UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXVIII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2024

APLICAÇÕES DE LEI DE NEWCOMB-BENFORD A CONJUNTOS DE DADOS REAIS

Bruno Rios Souza[1]; Ana Carla Percontini da Paixão[2]
1. Bolsista PIBIC/CNPq, Graduando em Licenciatura em Matemática, Universidade Estadual de Feira de Santana, e-mail: bruno.riosbrs@gmail.com
2. Orientador, Departamento de Ciências Exatas, Universidade Estadual de Feira de Santana, e-mail: anacarla@uefs.br

PALAVRAS-CHAVE: lei de newcomb-benford; análise de dados; sequências matemáticas.

INTRODUÇÃO

Identificada por Simon Newcomb em 1881 e formalizada por Frank Benford em 1938. A Lei de Newcomb-Benford, também conhecida como Lei dos Números Anômalos, descreve uma distribuição logarítmica para a frequência do primeiro dígito significativo em dados numéricos. De acordo com a lei, a probabilidade de um dígito d aparecer como primeiro algarismo é dada por $log_{10}(1 + \frac{1}{d})$.

A descoberta inicial de Newcomb não recebeu ampla atenção devido à falta de fundamentação matemática, mas a lei ganhou reconhecimento quando Benford demonstrou sua validade empírica com um estudo abrangente. Mas somente em 1996, Theodore Hill forneceu uma prova formal da lei, mostrando que ela é invariável em qualquer base numérica, o que consolidou a Lei de Newcomb-Benford como a única distribuição de probabilidade invariante de base. A lei tem aplicações em áreas como contabilidade e auditoria, ciências naturais e sociais, teoria dos números e estudos de caso. Exemplos incluem a análise de fraudes eleitorais e financeiras e a verificação da autenticidade de dados em diversos contextos.

Para analisar a aderência dos dados à Lei de Newcomb-Benford, utilizamos métodos estatísticos como o teste qui-quadrado e o Kolmogorov-Smirnov. O teste qui-quadrado compara a distribuição observada com a esperada e o Kolmogorov-Smirnov avalia a discrepância entre a função distribuição empírica e a teórica. Cada um desses métodos é aplicado para validar a aderência dos dados à lei e determinar a presença de discrepâncias significativas.

A lei é particularmente útil na detecção de fraudes, como em manipulações orçamentárias ou eleitorais, pois ajuda a identificar discrepâncias nos dados que podem indicar adulteração. Embora a Lei de Newcomb-Benford não forneça uma confirmação

definitiva de fraude, ela fornece indícios para dados com maior probabilidade de manipulação, dando uma orientação para investigação. Este estudo visa avaliar a validade da Lei de Newcomb-Benford e explorar métodos estatísticos para analisar a aderência de dados a essa lei, identificando os métodos mais eficazes e os tipos de dados mais adequados para sua aplicação.

METODOLOGIA

Esta análise foi dividida em três partes: análise de sequências matemáticas, análise de conjuntos de dados reais e teste de detecção de fraudes em dados alterados. A seguir, detalharemos todo o procedimento.

Para a análise das sequências matemáticas, foram considerados dez conjuntos com os primeiros três mil números de cada um. As sequências selecionadas foram: Números de Mersenne, Potências de 3 a 9, Sequência de Lucas e Números Triangulares. Sequências como a das potências de dois e a sequência de Fibonacci foram excluídas por já terem sido analisadas em estudos anteriores (ver Teixeira; 2016). As sequências foram geradas utilizando o Google Planilhas. Na qual:

- Para os Números de Mersenne, iniciamos escrevendo '1' em B_2 e aplicamos a fórmula $B_n=2*B_{n-1}+1$ para calcular os termos subsequentes.
- Para a potência de 3, iniciamos com '3' e aplicamos a fórmula $B_n=3*B_{n-1}$ para calcular os termos subsequentes. Para as demais potências, usamos uma fórmula similar.
- Para a sequência de Lucas, iniciamos com '1' na coluna B_{2} e B_{3} e aplicamos a fórmula $=B_{n-1}+B_{n-2}$ para os termos seguintes.
- Para os números triangulares, usamos $A_{2}=1$, $A_{n}=A_{n}+1$ e $B_n=A_{n}+B_{n-1}$.

Para os dados reais, foram analisados dezessete conjuntos extraídos de fontes governamentais, incluindo: Departamento de Informática do SUS (DataSus), Instituto Brasileiro de Geografia e Estatística (IBGE), Instituto de Pesquisa Econômica Aplicada (Ipea) e o Portal Transparência Bahia (Dados Abertos Bahia). E então colamos cada conjunto escolhido em uma planilha.

Já para testar a detecção de fraudes, selecionamos cinco conjuntos de dados reais entre os usados na pesquisa e pedimos para voluntários alterarem de forma aleatória e com diferentes métodos. Também colamos cada tabela adulterada em uma planilha distinta.

