Bulletin UASVM Agriculture, 66(1)/2009
Print ISSN 1843-5246; Electronic ISSN 1843-5386

Distribution Fitting 4. Benford test on a sample of observed genotypes number from running of a genetic algorithm

Lorentz JÄNTSCHI[1], Sorana D. BOLBOACĂ[2], Carmen E. STOENOIU[1], Mihaela IANCU[2], Monica M. MARTA[2], Elena M. PICĂ[1], Monica ŞTEFU[1], Adriana F. SESTRAŞ[3], Marcel M. DUDA[3], Radu E. SESTRAŞ[3], Ştefan ŢIGAN[2], Ioan ABRUDAN[1], Mugur C. BĂLAN[1]

[1] Technical University of Cluj-Napoca, 103-105 Muncii Bvd., 400641 Cluj, Romania
[2] Iuliu Haţieganu University of Medicine and Pharmacy, Cluj-Napoca, 400349 Cluj, Romania
[3] University of Agricultural Sciences and Veterinary Medicine of Cluj-Napoca, 3-5 Mănăştur St., 400372, Cluj, Romania
correspondence: lori@academicdirect.org

Abstract. A new designed genetic algorithm was run on experimental data obtained from measurements of octanol-water partition coefficients of a series of polychlorinated biphenils, in order to relate their structure with their activity. A family of molecular descriptors having all necessary ingredients to run a genetic algorithm on it characterized the structure. An experiment using different selection and survival strategies were conducted, when multiple runs were recorded and their results were analyzed. Total number and number of distinct of genotypes present in the generations leading to evolution were included in an analysis concerning the validity of data with Benford test, the methodology and the obtained results being given.

Keywords: distribution fitting; statistical agreement; genotypes number; genetic algorithm; survival strategy; selection strategy; Benford statistic

INTRODUCTION

The Benford test uses the normal distribution to check if for an array of numbers their digits follow the Benford distribution.

The hypothesis of the test is that the values of the observations measurements are often logarithmically distributed and thus the logarithm of the measurement set is uniform distributed. The distribution, test, and its statistic are called after the physician Frank BENFORD, who discovered first and it formulated intuitively (Benford, 1938), inspired in its survey by a short communication of Simon NEWCOMB (Newcomb, 1881). The proof of the distribution it comes later being given by Theodore P. HILL (Hill, 1995).

This intuitively result of counting digits occurrences of the numbers was found true to a large variety of datasets, including electricity bills (Christian and others, 1993), forensic and financial audits (de Marchi and Hamilton, 2006; Nigrini and Mittermaier, 1997), stock exchanges (Ley, 1996), river basin area, weights of chemicals and streets addresses (Benford, 1938), roundoff errors (Barlow and Bareiss 1985), population numbers (Sandron, 2002), ceasing rates (Leemis and others, 2000), physical and mathematical constants and a lot of processes described by power lows common in nature (Newcomb, 1881; Berger and Hill, 2007).

An important result is that the result (once observed when the number is expressed in a numeration base) is independent of the numeration base in which the numbers were expressed, even if the proportions of representation are changing (Pinkham, 1961; Hill, 1995).

A natural consequence of the existence of the Benford law is that this fact can be used to validate the reported data under the presumption of the altering (mystification) of them, the immediate approach being the comparing of the observed first digit frequencies with the theoretical ones (Diekmann, 2007; Günnel and Tödter, 2008).

The Benford test were run on the results giving numbers of alive genotypes obtained from runs of a genetic algorithm searching on structure-activity relationships between the structure of a series of 206 biphenyl polychlorinated compounds and their observed octanol-water partition coefficient, in order to test if these numbers follows the Bendford law.

MATERIALS AND METHODS

In forty-six independent runs for every combination of survival method and selection method from Deterministic, Tournament and Proportional types were recorded the evolutions of a genetic algorithm (Jäntschi, 2009). On these data, for twenty equal width classes of observation, from generation 1 to the generation 20000, the total number of genotypes and the number of distinct genotypes were grouped and the observed frequencies were obtained (Table 1).

