UNIVERSIDADE ESTADUAL DO CEARÁ

CENTRO DE CIÊNCIAS E TECNOLOGIA

MESTRADO PROFISSIONAL EM MATEMÁTICA EM REDE NACIONAL

JACINTO DA SILVA GOMES MATOS

LEI DE BENFORD: CONCEITO, EXPERIMENTOS E APLICAÇÕES

QUIXADÁ – CEARÁ

2020

JACINTO DA SILVA GOMES MATOS

LEI DE BENFORD: CONCEITO, EXPERIMENTOS E APLICAÇÕES

Dissertação apresentada ao Curso de Mestrado Profissional em Matemática em Rede Nacional do Programa de Pós-Graduação em Matemática do Centro de Ciências e Tecnologia da Universidade Estadual do Ceará, como requisito parcial à obtenção do título de Mestre em Matemática em Rede Nacional. Área de Concentração: Matemática Aplicada.

Orientador: Prof. Dr. Diego de Sousa Rodrigues

QUIXADÁ – CEARÁ
2020

Dados Internacionais de Catalogação na Publicação

Universidade Estadual do Ceará

Sistema de Bibliotecas

Matos, Jacinto Da Silva Gomes .
Lei de Benford: conceito, experimentos e aplicações [recurso eletrônico] / Jacinto Da Silva Gomes Matos. - 2020
Um arquivo no formato PDF do trabalho acadêmico com 70 folhas.

Dissertação (mestrado profissional) - Universidade Estadual do Ceará, Centro de Ciências e Tecnologia, Mestrado Profissional em Matemática em Rede Nacional, Fortaleza, 2020.
Área de concentração: Matemática Aplicada..
Orientação: Prof. Dr. Diego de Sousa Rodrigues.

1. Benford. 2. Estatística. 3. Teste qui-qudrado. I. Título.

JACINTO DA SILVA GOMES MATOS

LEI DE BENFORD: CONCEITO, EXPERIMENTOS E APLICAÇÕES

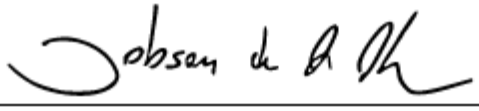
Dissertação apresentada ao Curso de Mestrado Profissional em Matemática em Rede Nacional do Programa de Pós-Graduação em Matemática do Centro de Ciências e Tecnologia da Universidade Estadual do Ceará, como requisito parcial à obtenção do título de Mestre em Matemática em Rede Nacional. Área de Concentração: Matemática Aplicada.

Aprovada em: 30 de outubro de 2020

BANCA EXAMINADORA

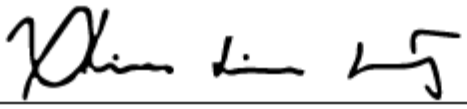
Prof. Dr. Diego de Sousa Rodrigues (Orientador)

Instituto Federal do Ceará – IFCE



Prof. Dr. Jobson de Queiroz Oliveira

Universidade Estadual do Ceará – UECE



Prof. Dr. Ulisses Lima Parente

Universidade Estadual do Ceará – UECE

Aos meus pais.

AGRADECIMENTOS

Inicialmente a Deus, por ter me amparado sempre, meus eternos agradecimentos!
Aos meus pais, e ao meu irmão, que tantas vezes me ajudaram para seguir em frente. Amo vocês!
A minha esposa e filha, lindas, porto todo carinho paciência e flexibilidade nos momentos de necessidades. Amo vocês!
Ao meu orientador, Professor Diego Rodrigues, pela paciência, atenção e dedicação oferecidas durante a construção deste trabalho.
Aos colegas do curso, vivemos bons momentos!
Aos professores do curso, por toda dedicação e compromisso.

Consagre ao Senhor tudo o que você faz,
e os seus planos serão bem-sucedidos.

(Provérbios 16:3)

RESUMO

A Lei de Newcomb-Benford é uma abordagem probabilística da distribuição do primeiro dígito em números escolhidos aleatoriamente, quaisquer que sejam suas origens. No decorrer do trabalho veremos definições, aspectos probabilísticos, estatísticos e fatores históricos que embasam a Lei de Newcomb-Benford (LNB). Forneceremos exemplos sobre a aplicabilidade e condições para o uso desta lei, inclusive com uma proposta pedagógica para a educação básica além de expor resultados obtido através de experimentos práticos. A Lei de Newcomb-Benford afirma que a probabilidade de cada número de 1 a 9 aparecer como primeiro dígito significativo em uma sequência não é uniforme, entretanto, seguem uma distribuição decrescente de acordo com o valor do dígito. Após a definição será abordada a generalização da LNB e algumas propriedades, tais como a invariância na mudança de base e na mudança de escala. Por fim, veremos que diversos conjuntos de números tendem a seguir a LNB e faremos experimentos com o PIB (produto interno bruto) e população das cidades brasileiras e casos de COVID-19, além dos resultados de um experimento pedagógico.

Palavras-chave: Benford. Estatística. Teste qui-qudrado.

ABSTRACT

The Newcomb-Benford Law is a probabilistic approach to the distribution of the first digit in numbers chosen at random, whatever their origins. In the course of the work, we will see definitions, probabilistic, statistical and historical factors that support the Newcomb-Benford Law (LNB). We will provide examples on the applicability and conditions for the use of this law, including a pedagogical proposal for basic education in addition to exposing results obtained through practical experiments. The Newcomb-Benford Law states that the probability that each number from 1 to 9 appears as the first significant digit in a sequence is not uniform, however, it follows a decreasing distribution according to the value of the digit. After the definition, the generalization of LNB and some properties, such as the invariance in the change of base and the change of scale, will be addressed. Finally, we will see that several sets of numbers tend to follow the LNB and we will do experiments with the GDP (gross domestic product) and population of Brazilian cities and cases of COVID-19, in addition to the results of a pedagogical experiment.

Keywords: Benford. Statistic. Chi-square test.

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE SIGLAS E ABREVIATURAS

BNCC	Base nacional curricular comum.
Fdp	Função densidade de probabilidade.
IBGE	Instituto brasileiro de geografia e estatística.
LBN	Lei de Newcomb-Benford
TCU	Tribunal de contas da união.
PIB	Produto interno brutos.
#A	Quantidade (número) de elementos do conjunto A.
E^C	Complementar de do conjunto E.
f_i	Frequência absoluta
f_r	Frequência relativa
f_{ia}	Frequência absoluta acumulada.
f_{ra}	Frequência relativa acumulada.
$\mathbb{N}$	Conjunto dos números naturais.
$\mathbb{Z}$	Conjunto dos números inteiros.
$\mathbb{R}+$	Conjunto dos números reais positivos.
$\mathbb{R}$	Conjunto dos números reais.
$\emptyset$	Conjunto vazio.
α	Lê-se: alfa.
Ω	Lê-se: Ômega. Letra maiúscula.
ω	Lê-se: Ômega. Letra minúscula.
φ	Lê-se: Phi.
μ	Lê-se: Mi.
σ	Lê-se: Sigma.
Γ	Lê-se: Gama.
χ^2	Lê-se: Qui-quadrado.

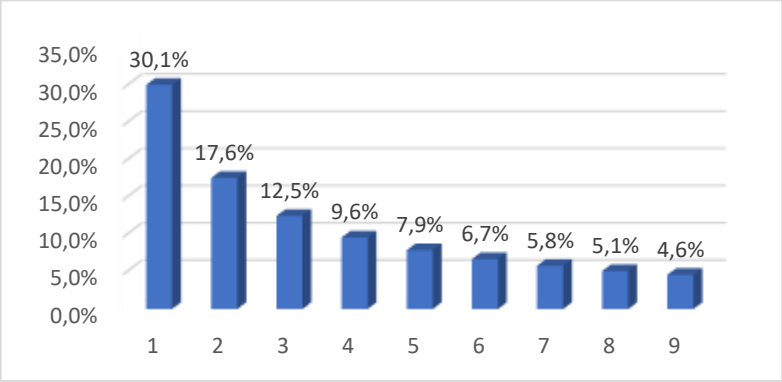
SUMÁRIO

1 INTRODUÇÃO

Este trabalho tem o objetivo de investigar e expor conceitos matemáticos e alguns fatores históricos inerentes a Lei de Newcomb-Benford (LNB), além de exemplificar a aplicabilidade e condições para o uso desta lei. Deste modo, esse trabalho é uma ferramenta para aqueles que pretendem aprender sobre o tema.

A Lei de Newcomb-Benford (também conhecida como, lei do primeiro dígito, lei de Benford ou lei dos números anômalos) refere-se à uma abordagem probabilística de distribuição do primeiro dígito em números escolhidos de maneira aleatória. Para iniciar o estudo sobre esse assunto façamos agora o seguinte exercício: Dada uma amostra unitária (um número) oriunda de um conjunto de dados numéricos qualquer, determine a probabilidade para cada valor de $d \in \{1, \ldots, 9\}$, ser primeiro dígito desta amostra. Intuitivamente podemos ser levados a afirmar que a probabilidade para qualquer que seja o valor de d será de $1/9$,ou seja, aproximadamente 11,11%. Mas se o conjunto citado no exercício obedecer a Lei de Benford, as probabilidades para os valores de *d* ficam muito distantes da regularidade proposta pela solução, neste caso a resposta seria semelhante a tabela abaixo:

Gráfico 1 - Distribuição dos primeiros dígitos conforme a LNB



Fonte: Elaborada pelo autor.

A Lei de Newcomb-Benford afirma que, ao contrário da homogeneidade que pode ser esperada, ao escolher aleatoriamente números quaisquer e calcular as frequências de seus primeiros algarismos significativos não nulos, ou seja, algarismos da extrema esquerda, a probabilidade de obtermos 1 no primeiro dígito é de 30,1%, e de obtermos 2 como primeiro dígito é de 17,6% e assim por diante conforme o gráfico anterior. A regra descrita por essa lei está presente em diversos conjuntos numéricos

tais como a extensão de rios do mundo e a quantidade de habitantes por cidade no mundo ou no Brasil. Essa lei também pode ser observada em conjuntos matemáticos, tais como o conjunto das potencias de 2 e a Sequência de Fibonacci.

Verificaremos a aproximação da Sequência de Fibonacci com a Lei de Newcomb-Benford. Sendo, $F_1 = 1$ e $F_2 = 1$ os respectivos valores iniciais da Sequência de Fibonacci, recursivamente podemos defini-la pela seguinte fórmula:

$$F_n = F_{n-1} + F_{n-2}.$$

Estes são os números iniciais da sequência:
1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, …
Observemos agora a seguinte tabela.

Tabela 1 - Comportamento dos 100 termos iniciais da sequência de Fibonacci perante a LNB

1° Dígito	Frequência absoluta	Frequência percentual
1	30	30 %
2	18	18 %
3	13	13 %
4	9	9 %
5	8	8 %
6	6	6 %
7	5	5 %
8	7	7 %
9	4	4 %

Fonte: Elaborada pelo autor.

Analisando a tabela é fácil perceber a similaridade com os percentuais propostos pela Lei de Benford, mas há algumas distorções bastante acentuadas, como no caso do algarismo 8. Agora analisaremos através da tabela abaixo o comportamento dos 1476 primeiros elementos da sequência perante a referida lei.

Tabela 2 - Comportamento dos 1476 termos iniciais da sequência de Fibonacci perante a LNB

1° Dígito	Frequência absoluta	Frequência percentual
1	301	30,15%
2	177	17,62%
3	125	12,47%
4	96	9,69%
5	80	7,93%
6	67	6,64%
7	56	5,76%
8	53	5,22%
9	45	4,54%

Fonte: Elaborada pelo autor.

Após a segunda análise é fato que o comportamento da frequência percentual no primeiro dígito é semelhante ao da Lei de Benford.

A Lei de Benford é realmente fascinante e desafia os instintos matemáticos, pode ser objeto didático para uma aula incrível, envolvendo probabilidade, estatística e diversos conceitos matemáticos a depender o nível da turma, mas ela vai além dos livros e da sala de aula. A lei do primeiro dígito é utilizada para detectar fraudes financeiras/contábeis, já foi usada como prova judicial, além de outras aplicações. Atualmente existem diversos trabalhos publicados sobre a aplicabilidade do tema. Mas se generalizarmos essa lei veremos que as probabilidades podem ir além do primeiro dígito, na verdade é possível ir até segundo, terceiro, enfim até o enésimo dígito.

1.1 A tabela das Frequências para o primeiro dígito

Inicialmente vamos verificar uma melhor maneira para definir o primeiro dígito significativo de um número, ou seja, o dígito não nulo mais à esquerda em sua representação decimal. Sejam $x \in \mathbb{R}^+$, podemos escrever cada número x como o produto de um $m \in [1,10)$ por uma potência de base 10, neste caso chamemos o número de m de mantissa de x. Isto é,

$$x=m10^n, \text{ com } n \in \mathbb{Z}.$$

Assim, temos que a parte inteira de m é o algarismo significativo de x, que é comumente denotada por $\lfloor m \rfloor$, e como anteriormente, a LNB trata justamente dos possíveis valores para $\lfloor m \rfloor$.

Com base na referida lei podemos afirmar que, para números coletados de forma aleatória, a frequência P(d) do primeiro dígito significativo d, com d = {1,...,9}, é determinada aproximadamente por

$$\log_{10}\left(1+\frac{1}{d}\right).$$

Ao substituir cada valor de d, obtemos a seguinte tabela:

Tabela 3 - Frequência P(d), para o primeiro dígito conforme a LNB

P	1	2	3	4	5	6	7	8	9
P(d)	0,301	0,1761	0,1249	0,0969	0,0792	0,0669	0,058	0,0511	0,0458

Fonte: Elaborada pelo autor.

Agora podemos afirmar que um conjunto satisfaz a lei de Benford se a probabilidade para a ocorrência do primeiro dígito significativo d, for a seguinte:

$$P(d) = \log_{10}(d+1) - \log_{10} d = \log_{10}\left(1+\frac{1}{d}\right)$$

Eis, o motivo do gráfico 1 apresentado na solução do exercício inicial.

1.2 A descoberta da Lei de Benford

1.2.1 (1881): Simon Newcomb e os livros contendo tábuas de logaritmos

O astrônomo e matemático américo-canadense, Simon Newcomb (1835-1909), observou que as primeiras páginas das tabelas de logaritmos, livros utilizados na época para realizar cálculos logarítmicos, eram muito mais utilizadas do que as últimas páginas. O fato é que as primeiras páginas eram correspondentes aos números com os primeiros algarismos significativos pequenos, como 1 ou 2. A partir dessa observação Newcomb propôs que em qualquer lista de números oriunda de um conjunto aleatório, a quantidade de números que tem “1” como primeiro algarismo significativo tende a ser maior. Através de suas pesquisas, Newcomb enunciou, por volta de 1881, a fórmula para P(d), que já vimos na introdução deste trabalho.

1.2.2 (1938): Os experimentos de Frank Albert Benford Jr.

Por volta de 1938, Frank Albert Benford Jr (1883-1948) coletou milhares de números de diversas origens, como área e superfície de lagos, tamanho de populações de 3259 locais dos EUA, constates físicas, dentre outros dados. O total de números utilizados foi superior a 20.000, distribuídos em diversos conjuntos e todos esses conjuntos seguiam a mesma distribuição proposta por Simon Newcomb . Com isto, essa regra ganhou o nome de lei de Benford.

1.2.3 (1995): Theodore P. Hill

A justificativa rigorosa para esta lei foi demonstrada por volta de 1995, com os trabalhos do matemático Theodore P. Hill, que conseguiu provar o comportamento matemático presente nas distribuições.

2 ASPECTOS ESTATÍSTICOS

Neste capítulo, abordaremos a estatística e probabilidade que servem como base para compreensão e aplicação da lei de Benford.

2.1 Conceitos iniciais

Pode se dizer que a **estatística** é o método científico utilizado para o trabalho de coleta, processamento, organização, análise e apresentação dos dados de determinada população de estudo, denominaremos esse trabalho de estudo estatístico.

Entende-se por **população** o conjunto composto por todos elementos que detêm caraterísticas que são de interesse de determinado estudo estatístico. O **estudo estatístico censitário** é aquele que a abrange diretamente todo elemento da população. Mas, quando não é possível estudar toda uma população, recorre-se a apenas uma parte representativa, que é denominada amostra. A **amostra** é um subconjunto da população usado no estudo estatístico.