A seguir, efetuamos os seguintes processos:

1. **Extração e Contagem dos Dígitos:**
    - O primeiro dígito dos números foi extraído usando a função `=ESQUERDA(B_n)`, e para os conjuntos de dados reais e adulterados, foi extraído também o segundo dígito.
    - Realizamos a contagem dos dígitos usando o comando `=CONT.SE ()` e comparamos com os valores esperados usando gráficos de barras e linhas.
    - Calculamos a frequência esperada e observada e gerou-se gráficos comparativos.
    - Copiamos a contagem dos primeiros dígitos e colamos em um txt.
2. **Análise Estatística:**
    - Aplicamos o teste qui-quadrado para o primeiro dígito e segundo dígito para verificar a adequação à distribuição uniforme.
    - Para o teste de Kolmogorov-Smirnov, foi necessário utilizar o RStudio e os txt com a contagem dos dígitos, através do comando `ks.test()` .

RESULTADOS E DISCUSSÃO

1. **Sequências Matemáticas**: Dentre as 10 sequências matemáticas analisadas, 9 seguiram a Lei de Newcomb-Benford. A exceção foi a sequência dos Números Triangulares, que não aderiu à lei. Esta discrepância pode ser atribuída à sua ordem de grandeza relativamente baixa se comparada com as outras sequências, exigindo uma maior atenção nisso na hora da escolha de um conjunto para análise.
2. **Dados Reais**: Para os 17 conjuntos de dados reais testados:
    - **Teste Qui-Quadrado**: Todos os 17 conjuntos de dados foram rejeitados pelo teste qui-quadrado, indicando que o teste é muito sensível e pode levar a conclusões de não aderência à lei, mesmo para conjuntos de dados que poderiam de fato segui-la.
    - **Teste Kolmogorov-Smirnov**: O teste Kolmogorov-Smirnov, sendo menos sensível, aprovou 14 dos 17 conjuntos de dados reais.
3. **Dados Adulterados**:
    - **Teste Kolmogorov-Smirnov**: Este teste rejeitou apenas um dos cinco conjuntos de dados adulterados. No entanto, o teste mostrou resultados próximos à rejeição para um segundo conjunto. Mesmo assim, observou-se que a alteração realizada pelos voluntários causou uma aproximação maior à rejeição do que antes das alterações. Indicamos uma alteração em seus parâmetros, considerando-se o p-valor igual à 1 e o D-valor igual à 0,1111 como critério para aprovação forneceria a rigidez necessária ao teste para validação.

Em síntese, o estudo demonstra que o teste Kolmogorov-Smirnov é mais adequado para analisar a aderência à Lei de Newcomb-Benford em dados reais e adulterados, ao passo que o teste qui-quadrado mostrou-se excessivamente sensível para esses fins. A análise

também destacou a importância de considerar a ordem de grandeza dos dados ao aplicar a Lei de Newcomb-Benford.

CONCLUSÃO

A sensibilidade do teste qui-quadrado e a robustez do teste Kolmogorov fornecem uma visão contrastante sobre a validade dos dados em relação à lei. A análise visual, mesmo que útil, não substitui a necessidade de testes estatísticos rigorosos.

Para futuras análises, sugerimos continuar com a procura de testes de aderência cada vez mais adequados e uma atenção quanto à magnitude para evitar problemas semelhantes aos encontrados com os conjuntos analisados.

O trabalho demonstrou-se bastante promissor, analisando 32 conjuntos de dados no total, proporcionando uma base sólida para futuras análises e validações de dados. O estudo sugere que o teste Kolmogorov-Smirnov é mais recomendável para a verificação de aderência em conjuntos de dados reais do que o qui-quadrado.

REFERÊNCIAS

Almeida, Daianne. DISTRIBUIÇÃO DE NEWCOMB-BENFORD: TEORIA E APLICAÇÕES NO PIB DA REGIÃO NORTE DO BRASIL, TRIBUTOS E CONSUMO DE ENERGIA NO ESTADO DO AMAPÁ. 2011. Trabalho de conclusão do curso - Licenciatura Plena em Matemática (graduação), Universidade Federal do Amapá. Macapá, 2011. Disponível em: < https://www.google.com/url?sa=t&source=web&rct=j&url=https://www2.unifap.br/matematica/files/2017/07/TCC-2011-Distribuicao.NB_.Daianne.pdf&ved=2ahUKEwi026jXlPz9AhV8A7kGHR8tCHEQFnoECBQQAQ&usg=AOvVaw3RMIQzYodxwb8gxKe0Xha7 >. Acesso: 3 de março de 2023.

Teixeira, A. C., & Kira, E. (2016). Lei de Benford e aplicações. Instituto de Matemática e Estatística da USP. Disponível em < https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.ime.usp.br/~act/Artigo_ACT_SIICUSP.pdf&ved=2ahUKEwiZ1LDRhfz9AhWOppUCHb3YCuAQFnoECBUQAQ&usg=AOvVaw0wLi8YaYgzgcpi8VXAonTx >. Acesso: 3 de março de 2023.