Universidade de São Paulo

Instituto de Física

Lei de NewcomB-Benford

Alisson Mendonça nº USP: 80658504
Fabio de Moraes Canedo nº USP:7994642
Mariana Morales Vilar nº USP:7580870
Thomas Andrioli Leick n°USP: 6452541
Victor Gomes da Costa Lobo n° USP: 7994405

São Paulo
Junho/2015

4300262/2015 – Métodos Estatísticos em Física Experimental

Sumário

Resumo

Através deste trabalho foi possível familiarizar-se com o que se conhece como a “Lei de Benford” para a distribuição de probabilidade dos dígitos de números de conjuntos de dados e foi possível verificar que a numeração das residências brasileiras registradas na lista telefônica não segue a Lei de NewcomB-Benford.

Introdução

Quando solicitamos a uma pessoa que diga qual é a probabilidade associada para o primeiro dígito de um conjunto de números, em geral, ela deve fornecer intuitivamente o valor 1/9, ou seja, a mesma probabilidade de ocorrência para cada um dos nove dígitos.

Newcomb[1] foi o primeiro a perceber que essa uniformidade não era válida em várias situações e concebeu uma distribuição de probabilidade para o primeiro dígito de números obtidos de várias fontes. Benford[2] mostrou formalmente a lei, e encontramos resultados e propriedades mais gerais da lei em Hill[3].

O Modelo teórico proposto por Benford para a frequência do Primeiro Dígito segue:

$$P(d)=\log\left(1+\frac{1}{d}\right), d=[1;9]\in\mathbb{Z}$$

A **Tabela 1** e o **Gráfico 1** ilustram a Lei de Benford:

Primeiro Dígito	1	2	3	4	5	6	7	8	9	Total
Porcentagem (%)	30,103	17,609	12,494	9,691	7,918	6,695	5,799	5,115	4,576	100,000

Tabela 1: Distribuição teórica do primeiro dígito segundo a lei de Benford.

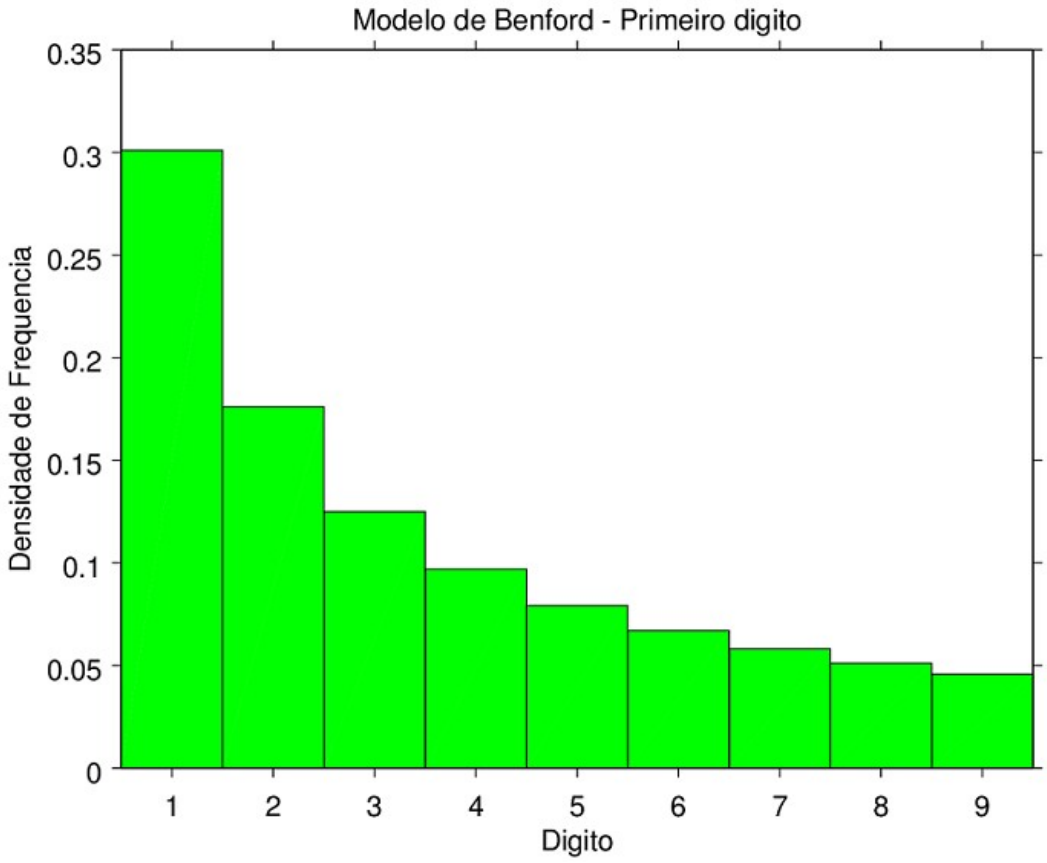

Gráfico 1: Distribuição teórica do primeiro dígito segundo a lei de Benford.

Tempos depois, Hill publicou seu artigo intitulado “The Significant-Digit Phenomenon” com uma generalização para as frequências dos Digitos:

$$P\left(\bigcap_{i=1}^{k}\{D_i=d_i\}\right)=\log_{10}\left[1+\left(\sum_{i=1}^{k} d_i*10^{k-i}\right)^{-1}\right]$$

Através desta generalização foi possível verificar a frequência dos algarismos para os segundos, terceiros e quartos dígitos e montar a **Tabela 2** e os 4 gráficos seguintes:

	Frequência			
Algarismo	1º Dígito	2º Dígito	3º Dígito	4º Dígito
0	-	0,11968	0,10178	0,10018
1	0,30103	0,11389	0,10138	0,10014
2	0,17609	0,10882	0,10097	0,10010
3	0,12494	0,10433	0,10057	0,10006
4	0,09691	0,10031	0,10018	0,10002
5	0,07918	0,09668	0,09979	0,09998
6	0,06695	0,09337	0,09940	0,09994
7	0,05799	0,09035	0,09902	0,09990
8	0,05115	0,08757	0,09864	0,09986
9	0,04576	0,08500	0,09827	0,09982

Tabela 2: Distribuição dos dígitos segundo a lei de Benford.

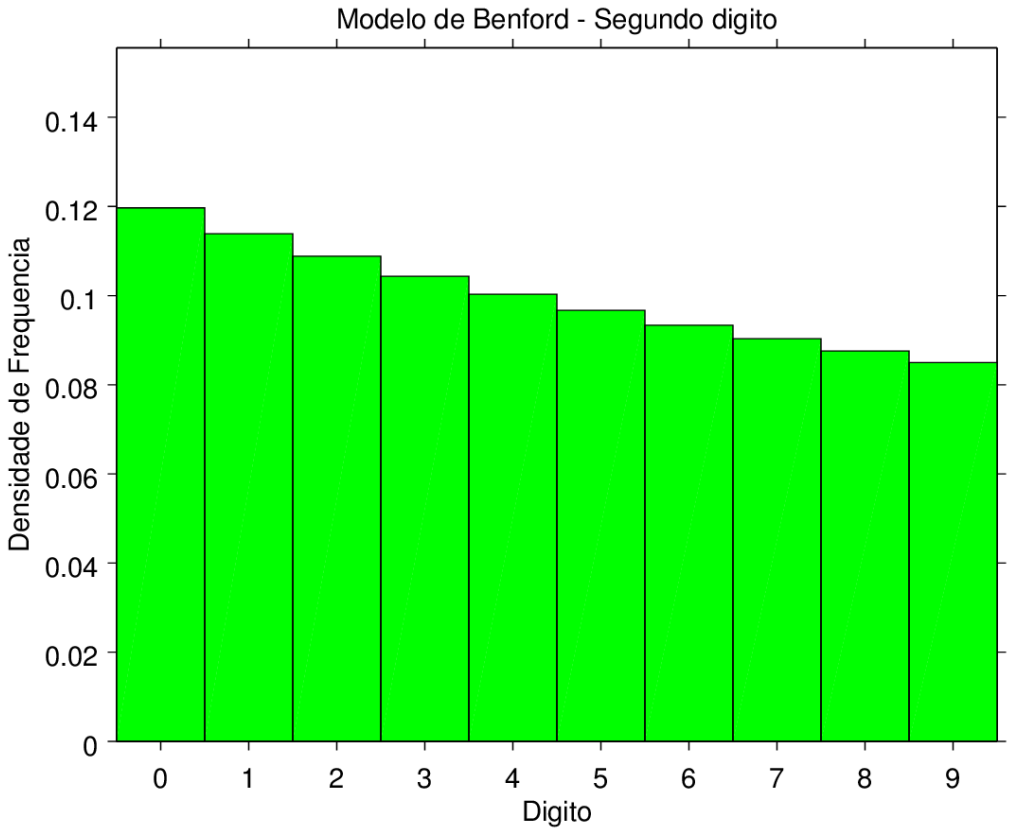

Gráfico 2: Distribuição teórica do segundo dígito segundo a lei de Benford.

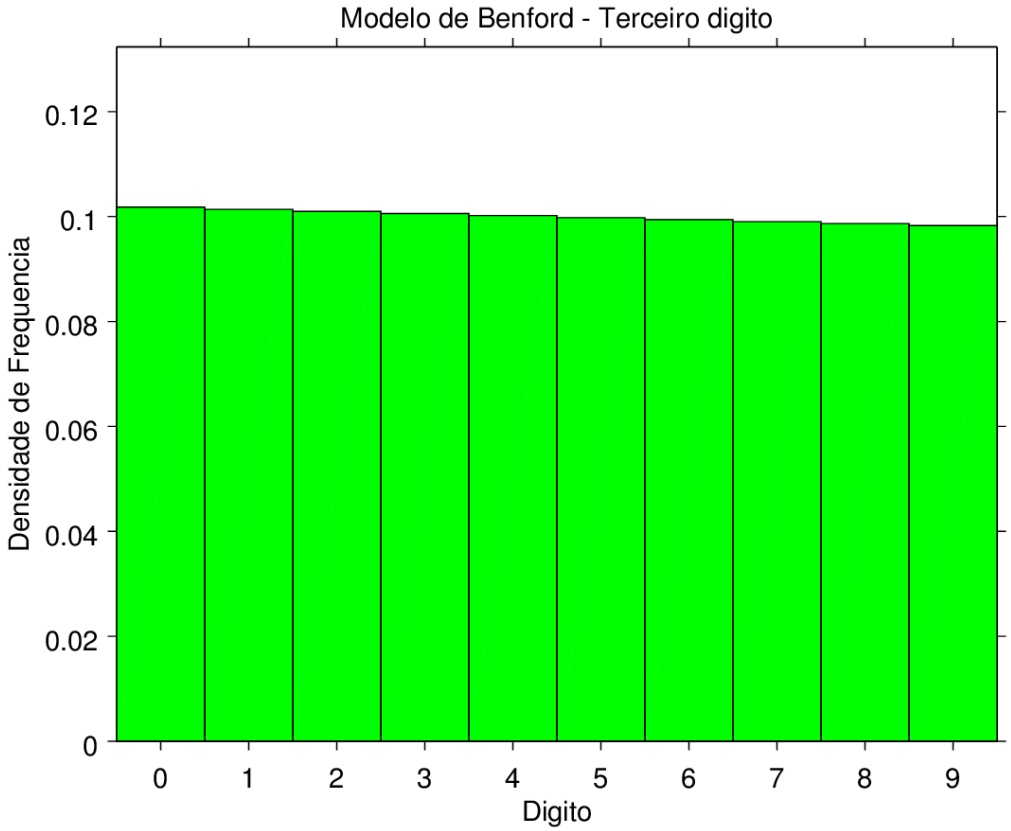

Gráfico 3: Distribuição teórica do terceiro dígito segundo a lei de Benford.

