

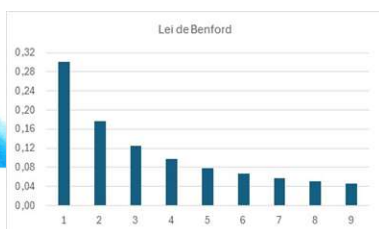
N.º 34 – Modelos de probabilidade discretos

A lei de Benford ou lei do primeiro dígito

Modelo uniforme discreto



Maria Eugénia Graça Martins
FCUL memartins@fc.ul.pt
Emília Oliveira
Escola Secundária de Tomaz Pelayo
ecmo.estp@gmail.com
Abril 2025



Em 1938, o físico americano Frank Benford publicou um artigo¹ no qual mostra um facto curioso que se passa com vários conjuntos de dados, no que diz respeito ao modo como se distribui a frequência do primeiro dígito. Este facto passou a ser designado por lei de Benford ou lei de Benford-Newcomb.

Modelo de probabilidade discreto – função massa de probabilidade

Considere-se um fenómeno aleatório cujo espaço de resultados tenha um número finito, n , de resultados possíveis, e seja X uma variável aleatória (v.a.) discreta, função que associa a cada resultado um número real x_i , $i, i = 1, 2, \dots, n$.

Construir um **modelo de probabilidade** para o fenómeno aleatório consiste em obter para cada valor da v.a. X , x_i , com $i = 1, \dots, n$, a probabilidade $p_i = P(X = x_i)$. Aos pares de valores (x_i, p_i) dá-se o nome de **função massa de probabilidade** (f.m.p.) de X .

O que é a lei de Benford?

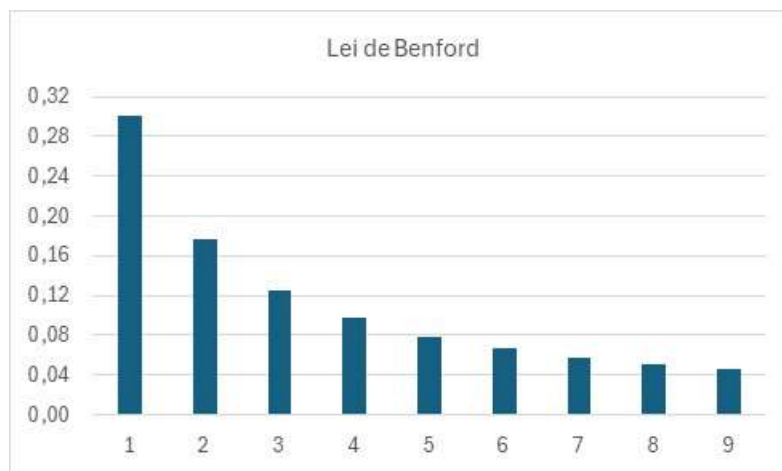
Um fenómeno curioso observado por Benford, mas que já havia sido observado pelo astrónomo Simon Newcomb², diz respeito à frequência do primeiro dígito de vários conjuntos de dados. Ao contrário do que seria de esperar, à distribuição das frequências do primeiro dígito não se ajusta um modelo de probabilidade uniforme, isto é, não se obtém a mesma frequência para os 9 dígitos. Na realidade, o fenómeno aleatório que consiste em verificar qual o primeiro dígito de cada número, de um conjunto vasto de números, pode ser bem

¹ Benford, F. “The Law of anomalous numbers”, Proceedings of the American Philosophical Society, 78, 551–572, 1938

² Newcomb, S., “Note on the frequency of use of different digits in natural numbers”, American Journal of Math. 4, 39–40, 1881

modelado por uma variável aleatória (v.a.) X , com a seguinte função massa de probabilidade (f.m.p):

$X = i$	1	2	3	4	5	6	7	8	9
$P(X = i)$	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046



As probabilidades foram obtidas a partir da função

$$P(X = a) = \log\left(\frac{a+1}{a}\right) \text{ com } a = 1, 2, \dots, 9.$$

Os valores da tabela satisfazem as condições para ser uma f.m.p., porque verificam as condições:

- i) $P(X = a) \geq 0$, para $a = 1, 2, \dots, 9$
- ii) $\sum_{i=1}^9 P(X = a) = 1$

O modelo de probabilidade cujas probabilidades seguem a lei de Benford tem aplicações importantes na modelação de grandes conjuntos de dados, como, por exemplo, dimensões de populações, comprimentos de rios, assim como na deteção de fraudes nos mundos financeiro e contabilístico, nos processos eleitorais, etc. Como exemplo da aplicação da lei de Benford na deteção de fraudes, sugere-se, por exemplo, a leitura dos objetivos dos trabalhos que se encontram em

<https://repositorio.ulisboa.pt/handle/10400.5/6134> e

<https://periodicos.iftm.edu.br/index.php/inova/article/view/694>.

