUERDA(`*pos*`)`.

3. **Conte a frequência de cada dígito:** Em outra região da planilha, usando fórmulas de planilha, como `=CONT.SE(`*d*`)`, conte quantas vezes cada dígito aparece como o primeiro dígito no conjunto de dados.

4. **Calcule as frequências esperadas:** De acordo com a Lei de Benford, a frequência esperada de cada dígito como o primeiro dígito segue a fórmula (2.1), p. 22. A frequência esperada

para o dígito 1 é de aproximadamente 30,1%, para o dígito 2 é de 17,6%, para o dígito 3 é de 12,5% e assim por diante.

5. Compare as frequências observadas e esperadas: Plote um gráfico comparando as frequências observadas (passo 3) com as frequências esperadas (passo 4) para cada dígito. Se as frequências observadas estiverem próximas das frequências esperadas, isso indica que o conjunto de dados segue a distribuição de Benford. Caso contrário, se houver grandes discrepâncias, isso pode indicar a presença de fraudes ou anomalias nos dados.

É importante ressaltar que a aplicação da Lei de Benford requer uma análise cuidadosa e consideração do contexto específico do conjunto de dados em questão. Além disso, é recomendado realizar análises estatísticas mais avançadas e investigações adicionais para confirmar qualquer suspeita de fraude ou anomalia nos dados.

3.2 Obtendo o primeiro dígito de um número

Usando logaritmos, podemos extrair a ordem de grandeza de um número. Essa é uma ótima aplicação dos logaritmos e da notação científica, que pode despertar os alunos para a utilidade dessas ferramentas matemáticas.

Vamos denotar a parte inteira de um número x por $\lfloor x \rfloor = k$ onde k é o inteiro em que $x \in [k, k+1)$. Vamos utilizar logaritmos na base 10.

A *notação científica* é o modo como ficou conhecida a técnica de escrever números reais[1] positivos muito pequenos ou muito grandes por meio do uso de uma potência de base dez, na forma

$$x = m \times 10^n, \quad 1 \leqslant m < 10, \quad n \in \mathbb{Z}. \tag{3.1}$$

Nessa disposição, m é chamado de *mantissa* e o inteiro n é chamado de *ordem de grandeza* de x.

Por exemplo, o raio da Terra é aproximadamente $R = 6\,378{,}1$ quilômetros, que, em notação científica, fica $R = 6{,}3781 \times 10^3$.

Podemos ver que o primeiro dígito de x, para a aplicação da lei de Benford, é o primeiro da mantissa m, que é sua parte inteira, isto é, $d = \lfloor m \rfloor$.

De (3.1), como $1 \leqslant m < 10$, temos que $1 \cdot 10^n \leqslant m \cdot 10^n < 10 \cdot 10^n$ e então

$$10^n \leqslant x < 10^{n+1},$$

e como o logaritmo é uma função crescente, preservando desigualdades, e $\log 10^k = k$,

$$\log 10^n \leqslant \log x < \log 10^{n+1} \Rightarrow n \leqslant \log x < n+1.$$

[1] Veja verbete *Notação científica*, portal Mundo Educação. <https://mundoeducacao.uol.com.br/matematica/notacao-cientifica.htm>, visitado em 12/02/2024.

Então n é a parte inteira de $\log x$, isto é,

$$n = \lfloor \log x \rfloor . \tag{3.2}$$

Por exemplo, voltando ao raio da Terra em quilômetros, $\log R \approx 3{,}80469$, e sua ordem de grandeza pode ser calculada por

$$n = \lfloor \log R \rfloor = \lfloor 3{,}80469 \rfloor = 3.$$

Uma vez calculada a ordem de grandeza n, podemos isolar m em (3.1):

$$m = \frac{x}{10^n} = x \times 10^{-n} \tag{3.3}$$

Agora podemos calcular o primeiro dígito:

$$d = \lfloor m \rfloor = \left\lfloor x \cdot 10^{-\lfloor \log x \rfloor} \right\rfloor , \tag{3.4}$$

que apesar de correta, por ser a composição de muitas funções, pode causar uma estranheza nos alunos, e pode ser melhor calcular d, numa planilha, parte por parte, isto é, numa coluna a ordem de grandeza, na próxima a mantissa e, na próxima, o primeiro dígito de x.