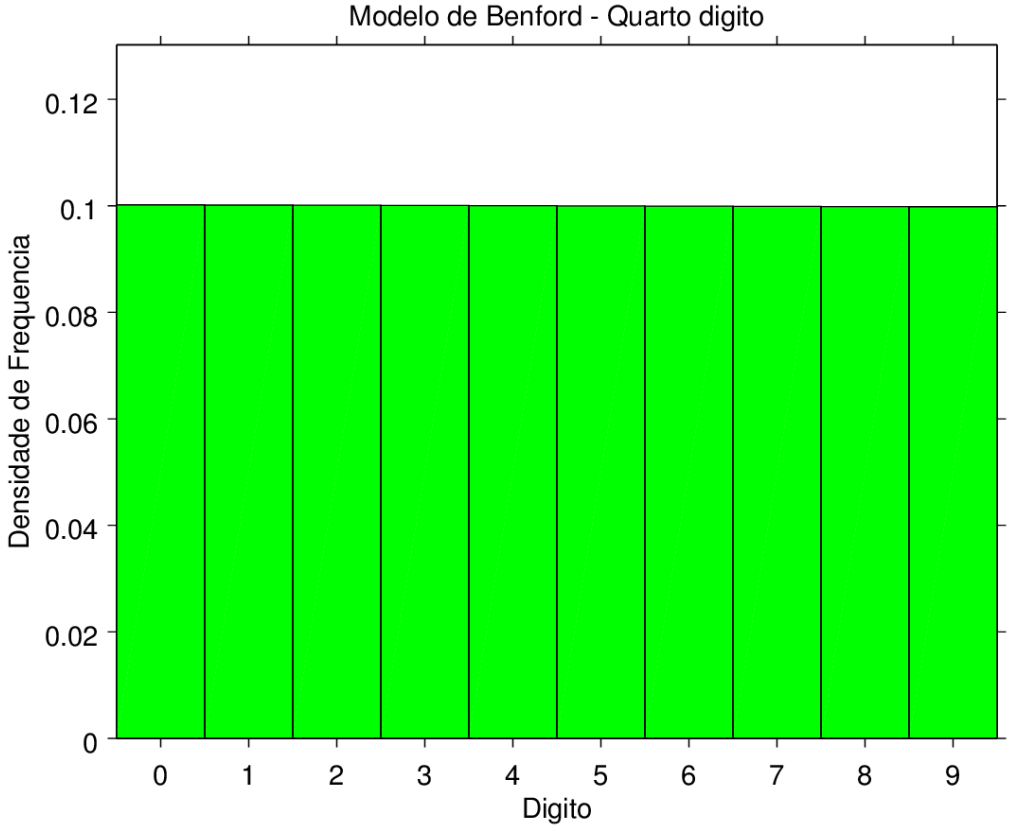

Gráfico 4: Distribuição teórica do quarto dígito segundo a lei de Benford.

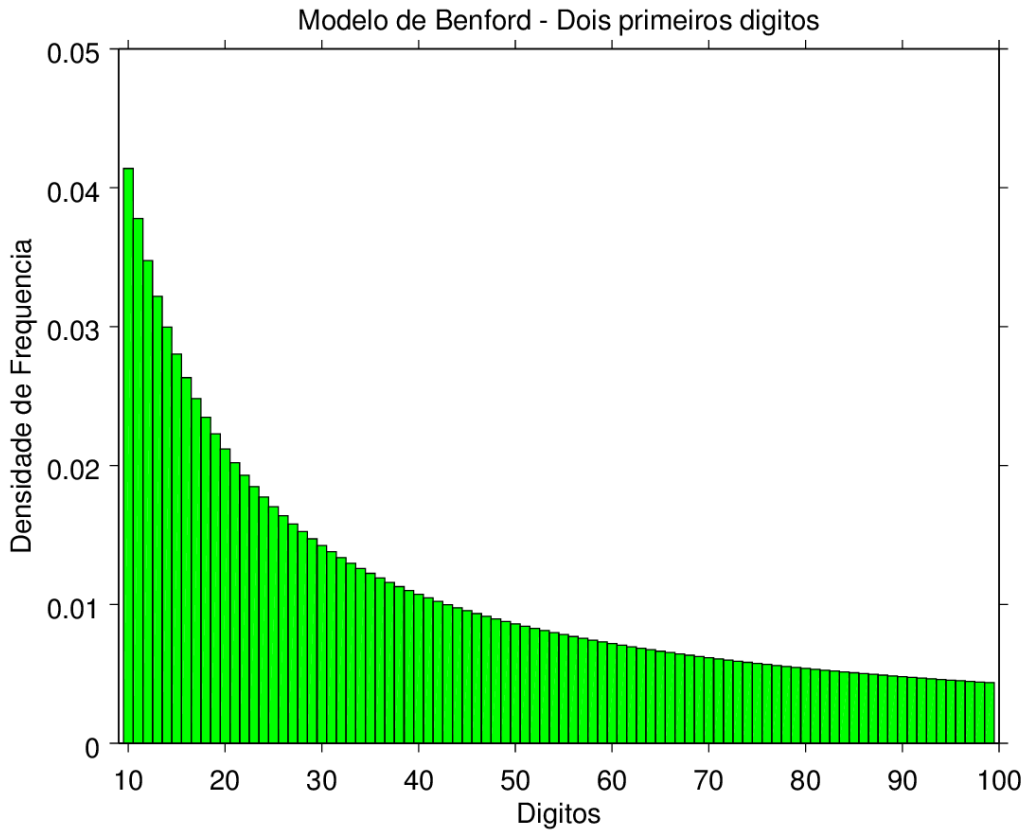

Gráfico 5: Distribuição teórica dos dois primeiros dígitos segundo a lei de Benford.

É conhecido que diversos conjuntos de dados seguem a distribuição de Benford, tais como a Sequência de Fibonacci, fatoriais, potências de 2, etc. Alguns processos contínuos também, como a multiplicação de bactérias e o decaimento radioativo.

Um determinado conjunto de dados **pode** seguir a distribuição de Benford caso a média seja maior que a mediana; os números sejam o resultado de alguma operação matemática; dados à nível de transação; distribuições com crescimento exponencial; etc.

Objetivo

Este trabalho tem por objetivo introduzir o que se conhece como a "Lei de Benford" para a distribuição de probabilidade dos dígitos de números de conjuntos de dados e verificar se os números das residências brasileiras listadas na lista telefônica seguem esta lei.

Metodologia

Para verificar a distribuição real dos dígitos em série de dados que seguem a "Lei de Benford" rigorosamente foram feitas simulações com um "Gerador de Dígitos Benford" (**GDB**) em Octave.

De forma resumida, este gerador de dígitos utiliza um gerador de números pseudo-aleatórios[4] Mersenne Twister[5] 64bits com coeficiente de Hurst[6] **H*** = 0.5 e intervalo de saída [0,1]. Um número **n** pseudo-aleatório é então convertido em um número inteiro **d** tal que $\boldsymbol{n}\in\left[P_a(\boldsymbol{d}-1), P_a(\boldsymbol{d})\right]$, onde P_a é a função Probabilidade Acumulada da Lei de Benford:

$$P_a(d)=\sum_{i=0}^{d}\{P(i)\}$$

Com este **GDB** foram realizadas diversas simulações para verificar o comportamento de séries com diferentes números de dados. As **Figuras** que seguem são composta por 4 gráficos referentes as simulações para o primeiro, segundo, terceiro e quarto dígitos e para os dois primeiros dígitos juntos:

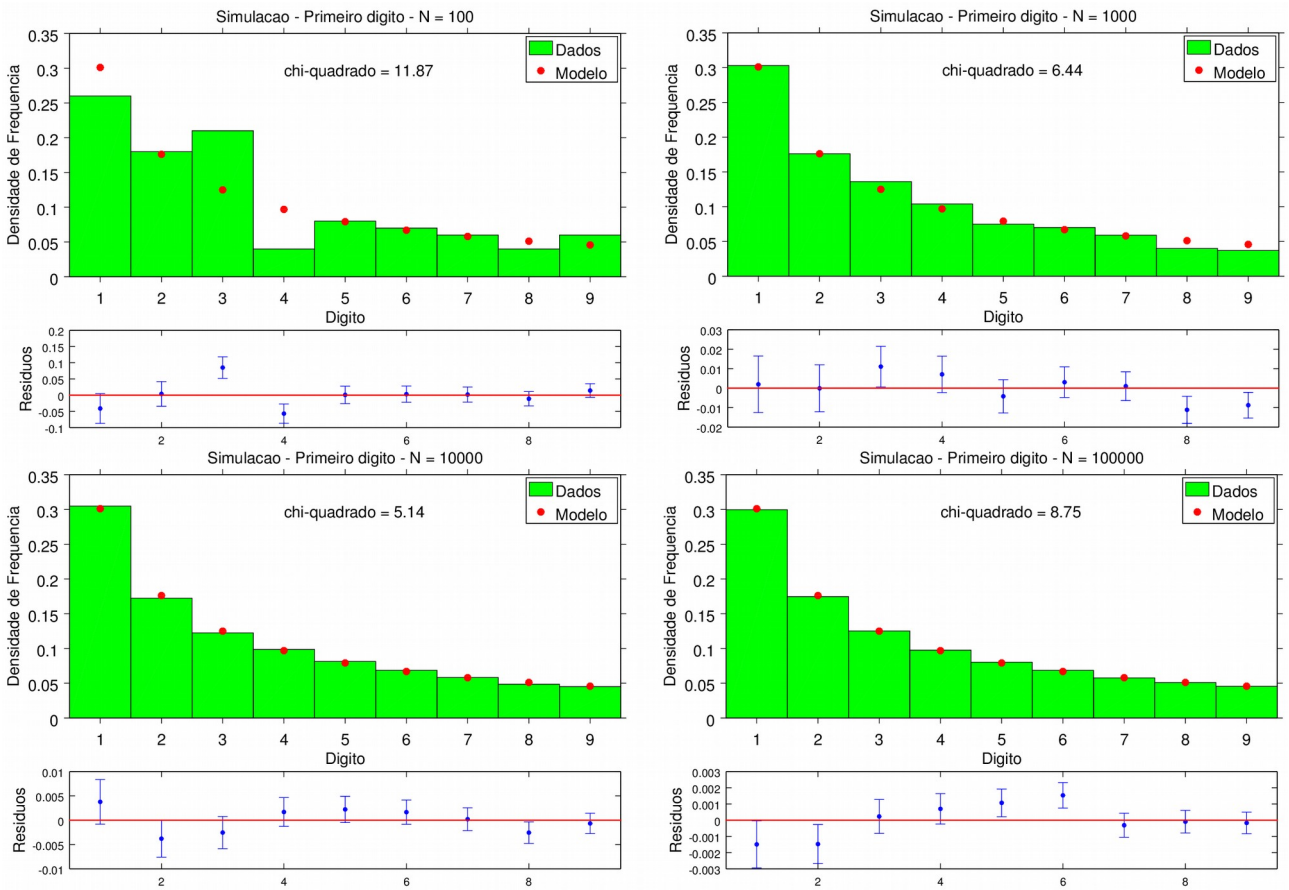

Figura 1: conjunto de gráficos com Simulações para o Primeiro Dígito.

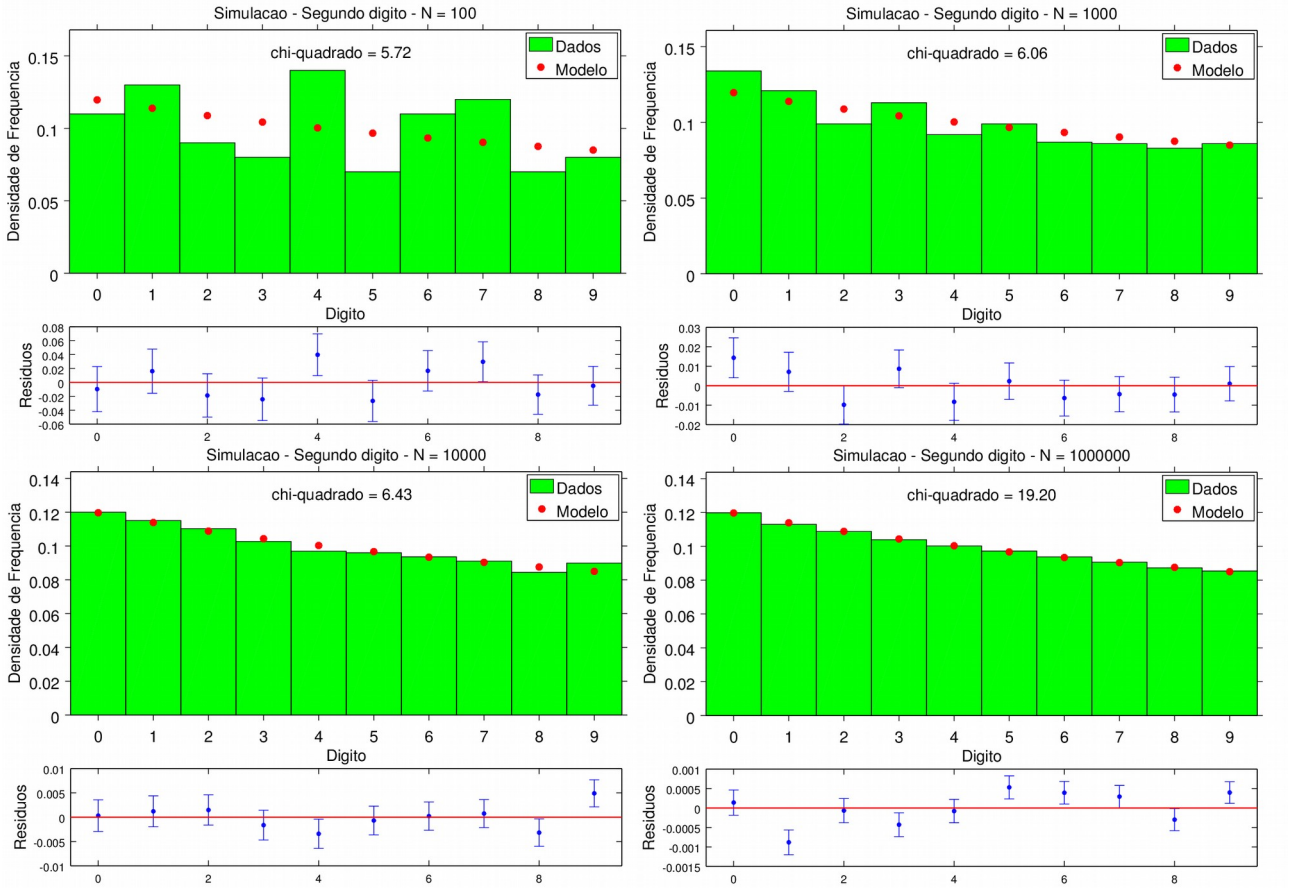

Figura 2: conjunto de gráficos com Simulações para o Segundo Dígito.

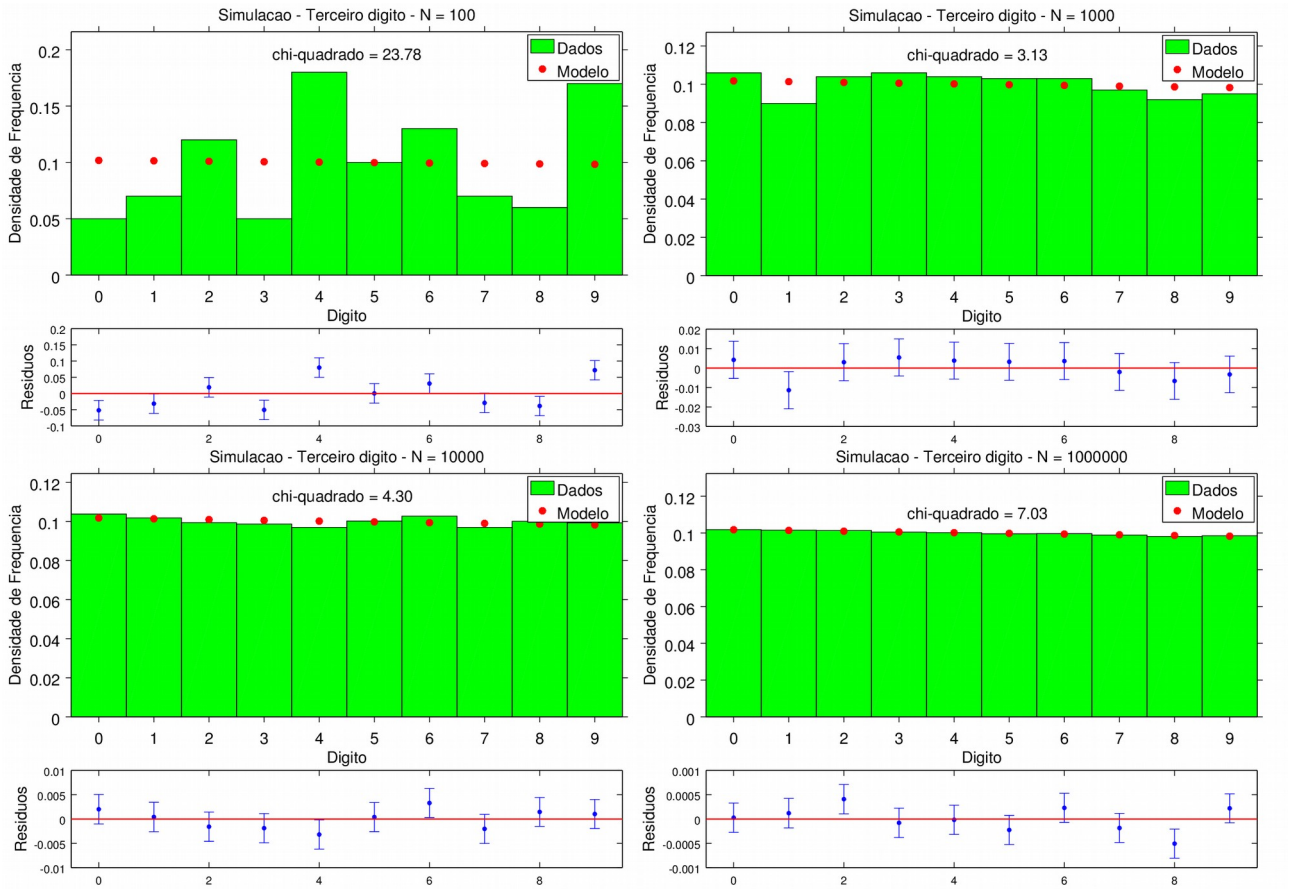

Figura 3: conjunto de gráficos com Simulações para o Terceiro Dígito.