De seguida, vai-se considerar a população constituída pelos tamanhos dos 278 municípios, à data dos Censos de 2021. Depois do estudo desta população, iremos ver se o conjunto dos dados é um exemplo da lei de Benford.

Modelo de probabilidade para a dimensão populacional dos municípios de Portugal Continental – nem sempre o modelo Uniforme discreto faz jus ao seu nome...!

1. População em estudo

A população em estudo é constituída pelos dados da seguinte tabela, inserida na folha de Cálculo Excel, que apresenta a **População residente (N.º) por Local de residência à data dos Censos 2021** (Consulta feita em 29 de janeiro de 2025 em www.ine.pt), em cada um dos 278 municípios de Portugal continental:

	Nome	Nº hab.	Nome	Nº hab.	Nome	Nº hab.	Nome	Nº hab.
1	Arc. de Valdevez	20718	70 Celorico da Beira	6583	140 Mogadouro	8301	210 S. M. de Penaguião	6100
2	Abrantes	34329	71 Celor. de Basto	17643	141 Moimenta da Beira	9410	211 Santarém	58662
3	Águeda	46119	72 Chamusca	8530	142 Moita	66255	212 Santiago do Cacém	27772
4	Aguiar da Beira	5231	73 Chaves	37590	143 Monção	17816	213 Santo Tirso	67709
5	Alandroal	5014	74 Cinfães	17730	144 Monchique	5462	214 São Brás de Alportel	11248
6	Alberg.-a-Velha	24840	75 Coimbra	140816	145 Mondim de Basto	6410	215 São João da Madeira	22143
7	Albufeira	44164	76 Condeixa-a-Nova	16732	146 Monforte	2992	216 S João da Pesqueira	6775
8	Alcácer do Sal	11112	77 Constância	3798	147 Montalegre	9261	217 São Pedro do Sul	15137
9	Alcanena	12472	78 Coruche	17355	148 Montemor-o-Novo	15799	218 Sardoal	3513
10	Alcobaça	54965	79 Covilhã	46455	149 Montemor-o-Velho	24571	219 Sátão	11030
11	Alcochete	19143	80 Crato	3225	150 Montijo	55682	220 Seia	21755
12	Alcoutim	2523	81 Cuba	4373	151 Mora	4135	221 Seixal	166507
13	Alenquer	44442	82 Elvas	20730	152 Mortágua	8963	222 Sernancelhe	5692
14	Alfândega da Fé	4324	83 Entroncamento	20141	153 Moura	13258	223 Serpa	13757
15	Alijó	10486	84 Espinho	31043	154 Mourão	2351	224 Sertã	14769
16	Aljezur	6045	85 Esposende	35132	155 Murça	5245	225 Sesimbra	52384
17	Aljustrel	8874	86 Estarreja	26213	156 Murtosa	10476	226 Setúbal	123496
18	Almada	177238	87 Estremoz	12680	157 Nazaré	14881	227 Sever do Vouga	11063
19	Almeida	5887	88 Évora	53577	158 Nelas	13119	228 Silves	37766
20	Almeirim	22012	89 Fafe	48497	159 Nisa	5952	229 Sines	14198
21	Almodôvar	6712	90 Faro	67622	160 Óbidos	11922	230 Sintra	385606
22	Alpiarça	6975	91 Felgueiras	55848	161 Odemira	29538	231 Sobral de M. Agraço	10540
23	Alter do Chão	3044	92 Fer. do Alentejo	7684	162 Odivelas	148034	232 Soure	17261
24	Alvaiázere	6238	93 Fer. do Zêzere	7800	163 Oeiras	171658	233 Sousel	4360
25	Alvito	2280	94 Figueira da Foz	58951	164 Oleiros	4904	234 Tábua	11160
26	Amadora	171454	95 Fig. de C Rodrigo	5148	165 Olhão	44614	235 Tabuaço	5034
27	Amarante	52116	96 Fig. dos Vinhos	5281	166 Oliv. de Azeméis	66175	236 Tarouca	7363