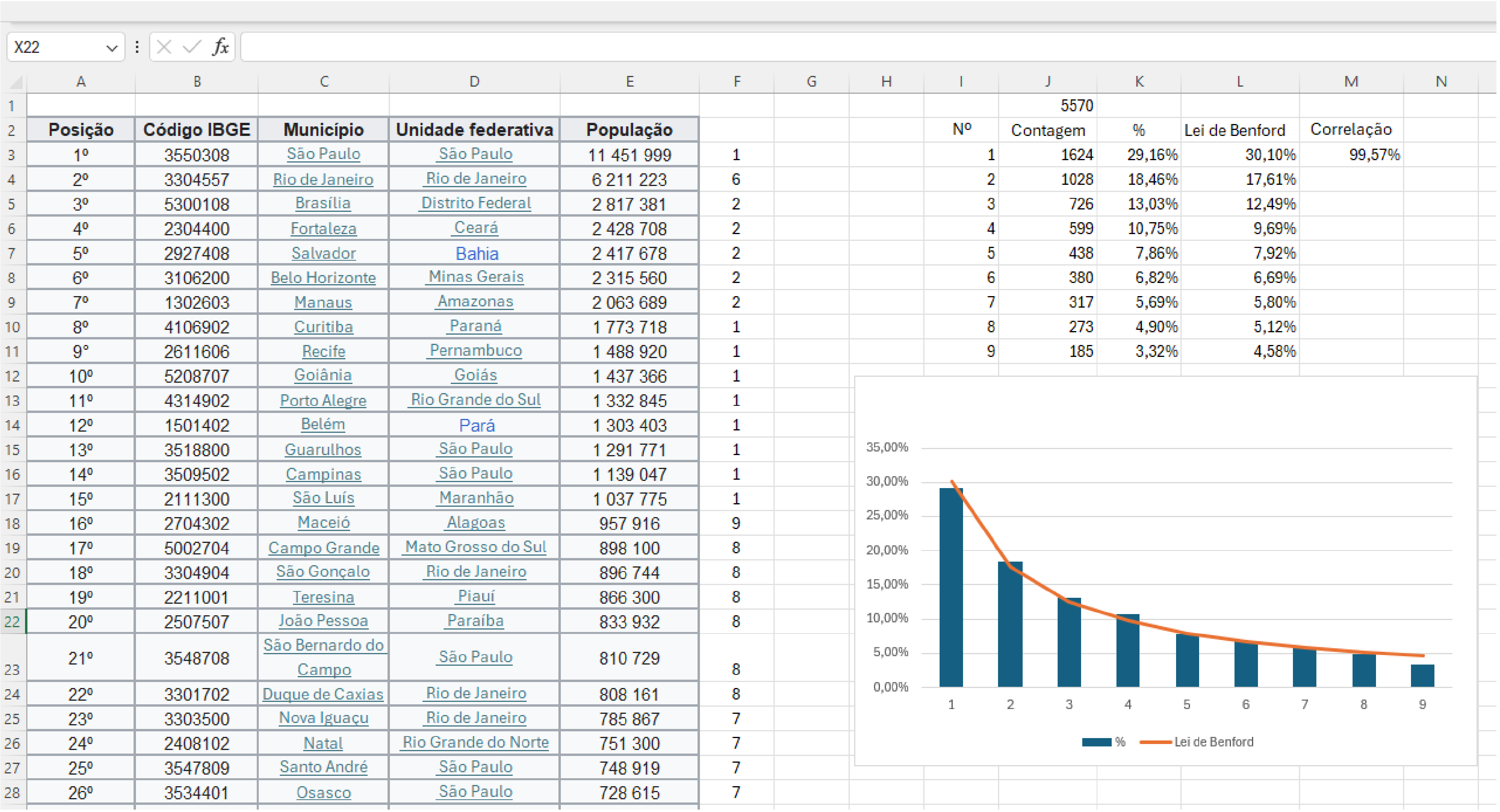
3.3 Exemplo 1 – população dos municípios no Brasil

Coletaremos os dados referente a população dos muncípios no Brasil, retirada do verbete “Lista de municípios do Brasil por população (2022)” da Wikipedia[2], que mostra uma compilação dos dados do Censo 2022 realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Esses dados serão organizados conforme mostra a Figura 2.