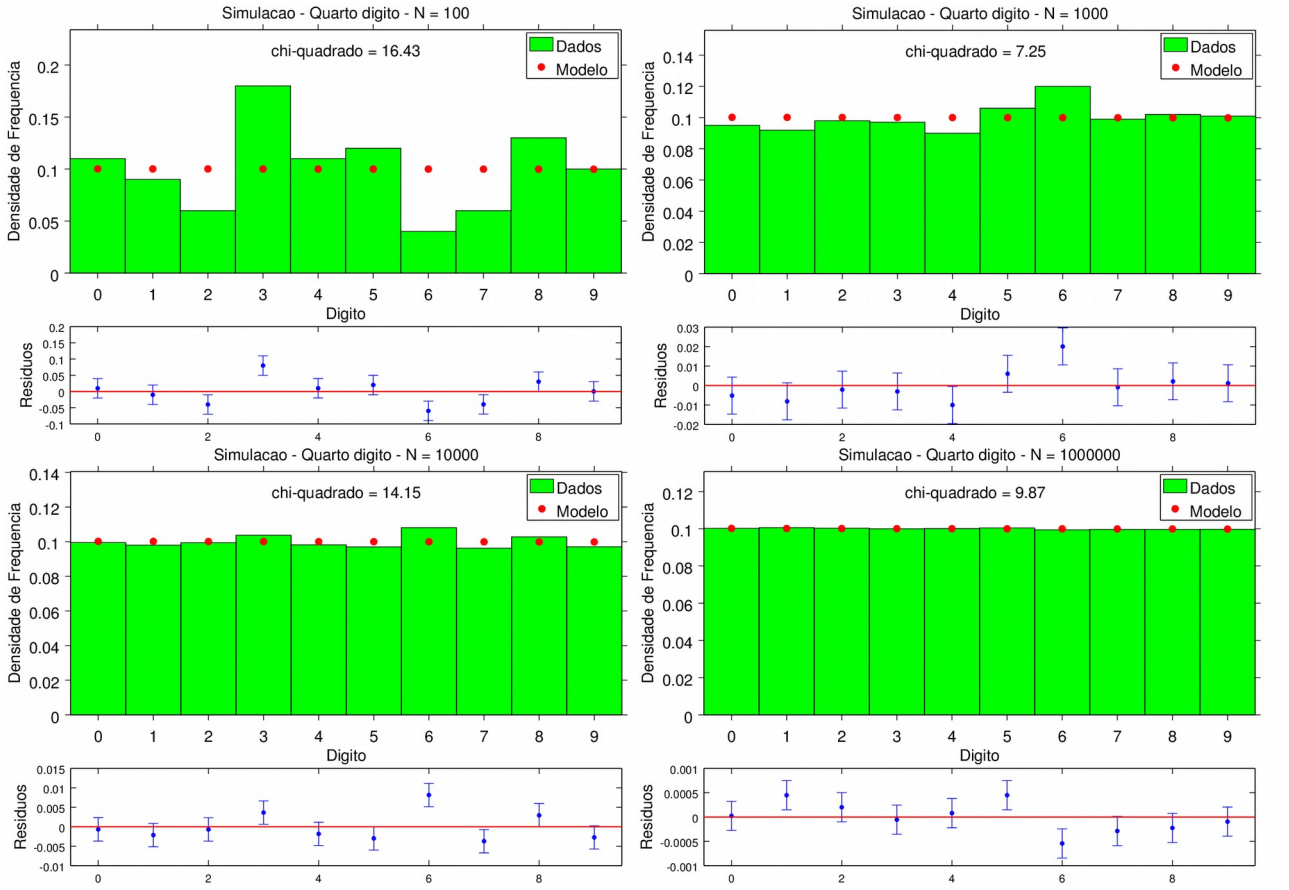

Figura 4: conjunto de gráficos com Simulações para o Quarto Dígito.

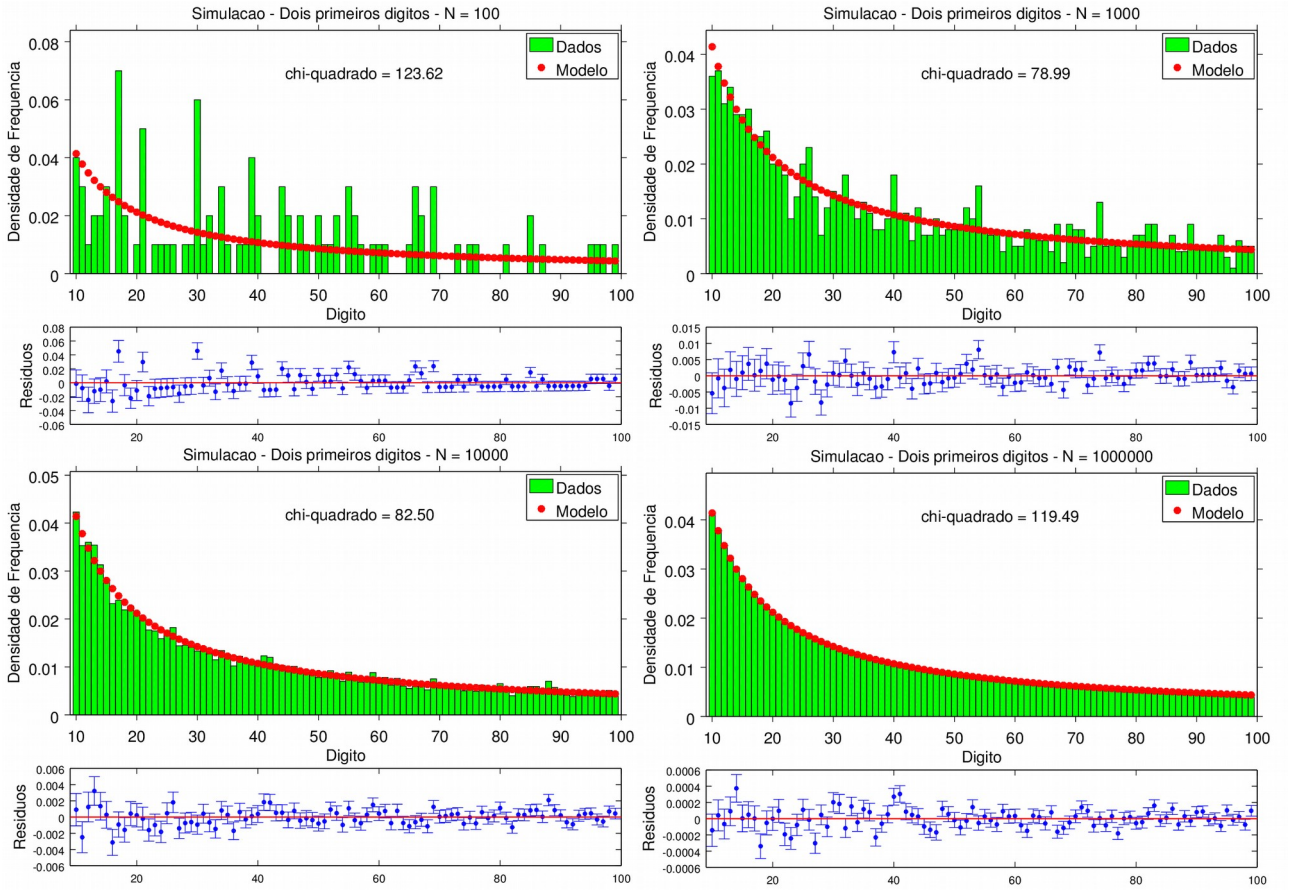

Figura 5: conjunto de gráficos com Simulações para os dois primeiros dígitos.

A Lei de Benford segue uma distribuição não-paramétrica discreta onde P(d) representa a porcentagem de sucessos de n tentativas independentes e cada tentativa resulta em apenas duas possibilidades: sucesso ou fracasso. Sendo assim, é fácil perceber que a Lei de Benford segue uma distribuição Binomial.

Partindo da função probabilidade binomial:

$$P_{N,p}(n) = \frac{N!}{(N-n)!\,n!}\, p^n (1-p)^{N-n}$$

fica fácil verificar a média do número de ocorrências $n_0 = nP$:

$$\langle n \rangle = \sum_{n=0}^{N} n \frac{N!}{(N-n)!\,n!}\, p^n (1-p)^{N-n} = nP$$

Sabendo também que $\sigma_n^2 = \left\langle (n-n_0)^2 \right\rangle$ e que podemos considerar $\left\langle (n-n_0)^2 \right\rangle = \langle n \rangle^2 - n_0^2$ então, com um pouco de álgebra, é possível mostrar que:

$$\sigma_n = \sqrt{Np(1-p)}$$

Para realizar os ajustes e verificar o χ^2 das simulações e dos dados obtidos utilizamos a incerteza calculada através da binomial mostrado acima. A **Tabela 3** contém o intervalo de χ^2 com confiança de 95% para cada um dos dígitos:

	χ2	
Dígito	mínimo	máximo
1º	2,70	19,02
2º	3,25	20,48
3º	3,25	20,48
4º	3,25	20,48
1º e 2º juntos	64,79	116,99

Tabela 3: Intervalo χ2 - confiança de 95%

Com essas simulações em mãos foi possível verificar o comportamento de dados que seguem a Lei de Benford e então, iniciar a tomada de dados referente aos números das residências brasileiras cadastradas na lista telefônica.

De início foram obtidas séries de dados coletados manualmente (entre 500 e 700 números) e então a coleta foi otimizada com um script em Python 3.4 com o auxílio de duas bibliotecas fundamentais: BeautifulSoup4[7] e Requests.

Para a separação dos dígitos foi feito um programa em C++ que, resumidamente, lê um arquivo de texto com dados, separa cada um dos dígitos e imprime um relatório com a contagem de cada algarismo para cada dígito.

RESULTADOS

Os dados brutos coletados não foram colocados neste relatório devido ao seu grande volume em cada série, portanto, não aparecerão tabelas de dados a partir daqui. Porém, as tabelas de frequência relativa de cada dígito para algumas séries de dados com suas respectivas incertezas sim; tais tabelas, junto com os histogramas e gráficos de resíduos que seguem ilustrarão de tal maneira que os dados brutos não farão falta.