Table 1. Observed frequencies of the viable genotypes in the generations leading to evolution from forty-six independent runs for different selection and survival strategy

SelSrv	Millennium	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
PP	num_obs	732	141	97	62	36	44	25	30	35	32	29	16	20	15	8	7	12	21	17	20
PP	sum_obs	8366	1580	1073	700	417	500	285	339	390	358	330	186	221	166	92	84	134	245	201	235
PT	num_obs	713	171	133	89	54	44	40	33	32	32	11	28	32	26	10	21	6	8	11	15
PT	sum_obs	8207	1960	1505	1013	609	505	455	383	363	369	127	313	362	279	111	239	69	90	128	161
PD	num_obs	748	159	101	85	60	71	24	25	28	21	22	33	24	25	10	15	15	22	14	18
PD	sum_obs	8825	1858	1166	990	702	829	273	286	326	243	257	384	278	292	116	171	176	256	162	207
TP	num_obs	655	159	110	88	57	31	37	33	34	27	25	23	17	9	7	8	9	12	10	13
TP	sum_obs	7470	1741	1202	992	639	331	410	363	372	300	278	262	191	100	79	92	99	134	115	147
TT	num_obs	740	178	92	62	56	37	38	31	21	23	19	25	14	17	18	25	9	1	7	12
TT	sum_obs	8475	1988	1025	700	625	424	427	352	237	248	218	269	159	195	199	278	102	12	82	136
TD	num_obs	757	110	98	83	61	66	49	39	21	31	29	25	31	16	5	4	14	9	11	16
TD	sum_obs	8969	1282	1145	963	704	769	569	453	244	364	333	291	364	187	59	48	165	105	129	188
DP	num_obs	422	94	44	33	43	28	15	43	31	15	15	17	12	17	16	9	15	17	6	5
DP	sum_obs	4739	987	450	349	467	289	150	451	344	167	166	183	116	163	157	99	150	190	68	57
DT	num_obs	431	110	65	51	48	39	38	45	36	43	14	25	29	13	8	16	8	10	11	2
DT	sum_obs	4883	1223	719	558	533	440	411	477	385	468	153	273	312	139	92	169	87	111	119	20
DD	num_obs	466	120	81	66	53	41	24	39	51	24	11	19	18	23	23	14	19	12	24	2
DD	sum_obs	5511	1402	949	772	627	486	283	459	596	283	130	225	212	271	270	166	222	143	285	24

SelSrv: 46 runs using Sel and Srv as selection and survival strategies;
Srv, Sel $\in$ {P, T, D}; P - Proportional strategy; T - Tournament strategy; D - Deterministic strategy;
Millennium = 1000 generations; num_obs: number of distinct genotypes; sum_obs: total number of genotypes;

First three digits of the numbers from Table 1 were included into the analysis of the agreement with Benford distribution.

The theoretical probabilities of the Benford distribution for first (d_0, Eq.1), second (d_1, Eq.2) and third (d_2, Eq.3) digits are given by following relationships:

$$p(d_0) = \log_b\left(1 + 1/d_0\right),\ d_0 = 1..(b-1) \quad (1)$$

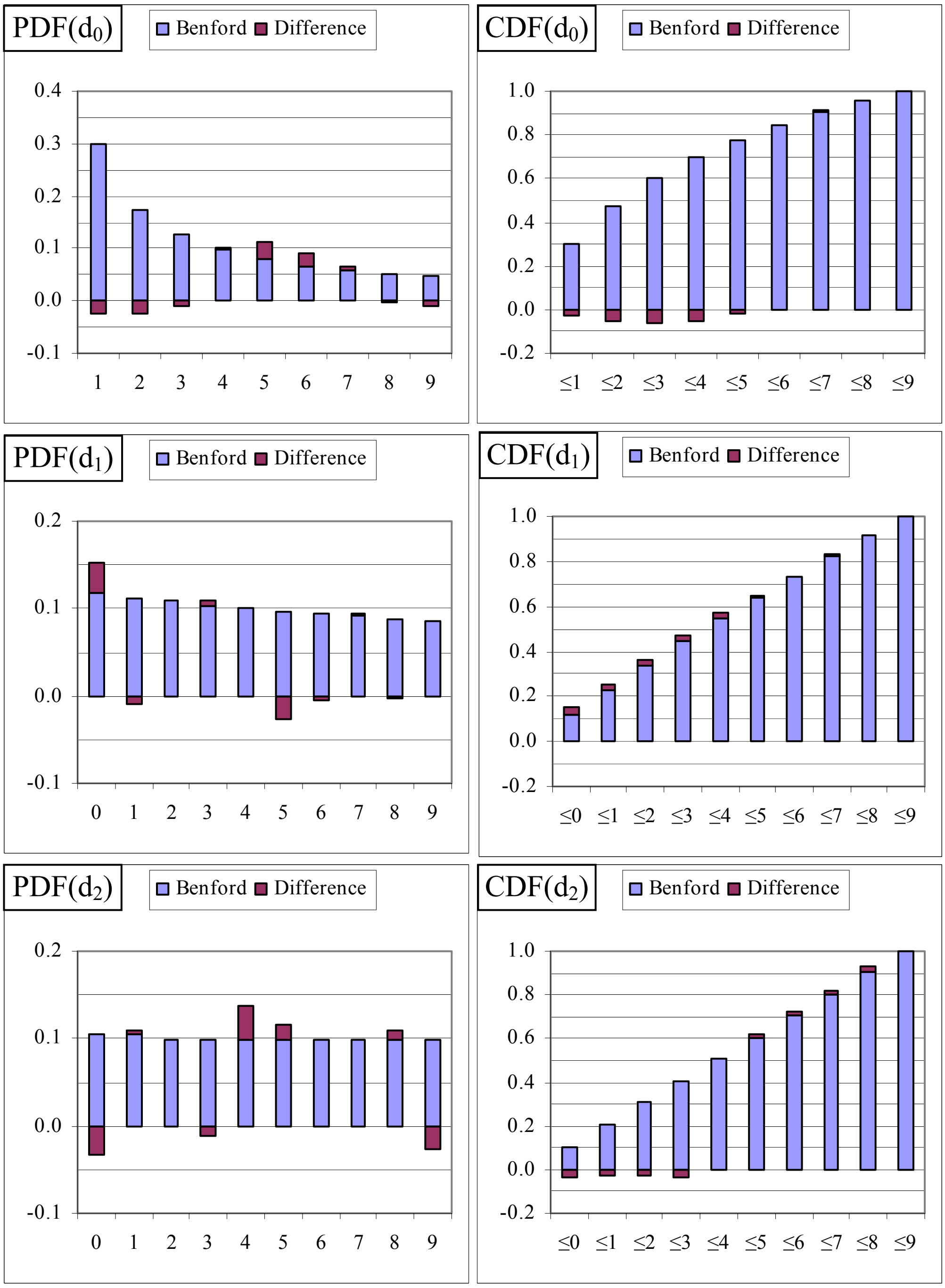
$$p(d_1) = \sum_{k=1}^{b-1} \log_b\left(1 + 1/(k \cdot b + d_1)\right),\ d_1 = 0..(b-1) \quad (2)$$

$$p(d_2) = \sum_{j=1}^{b-1}\sum_{k=0}^{b-1}\log_b\left(1+1/(j\cdot b^2+k\cdot b+d_2)\right),\ d_2 = 0..(b\text{-}1) \quad (3)$$

where b is the numeration base (b = 10 for the data given in Table 1).

RESULTS AND DISCUSSION

The theoretical (Benford) and the observed relative differences are given in Figure 1.



Benford: Theoretical distribution; Difference: observed difference relative to theoretical

Figure 1. Probability distribution functions (PDF) and cumulative distribution functions (CDF) for first three digits (d_0, d_1, and d_2 respectively) of the data from Table 1

The plots from Figure 1 show a small disagreement between observation and the model of the Benford distribution. In order to measure its observing probability, two statistics were involved: Chi Square (Table 2) and Kolmogorov-Smirnov (Table 3); their results are analyzed in the next.

Benford distribution does not have unknown (to be estimated) parameters. The numeration base (Eq.1-3) is known being 10, and thus the degrees of freedom are number of digits minus one (Table 2).

Table 2. χ^2 test on observed frequencies of the data from Table 1 (null hypothesis: first, second, and third digits follows Benford distribution)