Na coluna A, teremos a ordem dos municípios de acordo com a sua população, do maior para o menor; na coluna B, o código de IBGE; nas colunas C e D, o nome dos Estados e Municípios e na coluna E a população de cada um.

De todos os números que estão na coluna E, temos que tirar o primeiro dígito. Na célula E3, temos o número 1 como primeiro dígito, na E4 o 6 como primeiro dígito, na E5 o 2 como primeiro dígito e assim conforme o número de cada população do município. Na coluna F colocaremos os primeiros dígitos de todos os valores da coluna E. Poderíamos usar a fórmula (3.4), que na planilha pode ser escrita como `=TRUNCAR(E3*10^(-TRUNCAR(LOG10(E3))))`, onde `E3` é a célula e 1 é o primeiro número da posição. Ou melhor, fazer passo a passo, colocando numa coluna a ordem com a fórmula (3.2), que na planilha fica `=TRUNCAR(LOG10(E3))` por exemplo na posição F3, então calcular a mantissa, de acordo com (3.3) na coluna G com a fórmula `=E3*10^(-F3)` e então o primeiro dígito, de acordo com (3.4) na coluna H com `=TRUNCAR(G3)`.

[2] `https://pt.wikipedia.org/wiki/Lista_de_municípios_do_Brasil_por_população_-(2022)`



X22

Posição	Código IBGE	Município	Unidade federativa	População	
1º	3550308	São Paulo	São Paulo	11 451 999	1
2º	3304557	Rio de Janeiro	Rio de Janeiro	6 211 223	6
3º	5300108	Brasília	Distrito Federal	2 817 381	2
4º	2304400	Fortaleza	Ceará	2 428 708	2
5º	2927408	Salvador	Bahia	2 417 678	2
6º	3106200	Belo Horizonte	Minas Gerais	2 315 560	2
7º	1302603	Manaus	Amazonas	2 063 689	2
8º	4106902	Curitiba	Paraná	1 773 718	1
9°	2611606	Recife	Pernambuco	1 488 920	1
10º	5208707	Goiânia	Goiás	1 437 366	1
11º	4314902	Porto Alegre	Rio Grande do Sul	1 332 845	1
12º	1501402	Belém	Pará	1 303 403	1
13º	3518800	Guarulhos	São Paulo	1 291 771	1
14º	3509502	Campinas	São Paulo	1 139 047	1
15º	2111300	São Luís	Maranhão	1 037 775	1
16º	2704302	Maceió	Alagoas	957 916	9
17º	5002704	Campo Grande	Mato Grosso do Sul	898 100	8
18º	3304904	São Gonçalo	Rio de Janeiro	896 744	8
19º	2211001	Teresina	Piauí	866 300	8
20º	2507507	João Pessoa	Paraíba	833 932	8
21º	3548708	São Bernardo do Campo	São Paulo	810 729	8
22º	3301702	Duque de Caxias	Rio de Janeiro	808 161	8
23º	3303500	Nova Iguaçu	Rio de Janeiro	785 867	7
24º	2408102	Natal	Rio Grande do Norte	751 300	7
25º	3547809	Santo André	São Paulo	748 919	7
26º	3534401	Osasco	São Paulo	728 615	7

Nº	Contagem	%	Lei de Benford	Correlação
	5570			
1	1624	29,16%	30,10%	99,57%
2	1028	18,46%	17,61%	
3	726	13,03%	12,49%	
4	599	10,75%	9,69%	
5	438	7,86%	7,92%	
6	380	6,82%	6,69%	
7	317	5,69%	5,80%	
8	273	4,90%	5,12%	
9	185	3,32%	4,58%	

Figura 2 – Captura de tela da planilha com os dados da população dos municípios do Brasil.

Aconteceu dos dados não serem numéricos (E3 contém a string `11 451 999`). Poderíamos fazer a conversão, mas também podemos usar uma função pronta de planilha para o primeiro caractere de uma string, que funciona também com números. Então usaremos a função `ESQUERDA(E3;1)`.