As duas tabelas abaixo e os 2 gráficos são referentes ao Primeiro Dígito de duas séries de dados coletadas manualmente:

Digito	Obtido	Freq. Obt	Freq. Esp	Resíduo²	σ²	χ²
1	185	0,33214	0,30103	0,00097	0,00038	2,56146
2	77	0,13824	0,17609	0,00143	0,00026	5,50029
3	56	0,10054	0,12494	0,00060	0,00020	3,03322
4	51	0,09156	0,09691	0,00003	0,00016	0,18203
5	51	0,09156	0,07918	0,00015	0,00013	1,17098
6	32	0,05745	0,06695	0,00009	0,00011	0,80411
7	38	0,06822	0,05799	0,00010	0,00010	1,06719
8	30	0,05386	0,05115	0,00001	0,00009	0,08412
9	37	0,06643	0,04576	0,00043	0,00008	5,45013
Total	557	1,00000	1,00000	0,00381	0,00150	19,85353

Tabela 4: Primeiro Dígito, sobrenome Santos, coleta manual

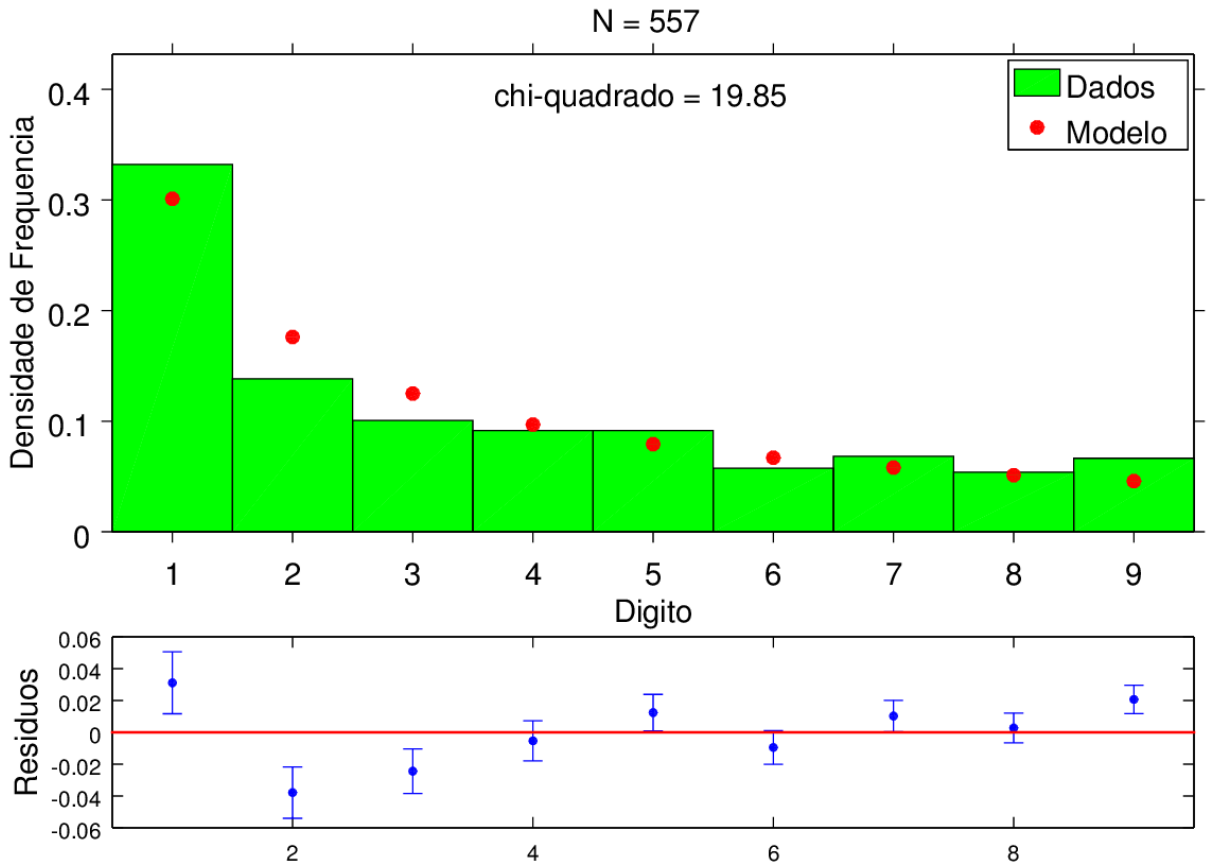

Gráfico 6: Primeiro Dígito, sobrenome Santos, coleta manual

Digito	Obtido	Freq. Obt	Freq. Esp	Resíduo²	σ²	χ^2
1	161	0,30492	0,30103	0,00002	0,00040	0,03806
2	95	0,17992	0,17609	0,00001	0,00027	0,05347
3	75	0,14205	0,12494	0,00029	0,00021	1,41329
4	56	0,10606	0,09691	0,00008	0,00017	0,50516
5	32	0,06061	0,07918	0,00035	0,00014	2,49864
6	29	0,05492	0,06695	0,00014	0,00012	1,22177
7	30	0,05682	0,05799	0,00000	0,00010	0,01332
8	33	0,06250	0,05115	0,00013	0,00009	1,40078
9	17	0,03220	0,04576	0,00018	0,00008	2,22365
Total	528	1,00000	1,00000	0,00121	0,00158	9,36813

Tabela 5: Primeiro Dígito, sobrenome Souza, coleta manual

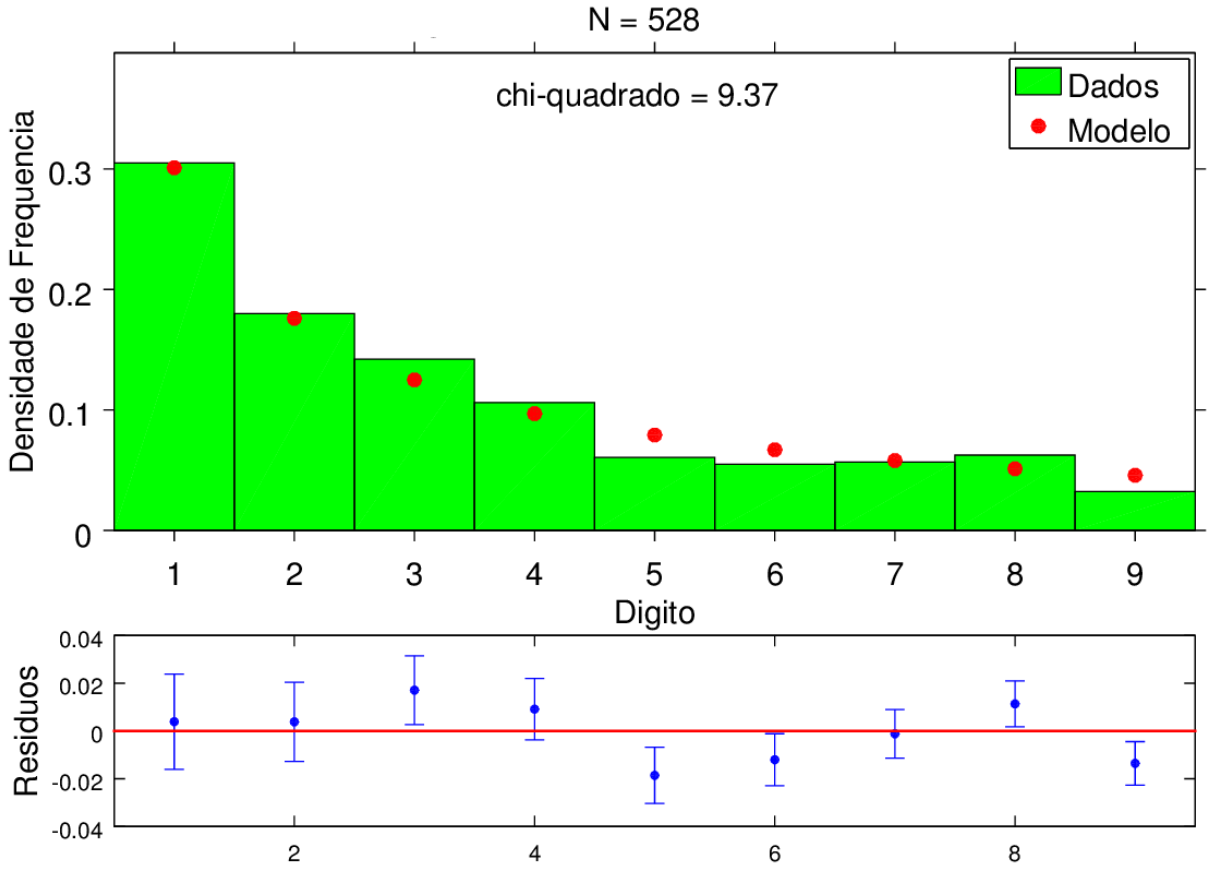

Gráfico 7: Primeiro Dígito, sobrenome Souza, coleta manual

Todos os dados referentes as tabelas e gráficos abaixo foram colhidos otimizadamente utilizando o script em Python já citado neste relatório. Para facilitar a visualização, os gráficos e tabelas não estarão dispostos seguindo um conjunto de gráficos para cada série de dados, serão conjuntos de gráficos para cada dígito.

As tabelas e gráficos abaixo são referentes ao Primeiro Dígito.

Série de dados com 1055 números de casas do no Estado de São Paulo, sobrenome Silva:

Digito	Obtido	Freq. Obt	Freq. Esp	Resíduo²	σ²	X²
1	307	0,2909953	0,3010300	0,0001007	0,0001994	0,5048891
2	184	0,1744076	0,1760913	0,0000028	0,0001375	0,0206135
3	130	0,1232227	0,1249387	0,0000029	0,0001036	0,0284148
4	112	0,1061611	0,0969100	0,0000856	0,0000830	1,0316724
5	102	0,0966825	0,0791812	0,0003063	0,0000691	4,4319264
6	74	0,0701422	0,0669468	0,0000102	0,0000592	0,1724504
7	46	0,0436019	0,0579919	0,0002071	0,0000518	3,9990316
8	53	0,0502370	0,0511525	0,0000008	0,0000460	0,0182204
9	47	0,0445498	0,0457575	0,0000015	0,0000414	0,0352427
Total	1.055	1,0000000	1,0000000	0,0007179	0,0007910	10,2424614

Tabela 6: Primeiro Dígito, sobrenome Silva, Estado de São Paulo, coleta otimizada