28	Amares	18595	97 For. de Algodres	4403	167 Oliveira de Frades	9506	237 Tavira	27523
29	Anadia	27532	98 Fr. de Esp Cinta	3216	168 Oliveira do Bairro	23132	238 Terras de Bouro	6358
30	Ansião	11642	99 Fronteira	2858	169 Oliv. do Hospital	19413	239 Tomar	36413
31	Arganil	11065	100 Fundão	26503	170 Ourém	44538	240 Tondela	25910
32	Armamar	5678	101 Gavião	3394	171 Ourique	4839	241 Torre de Moncorvo	6826
33	Arouca	21146	102 Góis	3811	172 Ovar	54953	242 Torres Novas	34111
34	Arraiolos	6606	103 Golegã	5400	173 Paços de Ferreira	55595	243 Torres Vedras	83072
35	Arronches	2789	104 Gondomar	164257	174 Palmela	68852	244 Trancoso	8413
36	Ar. dos Vinhos	13992	105 Gouveia	12222	175 Pamp. da Serra	4082	245 Trofa	38548
37	Aveiro	80954	106 Grândola	13822	176 Paredes	84354	246 Vagos	22886
38	Avis	3812	107 Guarda	40117	177 Paredes de Coura	8632	247 Vale de Cambra	21269
39	Azambuja	21421	108 Guimarães	156830	178 Pedrógão Grande	3390	248 Valença	13623
40	Baião	17534	109 Idanha-a-Nova	8355	179 Penacova	13113	249 Valongo	94672
41	Barcelos	116752	110 Ílhavo	39235	180 Penafiel	69629	250 Valpaços	14701
42	Barrancos	1438	111 Lagoa	23725	181 Pen. do Castelo	7333	251 Vendas Novas	11245
43	Barreiro	78345	112 Lagos	33494	182 Penamacor	4768	252 Viana do Alentejo	5318
44	Batalha	15557	113 Lamego	24312	183 Penedono	2738	253 Viana do Castelo	85778
45	Beja	33394	114 Leiria	128603	184 Penela	5440	254 Vidigueira	5175
46	Belmonte	6205	115 Lisboa	545796	185 Peniche	26429	255 Vieira do Minho	11955
47	Benavente	29709	116 Loulé	72332	186 Peso da Régua	14540	256 Vila de Rei	3279
48	Bombarral	12746	117 Loures	201590	187 Pinhel	8092	257 Vila do Bispo	5717
49	Borba	6428	118 Lourinhã	26240	188 Pombal	51170	258 Vila do Conde	80825
50	Boticas	5000	119 Lousã	17006	189 Ponte da Barca	11044	259 Vila Flor	6050
51	Braga	193324	120 Lousada	47364	190 Ponte de Lima	41164	260 Vila Franca de Xira	137529
52	Bragança	34582	121 Mação	6402	191 Ponte de Sor	15248	261 V. Nov. de Barquinha	7016
53	Cabec. de Basto	15558	122 M. de Cavaleiros	14251	192 Portalegre	22340	262 V. Nova de Cerveira	8921
54	Cadaval	13372	123 Mafra	86515	193 Portel	5747	263 V. Nov. de Famalicão	133534
55	Cald. da Rainha	50910	124 Maia	134977	194 Portimão	59845	264 V. Nova de Foz Côa	6304
56	Caminha	15797	125 Mangualde	18303	195 Porto	231800	265 Vila Nova de Gaia	303824
57	Campo Maior	8042	126 Manteigas	2909	196 Porto de Mós	23202	266 Vila Nova de Paiva	4662
58	Cantanhede	34212	127 M de Canaveses	49541	197 Póvoa de Lanhoso	21775	267 Vila Nova de Poiares	6803
59	Car. de Ansiães	5490	128 Marinha Grande	39024	198 Póvoa de Varzim	64255	268 Vila Pouca de Aguiar	11812
60	Carregal do Sal	9038	129 Marvão	3021	199 Proença-a-Nova	7167	269 Vila Real	49571
61	Cartaxo	23186	130 Matosinhos	172557	200 Redondo	6286	270 V. Real de S António	18824
62	Cascais	214124	131 Mealhada	19348	201 Reg. de Monsaraz	9871	271 Vila Velha de Ródão	3285