Normalmente o estudo estatístico coleta vários dados, pois há interesse em descobrir alguma característica ou tendência relacionada à população. Cada característica específica é denominada **variável**.

Uma variável é dita **quantitativa discreta** quando está associada à contagem de dados e neste caso sempre assume valores inteiros. Exemplo: quantidade de alunos aprovados, idade de cada aluno, ou quantas vezes o número 1 (um) aparece como primeiro dígito significativo em uma lista numérica aleatória que contem 1000 (mil) números.

Uma variável também pode ser classificada como **quantitativa contínua**, nesse caso, ela assume valores reais determinados por alguma unidade ou parâmetro de medida. Exemplo: altura, peso e preço.

2.2 Apresentação dos dados

A análise estatística dos dados coletados verifica a ocorrência de determinada variável, essa ocorrência é chamada de **frequência absoluta** ou

frequência relativa. Os dados do estudo estatístico são normalmente agrupados ou divididos em classes e apresentados por meio de tabela de frequência e gráficos.

- A frequência absoluta (f_i) é a quantidade, ou o número médio de vezes, que um elemento da população assume determinada caraterística no estudo estatístico, ou seja, a ocorrência de uma variável.
- A frequência relativa (f_r) é razão, fração ou porcentagem, que a frequência absoluta de cada variável representa em relação a somatória das frequências absolutas. Assim,

$$f_r = \frac{f_i}{\sum f_i}.$$

Exemplo 2.1. Considere a sequência A: $a_{n+1} = 2a_n\,, n \in \mathbb{Z}, n \geq 0, \text{e}\ \ a_1 = 1.$

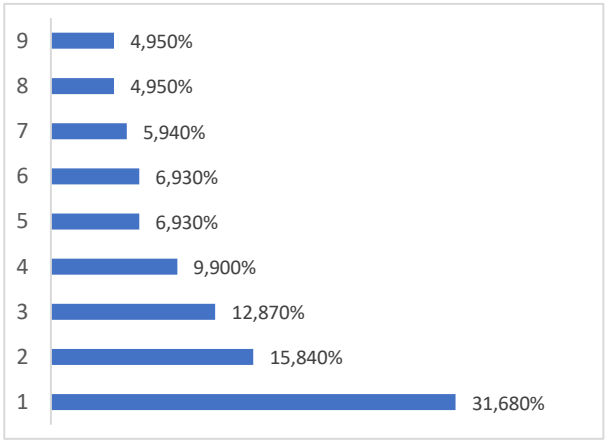
Com auxílio do software Excel, construiu-se a tabela 4 e o gráfico 2, das frequências do primeiro dígito significativo da sucessão dos 1024 termos iniciais de A.

Tabela 4 - Distribuição dos 1024 termos iniciais da sequência A perante a LNB

Primeiro dígito significativo	f_i	f_r
1	308	31,08%
2	181	17,68%
3	127	12,40%
4	100	9,77%
5	81	7,91%
6	70	6,84%
7	57	5,57%
8	55	5,37%
9	45	4,39%

Fonte: Elaborada pelo autor.

Gráfico 2 - Distribuição de frequência relativa da tabela 4



Fonte: Elaborada pelo autor.

Exemplo 2.2. Ilustraremos uma forma de representação com dados distribuídos em classes. Para isto temos a tabela de frequências e o histograma da área dos 184 municípios, incluindo a capital, do estado do Ceará em km^2 construída com base em dados expostos pelo IBGE (Instituto brasileiro de geografia e estatística).

Considerações iniciais.

i. A menor e maior área observada respectivamente é 72,68km^2 e 4262,30 km^2.

ii. O número de classes “n”, no caso n=13, foi definido pela parte inteira de $\sqrt{p}$, com p=184 (a quantidade da amostra estudada).

iii. A amplitude das classes “t”, foi definida por, t= (4262,30 -322,28)/13.

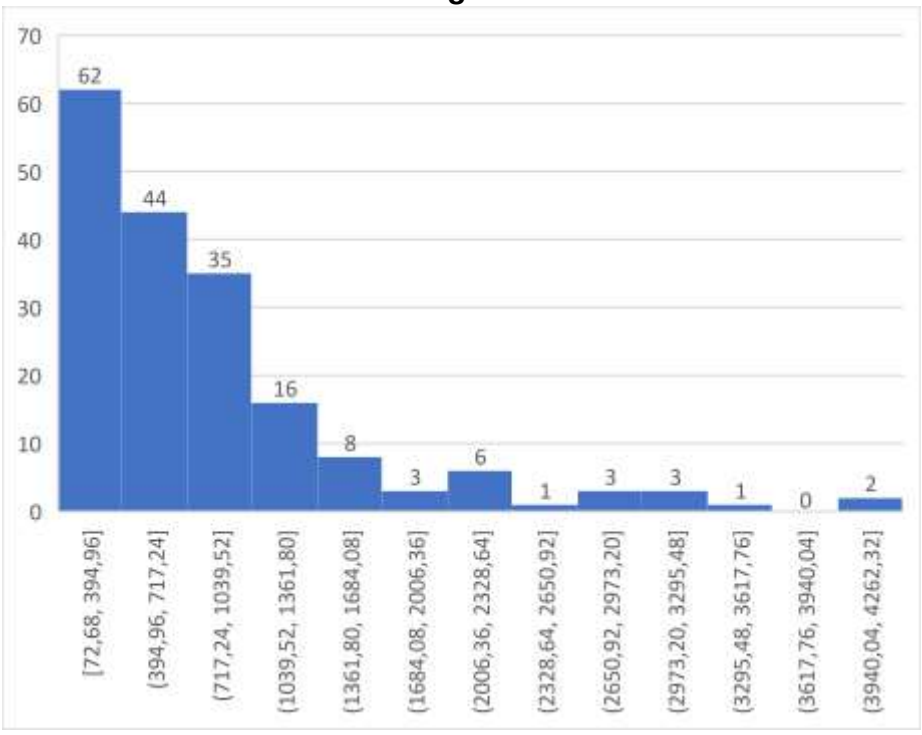
iv. Se x está contido na classe $a \vdash b$, então $a \leq x < b$.

Tabela 5 - Distribuição de frequência da área dos 184 municípios do estado do Ceará

Classes	f_i	f_r	f_{ia}	f_{ra}
72,68 ⊢ 394,96	62	33,70%	62	33,70%
394,96 ⊢ 717,24	44	23,91%	106	57,61%
717,24 ⊢ 1039,52	35	19,02%	141	76,63%
1039,52 ⊢ 1361,80	16	8,70%	157	85,33%
1361,80 ⊢ 1684,08	8	4,35%	165	89,67%
1684,08 ⊢ 2006,36	3	1,63%	168	91,30%
2006,36 ⊢ 2328,64	6	3,26%	174	94,57%
2328,64 ⊢ 2650,92	1	0,54%	175	95,11%
2650,92 ⊢ 2973,20	3	1,63%	178	96,74%
2973,20 ⊢ 3295,48	3	1,63%	181	98,37%
3295,48 ⊢ 3617,76	1	0,54%	182	98,91%
3617,76 ⊢ 3940,04	0	0,00%	182	98,91%
3940,04 ⊢ 4262,32	2	1,09%	184	100,00%

Fonte: Elaborada pelo autor.

Gráfico 3 -Histograma da tabela 5



Fonte: Elaborada pelo autor

3 PROBABILIDADE

3.1 Conceitos de medida

Definição 3.1 Dado um conjunto X, o conjunto das partes de X, $\mathcal{P}(X)$, é o conjunto dos subconjuntos de todo os subconjuntos de X.

Definição 3.2 Uma $\sigma - algebra$ de subconjuntos de X é uma família Σ de subconjuntos de X, isto é, $\Sigma \subset \mathcal{P}(X)$, tal que:

a) $\emptyset \subset \Sigma$;

b) $\text{para todo } E \in \Sigma, \text{seu complemento } E^c = X \backslash E \in \Sigma$;

c) $\text{para toda coleção } \{E_n\}_{n\in\mathbb{N}} \text{ em } \Sigma, \text{sua união } \bigcup_{n\in\mathbb{N}} E_n \in \Sigma$.

Os elementos de Σ são chamados de conjuntos mensuráveis.

Se Σ satisfaz a) e b) e, ao invés de c) satisfaz c') abaixo dizemos que é uma álgebra.

c') $\text{Dados } E, F \in \Sigma, \text{ sua união } E \cup F \in \Sigma$.

Exemplos 3.1

a) Existem duas $\sigma - \text{algebra}$ de subconjuntos de X que são canônicas são elas:
   $\Sigma = \{\emptyset, X\}, \text{a menor } \sigma - \text{algebra de X, e } P(X), \text{a maior } \sigma - \text{algebra de X.}$

b) Considere o conjunto X ={1,2,3,4}
   $\Sigma = \{\emptyset, \{1\}, \{2,3,4\}, X\}, \text{e } \Sigma = \{\emptyset, \{1,2\}, \{3,4\}, X\} \text{ são } \sigma - \text{algebra de X.}$

c) O conjunto $\Sigma = \{\emptyset, \mathbb{Q}, \mathbb{Q}^C, \mathbb{R}\}$, é uma $\sigma - \text{algebra de } \mathbb{R}$.

Definição 3.3 Um espaço de medida é uma tripla (X, Σ, μ), onde:

a) $X \text{ é um conjunto;}$

b) $\Sigma \subset P(X) \text{ é uma } \sigma - \text{algebra de subconjuntos de X;}$

c) $\mu \to [0, \infty] \text{ é função tal que:}$

   c' $\mu(\emptyset) = 0$;

   c" $(\sigma - \text{aditividade})$ $\text{Se } \{E_n\}_{n\in\mathbb{N}} \text{ é uma sequência disjuta em } \Sigma, \text{então}$

$$\mu\left(\bigcup_{n\in\mathbb{N}} E_n\right) = \sum_{n=0}^{\infty} \mu(E_n) .$$

μ é dita como uma medida de X.

Dizemos que a sequência $\{E_n\}_{n\in\mathbb{N}}$ é disjunta, se nenhum ponto pertence a mais do E_n. Ou seja, se

$$E_m \cap E_n = \emptyset, \text{para todo } m, n \in \mathbb{N}, m \neq n.$$

De modo análogo podemos afirmar que, se $\{E_i\}_{i\in I}$ se é uma família de conjuntos indexada por um conjunto arbitrário I .Deste modo, I é disjunto se:

$$E_i \cap E_j = \emptyset \text{ para todo } i, j \in \mathbb{N}, i \neq j.$$

3.2 Conceitos de probabilidade

É parte integrante do cotidiano real e prático das pessoas a convivência com algum tipo de incerteza, mesmo se reduzindo ao máximo o subjetivismo, os mais cautelosos se depararam com situações que podem não controlar totalmente. Mas, apesar das incertezas serem uma realidade, existe um ramo da matemática que nos ajuda a quantificar as possibilidades para o desfecho de determinadas situações com base nos seus possíveis resultados, essa área da matemática é a probabilidade.

O objetivo da probabilidade é estudar fenômenos que ocorrem ao acaso, ou seja, aqueles em que não há exatidão para o seu resultado. Neste capítulo iremos estudar alguns aspectos da probabilidade necessários no estudo da lei de Benford.

Definição 3.4 (Experimento aleatório) É uma experiência qualquer cujo resultado não é conhecido e não depende da quantidade de repetições desse experimento. Onde há pelo menos dois resultados possíveis e o máximo de informação disponível antes de sua realização são as chances para seus resultados.

Definição 3.5 (Espaço amostral) É o conjunto de todos os resultados de um experimento aleatório. É comumente designado por Ω. O espaço amostral pode ser discreto ou contínuo.

I. O espaço amostral discreto ocorre quando este é um conjunto enumerável. Por exemplo, o lançamento de uma moeda. Ω = {C, K}, onde C = CARA e K = COROA.
II. O espaço amostral contínuo é quando contém um ou mais intervalos. Por exemplo, o tempo de vida de uma lâmpada (célula, organismo, processo, etc...).

$$\Omega = \{x \in \mathbb{R} : x > 0\} = (0, +\infty) = \mathbb{R}^+.$$

Definição 3.6 (Evento) quaisquer subconjuntos do espaço amostral Ω. Normalmente é denotado por uma letra maiúscula. Podemos classificar um evento das seguintes maneiras. Um evento A é :

a. certo se A= Ω ou seja, $x \in A \leftrightarrow x \in \Omega$
b. simples quando A tem apenas um único elemento,
c. impossível quando A é conjunto vazio (Ø).

Por outro lado, dois eventos, A e B, são ditos incompatíveis (ou mutuamente exclusivos) se sua intersecção é vazia, A∩B = Ø.

Exemplo 3.2 Para um único lançamento de um dado honesto, com 6 faces numeradas de um a seis, verifica-se o número da face que ficará voltada para cima.

a. Nesse experimento o espaço amostral é Ω = {1,2,3,4,5,6}.
b. O evento é certo quando o subconjunto A de Ω for A={1,2,3,4,5,6}.
c. Evento: Obter um número primo. A={2,3,5} ⊂ Ω .
d. Evento: Obter um número que seja primo e também seja par. A={2} ⊂ Ω "Evento simples".
e. Evento: Obter um número maior que 6. A = Ø.

Definição 3.7 (Kolmogorov (1903-1987)): Sendo F uma σ-álgebra do espaço amostral Ω, probabilidade é uma função

$$P : F \rightarrow [0,1]$$

que satisfaz as seguintes condições:

I. Para todo evento A, $0 \leq P(A) \leq 1$, para qualquer $A \in F$.
II. $P(\Omega) = 1$,
III. Se $A = \{A_1, A_2, \dots, A_n\}$, n natural, $A_i \cap A_j = \emptyset$, com i, j naturais, então: $P(A) = \sum_{i=1}^{n} P(A_i)$.

A tripla (Ω,F,P) é chamado de espaço de probabilidade.

Proposição 3.1 Se A e B são eventos, então:

I. $(\Omega \backslash A) = A^C \subset A \longrightarrow P(A^C) = 1 - P(A)$.
Demonstração: $1 = P(\Omega) = P(A \cup A^C) = P(A) + P(A^C)$.
Daí, $P(A^C) = 1 - P(A)$.

II. $P(\emptyset) = 0$.
Demonstração: $P(\Omega) = P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset)$, pois $\Omega \cap \emptyset = \emptyset$.
Logo, $P(\emptyset) = 0$.

III. $P(A \backslash B) = P(A) - P(A \cap B)$.

Probabilidade de A ocorrer dado que B ocorreu. Demonstração: $P(A) = P[(A\backslash B) \cup (A \cap B)] = P(A\backslash B) + P(A \cap B)$,
pois $(A\backslash B) \cap (A \cap B) = \emptyset$.
Dest modo, $P(A\backslash B) = P(A) - P(A \cap B)$.

IV. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
Demonstração: $P(A \cup B) = P[(A\backslash B) \cup B] = P(A\backslash B) + P(B)$, pois $(A\backslash B \cap B) = \emptyset$. Como $P(A\backslash B) = P(A) - P(A \cap B)$, resulta que $P(A \cup B) = P(A) - P(A \cap B)$

V. Se $A \supset B$ então, $P(A) \geq P(B)$.
Demonstração: Como $P(A\backslash B) = P(A) - P(A \cap B)$, se $A \subset B$, então $P(A\backslash B) = P(A) - P(B)$. Como $P(A\backslash B) \geq 0$, temos $P(A) \geq P(B)$.

3.3 Interpretações da probabilidade

Subjetivamente, a probabilidade para ocorrência de um evento, é usado como indicativo para induzir ou medir o grau de crença na ocorrência deste evento. Mas veremos a seguir interpretações objetivas da probabilidade.

3.3.1 Interpretação de Laplace (1749-1827) “clássica”

Seja um experimento com espaço amostral finito onde todos os resultados são equiprováveis, ou seja, os eventos simples possuem a mesma importância. A interpretação clássica diz que a probabilidade para um evento A ocorrer é a razão entre o número de elementos de A, representado por #A, e o número de resultados possíveis e presentado por #Ω.

$$P(A) = \frac{\#A}{\#\Omega}.$$

Exemplo 3.3 Determinado setor de uma empresa tem 20 funcionários, sendo um gerente, dois coordenadores, sete analistas e dez assistentes administrativos. Ocorrerá um sorteio com participação exclusiva dos referidos funcionários. Sabendo que todos eles têm as mesmas chances, a probabilidade do gerente ser sorteado é

$P(1) = \frac{1}{20}$, já a probabilidade de um analista ser sorteado é $P(7) = \frac{7}{20}$ e a probabilidade de um coordenador ou assistentes administrativos ser sorteado é $P(12) = \frac{3}{5}$.