Na coluna I, escreveremos os algarismos de 1 a 9 e, na coluna J vamos realizar a contagem de quantas vezes cada algarismo aparece.Para que possamos contar o número de vezes que cada algarismo se repete vamos utilizar a função `CONT.SE`, para contar o número de células que atendem a um critério. Na forma mais simples, a função `CONT.SE(`*lugar*,*critério*`)` informa o número de ocorrências do valor *critério* nas células de *lugar*. Por exemplo, na célula I3 temos o dígito 1, e para contar esse dígito na coluna F, que é dada pelo intervalo `F:F`, podemos usar a fórmula de planilha `=CONT.SE(F:F;I3)` na célula J3. A partir disso, podemos colar essa fórmula nas células de J4 a J11 para fazer toda a contagem dos primeiros dígitos.

Após feito a contagem de cada algarismo, vamos para coluna K, onde cada célula dará a porcentagem, para o cálculo da porcentagem precisaremos do número total de de munícipios, para que possamos saber o número total de munícipios vamos utilizar a fórmula `=SOMA(J3:J11)`, representada na célula J1, dando um total de 5570 munícipios. Para o cálculo da porcentagem, iremos trabalhar na célula K3, onde o cáculo é feito da seguinda maneira, pegaremos o valor da célula J3 e dividiremos pelo total de munícipios que se encontra na célula J1, resultando o valor da porcentagem `=J3/$J1`, onde o valor é apresentado na célula K3, depois arrastamos da célula K3 até a K11 , o resultado é obtido em porcentagem, conforme podemos ver na planilha acima, o dígito 1 com uma porcentagem de 29,6%, o dígito 2 18,46%, dígito 3 13,03%, dígito 4 10,75%, dígito 5 7,86%, dígito 6 6,82%, dígito 7 5,69%,dígito 8 4,90% e dígito 9 3,32%. Note que a cada dígito a porcentam vai abaixando sucessivamente, isso mostra que para esse caso realmente funciona a lei de Benford.

Vamos calcular na coluna L a porcentagem utilizando a fórmula de Benford, representada por $P(d) = \log(1 + 1/d)$, onde d é o dígito a ser calculado, no caso L3 deve ter a fórmula `=LOG10(1+I3)`, onde o dígito 1 estána célula I3, e arrastamos a fórmuma da célula L3 até L11, para que todas possam calcular pela fórmula de Benford.

Nota-se que todos os valores bate extamente igual, ou seja, os valores da célula k3 é praticamente idêntico ao valor que apresenta na célula L3, se calcularmos a correlação, que é dada pela fórmula `=CORREL(K3:K11;L3:L11)`, e o valor está representado na célula M3, vamos notar que a porcentagem chega bem próximo a 100%.

Podemos representar essa porcentagem em um gráfico também, conforme está na planilha. Ele mostra que os valores que estão em barra são calculados sem a fórmula, e o que está em linha pela lei de Benford, nota se que pelo gráfico não há discrepância.

Para o cáculo do 2º dígito, utilizaremos os mesmo dados, número de Munícipios, só que agora calcularemos o 2º dígito também, considerando a coluna F para o 1º dígito e a coluna G para o 2º dígito, e utilizando a fórmula `=ESQUERDA(E3;1)`, para célula F3, obteremos o 1º dígito, basta arrastar para as próximas células da coluna F para que possa preencher com a fórmula.

Para a retirada do 2º dígito utlilizaremos a coluna G, conforme a Figura 3:

Aplicando a fórmula `=EXT.TEXTO(E3; 2;1)` na célula G1, onde `EXT.TEXTO` retorna um número específico de caracteres de uma cadeia de texto, começando na posição especificada, com base no número de caracteres especificado, ou seja, onde irá buscar na célula E3, o segundo dígito, que começa com 1. Para aplicar nas demais células basta arrastar até o final da coluna G.

Na coluna I e J , colocaremos a coluna para primeiro dígito e segundo dígito, sendo que na coluna I preencheremos da célula I3 ate I12 com o algarismo 1, I13 a I22 com o algarismo 2, assim sucessivamente e na na coluna J preencheremos J3 a J12 com os algarismo de 0 a 9, J13 a J22 com os algarismo de 0 a 9, assim sucessivamente, já que para o segundo dígito o algarismo zero pode sentar presente.

