Validação de Dados utilizando a Lei de Benford

Marcelo Otavio Milani

July 12, 2022

1 Lei de Benford

1.1 Breve Definição

A Lei de Benford, de forma resumida, pode ser interpretada como a lei que quantifica a probabilidade de ocorrência do primeiro dígito de números contidos em um determinado conjunto de dados.

Simon Newcomb em 1881 [New81], foi a primeira pessoa a perceber que em vários conjuntos de dados a ocorrência do primeiro dígito era distribuída de forma logarítmica em vez de uma forma uniforme. Mas o modelo teórico só foi demonstrado em 1938 por Frank Benford [Ben38].

Anos depois, em 1995, Theodore P. Hill [Hil95] publicou um estudo onde generalizou a Lei de Benford não só para o primeiro dígito mas para os demais dígitos também.

1.2 Modelo teórico para o 1º dígito

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

Em que:

- d = 1, 2, . . . , 9.
- P(d): Probabilidade de ocorrência do dígito d.

1.3 Demonstração gráfica da Lei de Benford para o 1º dígito

Aplicando a Equação 1, tem-se as seguintes probabilidades de ocorrência para cada possível primeiro dígito:

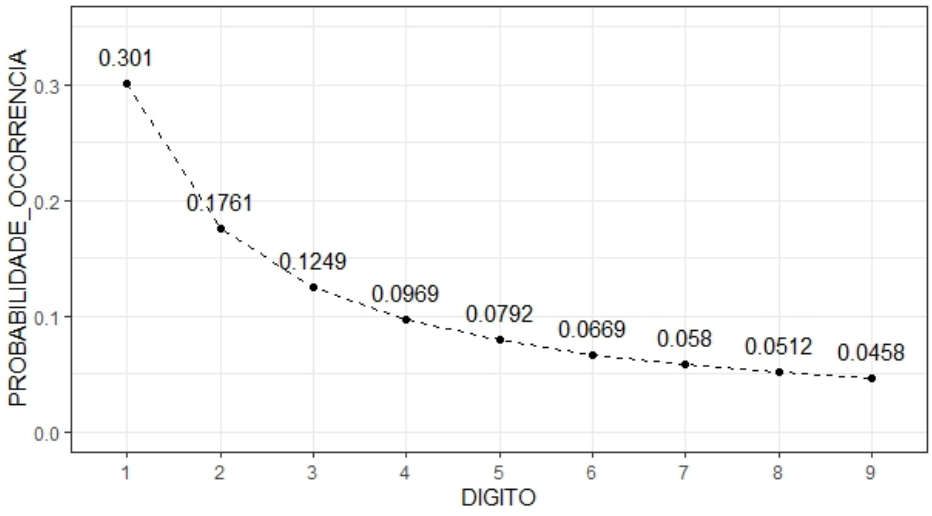

Figure 1: Probabilidade teórica da Lei de Benford.

1.4 Sequência de Fibonacci versus Números Primos

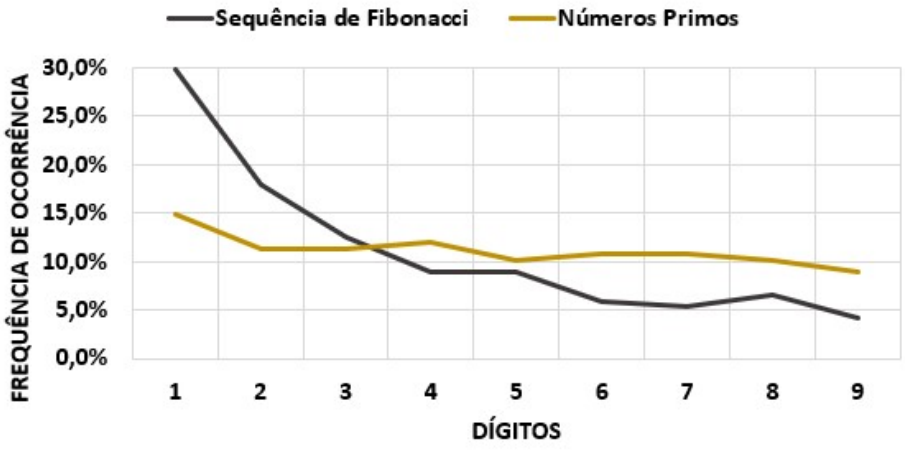

Figure 2: Sequência de Fibonacci versus Números Primos.