3.3.2 Interpretação Frequentista

A probabilidade de um evento com espaço amostral finito onde todos os eventos são equiprováveis, é o limite da frequência relativa deste evento quando o número de repetições deste experimento tende ao infinito.

$$\text{Sendo } f_n(A) \text{ o número de ocorrências do evento A então, } P(A) = \lim_{n\to\infty} \frac{f_{n(A)}}{n}.$$

Ainda sobre exemplo 3.2, com base na interpretação frequentista de probabilidade se fizéssemos 200 sorteios, cada funcionário teria a mesma chance de ser sorteado em cada um dos 200 sorteios. Deste modo, as probabilidades se manteriam para cada sorteio: P(gerente) = $\frac{1}{20}$, P(analista) = $\frac{7}{20}$, P(coordenador ou assistente administrativo) = $\frac{3}{5}$, e quanto maiores forem as repetições dos sorteios mais os resultados devem se aproximar destes apontamentos.

Exemplo 3.4 Tendo como base o exemplo 3.2, considere i, com $1 \leq i \leq 6$. Após um lançamento.

a) A probabilidade de um número i, estar na face voltada para cima é: $P(i) = \frac{1}{6}$.

b) Qual probabilidade de i não estar na face voltada para cima?

   Nesse experimento, Ω = {1,2,3,4,5,6}, portanto $P(\Omega) = \frac{6}{6} = 1$.

   Deste modo a probabilidade de i, não ocorrer é: $P(\Omega) - P(i) = 1 - \frac{1}{6} = \frac{5}{6}$.

c) Qual a probabilidade de ocorrer um i, de modo que i seja múltiplo de 2 ou de 3?

   Observe que os múltiplos de 2 e de 3 são os elementos dos conjuntos A e B respectivamente.

   $$A=\{2,4,6\};B=\{3,6\}.$$

   Note que $(A \cap B) \neq \emptyset$, então $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

   Deste modo;

   $$P(i) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{2}{3}.$$

d) A probabilidade de ocorrer um i divisor de 5 é menor que a probabilidade de ocorrer j, com $1 \leq j \leq 6$, impar?

Considere os conjuntos: A={1,5} e B ={1,3,5}, observe que $\mathrm{A} \subset \mathrm{B}$, $\mathrm{i} \in \mathrm{A}; \mathrm{j} \in \mathrm{B}$. Deste modo podemos perceber:

$$\mathrm{P(i)} = \frac{2}{6} \text{ e } \mathrm{P(j)} = \frac{3}{6} \rightarrow \mathrm{P(i)} < \mathrm{P(j)}.$$

Exemplo 3.5 Considere urna contendo 5 bolas e o seguinte experimento. Sortear duas bolas, uma após a outra, sem reposição, ou seja, após o primeiro sorteio ficam somente 4 bolas para o segundo sorteio, sabendo que as bolas estão numeradas de 1 a 5 e têm probabilidades idênticas. Qual a probabilidade da segunda bola sorteada conter um número primo?

Representemos: ocorrer uma bola contendo um número primo por *q* e ocorrer uma bola contendo um número não primo por ~*q*. Isto posto, a tabla abaixo mostra os dois cenários possíveis para *q* no segundo sorteio.

	$1º$ Sorteio	$2º$ Sorteio
Cenário 1 (C_1)	*q*	*q*
Cenário 2 (C_2)	~*q*	*q*

Para o $1º$ sorteio temos que: $\mathrm{P}(q) = \frac{3}{5}$ e $\mathrm{P}(\sim q) = \frac{2}{5}$. Deste modo, para ocorrer *q* no $2º$ sorteio é: $\mathrm{P}(\mathrm{C}_1) + \mathrm{P}(\mathrm{C}_1) = \frac{3}{5} * \frac{2}{4} + \frac{2}{5} * \frac{3}{4} = \frac{12}{20} = \frac{3}{5}$.

Observações:

- Há 5 possiblidades para o primeiro sorteio e 4 possibilidades para o segundo sorteio, deste modo o espaço amostral Ω é composto 20 eventos. $\Omega = \begin{Bmatrix} (1,2);(1,3);(1,4);(1,5) \\ (2,1);(2,3);(2,4);(2,5) \\ (3,1);(3,2);(3,4);(3,5) \\ (4,1);(4,2);(4,3);(4,5) \\ (5,1);(5,2);(5,3);(5,4) \end{Bmatrix}$.
- Há 3 possibilidades para acontecer *q* no primeiro sorteio e 2 possibilidades para ocorrer no segundo sorteio. Portanto há 6 eventos favoráveis ao (C_1). Conforme o conjunto E_1={(2,3);(2,5);(3;5);(3,2);(5,2);(5,3)}.
- Há 2 possibilidades para acontecer ~*q* no primeiro sorteio e 3 possibilidades para ocorrer *q* no segundo sorteio. Portanto há 6

eventos favoráveis ao (C_2). Conforme o conjunto E_2={(1,2);(1,3);(1;5);(4,2);(4,3);(4,5)}.

- Portanto:

$$P(C_1) = \frac{6}{20} = \frac{3}{10} \text{ e } P(C_2) = \frac{6}{20} = \frac{3}{10}$$

Deste modo, como $E_1 \cap E_2 = \emptyset$, probabilidade pedia é $P(C_1) + P(C_2) = \frac{3}{5}$.

Definição 3.8 Dados dois eventos A e B, com $P(A) \neq 0$ a probabilidade condicional de B na certeza de A é:

$$P(B\backslash A) = \frac{P(A \cap B)}{P(A)}.$$

Exemplo 3.6 Conforme o exemplo 3.5.

a) Qual a probabilidade de sair no segundo sorteio uma bola contendo um número primo, sabendo que no primeiro sorteio saiu uma bola contendo um úmero par?

Inicialmente vamos observar os seguintes pontos:

- Já vimos que o espaço amostral Ω é composto por 20 eventos.
- Há 8 eventos em que no primeiro sorteio ocorre uma bola contendo um úmero par conforme o conjunto A={$(2,1); (2,3); (2,4); (2,5); (4,1); (4,2); (4,3); (4,5)$}.
- Há 12 eventos em que no primeiro sorteio ocorre uma bola contendo um úmero par conforme o conjunto:
  $B = \{(1,2); (1,3); (1,5); (2,3); (2,5); (3,2); (3,5); (4,2); (4,3); (4,5); (5,2); (5,3)\}$.
- $(A \cap B) = \{(2,3); (2,5); (4,2); (4,3), (4,5)\}$.
  Deste modo temos que:

$$P(B\backslash A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{5}{20}}{\frac{8}{20}} = \frac{5}{8}.$$

b) Qual a probabilidade de sair uma bola contendo um número par e em seguida ocorrer uma bola contendo um número ímpar?
   Pela definição 3.8, temos que $P(A \cap B) = P(A)P(B\backslash A)$.
   Um número ou é par ou é ímpar, portanto, portanto para o primeiro sorteio queremos ocorra um número do conjunto P={2,4} e para o segundo sorteio queremos que ocorra um número do conjunto I={1,3,5}. Deste modo, a probabilidade é $\frac{2}{5} * \frac{3}{4} = \frac{3}{10}$, pois não há reposição de bolas.

Exemplo 3.7 (Profmat-MA12) Uma moeda com probabilidade de $\frac{1}{3}$ de dar cara, é lançada 40 vezes.

a) Explique por que a probabilidade p_k de se obter k caras nos 40 lançamentos é dada por

$$p_k = C_{40,k}\left(\frac{1}{3}\right)^k\left(\frac{1}{3}\right)^{40-k}$$

para $k = 0,1,2 \ldots ,40$.

Explicação: Há $C_{40,k} = \frac{40!}{k(40-k)!}$ com k caras em 40 lançamentos, cada uma com probabilidade $\left(\frac{1}{3}\right)^k\left(\frac{1}{3}\right)^{40-k}$.

b) Calcule para quais valores de k tem-se $p_{k+1} > p_k$.

$$p_{k+1} > p_k \leftrightarrow \frac{40!}{(k+1)(40-k-1)!}\left(\frac{1}{3}\right)^{k+1}\left(\frac{1}{3}\right)^{40-k-1} > \left(\frac{1}{3}\right)^k\left(\frac{1}{3}\right)^{40-k}\left(\frac{1}{3}\right)^k\left(\frac{1}{3}\right)^{40-k}.$$

Ou seja, se e somente se

$$\frac{1}{k+1} * \frac{1}{3} > \frac{1}{40-k} * \frac{2}{3}.$$

Portanto, $p_{k+1} > p_k \leftrightarrow 40 - k > 2k + 2 \leftrightarrow k < \frac{38}{3}$. Deste modo $k \leq 12$, pois $k \in \mathbb{Z}$.

c) Utilize b) para obter o valor de k para o qual a probabilidade de se obter k caras é máxima.

Por b) temos que

$k_0 < k_1 < k_2 < \cdots < k_{12} < k_{13}$, e que $k_{13} \geq k_{14} \geq k_{15} < \cdots \geq k_{39} \geq k_{40}$.

Embora valham, de fato, as desigualdades estritas, se for aplicado raciocínio análogo àquele feito em (b). O valor máximo, ocorre, portanto, em $k = 13$.

3.3 Medida de probabilidade

Definição 3.9 Dado um espaço de medida $(\mathrm{X}, \Sigma, \mu)$, dizemos que é um espaço de probabilidade se $\mu(\Omega) = 1$. Neste caso denotamos a media μ por P e dizemos que $(\mathrm{X}, \Sigma, \mathrm{P})$ é um espaço de probabilidade.

- Ω (espaço amostral).
- Os elementos da $\sigma - \mathrm{algebra}\ \Sigma$ são os eventos.
- A cada $\mathrm{A} \in \Sigma$ associamos sua probabilidade $\mathrm{P}(\mathrm{A})$.
- A função mensurável $\mathrm{X}\colon \Omega \to \mathbb{R}$ é chamada de varável aleatória.
- A integral $\int XdP$ chamada de esperança da variável aleatória X, $\mathrm{E}(\mathrm{X})$.

- Uma sequência $\{X_n\}_{n\in\mathbb{N}}$ de varável aleatória, refere-se um processo com variável aleatória discreta.
- Uma família $\{X_t\}_{t\in\mathbb{R}}$ de varável aleatória, refere-se um processo com variável aleatória contínua.

Exemplo 3.8

a) Considere o lançamento de dois dados no mesmo instante, um exemplo de processo probabilístico com variáveis discretas é processo variáveis discretas X_n igual a soma da face de ambos os dados que fica virada para cima em cada lançamento.

b) O valor de uma ação a cada instante de tempo é um exemplo de processo com variável aleatória contínua.

4 VARIÁVEL ALEATÓRIA

Definição 4.1 Dado um experimento aleatório, com o espaço de probabilidades (Ω,F,P), chamamos de variável aleatória X é qualquer função $X : \Omega \to \mathbb{R}$ de modo que:

$$X^{-1} = \{\omega \in \Omega : X(\omega) \in I\} \in F$$

para todo intervalo $I \subset \mathbb{R}$, ou seja, X é tal que a imagem inversa de qualquer $I \subset \mathbb{R}$ pertence a σ-Álgebra F.

Deste modo, podemos dizer que uma variável aleatória é uma função real, definida no espaço amostral Ω de um experimento aleatório, ou seja, é uma função que associa qualquer elemento de Ω um número real.

Exemplo 4.1 Considere o seguinte experimento: lançar uma moeda n vezes e verificar a face da moeda que fica virada para cima.

Podemos associar o resultado de cada lançamento aos números 1 e 0 da seguinte forma:

CARA → 1 COROA → 0.

Dessa forma o resultado do lançamento de uma moeda n vezes pode ser associado a um vetor com n entradas iguais a 1 ou 0, dependendo do resultado em cada lançamento. Por exemplo, o resultado de lançarmos uma moeda 5 vezes e obtermos a seguinte sequência:

CARA, CARA, COROA, CARA, COROA

é representado pelo vetor (1, 1, 0, 1, 0).

Temos então:

- Ω_n: Espaço amostral contendo todas as possibilidades de resultados, de acordo com a nossa identificação, $\Omega_n = \{0,1\}^n = \underbrace{\{0,1\} \times \cdots \times \{0,1\}}_{\text{n vezes}}$. Ou equivalente,

$$\Omega_n = \{\omega = (\omega_1, \dots, \omega_n) \text{ de modo que } \omega_i \in \{1,0\}, 1 \leq i \leq n\}.$$

- A_k: Evento de obter CARA k vezes. Obesrve que $A_k \subset \Omega_n$.

- *F*: Os conjuntos mensuráveis. No caso *F* é o conjunto das partes de Ω_n, ou seja, $F=2^{\Omega_n}$.

- *P*: A medida de probabilidade. Nesse caso dado, dado $\omega \in \Omega_n$, definimos:

$$P(\emptyset) = 0,\ P(\omega) = \frac{1}{2^n}\ , P(\Omega_n) = 1.$$

  Além disso, dado $A_k \in F$, consequentemente $A_k \in 2^{\Omega_n}$, temos

$$P(A_k) = \sum_{\omega \in A} P(\omega).$$

  Como A_k possui exatamente $\binom{n}{k}$ elementos, temos que

$$P(A_k) = \sum_{1}^{\binom{n}{k}} \frac{1}{2^n} = \frac{\binom{n}{k}}{2^n}.$$

- *X:* $\Omega_n \to \mathbb{R}$ é a variável aleatória definida por

$$X(\omega) = \sum_{i=1}^{n} \omega_i\ ,$$

  onde $\omega = (\omega_1, \dots, \omega_n)$ e *X* conta quantas vezes obtivemos CARA.

Observe que $I_m(X)=\{0,1,\dots,n\}$ e $X^{-1}(k)=A_k$, portanto a probabilidade de obtermos exatamente k CARAs em n lançamentos será

$$P(X=k) = P(A_k) = \frac{\binom{n}{k}}{2^n}.$$

Considerando determinado experimento aleatório, sendo X uma variável aleatória que pode assumir valores do conjunto $\{x_1, x_2, \cdots, x_n\}$ a probabilidade de ocorrência para cada valor de X é chamada de distribuição de probabilidade.

Se os valores de X formão um conjunto enumerável de pontos da reta, dizemos que X é uma variável aleatória discreta. Por outro lado, se X assume valores em um intervalo de números reais, X é uma variável aleatória contínua.

4.1 Definição e propriedades da Função de Distribuição

Definição 4.2 Função de Distribuição Acumulada.

Dada uma variável aleatória X, a função de distribuição acumulada de X é definida por:

$$f(x) = P(X \leq x) \in [0,1].$$

Proposição 4.1 Propriedades da Função de Distribuição

Seja f a função de distribuição de alguma variável aleatória X. Então f satisfaz as seguintes propriedades.

I. $\lim_{x\to\infty} f(X) = 1 \quad e \quad \lim_{x\to-\infty} f(X) = 0$

Demonstração: $\lim_{x\to\infty} f(X) = \lim_{x\to\infty} P(X \leq x) = P(X \leq \infty) = P(X \leq \infty) = P(\Omega) = 1.$

$\lim_{x\to-\infty} f(X) = \lim_{x\to\infty} P(X \leq x) = P(X \leq \infty) = P(X \leq -\infty) = P(\emptyset) = 0.$

II. f é tal que: $a < b \rightarrow f(A) \leq f(B)$ (f é não decrescente)

Demonstração: $a < b$, implica o evento $\{X - a\} \subset \{X - b\}$, com isto, $P(\{X - a\}) \leq P(\{X - b\})$, logo, $f(A) \leq f(B)$.

III. f é tal que: $f(b^+) = \lim_{h\to0} f(b + h) = f(b)$

Demonstração: $\lim_{h\to0} f(b + h) = f(b) = \lim_{h\to0} P(X \leq b + h)$

$= \lim_{h\to0} P(X < b \cup b \leq X \leq b + h)$

$= \lim_{h\to0} \{P(X < b) + P(b \leq X \leq b + h)\}$

$= P(X < b) + \lim_{h\to0} P(b \leq X \leq b + h)$

$= P(X < b) + P(X = b) = P(X \leq b) = f(b).$

Exemplo 4.2 Seja X uma variável aleatória e f a sua função de distribuição, com a,b $\in \mathbb{R}$ e a < b. Determine as seguintes probabilidades em termos de f.

a) P(X = a).