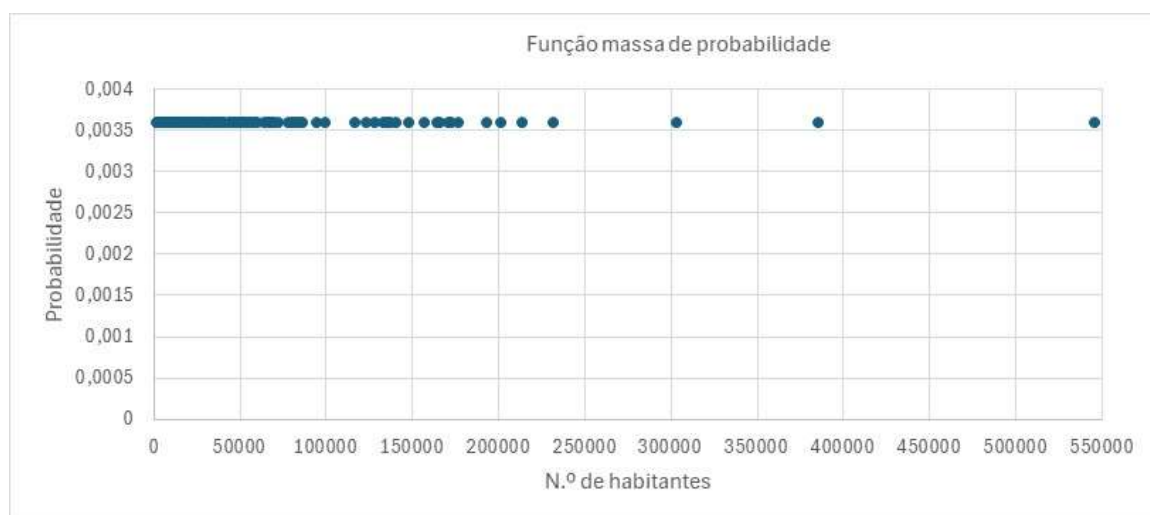


63	Cast. de Pêra	2645	132	Mêda	4630	202	Resende	10051	272	Vila Verde	46444
64	Castelo Branco	52272	133	Melgaço	7773	203	Ribeira de Pena	5884	273	Vila Viçosa	7387
65	Castelo de Paiva	15586	134	Mértola	6206	204	Rio Maior	21004	274	Vimioso	4149
66	Castelo de Vide	3116	135	Mesão Frio	3547	205	Sabrosa	5548	275	Vinhais	7768
67	Castro Daire	13736	136	Mira	12113	206	Sabugal	11280	276	Viseu	99551
68	Castro Marim	6439	137	Mir. do Corvo	12002	207	Salvat. de Magos	21607	277	Vizela	23896
69	Castro Verde	6873	138	Miran. do Douro	6463	208	Santa Comba Dão	10641	278	Vouzela	9580
			139	Mirandela	21384	209	S. Maria da Feira	136674			

Tabela 1 – N.º de habitantes, masculinos e femininos, em cada município de Portugal Continental

2. Função massa de probabilidade

A variável aleatória associada ao fenómeno aleatório, que consiste em selecionar um município ao acaso e verificar a sua dimensão populacional, assume os 278 valores constituídos pelas dimensões populacionais (todas diferentes) dos 278 municípios, cada um com probabilidade $\frac{1}{278}$, pelo que podemos dizer que segue o **modelo uniforme**, discreto, em 278 pontos. A f.m.p. é, assim, constituída pelas dimensões populacionais dos 278 municípios, cada um com probabilidade $\frac{1}{278}$ e com a seguinte representação gráfica:



Da representação anterior, pode-se concluir que as dimensões populacionais dos municípios se distribuem de forma extraordinariamente enviesada para a direita. Grande parte dos municípios tem dimensão inferior a 100 000 habitantes.

É usual utilizar barras para representar o valor das probabilidades. Tendo em conta o grande número de dados, optou-se, neste caso, por representar unicamente os pontos (Dimensão populacional do município, $\frac{1}{278}$), para os 278 municípios.