Para podermos fazer a contagem de dígitos, tanto do primeiro como o segundo, colocaremos na coluna L, aplicando a fórmula `CONT.SES`, na qual conta o número de células em um intervalo que atendem a determinados critérios, como temos que pegar dados de duas colunas, terá que especificar um intervalo, assim para o cálculo do 2º dígito, teremos que conta o número de células em um intervalo que atendem a determinados critérios, para o cálculo então do 2º dígito aplicaremos da seguinte maneira, `=CONT.SES(F:F;I3;G:G;J3)`, F será a coluna onde buscara o dígito, I3 será o dígito a ser contado, logo temo o intervalo devido ao 2º dígito, onde G é onde localiza o 2º dígito e J3 será o dígito a ser contado.

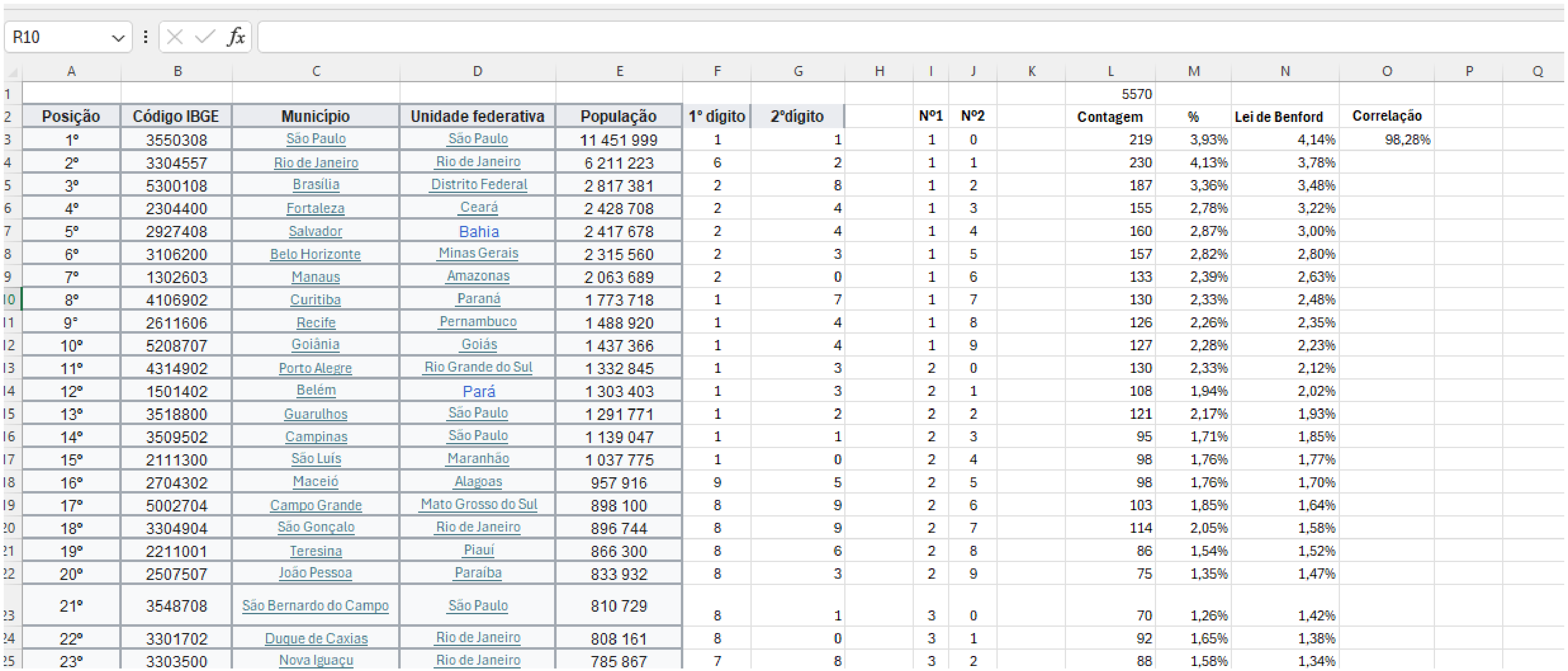
Para o cálculo da porcentagem, teremos a coluna M e N, sendo que M iremos calcular com os dados da coluna L dividido pelo número total de munícipios, um exemplo é se pegarmos a célula M3, o cálculo ficará L3/L1, dando uma porcentagem de 3,93%.

Para o cálculo da porcentagem na coluna N, aplicaremos a lei de Benford para o segundo dígito, `=LOG(1+(1/(I3&J3)))`, onde I3 são os dados do 1º dígito e J3 dados do 2º dígito, resultado um valor de 4,14%.