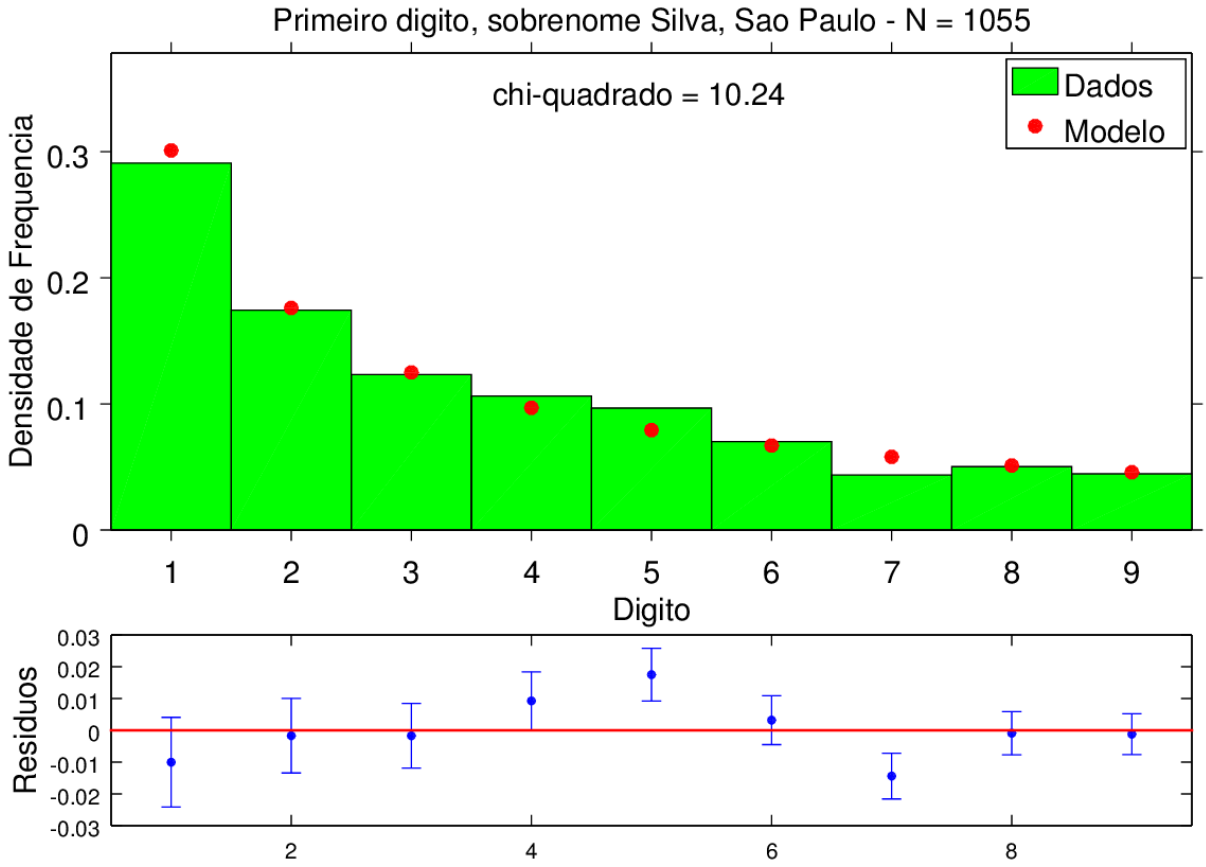

Gráfico 8: Primeiro Dígito, sobrenome Silva, Estado de São Paulo, coleta otimizada

Série de dados com 10536 números de casas do no Estado de São Paulo, sobrenome Silva:

Digito	Obtido	Freq. Obt	Freq. Esp	Resíduo²	σ²	X²
1	3.181	0,3019172	0,3010300	0,0000008	0,0000200	0,0394176
2	1.903	0,1806188	0,1760913	0,0000205	0,0000138	1,4886394
3	1.269	0,1204442	0,1249387	0,0000202	0,0000104	1,9467568
4	1.011	0,0959567	0,0969100	0,0000009	0,0000083	0,1094030
5	908	0,0861807	0,0791812	0,0000490	0,0000069	7,0796095
6	686	0,0651101	0,0669468	0,0000034	0,0000059	0,5689993
7	564	0,0535308	0,0579919	0,0000199	0,0000052	3,8384503
8	520	0,0493546	0,0511525	0,0000032	0,0000046	0,7017093
9	494	0,0468869	0,0457575	0,0000013	0,0000041	0,3077726
Total	10.536	1,0000000	1,0000000	0,0001192	0,0000792	16,0807578

Tabela 7: Primeiro Dígito, sobrenome Silva, Estado de São Paulo, coleta otimizada

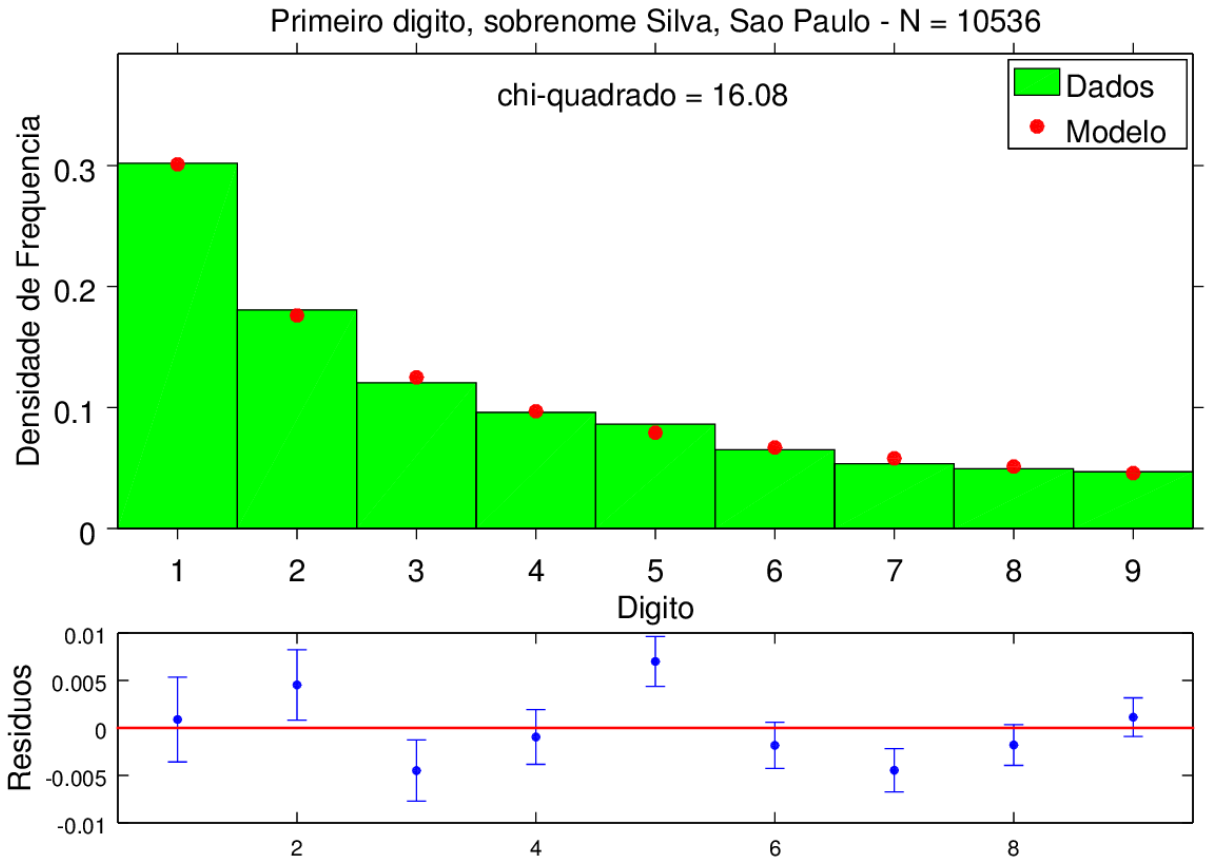

Gráfico 9: Primeiro Dígito, sobrenome Silva, Estado de São Paulo, coleta otimizada

Série de dados com 18022 números de casas em Curitiba, sobrenome Silva:

Digito	Obtido	Freq. Obt	Freq. Esp	Resíduo²	σ²	χ^2
1	5.091	0,2824881	0,3010300	0,0003438	0,0000117	29,4472226
2	3.288	0,1824437	0,1760913	0,0000404	0,0000081	5,0126180
3	2.298	0,1275108	0,1249387	0,0000066	0,0000061	1,0905298
4	1.824	0,1012096	0,0969100	0,0000185	0,0000049	3,8068292
5	1.492	0,0827877	0,0791812	0,0000130	0,0000040	3,2149056
6	1.264	0,0701365	0,0669468	0,0000102	0,0000035	2,9354134
7	1.057	0,0586505	0,0579919	0,0000004	0,0000030	0,1430911
8	913	0,0506603	0,0511525	0,0000002	0,0000027	0,0899612
9	795	0,0441128	0,0457575	0,0000027	0,0000024	1,1165451
Total	18.022	1,0000000	1,0000000	0,0004358	0,0000463	46,8571159

Tabela 8: Primeiro Dígito, sobrenome Silva, Curitiba, coleta otimizada

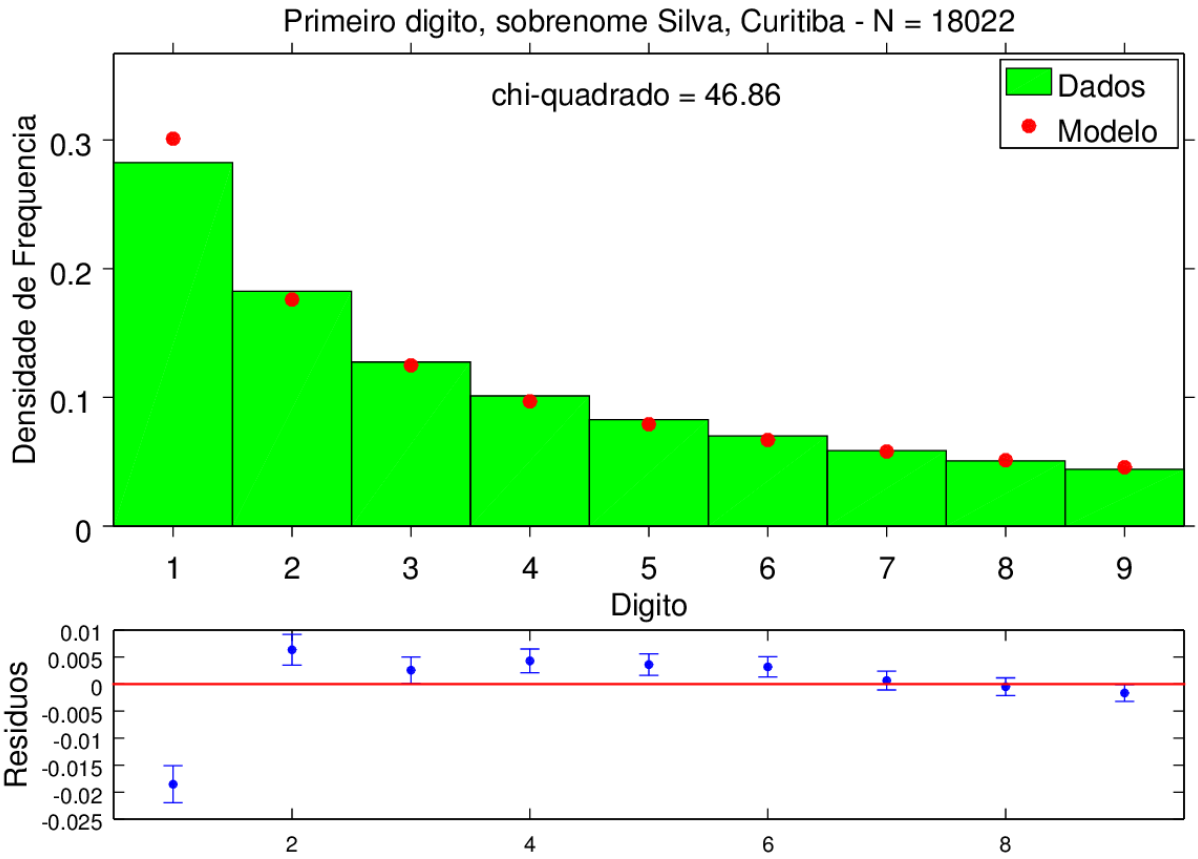

Gráfico 10: Primeiro Dígito, sobrenome Silva, Curitiba, coleta otimizada

Para facilitar a visualização e não estender ainda mais este documento, as tabelas para os gráficos seguintes foram omitidas porém, para a montagem de cada histograma e gráfico de resíduos foi feito exatamente os mesmos passos realizados nos gráficos anteriores.