Se observa na Figura 2, que a frequência de ocorrência dos primeiros dígitos dos valores da Sequência de Fibonacci se assemelham à Lei de Benford, ao passo que, para os Números Primos o comportamento é mais uniforme.

Para esse exemplo, considerou-se os primeiros 168 valores da Sequência de Fibonacci e dos Números Primos.

1.5 Condições para aplicabilidade da Lei

Como observado na Figura 2, para alguns conjuntos de dados a Lei de Benford se aplica e para outros não. Seguem as condições para aplicabilidade da lei:

- **Aleatoriedade:** os números do conjunto de dados não podem ser sequências ordenadas, como por exemplo: números de senhas de um serviço de atendimento;
- **Limites e Restrições:** os números não podem estar sujeitos a restrições prévias impostas no processo de escolha dos números;
- **Tamanho da Amostra:** por conta da aleatoriedade, quanto menor a quantidade de observações do conjunto de dados, maior a probabilidade de distorção da curva da Lei de Benford;
- **Grandeza dos Dados:** os números que compõem o conjunto de dados devem cobrir várias magnitudes de grandeza.

2 Exemplo de aplicação prática

2.1 Considerações Iniciais

- O objetivo da aplicação foi utilizar a Lei de Benford como instrumento para validação de uma base de dados real, que contém dados sobre o **patrimônio declarado de cada cliente ativo**;
- A base de dados se enquadra nas condições para aplicabilidade da Lei de Benford (Tópico 1.5);
- Foi utilizada a **Linguagem R** para implementar a aplicação (IDE utilizada: RStudio). Detalhes desse processo no GitHub: Validação de Dados - Lei de Benford - GitHub: MilaniMarcelo

2.2 PASSO 1: Carregar a base de dados

O Passo 1 consiste basicamente em carregar a base de dados a ser analisada.

Por causa da LGPD e razões de negócio, a base de dados não ficará disponível. Sendo assim, a integridade dos dados dos clientes fica preservada.

```
# PASSO 1: Carregar a base de dados (a base está no formato .csv)

#Carregar pacote de importação de dados
library(readr)
#Importar base de dados
BD_Patrimonio_Clientes <- read_csv("C:/Users/momil/OneDrive/Gestão - Marcelo Milani
```

Figure 3: Implementação do PASSO 1 na Linguagem R.

2.3 PASSO 2: Carregar pacote que analisa dados usando a Lei de Benford

Uma das razões pela qual a Linguagem R foi escolhida para a implementação desta aplicação, está na existência de um pacote chamado "benford.analysis" que facilita bastante o processo de análise dos resultados.

```
# PASSO 2: Carregar pacote que analisa dados usando a Lei de Benford

#Carregar pacote
library(benford.analysis)
```

Figure 4: Implementação do PASSO 2 na Linguagem R.

2.4 PASSO 3: Analisar os dígitos de primeira e segunda ordem pela Lei de Benford

```
# PASSO 3: Analisar os dígitos de primeira e segunda ordem pela Lei de Benford

#Comando para realizar a análise
bfd.cp <- benford(BD_Patrimonio_Clientes$Patrimonio_Cliente)
#Plotar resultados graficamente
plot(bfd.cp)
```

Figure 5: Implementação do PASSO 3 na Linguagem R.

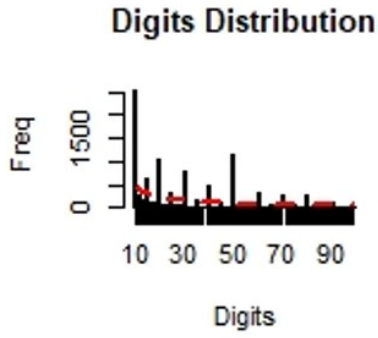

Figure 6: Contagem de observações em relação aos dois primeiros dígitos.

A Figura 6 exibe um gráfico gerado a partir do comando "plot(bfd.cp)". O gráfico se refere à contagem das observações em relação aos dois primeiros dígitos.

```
The 5 largest deviations:

  digits absolute.diff
1     10       2069.46
2     50       1070.01
3     20        827.90
4     30        653.03
5     40        351.05
```

Figure 7: 5 maiores desvios encontrados.

É possível observar que os dados do conjunto que começam com os dígitos "10", "50" e "20 apresentam os maiores desvios em relação à Lei de Benford. A Figura 7 traz os 5 maiores desvios encontrados na base de dados e complementa a observação gráfica da Figura 6.