Para mostrarmos que a uma relação, iremos fazer a correlação da porcentagem calculada na coluna M pela coluna N, e nota se que o valor da correlação chega a torno de 98,28%, bem próximo de 100%, isso mostra que a relação não tem discrepância.

3.4 Exemplo 2 – casos de COVID no Brasil

A aplicação da Lei de Benford na análise da distribuição dos dígitos dos dados relacionados à pandemia de COVID-19 no Brasil oferece uma oportunidade fascinante para demonstrar a eficácia desta teoria estatística em um contexto real e dinâmico.



Posição	Código IBGE	Município	Unidade federativa	População	1º dígito	2ºdígito	Nº1	Nº2	Contagem	%	Lei de Benford	Correlação
									5570			
1º	3550308	São Paulo	São Paulo	11 451 999	1	1	1	0	219	3,93%	4,14%	98,28%
2º	3304557	Rio de Janeiro	Rio de Janeiro	6 211 223	6	2	1	1	230	4,13%	3,78%	
3º	5300108	Brasília	Distrito Federal	2 817 381	2	8	1	2	187	3,36%	3,48%	
4º	2304400	Fortaleza	Ceará	2 428 708	2	4	1	3	155	2,78%	3,22%	
5º	2927408	Salvador	Bahia	2 417 678	2	4	1	4	160	2,87%	3,00%	
6º	3106200	Belo Horizonte	Minas Gerais	2 315 560	2	3	1	5	157	2,82%	2,80%	
7º	1302603	Manaus	Amazonas	2 063 689	2	0	1	6	133	2,39%	2,63%	
8º	4106902	Curitiba	Paraná	1 773 718	1	7	1	7	130	2,33%	2,48%	
9º	2611606	Recife	Pernambuco	1 488 920	1	4	1	8	126	2,26%	2,35%	
10º	5208707	Goiânia	Goiás	1 437 366	1	4	1	9	127	2,28%	2,23%	
11º	4314902	Porto Alegre	Rio Grande do Sul	1 332 845	1	3	2	0	130	2,33%	2,12%	
12º	1501402	Belém	Pará	1 303 403	1	3	2	1	108	1,94%	2,02%	
13º	3518800	Guarulhos	São Paulo	1 291 771	1	2	2	2	121	2,17%	1,93%	
14º	3509502	Campinas	São Paulo	1 139 047	1	1	2	3	95	1,71%	1,85%	
15º	2111300	São Luís	Maranhão	1 037 775	1	0	2	4	98	1,76%	1,77%	
16º	2704302	Maceió	Alagoas	957 916	9	5	2	5	98	1,76%	1,70%	
17º	5002704	Campo Grande	Mato Grosso do Sul	898 100	8	9	2	6	103	1,85%	1,64%	
18º	3304904	São Gonçalo	Rio de Janeiro	896 744	8	9	2	7	114	2,05%	1,58%	
19º	2211001	Teresina	Piauí	866 300	8	6	2	8	86	1,54%	1,52%	
20º	2507507	João Pessoa	Paraíba	833 932	8	3	2	9	75	1,35%	1,47%	
21º	3548708	São Bernardo do Campo	São Paulo	810 729	8	1	3	0	70	1,26%	1,42%	
22º	3301702	Duque de Caxias	Rio de Janeiro	808 161	8	0	3	1	92	1,65%	1,38%	
23º	3303500	Nova Iguaçu	Rio de Janeiro	785 867	7	8	3	2	88	1,58%	1,34%	

Figura 3 – Captura de tela da planilha com os dados do 2º dígito para a população dos municípios do Brasil.

Contextualização

A COVID-19, ao se propagar globalmente, gerou uma enorme quantidade de dados, desde casos confirmados e óbitos, até testes realizados e outras métricas relacionadas à saúde pública. Este estudo focaliza a distribuição dos dígitos nos registros diários de novos casos de COVID-19 no Brasil desde o início até o dia 19 de janeiro de 2024, quando coletamos esses dados do Portal DataSUS[3] do Ministério da Saúde do Brasil.

Aplicando a análise dos dados numa planilha

Coletando os dados de Covid-19 em uma planilha do execel, iremos organizar os dados de forma que objetivo será a retirada do 1º dígito do números apresentados, e aplicando a Lei de Benford para verificar se a fraude.