A quantidade total de números, sobrenome coletado e Estado/Cidade foram colocados no título de cada Gráfico e o $\chi 2$ para cada um deles está localizado dentro da grade do histograma.

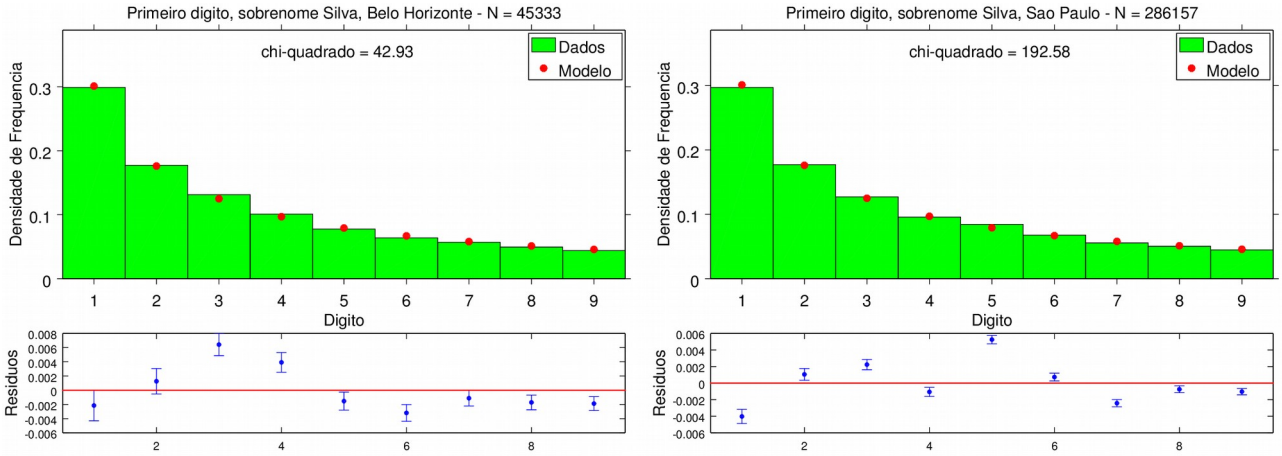

Figura 6: Primeiro Dígito, sobrenome Silva, coleta otimizada

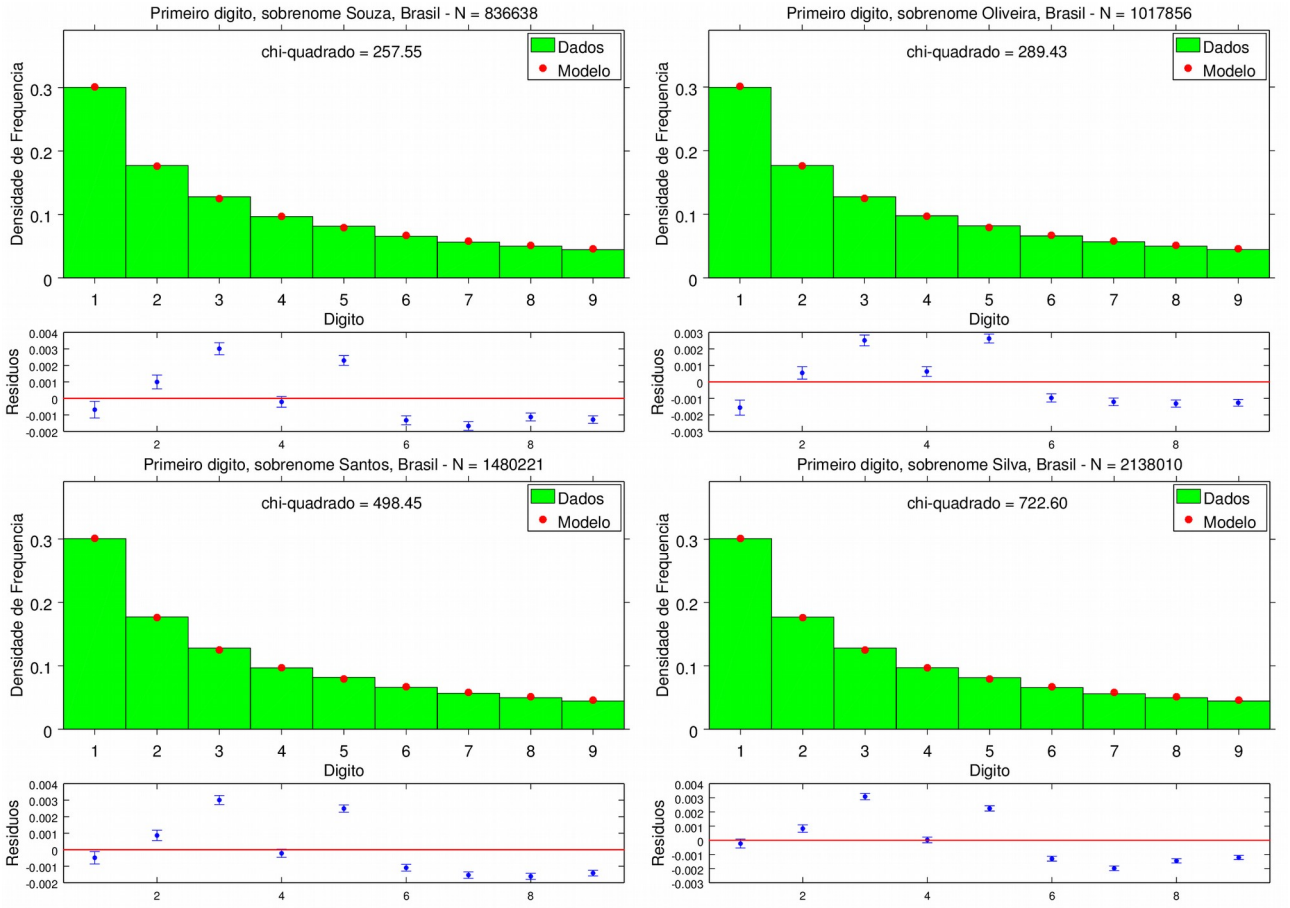

Figura 7: Primeiro Dígito, todo o território nacional, coleta otimizada

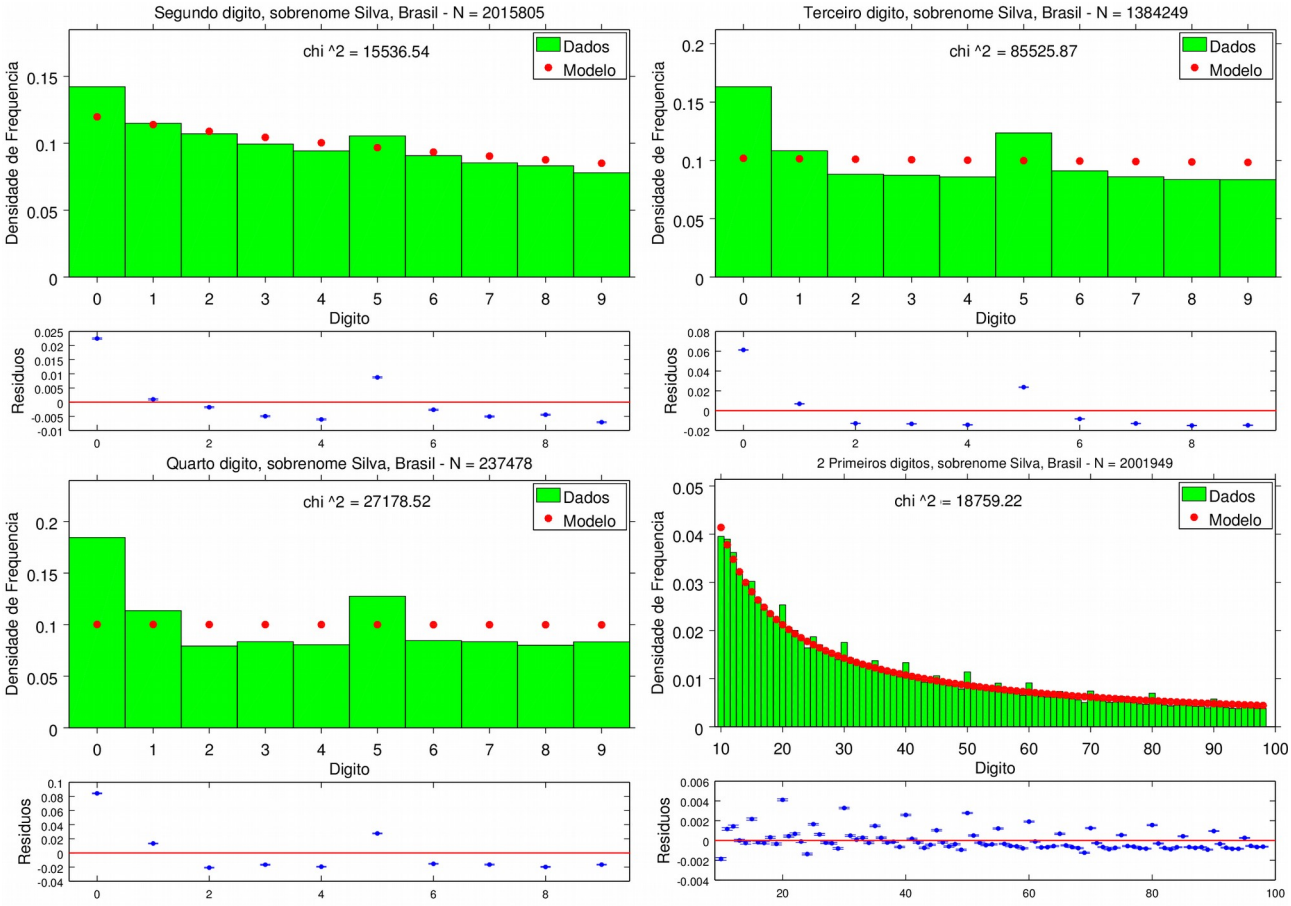

Figura 8: Conjunto de Gráficos referentes ao nome SIlva em todo o Brasil, coleta otimizada