No "PASSO 4", vamos verificar estatisticamente a aderência dos dados de patrimônio dos clientes à Lei de Benford.

2.5 PASSO 4: Validar os resultados obtidos

```
# PASSO 4: Validar os resultados obtidos
#         (Testes estatísticos: Qui-Quadrado e Desvio Absoluto Médio)

#Exibir os resultados da análise realizada
bfd.cp
```

Figure 8: Implementação do PASSO 4 na Linguagem R.

```
        Pearson's Chi-squared test

data:  BD_Patrimonio_Clientes$Patrimonio_Cliente
X-squared = 36476, df = 89, p-value < 2.2e-16


        Mantissa Arc Test

data:  BD_Patrimonio_Clientes$Patrimonio_Cliente
L2 = 0.016847, df = 2, p-value < 2.2e-16

Mean Absolute Deviation (MAD): 0.01291432
MAD Conformity - Nigrini (2012): Nonconformity
Distortion Factor: -22.11371

Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!:
```

Figure 9: Resultado dos testes estatísticos.

O resultado do Teste de Pearson obteve um *value-p* menor que 5% (IC 95%), isso sinaliza que a base de dados com o patrimônio declarado dos clientes não adere à Lei de Benford.

O teste MAD (*Mean Absolute Deviation*) também resulta em uma não conformidade dos dados à Lei de Benford.

Contudo, mais importante do que saber se os dados como um todo aderem ou não à Lei de Benford, é saber o tamanho do desvio e quais são os dígitos suspeitos. O próprio resultado da análise deixa isso claro na seguinte frase: *"Remember: Real data will never conform perfectly to Benford's Law. You should not focus on p-values!"*.

Sendo assim, o "PASSO 5" traz a parte mais importante da aplicação.

2.6 PASSO 5: Gerar lista com os dados que apresentam distorção (Dados suspeitos)

Com a implementação do "PASSO 5" (Figura 10), é possível gerar uma lista com os dados "suspeitos".

```
# PASSO 5: Gerar lista com os dados que apresentam distorção (Dados suspeitos)
Dados_Suspeitos <- getSuspects(bfd.cp, BD_Patrimonio_Clientes)
```

Figure 10: Implementação do PASSO 5 na Linguagem R.

	Cliente	Patrimonio_Cliente
1	Cliente_ID_3	10000.00
2	Cliente_ID_8	500000.00
3	Cliente_ID_12	1000.00
4	Cliente_ID_14	50000000.00
5	Cliente_ID_17	100000.00
6	Cliente_ID_18	50000.00
7	Cliente_ID_19	50000.00
8	Cliente_ID_21	100000.00
9	Cliente_ID_23	10000.00
10	Cliente_ID_27	100000.00

Figure 11: Lista dos clientes que se enquadram como "suspeitos" em relação ao Patrimônio declarado.

A lista é formada com os dados dos 2 grupos de dígitos com maior desvio (diferença absoluta) em relação à Lei de Benford. Nesta aplicação os 2 grupos são os dígitos "10" e "50", conforme mostrado na Figura 7.

Sendo assim, a Figura 11 exibe alguns clientes que possuem um patrimônio declarado sob suspeita de divergir da realidade.

Por uma questão de limitação de imagem, esses são apenas alguns exemplos, mas no total foram encontrados 3.680 dados suspeitos, que representam cerca de 34% da base de dados.

Ao analisar os exemplos da Figura 11 é possível perceber um comportamento curioso. Todos os 10 valores são discretos, sendo que para valores monetários é mais comum encontrar valores contínuos. Esse comportamento pode ser oriundo de arredondamentos por parte dos clientes, que por uma série de razões não compartilham seus verdadeiros patrimônios.

Essa lista de dados suspeitos é um *input* valioso para qualquer processo de análise e modelagem que envolva essa base de dados. Com ela pode-se evitar o famoso *"Garbage in, garbage out"*.

References

[Ben38] Frank Benford. The law of anomalous numbers. *Proceedings of The American Philosophical Society*, 78:551–572, 1938.

[Hil95] T. P. Hill. The significant-digit phenomenon. *The American Mathematical Montly*, 102:322–327, 1995.

[New81] Simon Newcomb. Note on the frequency of use of the different digits in natural numbers. *Amer. Journ. of Math*, 4:39–40, 1881.