$P(X = a) = P(X \leq a) - P(X < a) = f(a) - \lim_{h\to0^+} P(X \leq a+h) = f(a) - \lim_{h\to0^+} f(X \leq a + h).$

b) P(a ≤X ≤ b)

$P(a \leq X \leq b) = P(X \leq b) - P(X < a) = P(X \leq b) - \big(P(X \leq a) - P(X = a)\big)$

$= f(b) - f(a) + P(X = a).$

4.2 Função Densidade de probabilidade

Definição 4.3 Uma variável aleatória contínua X com função de distribuição F é dita como absolutamente contínua quando existe uma função não negativa φ tal que

$$1 = \int_{-\infty}^{+\infty} \varphi\,(t)dt\,, \text{para todo } x \in \mathbb{R}$$

Nesse caso a função φ é denominada função densidade da variável aleatória X.

Proposição 4.2 Seja X uma variável aleatória contínua com função densidade φ e $a, b \in \mathbb{R}, a \leq b$. Então,

$$P(a \leq X \leq b) = \int_{a}^{b} \varphi\ (t)dt.$$

Demonstração:

$$\text{Quando } a = b, \text{temos } x = a, \text{logo } P(a \leq X \leq b) = \int_{a}^{a} \varphi\ (t)dt = 0.$$

$$\text{Isto posto, considerando o exemplo 4.2 temos que: } P\ (a \leq X \leq b) = f(a) - f(b).$$

$$\text{Por outro lado, } f(x) = \int_{-\infty}^{x} \varphi(t)dt\,, \text{para todo } x \in \mathbb{R}\,, \text{com isto:}$$

$$f(a) = \int_{-\infty}^{a} \varphi(t)dt \quad \text{e} \quad f(b) = \int_{-\infty}^{b} \varphi(t)dt.$$

Agora temos que;

$$P(a \leq X \leq b) = f(b) - f(a) = \int_{-\infty}^{b} \varphi(t)dt - f(b) = \int_{-\infty}^{a} \varphi(t)dt = \int_{a}^{b} \varphi(t)dt.$$

Definição 4.4 (Esperança de Variável Aleatória Contínua) Seja X uma variável aleatória contínua com função densidade de probabilidade φ. Define-se a esperança (ou média ou valor esperado) de X como:

$$E(X) = \int_{-\infty}^{+\infty} X\varphi\,(T)d_X.$$

Definição 4.5 Variância de Variável Aleatória Contínua. Seja X variável aleatória contínua tal que $E(X) \in \mathbb{R}$. A variância de X é definida como

$$\nu(X) = E\left[\left(X - E(X)\right)^2\right] = \int_{-\infty}^{+\infty} \left(X - E(X)\right)^2 \varphi\,(T)dx.$$

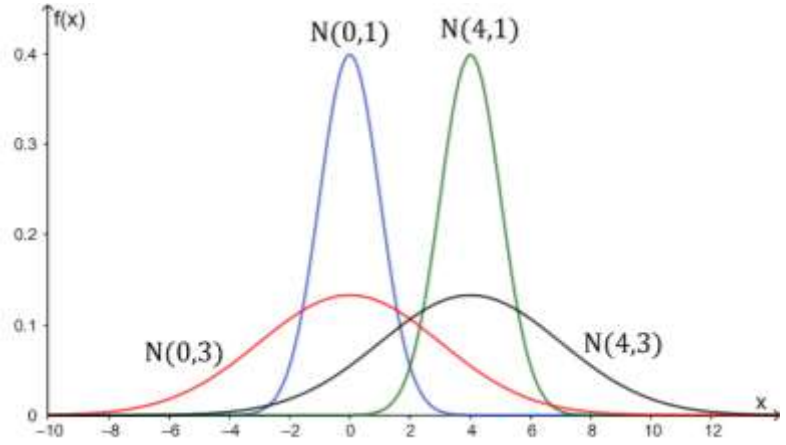
4.3 Distribuição Normal

A distribuição normal é caracterizada pela seguinte Função de Densidade de Probabilidade cujo gráfico em forma de sino (uma curva):

$$\varphi(\mathrm{x}) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{\mathrm{x}-\mu}{\sigma}\right)^2}, -\infty < x < \infty \cdot$$

A Curva normal é especificada usando dois parâmetros: a média da distribuição μ , e o desvio padrão σ, que é raiz quadrada da variância σ^2.Deste modo denota-se por N((μ, σ). O gráfico abaixo mostra alguns exemplos.

Gráfico 4 - Função densidade de Distribuição Normal com alguns N(μ, σ)



Fonte: Elaborada pelo autor.

Observações importantes:

I. A área total sob a curva é igual a 1.

II. A média μ refere-se ao centro da distribuição (existe simetria em torno de μ) já desvio padrão σ relaciona-se ao espalhamento (ou achatamento) da curva.

III. Seja Z a variável com distribuição normal com média = 0 e variância = 1, neste caso temos a distribuição normal padrão de probabilidade:

$$\varphi(\mathrm{x}) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\mathbf{z}^2}, -\infty < \mathrm{x} < \infty$$

IV. Qualquer distribuição normal com média μ e desvio padrão d pode ser transformada, para efeito de cálculo de probabilidades, na distribuição normal padrão, através seguinte mudança de variável:

$$Z = \frac{X - \mu}{\sigma}.$$

4.4 Distribuição qui-quadrado χ^2

Karl Pearson por volta de 1900, apresentou o teste qui-quadrado simbolizado por χ^2. Este teste não depende de parâmetros populacionais como média e variância, trata-se basicamente de comparar as frequências esperadas (F_E) e frequências observadas (F_O), em busca de possíveis divergências.

O teste χ^2 permite verificar se certos dados seguem determinada distribuição de probabilidade. Logo é viável utilizar esse teste para verificar se n números seguem a lei de Benford (Rousseau e Saint-Aubin,2015).

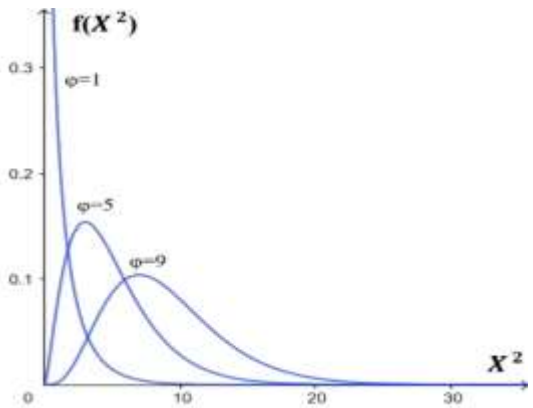
Definição 4.6. Uma variável aleatória contínua X tem distribuição qui-quadrado com φ graus de liberdade se sua função densidade for dada por:

$$f(x) = \frac{1}{2^{\frac{\varphi}{2}}\ \Gamma\left(\frac{\varphi}{2}\right)} x^{\left(-\frac{\varphi}{2}\right)-1} e^{\left(-\frac{x}{2}\right)};\ \varphi > 0, x > 0,\ \ sendo\ \Gamma(\omega) = \int_0^{\infty} x^{\omega-1} e^{x} dx,\ \omega > 0.$$

Chamemos X de χ^2_{φ} (Qui-quadrado).

Exemplo 4.3 A função densidade de probabilidade (fdp) qui-quadrado está representada graficamente para alguns valores de φ:

Gráfico 5 - Função densidade de probabilidade representada graficamente para alguns valores de φ.



Fonte: Elaborada pelo autor.

Definição 4.7 Se as variáveis aleatórias independentes x_i com i={1,2,..., φ} e $\varphi \in \mathbb{N}$ admitem; $(\mu_1, \mu_2, \dots, \mu_\varphi)$ e $(\sigma_1^2, \sigma_2^2, \dots, \sigma_\varphi^2)$ como suas distribuições normais de média, e variância respectivamente, então,

$$\mathbf{x}_\varphi^2 = \sum_{\mathbf{i=1}}^{\varphi} \left(\frac{\mathbf{x_i} - \mu_i}{\sigma_i}\right)^{\mathbf{2}} = \sum_{\mathbf{i=1}}^{\varphi} Z^2$$

segue uma distribuição qui-quadrado com φ graus de liberdade.

Considere um conjunto de dados C, graus de liberdade é a quantidade de elementos de C que podem variar após a inserção de determinadas restrições a todo $c \in C$. Por exemplo: considere a seguinte de um conjunto composto por n números inteiros: $\{x_1, \cdots, x_n\}$ de modo que $\frac{x_1, \cdots, x_n}{n} = S$ .Agora, observe que, se $n-1$ números variassem de forma independente seria possível atribuir ao x_i que não variou independentemente um valor numérico de modo que não houvesse variação em S. Ou seja, se restringíssemos a este contexto um S constante teríamos $\varphi = n - 1$ graus de liberdade (quantidade de números que podem vaiar independentemente).

Além do que é exemplificado nesse trabalho há diversas aplicabilidades e interpretações para grau de liberdade φ , tanto na matemática como na estatística, e conforme o contexto da aplicação existem procedimentos para determina-lo.

Para o cenário da LNB (verificar se há discrepância significativa entre determinado conjunto de dados e a LNB), podemos determinar $\varphi = K - 1$, onde K é a quantidade de dígitos que que podem assumir a determinada posição que está sendo analisada. Por exemplo, para conjuntos de números na base 10, há K=9 nove possibilidades para o primeiro dígito e K=10 possiblidades para o segundo dígito, portanto nesse contexto temos $\varphi = 8$ e $\varphi = 9$ respctivamente.

Exemplo 4.4 Se *X* é uma variável aleatória com distribuição normal padronizada, então, X^2 segue uma distribuição χ_φ^2 com φ=1.

Pondo X^2=A, temos que:

$$P(A \leq a) = P\left(-\sqrt{a} \leq X\sqrt{a}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{a}}^{\sqrt{a}} e^{\frac{-x^2}{2}} dx \, .$$

Como a função normal padronizada é simétrica em torno do eixo vertical 0, pois μ =0, temos que:

$$\frac{1}{\sqrt{2\pi}} \int_{-\sqrt{a}}^{\sqrt{a}} e^{\frac{-x^2}{2}} dx \; = \frac{2}{\sqrt{2\pi}} \int_{0}^{\sqrt{a}} e^{\frac{-x^2}{2}} dx \, .$$

Agora, pondo $y = \sqrt{a}$ e temos que $\frac{dy}{dx} = \frac{y^{-\frac{1}{2}}}{2}$ e que $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, deste modo

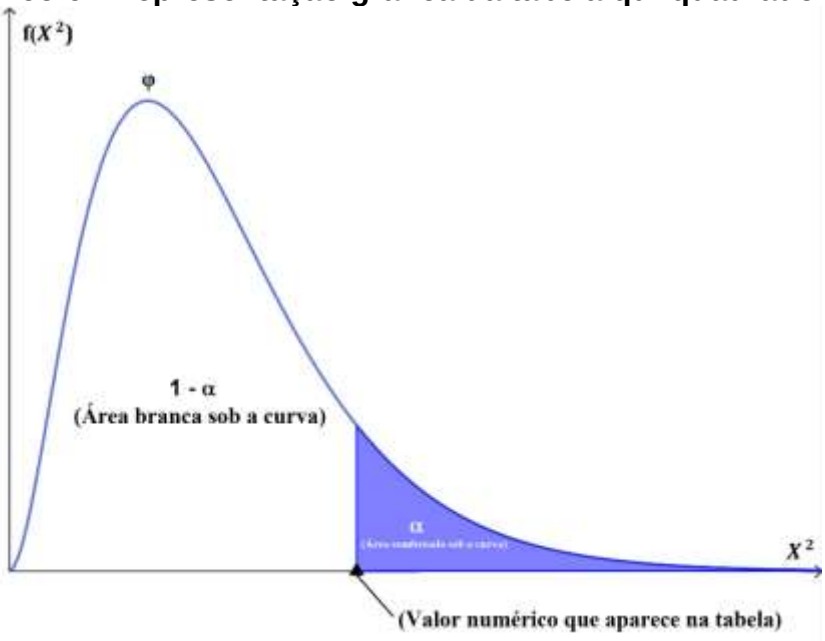
$$P(A \leq a) = \frac{1}{\sqrt{2\pi}} \int_0^a y^{-\frac{1}{2}} e^{-\frac{1}{2}} dy = \frac{1}{2^{\frac{1}{2}}\,\Gamma\left(\frac{1}{2}\right)} y^{-\frac{1}{2}} e^{-\frac{1}{2}} dt \,.$$

Portanto X^2 segue uma distribuição χ_1^2.

4.4.1 Tabela de distribuição qui-quadrado $\boldsymbol{\chi^2}$.

Apresentaremos agora a tabela qui-quadrado, nela veremos o $\boldsymbol{\chi^2_{tab}}$ (qui-quadrado tabelado) que está relacionado ao grau de liberdade φ e a significância α. Entende-se por α a área caudal a direita sob a curva conforme a gráfico 6.

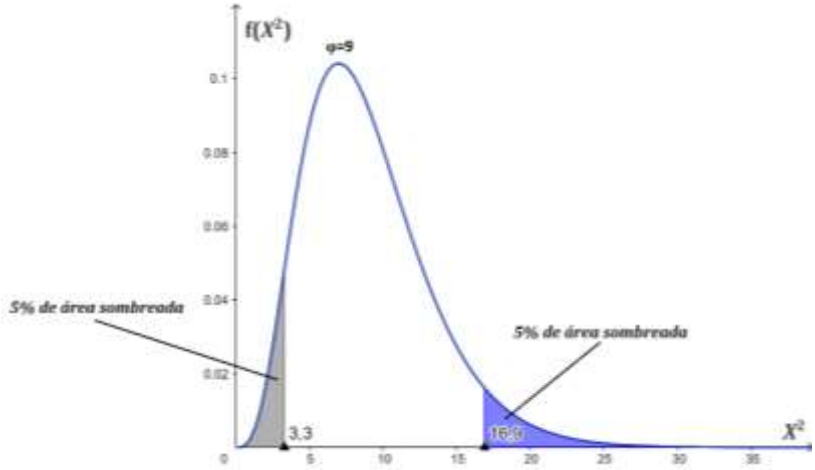
Gráfico 6 - Representação gráfica da tabela qui-quadrado.



Fonte: Elaborada pelo autor.

Exemplo 4.5 Admitindo: $\varphi = 9$ e $\alpha = 5\%$.

Gráfico 7 - Representação gráfica da Fdp qui-quadrado com $\varphi = 9$ e $\alpha = 5\%$.



Fonte: Elaborada pelo autor.

Observe que P(X ≤ x) trata-se da área na cinza à esquerda da distribuição, isto posto, temos que: P(X ≤ x) = 1 – P(X ≥ x) = 1 – α. Façamos então as seguintes considerações: o valor da abscissa à direita é chamado qui-quadrado superior ($x^2 = 16{,}9$) e está tabelado com as seguintes bases: $\varphi = 9$, $\alpha = 5\%$. Já o valor da abscissa à esquerda, chamado qui-quadrado inferior ($x^2 = 3{,}3$) está tabelado com as bases: $\varphi = 9$ e $\alpha = 1 - 0{,}05 = 0{,}95 = 95\%$.

4.4.1.1 Interpretação da Tabela de distribuição qui-quadrado (χ^2)

Tabela 6 - Distribuição qui-quadrado (χ^2) para os valores de x para $\alpha = P(X \geq x)$, com variável aleatória X

φ \ α	.....
1 ⋮ ⋮ ⋮ ⋮ ⋮	O χ^2_{tab} de uma variável aleatória X que deixa à sua direita determinar área α, para $\alpha = P(X \geq x)$, com φ graus de liberdade.

Fonte: Elaborada pelo autor.

A seguir temos uma tabela de distribuição qui-quadrado (χ^2) que será útil para verificarmos se um conjunto está em conformidade com a Lei de Benford

Tabela 7 - Valores tabelados de χ2, para alguns valores de α e φ

Grau de liberdade: φ	$\alpha = \mathbf{P}(\mathbf{X} \geq \mathbf{x})$		
	10%	**5%**	**1%**
1	2,71	3,84	6,64
2	4,60	5,99	9,21
3	6,25	7,82	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09
6	10,64	12,59	16,81
7	12,02	14,07	18,48
8	13,36	15,51	20,09
9	14,68	16,92	21,67
10	15,99	18,31	23,21

Fonte: Elaborada pelo autor.