Cálculos

A Frequência relativa foi calculada dividindo a quantidade de números que começam com um dígito específico n_d pelo número total de dados daquela série **N**. Como os histogramas foram feitos utilizando a frequência relativa, logo a incerteza para cada "bin" do histograma também foi normalizada pelo número total de dados da série:

$$\sigma_d = \sqrt{\frac{p_{(d)}(1-p_{(d)})}{N}}$$

Como os gráficos de resíduos não apresentaram nenhum vício aparente foi realizado o teste de $\chi 2$ para verificar a adequação dos modelos:

$$\chi^2 = \sum_{d=0}^{9} \frac{(O_d - E_d)^2}{\sigma_d^2}$$

Onde O_d é a frequência relativa obtida para o dígito d e E_d é a frequência relativa esperada para o dígito d. Para facilitar a visualização de como o $\chi 2$ se comporta de acordo com o número total de dados **N**, foi feito então, o gráfico que segue:

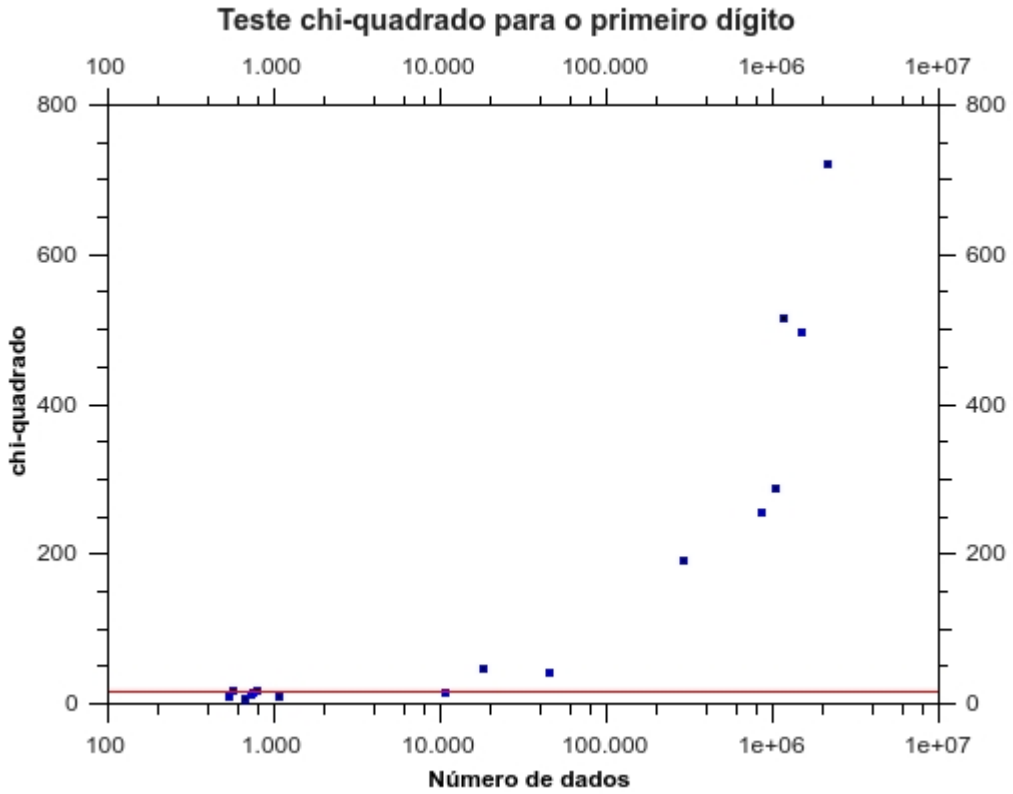

Gráfico 11: $\chi 2$ em função do número de dados para o Primeiro Digito

Conclusões

Após realizar as simulações com o **GDB**, foi possível entender o comportamento de séries de dados que seguem a Lei de Benford. Essas simulações foram extremamente importantes tanto para guiar os estudos durante a coleta de dados quanto para comparações após as coletas.

Após analisar o $\chi 2$ de cada série de dados e com ajuda do **Gráfico 11** fica fácil perceber que o modelo funciona para até 11mil dados aproximadamente. Isso mostra a importância de se coletar diversos dados e como a automação foi fundamental para este estudo pois, sem o script em Python para colher os dados automaticamente a conclusão deste trabalho seria de que os números das residências no Brasil seguem a Lei de Benford, o que é um grande equívoco!

Conforme se aumenta o número de dados de uma determinada amostra, mais estes dados se aproximam da frequência relativa real da população, consequentemente a incerteza dos dados amostrais diminui. No caso das séries de dados deste estudo, as incertezas dos dados convergem para zero muito mais rápido de que os dados amostrais convergem para a Lei de Benford. Isso implica em uma não adequação do modelo.

Para o Segundo, Terceiro e Quarto dígitos não foi preciso fazer um estudo mais aprofundado pois, apenas de se olhar o histograma já é o suficiente para perceber que o modelo não é adequado aos dados. Isso já era previsto uma vez que, ao escolher o número das casas as pessoas tendem a priorizar um final com 0, 1 ou 5.

Discussão Final

Foram retirados dados de cidades planejadas como Brasília, Belo Horizonte e Curitiba para verificar se existiria alguma diferença na distribuição dos dados em relação a cidades não planejadas. Essa diferença não existiu e, portanto, alguns destes dados foram retirados deste documento.

A Lei de Benford se aplicou muito bem as séries com até 11mil dados e não se aplica para mais do que isso, isso fez com que surgissem alguns questionamentos de porque isto estava ocorrendo. O **Gráfico 12** é um histograma dos números das casas (dados brutos):

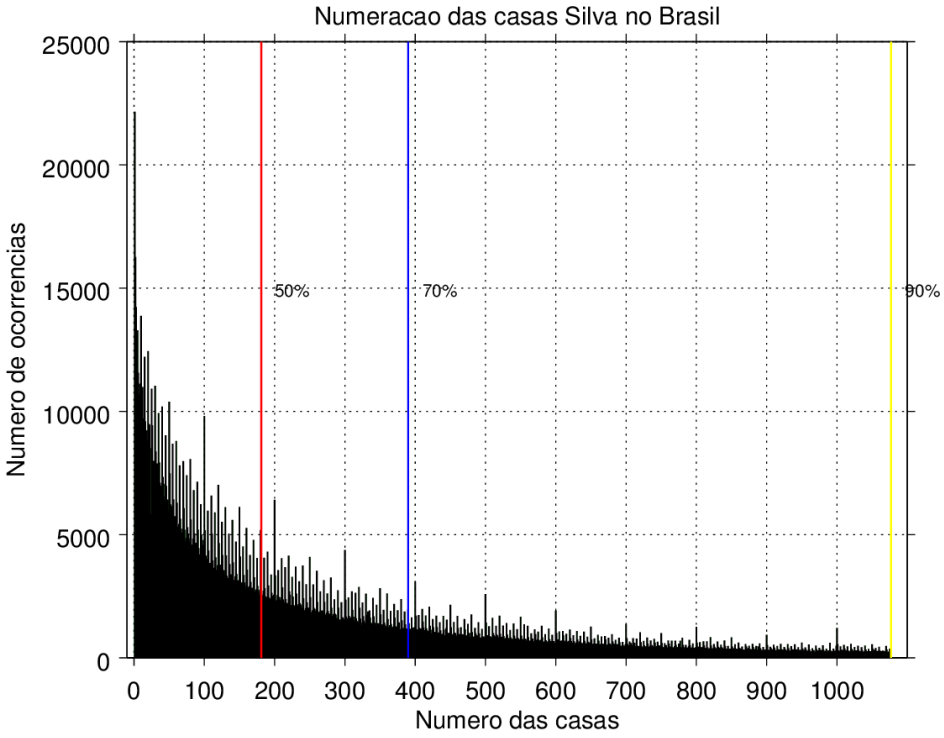

Gráfico 12: Histograma dos números das residências

A Lei de Benford funciona melhor quanto maior for a distância entre o mínimo e o máximo dos dados. Essa distribuição do **Gráfico 12** se mantém praticamente constante para todas as séries de dados, onde 50% dos dados estão concentrados, a grosso modo, no intervalo]0,200[, 70% dos dados no intervalo]0,400[e 90% dos dados no intervalo]0,1100[.

O Fato de aumentarmos o número de dados e mantermos o alcance dos dados praticamente o mesmo pode ser um dos fatores que inviabilizam a Lei de Benford aplicada aos números das residências.

Referências

[1] Newcomb, Simon, “Note on the frequency of use of the different digits in natural numbers”, Amer. Journ. of Math. 4, p.39-40, 1881.
[2] Benford, Frank, “The law of anomalous numbers”, Proceedings of The American Philosophical Society, vol. 78, p.551-572, 1938.
[3] Hill, T. P., “The Significant-Digit Phenomenon”, The American Mathematical Montly, Vol. 102, nº 4, p.322-327, 1995.
[4] https://www.inf.pucrs.br/~smusse/Simulacao/PDFs/GeradorAleatorios.pdf
[5] Journal ACM Transactions on Modeling and Computer Simulation (TOMACS) - Special issue on uniform random number generation TOMACS Homepage archive Volume 8 Issue 1, Jan. 1998 Pages 3-30.
[6] Hurst, H.E.; Black, R.P.; Simaika, Y.M. (1965). *Long-term storage: an experimental study.* London: Constable.
[7] http://www.crummy.com/software/BeautifulSoup/bs4/doc/