Modelo Uniforme discreto (ou Distribuição Uniforme discreta)

Dada uma variável aleatória X , discreta, assumindo os valores x_i , com $i = 1, 2, \dots, n$, diz-se que X segue o modelo Uniforme, ou tem distribuição Uniforme, nos n pontos, se e só se a função massa de probabilidade é dada por

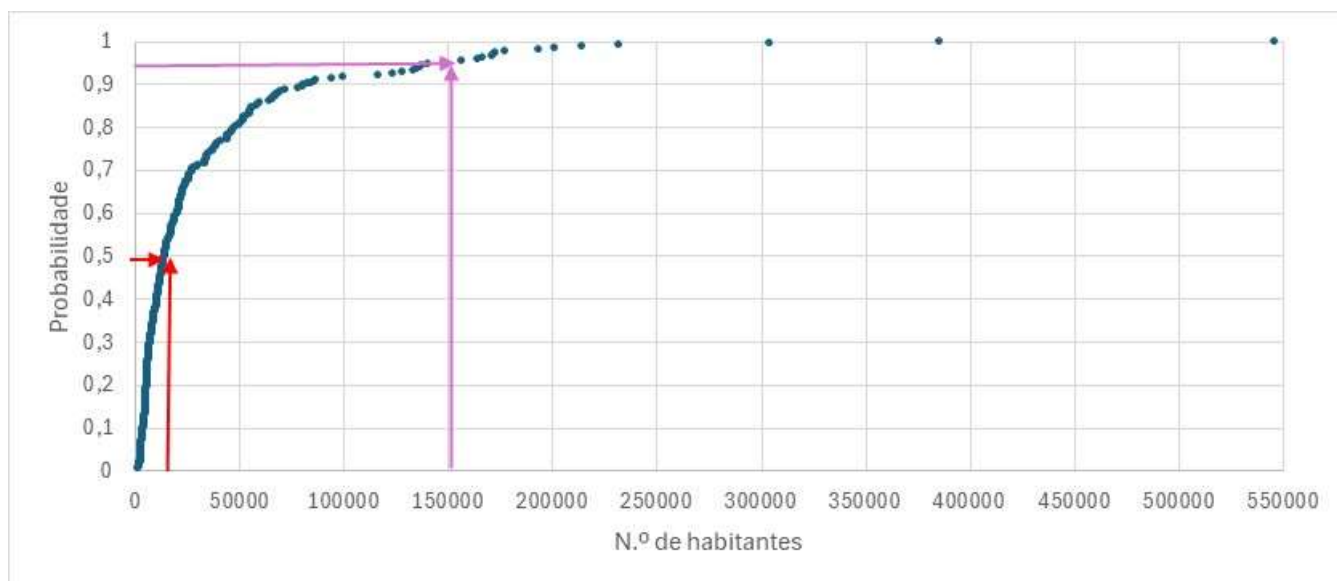
$$P(X = x_i) = \frac{1}{n} \quad i = 1, 2, \dots, n.$$

Um caso especial do Modelo Uniforme, discreto, é aquele em que os pontos x_i , estão igualmente afastados, ou seja, *uniformemente* distribuídos.

O exemplo dos municípios não é o caso mais usual do **modelo uniforme**. Embora este seja definido como o modelo matemático adequado para representar fenómenos aleatórios, cujo espaço de resultados tenha um número finito de resultados x_i com $i = 1, 2, \dots, n$, cada um com probabilidade $p_i = \frac{1}{n}$, na literatura surgem, sistematicamente, exemplos em que os resultados, quando numéricos, estão igualmente espaçados. O exemplo mais frequente é o associado ao lançamento do dado equilibrado de n faces e à modelação do fenómeno aleatório que consiste em observar o número de pintas i , $i = 1, 2, \dots, n$, da face que fica voltada para cima, atribuindo-lhe probabilidade $p_i = \frac{1}{n}$.

3. Função de probabilidades acumuladas

Uma representação gráfica que permite visualizar melhor a forma como, neste caso, os dados se distribuem, é a função das probabilidades acumuladas:

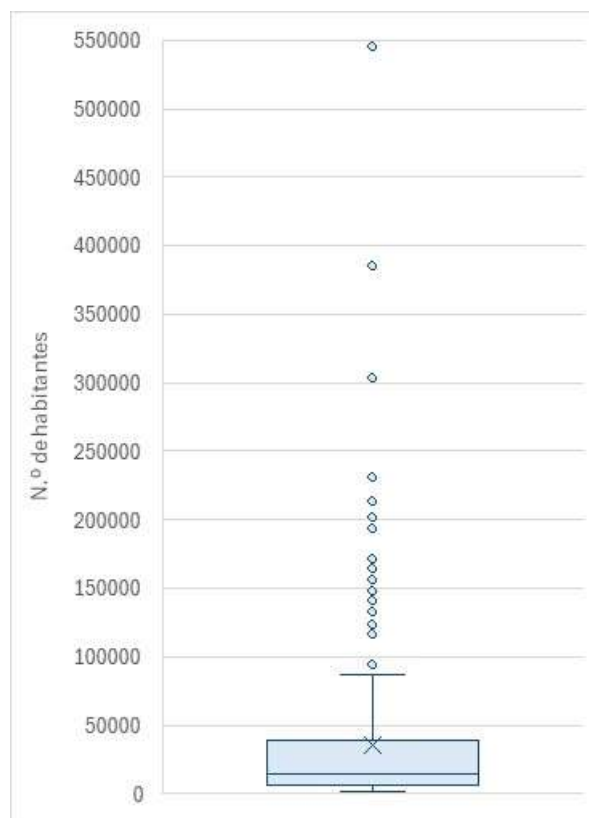


Algumas conclusões obtidas da representação anterior para a distribuição dos municípios, quanto ao número de habitantes:

- a) Cerca de 95% dos municípios têm um número de habitantes inferior a 150 000.
- b) 50% dos municípios têm dimensão populacional inferior a cerca de 15 000 habitantes.

4. Representação em caixa de bigodes

As conclusões anteriores podem ser confirmadas na representação em caixa de bigodes seguinte:



Esta representação permite concluir, como se esperava, que:

- a) A média – assinalada com x, é superior à mediana;
- b) Existem vários *outliers*, para valores grandes;
- c) Na parte inferior dos dados, a distribuição dos municípios é muito concentrada;
- d) A distribuição dos municípios tem enviesamento para a direita, tanto na parte central dos dados, como na cauda direita, sendo este bastante acentuado.

Estudo da distribuição do primeiro dígito da dimensão populacional dos municípios de Portugal Continental

1. População em estudo

A partir de uma tabela em Excel, com o número de habitantes de cada um dos municípios, obtida da Tabela 1, utilizou-se a função do Excel $Find("i";célula\ com\ dados)$, com $i = 1, 2, \dots, 9$. Por exemplo, $Find("4";B23)$, devolve a posição do dígito 4, no número da célula B23. Se o número não tiver o dígito 4, a função devolve a mensagem #VALUE!:

N.º	Dígito								
	1	2	3	4	5	6	7	8	9
20718	4	1	#VALUE!	#VALUE!	#VALUE!	#VALUE!	3	5	#VALUE!
34329	#VALUE!	4	1	2	#VALUE!	#VALUE!	#VALUE!	#VALUE!	5
46119	3	#VALUE!	#VALUE!	1	#VALUE!	2	#VALUE!	#VALUE!	5
5231	4	2	3	#VALUE!	1	#VALUE!	#VALUE!	#VALUE!	#VALUE!
5014	3	#VALUE!	#VALUE!	4	1	#VALUE!	#VALUE!	#VALUE!	#VALUE!
24840	#VALUE!	1	#VALUE!	2	#VALUE!	#VALUE!	#VALUE!	3	#VALUE!
44164	3	#VALUE!	#VALUE!	1	#VALUE!	4	#VALUE!	#VALUE!	#VALUE!
11112	1	5	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
12472	1	2	#VALUE!	3	#VALUE!	#VALUE!	4	#VALUE!	#VALUE!
54965	#VALUE!	#VALUE!	#VALUE!	2	1	4	#VALUE!	#VALUE!	3
19143	1	#VALUE!	5	4	#VALUE!	#VALUE!	#VALUE!	#VALUE!	2
2523	#VALUE!	1	4	#VALUE!	2	#VALUE!	#VALUE!	#VALUE!	#VALUE!
44442	#VALUE!	5	#VALUE!	1	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
4324	#VALUE!	3	2	1	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
10486	1	#VALUE!	#VALUE!	3	#VALUE!	5	#VALUE!	4	#VALUE!