Na coluna B estarão todos os números das pessoas contanimadas, onde cada célula de B representa uma cidade, na coluna A vamos retirar o primeiro dígito de cada número da coluna B.

Para retirada do primeiro dígito, de acordo com (3.4), vamos utilizar a fórmula `=TRUNCAR(B5*10^(-TRUNCAR(LOG10(B5))))`, onde truncar é retirar o primeiro dígito, iremos aplicar essa fórmula na célula A5, e para as demais iremos somente arrastar a fórmula para as outras células, logo vemos que para a célula B5 que representa o número 148, seu 1º dígito será 1.

Retirado todos os primeiro dígitos, iremos calcular o total de algarismo 1,2,3,4,5,6,7,e,9 que apresenta como primeiro dígito que está representado nas colunas D4 a L4.

Com a retirada dos primeiros dígitos, iremos fazer a soma de quantas vezes os algarismos de 1 a 9 aparecem na primeira posição. Na célula D1 é calculado a quantidade de vezes que aparece o algarismo 1, sendo um total de 199, na célula D2 representada pela quantidade de algarismo 2, D3 a quantidade de algarismos 3 e assim sucessivamente.

Na célula D2, aplicaremos a fórmula para o cálculo da porcentagem, `=D1/645` e, arrastando de D2 até a L2, sendo o número total de contaminados igual a 645.

Com o cálculo de porcentagem de cada algarismo, notamos que para o 1 dígito uma porcentagem de 30,9%, para 2 dígito 17,5, para o 3 dígito 11,3%, para o 4 dígito 9,6%, para 5 dígito 8,5%, para o 6 dígito 5,1%, para o 7 dígito 6,8%, para o 8 dígito 5,5% e para o 9 dígito 5,0.

Isso mostra que os dados coletados se aplica a lei de Benford, ou seja, não havendo fraudes.

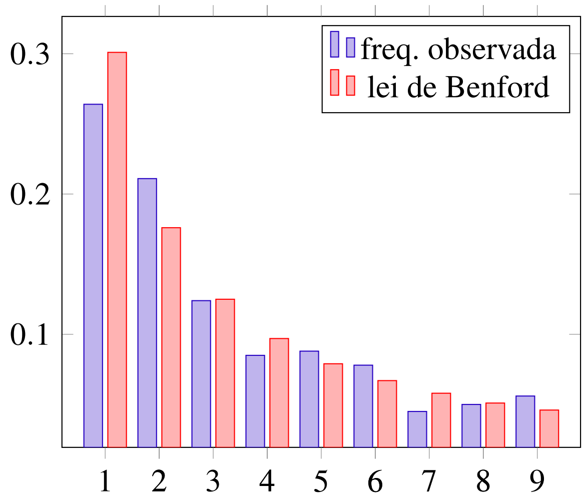
[3] link dos dados

Aplicação da Lei de Benford

Extraímos os dados de 645 municípios do Estado de São Paulo. Abaixo está a compilação dos resultados dos municípios coletados, com o número total de cada dígito apresentado, frequência observada e a frequência utilizando a lei de Benford.

dígito	1	2	3	4	5	6	7	8	9
contagem	170	136	80	55	57	50	29	32	36
freq. observada	0,264	0,211	0,124	0,085	0,088	0,078	0,045	0,050	0,056
freq. lei de Benford	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046

Essas frequências podem ser visualizadas no gráfico abaixo



Ao aplicarmos a Lei de Benford a esses dados, observamos a distribuição dos primeiros dígitos significativos próxima da esperada pela Lei de Benford.

Resultados e Observações

Essa análise não apenas valida a utilidade da Lei de Benford na detecção de padrões, mas também pode apontar para áreas de investigação mais detalhada. Variações significativas em relação à distribuição esperada podem sugerir irregularidades nos dados ou, possivelmente, características únicas na coleta de informações relacionadas à COVID-19.

Entendemos que a aplicação da Lei de Benford em dados de saúde pública tem suas limitações, como possíveis flutuações sazonais, mudanças na política de testagem, e outros fatores externos que podem influenciar a distribuição dos dígitos.