Segue no anexo 2 uma tabela de distribuição (χ^2) com 100 graus de liberdade.

4.4.2 Procedimentos do teste χ^2: uma abordagem voltada à LNB

Seja $\mathcal{E}$ um experimento aleatório realizado n vezes dotado de (E_1,E_2,..., E_K), K eventos, de modo que para cada evento E_i haja uma expectativa de frequência já determinada, ou seja, há (F_1,F_2,...,F_K) frequências esperadas. Deste modo o do teste χ^2 serve para verificar se há discrepância entre fraqueias (f_1,f_2,...,f_K) observadas no experimento para cada evento E_i e as frequências esperadas. Em outras palavras, para cada E_i temos as frequências F_i (esperada), f_i (observada no experimento) e o teste serve para verificar f_i é estatisticamente condizente Fi. Andrade Martins (2010), no livro Estatística Geral e Aplicada, explicita os procedimentos que usaremos para a aplicação do teste χ^2.

1. Enunciar as hipóteses H_0 (não há discrepância entre as frequências observadas e esperadas) e H_1(há discrepância entre as frequências observadas e esperadas).
2. Escolha α, o nível de significância do teste (podemos determinar α=5%, este valor trata de uma probabilidade de erro). Determinar o grau de liberdade: φ = K-1. No caso da LNB, se a base é 10 então φ = 9-1=8.
3. Determinar x_{cal}^2 conforme a equação abaixo:

$$x_{cal}^2 = \sum_{i=1}^{K} \frac{(\mathrm{f_K} - \mathrm{F_K})^2}{\mathrm{F_K}}.$$

4. Verificar na tabela de destruição Qui-Quadrado, o valor tabelado conforme α e φ determinados (chamaremos esse valor de x_{tab}^2).
5. Conclusão: Se $x_{cal}^2 < x_{tab}^2$ não rejeitamos Ho (não há discrepância entre as frequências observadas e esperadas). Por outro lado, Se $x_{cal}^2 > x_{tab}^2$ rejeitamos Ho (conclui-se com, com risco α, que há discrepância entre as frequências observadas e esperadas, não há adequação).

Exemplo 4.6 Verificar a adequação dos valores da tabela 4, do exemplo 2.1, com a LNB, pondo: α=5% e observando que φ = 8, construímos a seguinte tabela.

Tabela 8 - Distribuição dos primeiros dígitos dos 1024 termos iniciais da sequência A e teste qui-quadrado.

Dígito	f_{AO}	f_{RO}	F_{AE}	F_{RE}	x_{cal}^2
1	308	30,08%	308,22	30,10%	0,00016
2	181	17,68%	180,33	17,61%	0,00252
3	127	12,40%	127,90	12,49%	0,00630
4	100	9,77%	99,23	9,69%	0,00604
5	81	7,91%	81,10	7,92%	0,00013
6	70	6,84%	68,51	6,69%	0,03260
7	57	5,57%	59,39	5,80%	0,09634
8	55	5,37%	52,33	5,11%	0,13661
9	45	4,39%	46,90	4,58%	0,07691
Totais	1024	1	1024	1	0,35760

Fonte: Elaborada pelo autor

f_{AO}: Frequência absoluta observada.
f_{RO}: Frequência relativa observada.
F_{AE}: Frequência absoluta esperada pela LNB.
F_{RE}: Frequência relativa esperada pela LNB.

Observando a distribuição qui-quadrado na tabela 7 temos que: $x_{cal}^2 < x_{tab}^2$ portanto não há discrepância entre as frequências observadas e esperadas.

5 A LEI DE NEWCOMB-BENFORD

Como já vimos na introdução deste trabalho a LNB afirma que há uma frequência aproximada para os primeiros dígitos significativos dos elementos de um conjunto de números coletados de forma aleatória, conforme as tabelas 1 e 3. Também vimos que um conjunto satisfaz a lei de Benford se a probabilidade para a incidência do primeiro dígito significativo d, com d ∈ {1, ..., 9}, for:

$$P(d) = \log_{10}(d+1) - \log_{10} d = \log_{10}\left(1+\frac{1}{d}\right).$$

Com base em $P(d)$ temos que os menores dígitos tem maiores probabilidades. E considerando $d \in \mathbb{N}, \text{e } 1 \leq d \leq 9$ observa-se que:

$$\sum_{\mathrm{d}=1}^{9} \log_{10}\left(1+\frac{1}{\mathrm{d}}\right) = \sum_{\mathrm{d}=1}^{9} \log_{10}\left(\frac{1+\mathrm{d}}{\mathrm{d}}\right) = \sum_{\mathrm{d}=1}^{9} (\log_{10}(\mathrm{d}+1) - \log_{10}\mathrm{d}) =$$

$$\log_{10} 10 - \log_{10} 1 = 1, \text{ isto posto, temos que;}$$

$$P(D \leq d) = \sum_{1 \leq \theta \leq d} P(\theta) = \sum_{1 \leq \theta \leq d} P\left(1+\frac{1}{\theta}\right) = \log\left(\prod_{1 \leq \theta \leq d}\left(1+\frac{1}{\theta}\right)\right) = \log(d+1).$$

Sendo M a mantissa, podemos determinar que de fato, a LNB expressa uma situação de probabilidade contínua, uma vez que a probabilidade para m=i trata da probabilidade de m pertencer ao intervalo [i,i+1), conforme tabela a seguir:

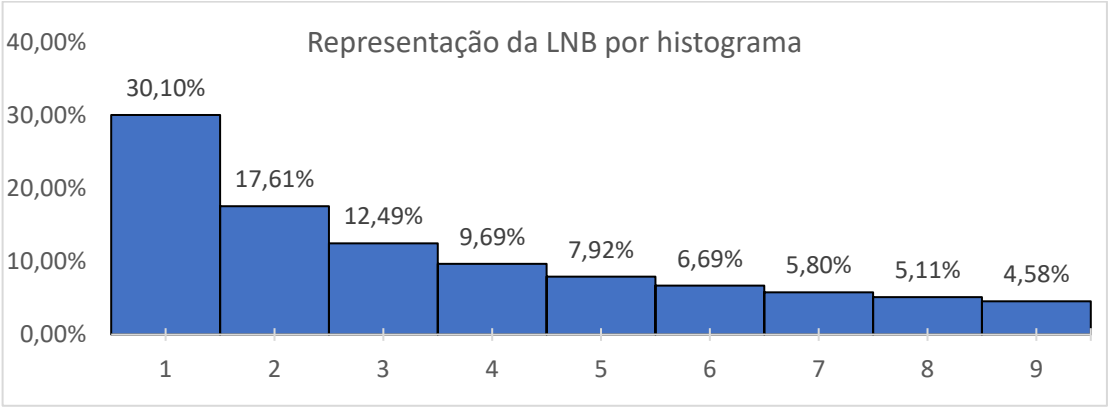
Tabela 9 - Classes e frequência relativa conforme a LNB.

Classes	Freq. Relativa
1⊢2	30,10%
2⊢3	17,61%
3⊢4	12,49%
4⊢5	9,69%
5⊢6	7,92%
6⊢7	6,69%
7⊢8	5,80%
8⊢9	5,11%
9⊢10	4,58%

Fonte: Elaborada pelo autor

Segue a seguir um histograma com a representação da tabela anterior.

Gráfico 8 - Histograma da frequência relativa conforme a LNB.

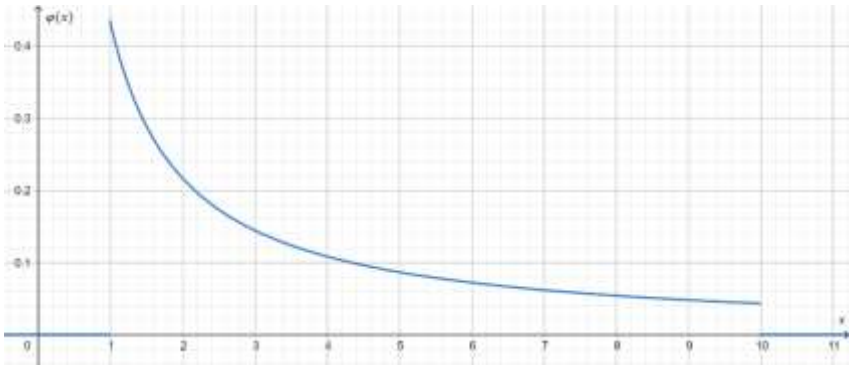


Fonte: Elaborado pelo autor.

Definição 5.1 Uma variável aleatória M com valores em [1,10) segue a lei de Newcomb-Benford se sua função densidade é dada por:

$$\varphi(x) = \begin{cases} \dfrac{1}{x \log 10}, x \in [1,10), \\ 0, \qquad\quad x \notin [1,10). \end{cases}$$

Gráfico 9 - Função densidade da lei de Newcomb-Benford.



Fonte: Elaborado pelo autor.

Como $a \leq M < b$, com base nas definições 5.1 e 4.3 temos que;

$$P(a \leq M < b) = \int_a^b \varphi(x)dx\,.$$

Agora, observe que a probabilidade;

$$P(i) = P(i \leq M < 1+i) = \int_i^{1+1} \frac{1}{x\ln 10}\,dx = \frac{1}{\ln 10}(\ln(i+1) - \ln(i))$$

$$= \frac{1}{\ln 10}\left(\ln\frac{1+i}{i}\right) = \frac{\ln\left(1+\frac{1}{i}\right)}{\ln 10} = \log_{10}\left(1+\frac{1}{i}\right).$$

Portanto a definição 5.1 de fato generaliza a LNB, deste modo é possível determinar a probabilidade da ocorrência de determinado dígito b, com $b \in \{0,1,2 \dots, 9\}$ como segundo, terceiro ou enésimo algarismo significativo de uma mantissa M. Logo, podemos , de modo discreto, verificar que a probabilidade para $b_1, b_1, \cdots, b_k$ estarem respectivamente como os primeiros k-ésimos dígitos significativos é definida por;

$$\log_{10}\left(1+\frac{1}{b_1,b_1,\cdots,b_k}\right).$$

Exemplo 5.1 A probabilidade de um número começar com os dígitos 4, 5 e 6 respectivamente é

$$\log_{10}\left(1+\frac{1}{456}\right)=\log_{10}\left(\frac{457}{456}\right)\cong 0{,}00095=0{,}095\%\,.$$

Exemplo 5.2 Qual a probabilidade para o dígito 0 ser o segundo algarismo significativo de uma mantissa M?

Como já conhecemos a generalização da LNB, para determinar tal probabilidade basta calcular o conjunto,

P=[1;1,1)∪[2;2,1)∪[3;3,1)∪ $[4;4{,}1) \cup [5;5{,}1) \cup [6;6{,}1) \cup [7;7{,}1) \cup [8;8{,}1)$ ∪[9;9,1)

portanto temos que,

$$\mathrm{P} = \log_{10}\left(1+\frac{1}{10}\right)+\log_{10}\left(1+\frac{1}{20}\right)+\cdots+\log_{10}\left(1+\frac{1}{90}\right)=0{,}11968.$$

Logo, podemos afirmar que a probabilidade para o dígito 0 ser o segundo algarismo é $0{,}11968$, pois tal fato ocorre quando a mantissa M pertence ao conjunto P.

5.1 Invariância da mudança de escala e mudança de base

O matemático estadunidense Theodore P. Hill, por volta do ano de 1995, formalizou e justificou rigorosamente a LNB. Com base no trabalho de Hill é possível afirmar que de fato a LNB é invariante na escala, sendo na verdade a única lei de probabilidade com essa propriedade, além de ser a única lei de probabilidade não trivial que é invariante na mudança de bases. Os argumentos de (Rousseau e Saint-Aubin,2015) foram usados como base para os exemplos deste capítulo,

Definição 5.2 Seja M uma mantissa, então o espaço de probabilidade sobre M é $(\mathbb{R}^+, M, \mathrm{P})$, onde $m \in M$.

Definição 5.3 Um espaço de probabilidade P em $(\mathbb{R}^+, M, \mathrm{P})$ é invariante na escala se

$$\forall s \in \mathbb{R}, \forall \mathrm{S} \in \mathrm{M}, \mathrm{P}(\mathrm{S}) = \mathrm{P}(\mathrm{Ss})$$

As seguintes propriedades mostram que a LNB é de fato invariante na mudança de bases e de escalas.

Propriedade 5.1 Um espaço de probabilidade P em $(\mathbb{R}^+, M, \mathrm{P})$ é invariante na escala se e somente se P satisfaz a LNB.

Já vimos que a LNB generalizada trata-se de uma distribuição de probabilidade atuante sobre a mantissa M , logo descrita em termos uma função densidade no intervalo $[1,10)$. Isto posto, considere um número real positivo c, uma variável aleatória X e a variável aleatória $\mathrm{Y} = c\mathrm{X}$. Deste modo, esta propriedade afirma que, se X segue LNB, então Y também segue a LNB. Em outras palavras, consideremos K um conjunto de números que segue a LNB, então se multiplicarmos cada $k \in K$ por c, o conjunto formado pelos números oriundos deste produto também seguirá a LNB.

Exemplo 5.3 Em relação ao conjunto K no contexto exposto anteriormente, considere uma mudança de escala simples obtida pelo produto de cada $k \in K$ por c=2.

Neste cenário, os números com $\lfloor m \rfloor = 1$, serão transformados em números com $\lfloor m \rfloor = 2$ ou $\lfloor m \rfloor = 3$, expressemos esta transformação da seguinte forma:

$$B(1) = B(2) + B(3).$$

De fato,

$$B(1) = B(2) + B(3) = \log_{10}\left(1 + \frac{1}{2}\right) + \log_{10}\left(1 + \frac{1}{3}\right) = \log_{10} 2$$

De modo análogo temos que:

$$B(2) = B(4) + B(5) = \log_{10}\left(1 + \frac{1}{4}\right) + \log_{10}\left(1 + \frac{1}{5}\right) = \log_{10}\frac{2}{3}$$

$$B(3) = B(6) + B(7) = \log_{10}\left(1 + \frac{1}{6}\right) + \log_{10}\left(1 + \frac{1}{7}\right) = \log_{10}\frac{4}{3}$$

$$B(4) = B(8) + B(9) = \log_{10}\left(1 + \frac{1}{8}\right) + \log_{10}\left(1 + \frac{1}{9}\right) = \log_{10}\frac{5}{4}$$

Já os números com $\lfloor m \rfloor = 1$ são gerados a partir dos números com

$$\lfloor m \rfloor = 5; \lfloor m \rfloor = 6; \lfloor m \rfloor = 7; \lfloor m \rfloor = 8; \lfloor m \rfloor = 9.$$

Observe que;

$$B(5) + B(6) + B(7) + B(8) + B(9) = B(1).$$

Exemplo 5.4 Se uma variável aleatória X segue a LNB, então a variável aleatória $Y = cX$ também segue a LNB para $c \in \left[\frac{1}{10}, 1\right)$

Como já definimos a generalizamos a LNB, incialmente, para este exemplo, iremos considerar os seguintes pontos;

1. A função de distribuição acumulada de uma variável aleatória M é definida por

$$\mathrm{f(x)=P}(M \leq \mathrm{x}). \qquad (1)$$

2. Se X obedece a LNB função de distribuição acumulada é dada por

$$\mathrm{f(x)} = \begin{cases} 0, \mathrm{x} < 1, \\ \log_{10} \mathrm{x}, \\ 1, \mathrm{x} \geq 1. \end{cases} \qquad (2)$$

Observe que podemos escrever $c = m10^n, \mathrm{com}\ m\ (\mathrm{mantissa\ de}\ c) \in [1{,}10)$, assim vemos que a mantissa de cX é a mesma de mX . Deste modo vamos mostrar que: se uma varável aleatória X segue a LNB então, a mantissa M de cX para $c \in \left[\frac{1}{10}, 10\right)$ obedece (2).

Vamos calcular P($M \leq q$) para $q \in [1{,}10]$.

M é mantissa de cX, portanto $M \in [c, 10c)$. Deste modo,

$$M = \begin{cases} cX, & \geq 1 \\ 10cX, & < 1\,. \end{cases}$$

Então,

$$\mathrm{P}(M \leq q) = \mathrm{P}(1 \leq cX \leq q) + \mathrm{P}(1 \leq 10cX \leq q)$$
$$= \mathrm{P}\left(\frac{1}{c} \leq X \leq \frac{q}{c}\right) + \mathrm{P}\left(\frac{1}{10c} \leq X \leq \frac{q}{10c}\right).$$

Observe que há duas situações a se considerar, $\frac{q}{c} > 10\ ou\ \frac{q}{c} < 10$.