Depois de aplicar a função Find a todos os municípios, utilizou-se a função Countif("1";coluna), para obter a frequência absoluta com que cada dígito ocupa a primeira posição, tendo-se obtido os seguintes valores:

Frequência absoluta do dígito i , $i = 1, 2, \dots, 9$, na primeira posição

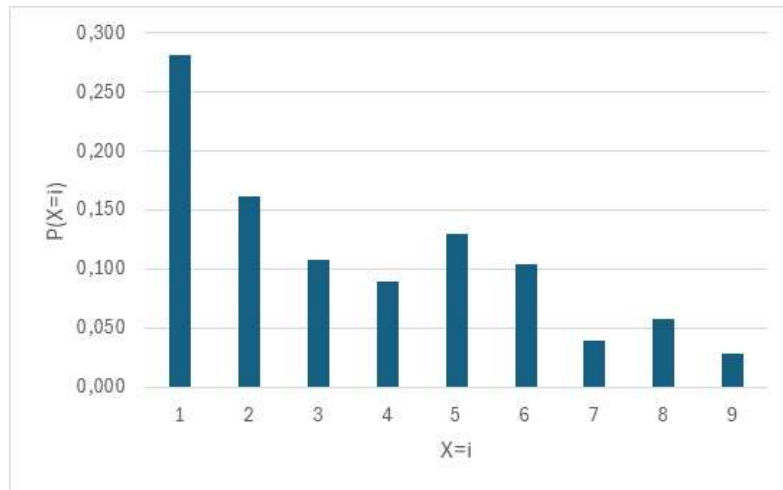
1	2	3	4	5	6	7	8	9
78	45	30	25	36	29	11	16	8

Tabela 2

2. Função massa de probabilidade

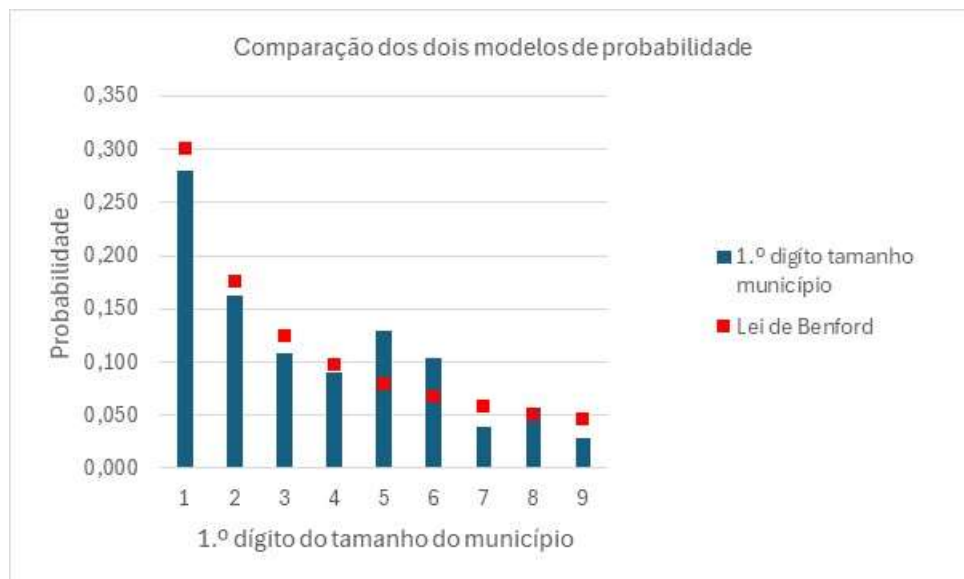
Representando por X a v.a. que representa o dígito na primeira posição, no número de habitantes de cada município, a partir da tabela 2, que apresenta o número de casos favoráveis para cada um dos valores de X , e tendo em consideração que o número de casos possíveis são 278, tem-se a seguinte f.m.p. para X :

$X=i$	1	2	3	4	5	6	7	8	9
$P(X=i)$	0,281	0,162	0,108	0,090	0,129	0,104	0,040	0,058	0,029



Da representação anterior ressalta que a probabilidade do dígito 1 é muito superior à do dígito 2, que é o segundo com maior frequência. Existem algumas semelhanças com o modelo de Benford, embora não se consiga um ajustamento perfeito – lembremo-nos de que, no modelo de Benford, as probabilidades são todas decrescentes, por serem dadas pela função logaritmo, que decresce com o aumento do argumento.

No gráfico seguinte, são apresentadas as diferenças entre os dois modelos:



tamanho \approx dimensão populacional