CAPÍTULO

4

UM PLANO DE AULA

Neste contexto, explorar a Lei de Benford em sala de aula não apenas enriquece a compreensão dos alunos sobre padrões matemáticos, mas também os prepara para interpretar e analisar dados de forma crítica e informada. Ao investigar a Lei de Benford, os estudantes têm a oportunidade de mergulhar em um universo matemático intrigante, repleto de aplicações práticas e desafios estimulantes.

Pensando nisso vamos explorar os conceitos da Lei de Benford e suas aplicações, alinhando-se com os princípios da Base Nacional Comum Curricular (BNCC) que incentivam a abordagem da interdisciplinaridade em sala de aula e a conexão entre teoria e prática.

4.1 Plano de Aula

Disciplinas: Matemática/Tecnologia e Informação

Unidade Temática: Estatística

Turma: 1º Ano Ensino Médio

Objetivo Geral: Introduzir os alunos à Lei de Benford e explorar sua aplicação em diferentes contextos, promovendo a compreensão dos padrões matemáticos subjacentes.

Objetivos Específicos: Compreender o conceito e os princípios da Lei de Benford, analisar como a Lei de Benford é aplicada em áreas como contabilidade, auditoria e detecção de fraudes, e realizar atividades práticas para verificar a conformidade de conjuntos de dados com a Lei de Benford.

Abordagem Pedagógica: A Lei de Benford, sendo uma observação estatística sobre a distribuição de dígitos em conjuntos de dados, pode ser abordada pedagogicamente em várias disciplinas, dependendo do contexto e dos objetivos educacionais.

Recursos: Quadro ou projetor para apresentação, conjuntos de dados para análise, calculadora ou planilha, e o acesso à internet para pesquisa adicional.

Quantidade de aulas previstas: 4 aulas

Metodologia: Apresentação do conceito da Lei de Benford e sua relevância em diferentes áreas, discussão sobre a distribuição dos dígitos em conjuntos de dados reais e a expectativa de conformidade com a Lei de Benford.

Desenvolvimento da Aula: Análise de exemplos práticos de conjuntos de dados e verificação da conformidade com a Lei de Benford, exploração de como a Lei de Benford é utilizada na prática, destacando sua importância na detecção de anomalias e fraudes.

Divisão dos alunos em grupos para analisar conjuntos de dados fornecidos.

Cada grupo deve verificar se os dados seguem a distribuição esperada pela Lei de Benford e apresentar os resultados.

Compartilhamento dos resultados das análises realizadas pelos grupos.

Reflexão sobre a importância da Lei de Benford e suas aplicações práticas.

Avaliação: Participação dos alunos durante as discussões e atividades e compreensão demonstrada na análise dos conjuntos de dados em relação à Lei de Benford.

REFERÊNCIAS

BENFORD, F. The law of anomalous numbers. **Proceedings of the American Philosophical Society**, v. 78, n. 4, p. 551–572, Mar. 1938. Disponível em: <https://www.jstor.org/stable/984802>.

BRASIL. Ministério da Educação. **Base Nacional Comum Curricular**. Brasília, 2018.

BREUNIG, C.; GOERRES, A. Searching for electoral irregularities in an established democracy: Applying Benford's Law tests to Bundestag elections in Unified Germany. **Electoral Studies**, v. 30, n. 3, p. 534–545, 2011.

BUSSAB, W.; MORETTIN, P. **Estatística básica**. 6. ed. São Paulo: Editora Saraiva, 2010.

GEPP, A.; KUMAR, K.; BHATTACHARYA, S. Taking the hunch out of the crunch: A framework to improve variable selection in models to detect financial statement fraud. **Accounting & Finance**, Oct. 2023.

HILL, T. P. 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**, v. 86, n. 4, p. 358–363, Jul.-Ago. 1998.

RODRIGUES, L.; MIRANDA, C.; MUSIAL, N. A Lei de Newcomb-Benford como ferramenta de auditoria: Uma análise das despesas orçamentárias nos municípios paranaenses. **Revista do TCU**, v. 1, n. 152, 2023.

SILVA, A. S. **The application of Benford's Law in detecting accounting fraud in the Financial Sector**. Dissertação (Mestrado em Economia) — Lisboa School of Economics & Management, Lisboa, 2013.

TILDEN, C.; JANES, T. Empirical evidence of financial statement manipulation during economic recessions. **Journal of Finance and Accountancy**, v. 10, Jun. 2012.