1ª: $\frac{q}{c} < 10 \rightarrow \mathrm{P}\left(\frac{1}{c} \leq X \leq \frac{q}{c}\right) = \log_{10}\frac{q}{c} - \log_{10}\frac{1}{c} = \log_{10} q, e\ \mathrm{P}\left(\frac{1}{10c} \leq X \leq \frac{q}{10c}\right) = 0.$

Portanto, P($M \leq q$) $= \log_{10} q$.

2ª: $\frac{q}{c} > 10 \rightarrow \mathrm{P}\left(\frac{1}{c} \leq X \leq \frac{q}{c}\right) = \mathrm{P}\left(\frac{1}{c} \leq X \leq 10\right)\ e\ \mathrm{P}\left(\frac{1}{10c} \leq X \leq \frac{q}{10c}\right) = \mathrm{P}\left(1 \leq X \leq \frac{q}{10c}\right)$

Portanto, P($M \leq q$) $= \mathrm{P}\left(\frac{1}{c} \leq X \leq 10\right) + \mathrm{P}\left(1 \leq X \leq \frac{q}{10c}\right)$

$$= \left(\log_{10} 10 - \log_{10}\frac{1}{c}\right) + \left(\log_{10}\frac{q}{10c} - \log_{10} 1\right)$$
$$= \left(1 - \log_{10}\frac{1}{c}\right) + \left(\log_{10}\frac{1}{c} + \log_{10} q - \log_{10} 10\right)$$
$$= \log_{10} q.$$

Propriedade 5.2 Se uma medida de probabilidade P em $(\mathbb{R}^+, M, P(m))$ que segue a LNB então ela é invariante de base.

Em uma base $b, com\ b > 1$, há $b - 1$ algarismos que podem ser o primeiro dígito significativo de um número, pois neste contexto os dígitos não nulos são os elementos do conjunto $A = \{1{,}2, \ldots, b - 1\}$. Deste modo, tal propriedade afirma que a frequência do primeiro dígito significativo i em uma base b é

$$B_b(i) = \log_b\left(1 + \frac{1}{i}\right),$$

Por exemplo, para $b = 4$, temos $A = \{1,2,3\}$ e $B_4(i) = \log_4\left(1 + \frac{1}{i}\right)$, para cada $i \in A$. Indo além a propriedade também diz que se um conjunto obedece a LNB em determinada base ele continuará obedecendo a LNB mesmo se alterarmos a ase desse conjunto. No exemplo 6.1, verificaremos que de fato o conjunto da sequência de Fibonacci segue a LNB em qualquer base b.

No trabalho Benford's Law (2011) de Adrien Jamain as demonstrações destas propriedades explícitas de forma organizada.

6 EXPERIMENTOS E APLICAÇÕES COM A LNB

6.1 Experimento envolvendo a LNB, população e PIB

A seguir apresentaremos o resultado da análise do comportamento de dados reais obtidos a partir do site do IBGE (Instituto Brasileiro de Geografia e Estatística) perante a LNB, com o intuito de verificar se a distribuição observada dos dois primeiros dígitos tem assemelhas com à distribuição esperada da Lei de Newcomb-Benford.

A seguir temos as tabelas com as frequências observadas e esperadas para o primeiro e segundo dígito e resultado do teste qui-quadrado, para população de cada município brasileiro.

Tabela 10 - Análise da população dos municípios brasileiros perante a LNB

ANÁLISE DO PRIMEIRO DÍGITO PERANTE A LNB PARA POPULAÇÃO DE CADA MUNICÍPIO (Nº de habitantes).					
DÍGITO	F_{AO}	F_{RO}	F_{AE}	F_{RE}	x^2_{cal}
1	1662	29,84%	1676,57	30,1%	0,13
2	1022	18,35%	980,88	17,6%	1,72
3	722	12,96%	696,25	12,5%	0,95
4	585	10,50%	538,06	9,7%	4,09
5	449	8,06%	441,14	7,9%	0,14
6	370	6,64%	373,19	6,7%	0,03
7	334	6,00%	323,06	5,8%	0,37
8	262	4,70%	284,63	5,1%	1,80
9	164	2,94%	256,22	4,6%	33,19
Totais	5570	100,00%	5570	100,00%	42,43
ANÁLISE DO SEGUNDO DÍGITO PERANTE A LNB PARA POPULAÇÃO DE CADA MUNICÍPIO (Nº de habitantes).					
DÍGITO	F_{AO}	F_{RO}	F_{AE}	F_{RE}	x^2_{cal}
0	656	11,78%	666,6176	11,97%	0,17
1	678	12,17%	634,3673	11,39%	3,00
2	586	10,52%	606,1274	10,88%	0,67
3	565	10,14%	581,1181	10,43%	0,45
4	544	9,77%	558,7267	10,03%	0,39
5	531	9,53%	538,5076	9,67%	0,10
6	526	9,44%	520,0709	9,34%	0,07
7	505	9,07%	503,2495	9,04%	0,01
8	509	9,14%	487,7649	8,76%	0,92
9	470	8,44%	473,45	8,50%	0,03
Totais	5570	100,00%	5570	100,00%	5,63

Fonte: Elaborado pelo autor.

O teste qui-quadrado aponta que há diferença estatisticamente significativa entre as frequências conservadas e esperadas para o primeiro dígito, considerando α=5% e observando que φ = 8 conforme a tabela 7. Porem ao observar as distribuições verifica-se que de fato há semelhanças entre a frequência observada com àquela esperada pela LNB, deste modo, se analisarmos a tabela é fácil verificar que os valores para o dígito 9 são responsáveis pela discrepância apontadas pelo teste de inferência estatística. Ou seja, se tivéssemos que investigar o motivo a discrepância seria uma alternativa analisarmos inicialmente os valores apontados para o 9. Já para o segundo dígito, considerando α=5% e observando que φ = 9 conforme a tabela 7, o teste qui-quadrado não aponta discrepância entre os valores observados e valores esperado.

Vejamos agora as tabelas com as frequências observadas e esperadas para o primeiro e segundo dígito e resultado do teste qui-quadrado, para o PIB de cada município brasileiro.

Tabela 11 - Análise do PIB dos municípios brasileiros perante a LNB.

ANÁLISE DO PRIMEIRO DÍGITO PERANTE A LNB PARA PIB DE CADA MUNICÍPIO A PREÇOS CORRENTES.					
DÍGITO	F_{AO}	F_{RO}	F_{AE}	F_{RE}	x^2_{cal}
1	1615	28,99%	1676,57	30,1%	2,26
2	897	16,10%	980,88	17,6%	7,17
3	688	12,35%	696,25	12,5%	0,10
4	564	10,13%	538,06	9,7%	1,25
5	486	8,73%	441,14	7,9%	4,56
6	406	7,29%	373,19	6,7%	2,88
7	333	5,98%	323,06	5,8%	0,31
8	308	5,53%	284,63	5,1%	1,92
9	273	4,90%	256,22	4,6%	1,10
Totais	5570	100,00%	5570,00	100,00%	21,55
ANÁLISE DO SEGUNDO DÍGITO PERANTE A LNB PARA PIB DE CADA MUNICÍPIO A PREÇOS CORRENTES.					
DÍGITO	F_{AO}	F_{RO}	F_{AE}	F_{RE}	x^2_{cal}
0	670	12,03%	666,62	11,97%	0,02
1	654	11,74%	634,37	11,39%	0,61
2	596	10,70%	606,13	10,88%	0,17
3	572	10,27%	581,12	10,43%	0,14
4	546	9,80%	558,73	10,03%	0,29
5	504	9,05%	538,51	9,67%	2,21
6	530	9,52%	520,07	9,34%	0,19
7	498	8,94%	503,25	9,04%	0,05

8	484	8,69%	487,76	8,76%	0,03
9	516	9,26%	473,45	8,50%	3,82
Totais	5570	100,00%	5570,00	100,00%	7,54

Fonte: Elaborado pelo autor.

Assim como na situação anterior, o teste qui-quadrado fornece diferenças significativas entre as frequências observadas e esperadas para o primeiro dígito, considerando α=5% e observando que φ = 8 conforme a tabela 7. Porém, novamente, vemos que de fato há semelhanças entre a frequência observada com àquela esperada pela LNB, o responsável pela discrepância neste caso são os valores do dígito. Já para o segundo dígito, considerando α=5% e observando que φ = 9 conforme a tabela 7, o teste qui-quadrado não aponta discrepância entre os valores observados e valores esperado.

6.2 Algumas aplicações da LNB

6.2.1 Conjuntos que seguem a LNB

Inicialmente vamos destacar que a LNB não é aplicável a todos os conjuntos numéricos, por exemplo, quando há uma limitação clara para o primeiro dígito como no caso da altura das pessoas cujo primeiro dígito tende a ser 1 ou 2, fica evidente a não aplicabilidade da LNB.

Com base em trabalhos de (Rousseau e Saint-Aubin,2015) e Hill(1995) podemos verificar que conjuntos compostos por muitos elementos de dados numéricos construídos de forma natural e/ou oriundos das mais diversas fontes, sem que haja uma ordem predefinida para os primeiros dígitos e conjuntos contendo números com várias ordens de magnitude ($x=m10^n$, $m \in [1,10)$ e $n \in \mathbb{Z}$, n é a magnitude), tendem a seguir a LNB. Além dos casos citados, objetivamente, temos que sequencias do tipo:

$$\{a^n\}, n \geq 1, \text{seguem LNB em base b desde que o } \log_b a \text{ seja iracional.}$$

Um exemplo deste caso é a sequência de números do tipo números 2^n, que de fato segue a LNB, conforme verificamos no exemplo 4.6.

Exemplo 6.1 A Sequência de Fobonacci segue a LNB em toda as bases.
A Sequência de Fobonacci já foi apresentada neste trabalho de forma recursiva, más para este exemplo vamos usar um resultado já conhecido para os números dessa sequência.

$$F_n = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right).$$

Observe que, quanto maior for o valor de n mais a potência $\left(\frac{1-\sqrt{5}}{2}\right)^n$ se aproxima do ZERO. Deste modo, para o contexto da LNB, podemos observar que parte inteira da mantissa de F_n tem o mesmo comportamento da parte inteira da mantissa de $\frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n$. Portanto basta mostrar que $\frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n$ segue LNB, faremos isto e duas etapas.

(a) O produto por $\frac{1}{\sqrt{5}}$ trata-se de uma mudança de escala e como já vimos a LNB é invariante na mudança de escala.

(b) Considerando (a), devemos mostrar que a sequência $\left\{\left(\frac{1+\sqrt{5}}{2}\right)^n\right\}$ segue a LNB em base b. Para isto, basta provar que $\log_b \frac{1+\sqrt{5}}{2}$ é irracional para todo b inteiro. Portanto, suponhamos o contrário, ou seja,

$$\log_b \frac{1+\sqrt{5}}{2} = \frac{p}{q},\ com\ p, q\ \in \mathbb{Z}\ e\ q \neq 0.$$

Deste modo,

$$\log_b \frac{1+\sqrt{5}}{2} = \frac{p}{q} \Leftrightarrow \frac{1+\sqrt{5}}{2} = b^{\frac{p}{q}} \Leftrightarrow \left(\frac{1+\sqrt{5}}{2}\right)^q = b^p$$

Chegamos a uma contradição na última igualdade, pois os $\left(\frac{1+\sqrt{5}}{2}\right)^q$ e b^p são irracionais e racionais respectivamente.

6.2.2 Aplicação da LNB para fraudes financeiras

Há diversos trabalhos que apontam a LNB como uma ferramenta que pode detectar indícios de fraudes financeira. Negrini (1995 ,2012), aponta que uma discrepância exacerbada entre a LNB e um conjunto de dados financeiro pode sugerir indício de fraude ou dados fabricados. Contudo, é importante ressaltar que a discrepância entre os dados analisado e a LNB não prova que de fato houve fraude,

a discrepância apenas indica uma possível anormalidade uma situação distinta da esperada.

6.2.3 Análise de dados do genoma

Friar, Goldman, e Pérez-Mercader (2010) fizeram analises com dados disponíveis de mais de mil genomas, e fizeram uso da LNB para testar a seguinte observação: O número de quadros de leitura abertos1 e sua relação com o tamanho do genoma difere entre eucariontes e procariontes, sendo que o primeiro apresenta uma relação log-linear e o segundo, uma relação linear.

BROWN (2010), “Em genética, uma fase de leitura aberta (ORF, de open reading frameem inglês) é uma sequência de DNA entre um códon de início (normalmente uma metionina, ATG) e um códon de parada (TAA, TAG ou TGA)”.

6.2.4 Dados Macroeconômicos

De acordo com CUNHA (2013) os dados macroeconômicos reportados ao Gabinete de Estatísticas da União Europeia - Eurostat pelos 27 países membros da UE, foram analisados perante a LNB. O país que teve o maior desvio foi a Grécia, posteriormente a suspeita de manipulação dos dados deste país foi oficialmente confirmada pela Comissão Europeia. Tal fato evidenciou a efetividade da Lei de Benford na detecção de supostas irregularidades e manipulações de dados macroeconômicos.

6.2.5 Conjuntos que não obedecem a Lei de Newcomb-Benford

Como já destacamos há diversos conjuntos que não seguem a LNB portanto exemplificaremos alguns casos nesta seção.

1. Conjuntos numéricos que tendem a obedecer a valores definidos para máximo e mínimo não seguem a LNB, pois nestes casos a frequência dos primeiros dígitos seguem os valores indicados para máximos e mínimo. Por exemplo, a altura das pessoas adultas em metros que geralmente é maior 1m e menor que 2m.

2. Resultado de sorteios numéricos (loterias, bingos e etc) não seguem a LNB. Ao sortear um número de um conjunto M={m_1 , m_2 ,..., m_j} com j $\in \mathbb{N}$, considerando que cada m_i tem a mesma probabilidade de ser sorteado, a probabilidade P(d) do número sorteado ter o algarismo d como primeiro dígito significativo é dada pela razão entre a quantidade de números pertencentes a M, cujo primeiro dígito significativo é o algarismo *d* e a quantidade de elementos do conjunto M. Portanto, o cálculo das probabilidades para os resultados de sorteios não segue a LNB, a não ser que haja alguma manipulação das probabilidades, por exemplo, se o percentual de elementos do conjunto M cujo primeiro dígito significativo é d for semelhante ou idêntico ao apresentado pela LNB.
3. Conjuntos usados para identificar ou catalogar objetos não seguem a LNB. Para catalogar ou identificar numericamente objetos ou elementos quaisquer se faz necessário impor determinada ordem numérica, por exemplo, o CPF (cadastro de pessoa física) é composto por números fixos conforme a região onde o cadastrado nasceu e por números aleatórios (portanto uma espécie de sorteio), logo não segue a LNB.

6.3 Análise de casos de COVID-19 perante a Lei de Newcomb-Benford.

Nesta etapa do trabalho apresentaremos uma análise perante a LNB dos casos confirmados e notificados de COVID-19 (doença causadora de pandemia em 2020) em cada município dos estados do Ceará e São Paulo. Sobre a COVID-19 “é uma doença causada pelo coronavírus, denominado SARS-CoV-2, que apresenta um espectro clínico variando de infecções assintomáticas a quadros graves.” (Ministério da Saúde,2019).

> “COVID-19 é a doença infecciosa causada pelo novo coronavírus, identificado pela primeira vez em dezembro de 2019, em Wuhan, na China. A Organização Mundial da Saúde (OMS) declarou, em 30 de janeiro de 2020, que o surto da doença causada pelo novo coronavírus (COVID-19) constitui uma Emergência de Saúde Pública de Importância Internacional – o mais alto nível de alerta da Organização, conforme previsto no Regulamento Sanitário Internacional. Em 11 de março de 2020, a COVID-19 foi caracterizada pela OMS como uma pandemia.” (Escritório Regional para as Américas da Organização Mundial da Saúde © Organização Pan-Americana da Saúde,2020).

As fontes dos dados analisados dos respectivos estados são os sites: Integra SUS – Transparecia da Saúde do Estado do Ceará; e SP CONTRA O NOVO CORONAVÍRUS. Estes são sites oficiais, ligados aos respectivos governos e apresentam diariamente o número de casos confirmados da doença além outros indicadores. Trabalhamos com quantidades de casos confirmados, atualizados até o dia 23 de agosto 2020. Ao todo analisou-se 957.647 distribuídos entre os 829 municípios dos estados citados.

Para a análise verificou-se a quantidade de casos, confirmados e publicados, em cada município até a data de citada. Em seguida as foram tabeladas as frequências relativas e absolutas de cada dígito como primeiro e segundo algarismo significativo das quantidades. As frequências observadas forma comparadas com as frequências propostas pela LNB e o teste qui-quadrado foi usado para verificar se há discrepância estatisticamente significativa entre os resultados observados e esperados pela LNB. Conforme a tabelas a seguir.

Tabela 12 - Análise do 1º dígito perante a LNB de casos confirmados de COVID-19 por município do Ceará e São Paulo (23/08/2020).

Dígito	Fao	Fro	Fae	Fre	x^2_{cal}
1	240	28,9%	30,1%	249,83	0,39
2	150	18,1%	17,6%	146,163	0,10
3	109	13,1%	12,5%	103,75	0,27
4	88	10,6%	9,7%	80,178	0,76
5	62	7,5%	7,9%	65,736	0,21
6	56	6,7%	6,7%	55,61	0,00
7	49	5,9%	5,8%	48,14	0,02
8	38	4,6%	5,1%	42,413	0,46
9	37	4,5%	4,6%	38,18	0,04
Somas	829	100%	100%	829	2,24

Fonte: Elaborado pelo autor.

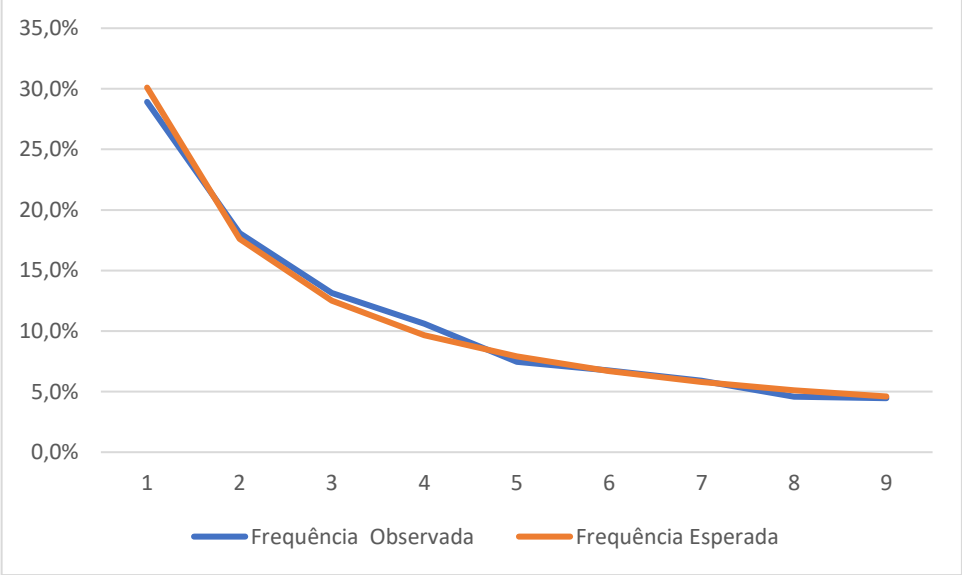
Tabela 13 - Análise do 2º dígito perante a LNB de casos confirmados de COVID-19 por município do Ceará e São Paulo (23/08/2020)

Dígito	Fao	Fro	Fae	Fre	x^2_{cal}
0	96	11,6%	99,33	11,97%	0,11
1	94	11,3%	94,53	11,39%	0,00
2	101	12,2%	90,32	10,88%	1,26
3	98	11,8%	86,59	10,43%	1,50
4	94	11,3%	83,26	10,03%	1,39
5	82	9,9%	80,24	9,67%	0,04
6	69	8,3%	77,50	9,34%	0,93
7	75	9,0%	74,99	9,04%	0,00
8	60	7,2%	72,68	8,76%	2,21
9	60	7,2%	70,55	8,50%	1,58
Somas	829	100%	829	100%	9,03

Fonte: Elaborado pelo autor.

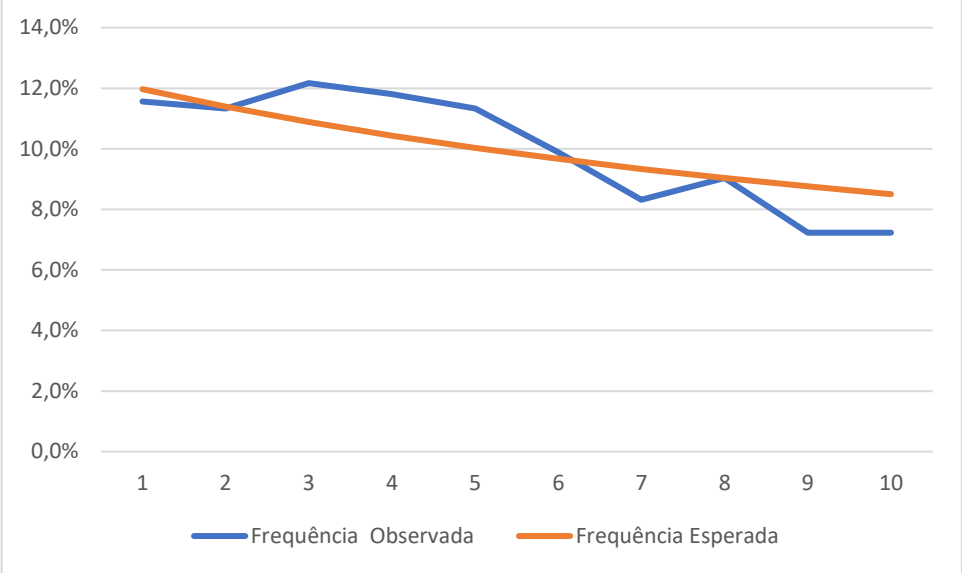
Portanto, considerando: α=5% e observando que φ = 8 e φ = 9, conforme a tabela 7, para as análises dos primeiros e segundos dígitos respectivamente, temos que o teste qui-quadrado aponta que não há, em ambos os casos tabelados, diferença estatisticamente significativa entre as frequências observadas e esperadas. A seguir esse resultado é corroborado graficamente.

Gráfico 10 - Análise do 1º dígito perante a LNB de casos confirmados de COVID-19 conforme tabela 12



Fonte: Elaborado pelo autor.

Gráfico 11 - Análise do 2º dígito perante a LNB de casos confirmados de COVID-19 conforme tabela 13



Fonte: Elaborado pelo autor.

De fato, os gráficos, assim como o teste qui-quadrado, apontam que há semelhanças entre as frequências observadas e esperadas. Portanto podemos concluir que os dados analisados seguem a LNB.

Tal análise foi realizada novamente, usando quantidades de casos confirmados, atualizados até o dia 1 de novembro de 2020. Neste cenário analisou-se mais de 1,3 milhões de casos distribuídos entre os 829 municípios dos estados citados, e verificou-se que os dados continuavam a não apresentar discrepância estatisticamente significativa perante a LNB. A seguir apresentamos tabelas e gráficos da distribuição dos dígitos e o teste qui.quadrado (manteve-se α=5%) referente a análise.

Tabela 14 - Análise do 1º dígito perante a LNB de casos confirmados de COVID-19 por município do Ceará e São Paulo (01/11/2020).

Dígito	Fao	Fro	Fae	Fre	x^2_{cal}
1	235	28,28%	250,16	30,10%	0,92
2	137	16,49%	146,33	17,61%	0,60
3	108	13,00%	103,82	12,49%	0,17
4	66	7,94%	80,53	9,69%	2,62
5	76	9,15%	65,80	7,92%	1,58

6	58	6,98%	55,63	6,69%	0,10
7	64	7,70%	48,19	5,80%	5,19
8	44	5,29%	42,51	5,12%	0,05
9	43	5,17%	38,02	4,58%	0,65
Somas	831	100%	831,00	100%	11,88

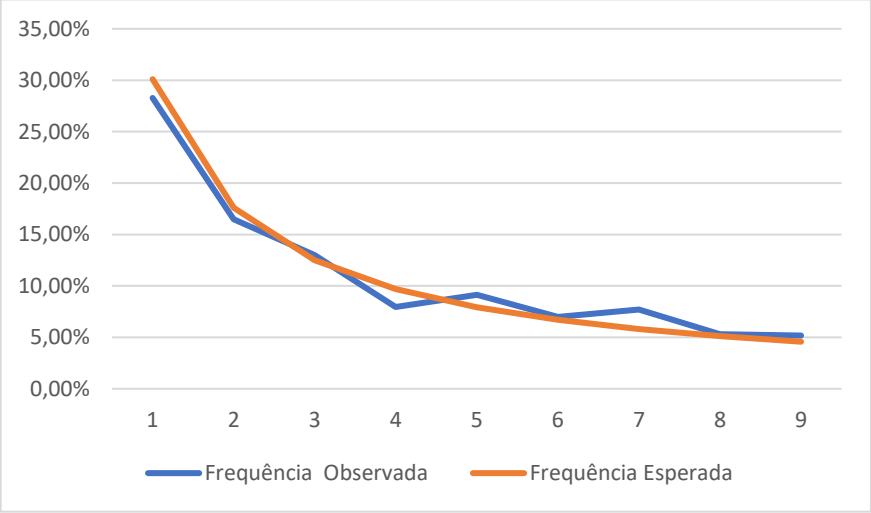
Fonte: Elaborado pelo autor.

Tabela 15 - Análise do 2º dígito perante a LNB de casos confirmados de COVID-19 por município do Ceará e São Paulo (01/11/2020).

0	104	12,52%	99,45	11,97%	0,21
1	92	11,07%	94,64	11,39%	0,07
2	85	10,23%	90,43	10,88%	0,33
3	82	9,87%	86,70	10,43%	0,25
4	84	10,11%	83,36	10,03%	0,00
5	84	10,11%	80,34	9,67%	0,17
6	87	10,47%	77,59	9,34%	1,14
7	70	8,42%	75,08	9,04%	0,34
8	62	7,46%	72,77	8,76%	1,59
9	81	9,75%	70,64	8,50%	1,52

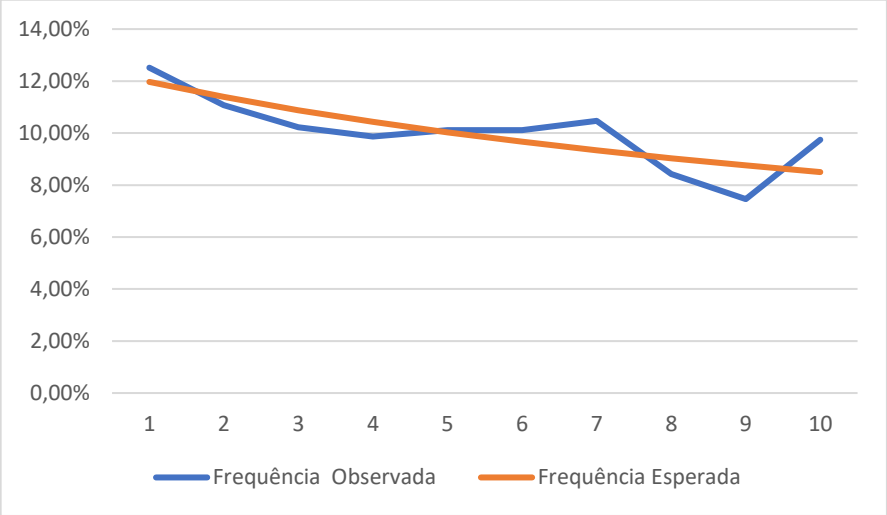
Fonte: Elaborado pelo autor.

Gráfico 12 - Análise do 1º dígito perante a LNB de casos confirmados de COVID-19 conforme tabela 14



Fonte: Elaborado pelo autor.

Gráfico 13 - Análise do 2º dígito perante a LNB de casos confirmados de COVID-19 conforme tabela 15



Fonte: Elaborado pelo autor.

6.4 Lei de Newcomb-Benford, uma abordagem para o ensino médio

Nessa seção apresentaremos uma proposta didática para aplicação da LNB ao ensino médio. Pois a LNB engloba diversos assuntos presente no currículo da matemática da educação básica, como aspectos da estatística e da probabilidade por exemplo.

Previamente acreditou-se que a Lei de Newcomb-Benford, através do trabalho pedagógico, conseguiria fornecer significados práticos a conteúdos tradicionais do currículo. Portanto esta aplicação, teve duas bases bem acentuadas e interligadas; revisar e praticar diversos conteúdos afins e estimular nos alunos curiosidade sobre os saberes matemáticos.

Instigar a curiosidade é uma grade conquista, pois quando o aluno sente interesse por algo ele tende a questionar e pesquisar sobre assunto e acaba adquirindo conhecimento de forma autônoma. Ou seja, o sucesso de uma aula cujo objetivo é instigar curiosidades pode agregar mais atenção e vontade por parte dos alunos nas aulas seguintes.

> O exercício da curiosidade convoca a imaginação a intuição, as emoções, a capacidade de conjecturar, de comparar, na busca da perfilização do objeto ou do objeto ou achado de sua razão de ser. Um ruído, por exemplo, pode provocar minha curiosidade. Observo o espaço onde parece que se está verificando. Aguço o ouvido. Procuro comparar com outo ruído cuja razão de ser já conheço. Investigo o espaço. Admito hipóteses várias em torno da

> possível origem do ruído. Elimino algumas até que chegou a sua explicação.(FREIRE, 2006, p. 88)

A turma escolhida para a aplicação do conteúdo foi, a turma que do 2º ano do ensino médio, juntamente com o curso de financias na Escola Estadual de Educação Profissional José Vidal Alves, situada no município de Canindé-CE.

6.4.1 A proposta didática

6.4.1.1 Objetivos

a. Objetivos gerais.

A abordagem da LNB usou como base conteudista assuntos vinculados a Estatística e Probabilidade do ensino médio, e teve como objetivo geral; apresentar, definir e generalizar a lei LNB, mostrar o teste qui-quadrado e sua aplicação no contexto da LNB e, consequentemente, apresentar aos alunos de forma teórica e com experimentos práticos uma aplicação direta dos conteúdos estudados em sala de aula a fim de aproximá-los de procedimentos matemáticos, fazendo-os perceber que existe significado real para os conteúdos estudados e instigá-los a buscar essa significância matemática de forma mais autônoma estando dentro ou fora do ambiente escolar.

b. Objetivos específicos.

Alinhado com a BNCC (Base nacional curricular comum) que está em vigor, são proposto os seguintes objetivos específicos:

1. Analisar tabelas, gráficos e amostras de pesquisas estatísticas apresentadas em relatórios divulgados por diferentes meios de comunicação e procurar identificar padrões.
2. Planejar e executar pesquisa amostral sobre questões relevantes, usando dados coletados diretamente ou em diferentes fontes.
3. Construir e interpretar tabelas e gráficos de frequências com base em dados obtidos em pesquisas por amostras estatísticas, incluindo o uso dos softwares EXCEL, LibreOffice Calc ou semelhantes.

6.4.2 Recursos didáticos e processos metodológicos

Tanto professor como alunos fizeram uso de computadores com softwares como EXCEL, LibreOffice Calc ou semelhantes, para melhor exposição do conteúdo o professor fez uso de retroprojetor.

Antes de começar as aulas sobre o assunto o professor coletou dados numéricos: a quantidade de habitantes e PIB de cada município brasileiro (com base no site o IBGE) e uma lista contendo números do tipo a^n com a,n naturais e outras sequências matemáticas como a de Fibonacci (com uso dos softwares citados é possível fazer essas listas com dezenas de centenas de termos no momento da aula). Estes dados foram usados como objeto pedagógico.

A abordagem inicial foi feita através de indagações e situação problema com intuito de explicar a proposta probabilística feita pela LNB (semelhante ao que vimos nos parágrafos iniciais da introdução deste trabalho), seguida de comentários sobre fatores históricos relacionados a descoberta desta lei. Após essas etapas, houve a definição e generalização da LNB. Como a abordagem foi com alunos do ensino médio a argumentação matemática esteve nos termos da educação básica.

Após exposição, clara e objetiva, os alunos usaram os softwares citados para construir tabelas de frequência para o primeiro dígito de conjuntos previamente determinados pelo professor (a turma trabalhou com diversos conjuntos). Com base no capítulo 2 o professor exemplificou o processo de construção das tabelas.

Após a construção da tabela os alunos puderam comparar os resultados obtidos com aqueles propostos pela LNB e aplicaram do teste qui-quadrado.

Para finalizar a exposição sobre assunto, discutiu-se sobre conjuntos que não se adequam a LNB e sobre a aplicações realizadas com a LNB. No anexo é mostrado o planejamento das aulas que foram ministradas sobre esse tema voltadas a pesquisa apresentada neste trabalho.

6.4.3 Resultados obtidos

Ao final das aulas os alunos responderam um questionário sobre o assunto. Nas respostas foi unânime que o conteúdo é muito interessante e que despertou um olhar diferente para o significado daquilo que é estudado em sala de aula.

Os alunos também fizeram muitas indagações sobre as aplicações de outros conteúdos, como contagem e geometria por exemplo, eles realmente ficaram curiosos pra saber onde outros conteúdos podem ser aplicados.

Portanto avalia-se de forma positiva as aulas com o assunto LNB pois, além revisar e intensificar conteúdos convencionais ligados estatística e probabilidade, desperta nos alunos a curiosidade sobre a significância por trás dos conteúdos estudados nas aulas de matemática.

7 CONCLUSÃO

A Lei de Newcomb-Benford, afirma que em um conjunto de números coletados aleatoriamente, em grande quantidade, os primeiros dígitos de cada número não têm distribuição uniforme, aliás é bem distante da uniformidade que possa ser esperada. A lei afirma que a distribuição para os primeiros dígitos ocorre conforme a formulação seguinte:

$$\boldsymbol{P(d) = \log_{10}\left(1 + \frac{1}{d}\right)}.$$

Destacamos o êxito dessa pesquisa visto que todos os objetivos foram alcançados. Uma vez que, usando conceitos matemáticos e estatísticos no âmbito discreto e contínuo foi possível conceituar e generalizar e realizar experimentos com a LNB. À vista disso, o estudo mostra que a LNB é fascinante e tem aspectos únicos, como o fato de ser a única lei de probabilidade invariante na mudança de escala, a única lei de probabilidade não trivial invariante na mudança de bases. Indo além, o estudo apresentou sequências numéricas que seguem a LNB, apontou que é aceitável afirmar que conjuntos com muitos elementos, estes oriundos de diversas fontes, e ordens de magnitude, gerados de forma natural tendem a seguir a LNB e também se observou que a referida lei pode ser usada para detectar indícios de fraudes, ou seja dados adulterados.

Por outro lado, também foi possível verificar que conjuntos numéricos que tendem a obedecer a valores definidos para máximo e mínimo, resultado de sorteios numéricos (loterias, bingos e etc) e conjunto usados para identificar ou catalogar objetos não seguem a LNB.

Além da pesquisa bibliográfica, foram realizados experimentos com os dados reais do PIB (produto interno bruto) e população de cada município brasileiro, os resultados destes experimentos apontaram que de fato há muitas semelhanças entre a frequências observadas e esperadas. Em ambos os casos, o segundo dígito não apresentou discrepância significativa com aquilo que é proposto pela LNB. Já para primeiro dígito, o estudo apontou uma diferença estatisticamente significativa entre as frequências. Porem conforme aponta a LNB, os menores dígitos sempre ocorreram mais vezes, mesmo se comparados dois a dois, e a discrepância foi ocasionada por apenas um dígito em cada conjunto analisado. Também houve um experimento com os casos de COVID-19 em todos os municípios de dois estados

brasileiros (Ceará e São Paulo) o resultado deste experimento não apresentou discrepância significativa para o primeiro e segundo dígito perante a LNB.

Contudo, há um destaque especial para o experimento pedagógico realizados com a LNB, pois ela mostrou ser uma boa ferramenta para trabalhar conteúdos tradicionais do currículo como; porcentagem, logaritmo, tabelas de frequência, interpretação e construção gráfica, probabilidade, e outros, além disso o experimento em sala de aula mostrou outra consequência muito positiva, verificou-se que as atividades realizadas em sala conseguem cativar os alunos e provocar curiosidade não apenas pela LNB, más sobre a matemática de modo geral. Consequentemente o debate sobre a aplicabilidade da LNB se estende até as aplicabilidades de outras áreas da matemática. Deste modo, através da LNB é possível trabalhar diversos conteúdos e cativar os alunos para matemática.

REFERÊNCIAS

BARBETTA, P. A. **Estatística aplicada às Ciências Sociais**. 7. ed. Florianópolis: UFSC, 2010.

BRASIL. Fundação SAEDE. **SP contra o novo coronavírus.** Página inicial. Disponível em: https://www.seade.gov.br/coronavirus/. Acesso em: 24 ago. 2020.

BRASIL. Ministério da Educação. **Base nacional curricular comum:** educação é base. Disponível em: http://basenacionalcomum.mec.gov.br/images/BNCC_EI_EF_110518_versaofinal_site.pdf.Acesso em: 11 dez. 2019.

BRASIL. Ministério da Saúde. **Coronavírus.COVID-19**. Disponível em: https://coronavirus.saude.gov.br/sobre-a-doenca#o-que-e-covid. Acesso em: 24 ago. 2020.

BROWN, Terry.**Gene cloning and DNA analysis:** an introduction. [S.l.]: John Wiley & Sons, 2010. Disponível em :https://bmm0586.fandom.com/pt br/wiki/Fase_de_Leitura_Aberta. Acesso: 02 nov. 2020.

CABRAL, Marco A.P. **Introdução à Teoria da Medida e Integral de Lebesgue**. Rio de Janeiro: Universidade Federal do Rio de Janeiro Rio de Janeiro, 2016.

CEARÁ. Secretaria da Saúde do Estado. **Integra SUS:** Transparecia da Saúde do Estado do Ceará. Disponível em: https://integrasus.saude.ce.gov.br/. Acesso em: 24 ago. 2020.

CUNHA, F. **Aplicações da Lei NEWCOMB-BENFORD à auditoria de obras públicas.** 2013. 121f. Dissertação (Mestrado profissional em Regulação e Gestão de Negócios) - Instituto de Ciências Humanas, Universidade de Brasília, Brasília, 2013.

FREIRE, Paulo. **Pedagogia da esperança:** um reencontro com a Pedagogia do oprimido. São Paulo: Paz e Terra, 2006.

FREIRE, Paulo. **Pedagogia do oprimido.** 17. ed. Rio de Janeiro: Paz e Terra, 1987.

FRIAR, J. L.; GOLDMAN, T.; PÉREZ–MERCADER, J. Genome Sizes and theBenford Distribution. **PLoS ONE,** v. 7, n. 5, 2012. Disponível em: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0036624&type=printable.Acesso em: 02 nov. 2020.

HILL, Theodore P. A Statistical Derivation of the Significant-Digit Law. **Statistical Science,** v. 10, n. 4, p. 354-363,1995.

JAMAIN, Adrien. **Benford`s Law**. London: Imperial College, 2001.

MARTINS, G. Andrade. **Estatística Geral e Aplicada**. São Paulo: Atlas, 2010.

MORGADO, A. César; carvalho, P. C. Pinto. **Matemática Discreta**. 2. ed. Rio de Janeiro: [s.n.], 2015. (Coleção PROFMAT).

NEGRINI, Mark; MITTERMAIER, Linda J. The use of Benford's Law as an aid in analytical procedures. **Auditing.,** v.16, 2017.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE. **Escritório Regional para as Américas da Organização Mundial da Saúde**. Disponível em: https://www.paho.org/pt. Acesso em: 24 ago. 2020.

PORTAL ACTION. **Estatcamp - Consultoria Estatística e Qualidade.** Disponível em: http://www.portalaction.com.br/probabilidades/63-distribuicao-qui-quadrado. Acesso em: 20 jun. 2020.

ROUSSEAU, Christiane; SAINT-AUBIN, Yvan. **Matemática e Atualidade.** Rio de Janeiro: ISBM, 2015.

TRIBUNAL DE CONTAS DA UNIÃO. **Portal TCU.** Disponível em: https://portal.tcu.gov.br/colab-i/noticias/aplicacoes-da-lei-de-benford-a-auditoria-de-obras-publicas.htm. Acesso em: 15 nov. 2019.

ANEXOS

ANEXO A- PLANEJAMENTO – LEI DE NEWCOMB-BENFORD (LNB)

Quantidade de aulas: 4 de 50 minutos.

Objetivo geral da proposta.

A abordagem da LNB usará como base conteúdos de Estatística e Probabilidade que são estudados no ensino médio. Logo, citamos que o objetivo geral é apresentar, definir e generalizar a lei LNB, mostrar o teste qui-quadrado e sua aplicação no contexto da LNB e consequentemente apresentar aos alunos de forma teórica e com experimentos práticos uma aplicação direta dos conteúdos estudados em sala de aula e com isto aproximar o educando de procedimentos matemáticos, fazendo-o perceber que existe significado real para cada conteúdo de estudado e instigá-lo a buscar essa significância matemática de forma mais autônoma estando dentro ou fora do ambiente escolar.

Objetivos específicos.

Com base na BNCC (Base nacional curricular comum) que está vigor são propostos os seguintes objetivos específicos:

1. Analisar tabelas, gráficos e amostras de pesquisas estatísticas apresentadas em relatórios divulgados por diferentes meios de comunicação e procurar identificar padrões.
2. Planejar e executar pesquisa amostral sobre questões relevantes, usando dados coletados diretamente em diferentes fontes.
3. Construir e interpretar tabelas e gráficos de frequências com base em dados obtidos

em pesquisas por amostras estatísticas, incluindo ou não uso dos softwares EXCEL, LibreOffice Calc ou semelhantes.

Sequência didática da 1ª e 2ª aula.

Objetivos:

4. Apresentar e definir e generalizar a LNB.
5. Mostrar o teste qui.quadrado e sua aplicação no contexto da LNB.

6. Analisar tabelas, gráficos e amostras de pesquisas estatísticas apresentadas em relatórios divulgados por diferentes meios de comunicação, identificar e analisar padrões.

Detalhamento da aula/abordagem didática:

1. Aula expositiva com incitação ao debate em torno do tema proposto e a valorização da interação entre os presentes na aula.
2. Com auxílio do retroprojetor faz-se uso dos exercícios, exemplos, fatores históricos e considerações postas na introdução desse trabalho acorrerá a apresentação e definição da LNB. Em seguida, com base no tópico 4.4 em especial o exemplo 4.5, será abordado o teste qui.quadrado no contexto da LNB.

   Observações:

   Para os fins dessa aula usaremos a seguinte definição: A Lei de Benford define que a probabilidade do primeiro dígito de um número segue a tabela 1.

   É importante deixar claro que existem definições mais formais.
3. Com auxílio do retroprojetor, o professor explica sobre generalização da LNB, faz isso usando a tabela 9 o contexto em torno desta e os exemplos 5.1 e 5.2.

Recursos didáticos:

Lousa, pincéis, computador, retroprojetor e softwares EXCEL, LibreOffice Calc ou semelhantes.

Sequência didática da 3ª e 4ª aula.

Objetivos:

1. Apresentar aos alunos aplicações dos conteúdos estudados.
2. Planejar e executar pesquisa amostral sobre questões relevantes, usando dados coletados diretamente em diferentes fontes.
3. Construir e interpretar tabelas e gráficos de frequências com base em dados obtidos em pesquisas por amostras estatísticas, incluindo ou não uso dos softwares EXCEL, LibreOffice Calc ou semelhantes.

Detalhamento da aula/abordagem didática:

1. Aula expositiva com incitação ao debate em torno do tema proposto, da aplicabilidade da matemática e aplicação de exercício que colocam em prática o conteúdo estudado.
2. Com os dados previamente coletados pelo professor, os alunos fazem experimentos com a LNB conforme o exposto no capítulo 6. Para essa atividade os alunos necessitam de computador com softwares com EXCEL, LibreOffice Calc ou semelhantes

   Observações:

   A aula pode ser realizada em laboratório de informática, de forma individual, em duplas ou equipes com 3 ou 4 integrantes a depender da quantidade maquinas disponíveis.

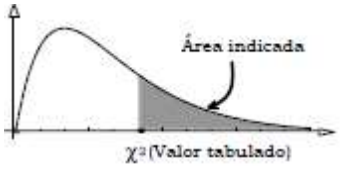
   É importante que os dados utilizados no experimento já estejam nos computadores antes de começar a aula.
3. Após os experimentos, o professor direciona o debate aos tópicos 6.2.1, 6.2.2 e 6.2.3.

Recursos didáticos:

Lousa, pincéis, retroprojetor, pen drive, e computador com softwares EXCEL, LibreOffice Calc ou semelhantes (para os alunos e para o professor).

ANEXO B - TABELA DE DISTRIBUIÇÃO QUI-QUADRADO COM gl GRAU DE LIBERDADE.

TABELA DE DISTRIBUIÇÃO QUI-QUADRADO COM gl GRAU DE LIBERDADE.



	Área na cauda superior								
gl	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	1,32	2,71	3,84	5,02	6,63	7,88	9,14	10,83	12,12
2	2,77	4,61	5,99	7,38	9,21	10,60	11,98	13,82	15,20
3	4,11	6,25	7,81	9,35	11,34	12,84	14,32	16,27	17,73
4	5,39	7,78	9,49	11,14	13,28	14,86	16,42	18,47	20,00
5	6,63	9,24	11,07	12,83	15,09	16,75	18,39	20,51	22,11
6	7,84	10,64	12,59	14,45	16,81	18,55	20,25	22,46	24,10
7	9,04	12,02	14,07	16,01	18,48	20,28	22,04	24,32	26,02
8	10,22	13,36	15,51	17,53	20,09	21,95	23,77	26,12	27,87
9	11,39	14,68	16,92	19,02	21,67	23,59	25,46	27,88	29,67
10	12,55	15,99	18,31	20,48	23,21	25,19	27,11	29,59	31,42
11	13,70	17,28	19,68	21,92	24,73	26,76	28,73	31,26	33,14
12	14,85	18,55	21,03	23,34	26,22	28,30	30,32	32,91	34,82
13	15,98	19,81	22,36	24,74	27,69	29,82	31,88	34,53	36,48
14	17,12	21,06	23,68	26,12	29,14	31,32	33,43	36,12	38,11
15	18,25	22,31	25,00	27,49	30,58	32,80	34,95	37,70	39,72
16	19,37	23,54	26,30	28,85	32,00	34,27	36,46	39,25	41,31
17	20,49	24,77	27,59	30,19	33,41	35,72	37,95	40,79	42,88
18	21,60	25,99	28,87	31,53	34,81	37,16	39,42	42,31	44,43
19	22,72	27,20	30,14	32,85	36,19	38,58	40,88	43,82	45,97
20	23,83	28,41	31,41	34,17	37,57	40,00	42,34	45,31	47,50
21	24,93	29,62	32,67	35,48	38,93	41,40	43,77	46,80	49,01
22	26,04	30,81	33,92	36,78	40,29	42,80	45,20	48,27	50,51
23	27,14	32,01	35,17	38,08	41,64	44,18	46,62	49,73	52,00
24	28,24	33,20	36,42	39,36	42,98	45,56	48,03	51,18	53,48
25	29,34	34,38	37,65	40,65	44,31	46,93	49,44	52,62	54,95
26	30,43	35,56	38,89	41,92	45,64	48,29	50,83	54,05	56,41
27	31,53	36,74	40,11	43,19	46,96	49,65	52,22	55,48	57,86
28	32,62	37,92	41,34	44,46	48,28	50,99	53,59	56,89	59,30
29	33,71	39,09	42,56	45,72	49,59	52,34	54,97	58,30	60,73
30	34,80	40,26	43,77	46,98	50,89	53,67	56,33	59,70	62,16
35	40,22	46,06	49,80	53,20	57,34	60,27	63,08	66,62	69,20
40	45,62	51,81	55,76	59,34	63,69	66,77	69,70	73,40	76,10
45	50,98	57,51	61,66	65,41	69,96	73,17	76,22	80,08	82,87
50	56,33	63,17	67,50	71,42	76,15	79,49	82,66	86,66	89,56
100	109,1	118,5	124,3	129,6	135,8	140,2	144,3	149,4	153,2

Nota: A coluna em destaque é a mais usada.

Fonte: BARBETTA, P. A. - Estatística aplicada às Ciências Sociais. 7 ed. Florianópolis: Editora da UFSC, 2010