Digit (i)			Expected frequency (E_i)			Observed frequency (O_i)			$(\lvert O_i-E_i\rvert-0.5)^2$			$(\lvert O_i-E_i\rvert-0.5)^2/E_i$		
d_0	d_1	d_2	d_0	d_1	d_2	d_0	d_1	d_2	d_0	d_1	d_2	d_0	d_1	d_2
0	0	0	-	40	19	-	28	25	-	144	36	-	3.31	1.59
1	1	1	108	38	19	117	41	18	81	9	1	0.67	0.16	0.01
2	2	2	63	37	18	72	37	18	81	0	0	1.15	0.01	0.01
3	3	3	45	35	18	48	33	20	9	4	4	0.14	0.06	0.13
4	4	4	35	34	18	33	34	11	4	0	49	0.06	0.01	2.35
5	5	5	29	33	18	17	42	15	144	81	9	4.56	2.19	0.35
6	6	6	24	32	18	16	34	18	64	4	0	2.34	0.07	0.01
7	7	7	21	31	18	18	30	18	9	1	0	0.30	0.01	0.01
8	8	8	18	30	18	19	31	16	1	1	4	0.01	0.01	0.13
9	9	9	17	29	18	20	29	23	9	0	25	0.37	0.01	1.13
Σ (df=8)	Σ (df=9)	Σ (df=9)	360	339	182	360	339	182	402	244	128	9.60	2.53	4.12
$X^2(d_0)$=9.6; $p\chi^2$(9.6,8)=29.4%; $X^2(d_1)$=2.53; $p\chi^2$(2.53,9)=98.0%; $X^2(d_2)$=4.12; $p\chi^2$(4.12,9)=90.3%;														

Table 3. Kolmogorov-Smirnov test on observed cumulative frequencies of the data from Table 1 (null hypothesis: first, second, and third digits follows Benford distribution)

Digit			Expected (d_0e, d_1e, d_2e) and Observed (d_0o, d_1o, d_2o)						Difference			| Difference |		
d_0	d_1	d_2	d_0a	d_1a	d_2a	d_0o	d_1o	d_2o	d_0	d_1	d_2	d_0	d_1	d_2
0	0	0	0	40	19	0	28	25	0	12	-6	0	**12**	6
1	1	1	108	78	38	117	69	43	-9	9	-5	9	9	5
2	2	2	171	115	56	189	106	61	-18	9	-5	18	9	5
3	3	3	216	150	74	237	139	81	-21	11	-7	**21**	11	**7**
4	4	4	251	184	92	270	173	92	-19	11	0	19	11	0
5	5	5	280	217	110	287	215	107	-7	2	3	7	2	3
6	6	6	304	249	128	303	249	125	1	0	3	1	0	3
7	7	7	325	280	146	321	279	143	4	1	3	4	1	3
8	8	8	343	310	164	340	310	159	3	0	5	3	0	5
9	9	9	**360**	**339**	**182**	360	339	182	0	0	0	0	0	0
Σ	Σ	Σ	-	-	-	-	-	-	-66	55	-9	82	55	37
$D(d_0)$=21; $K(d_0)=21\sqrt{9}/360$; pKS(9,$21\sqrt{9}/360$)=90.8%; $D(d_1)$=12; $K(d_1)=12\sqrt{10}/339$; pKS(10, $12\sqrt{10}/339$)=95.2%; $D(d_1)$=7; $K(d_1)=7\sqrt{10}/182$; pKS(10, $7\sqrt{10}/182$)=94.6%;														

High observation probabilities results from Chi Square test. A geometric mean of 64% of all three probabilities gives 64% probability to observe a worst agreement between the observed data following the Benford distribution and the theoretical Benford distribution under Chi-Square test hypothesis. Thus the hypothesis of Benford distribution of the digits from Table 1 cannot be rejected by using Chi Square statistic. Even higher probabilities results from Kolmogorov-Smirnov test. A geometric mean of 94% of all three probabilities gives 94% probability to observe a worst agreement between the observed data following the Benford distribution and the theoretical Benford distribution under Kolmogorov-Smirnov test hypothesis. Thus the hypothesis of Benford distribution of the digits from Table 1 cannot be rejected by using Kolmogorov-Smirnov statistic.

The only one remained question about the analysis of the data from table 1 is regarding the difference between observation probabilities given by Chi Square test (64%) and Kolmogorov-Smirnov test (94%).

It's well known () that the Chi Square test assumes a normality distribution of the errors (squared in Table 2, $(|O_i-E_i|-0.5)^2$ column). A measure of the departure from normality may be given by the Jarque-Bera statistic (Jarque and Bera, 1981), as a measure of the sufficiency (Fisher, 1922) for Chi Square and Kolmogorov-Smirnov statistics. Under assumption of the Gauss distribution of the differences (when Chi Square has highest accuracy) the expected population kurtosis is 3, and is 6 under assumption of Laplace distribution of the differences (skewness expectation being zero). The kurtosis, the skewness and the Jarque-Bera statistic under two assumptions (Gauss and Laplace) for the differences from Table 2 are given in Table 4.

Table 4. Kurtosis and skewness of the disagreement between observed and the model

Disagreement	d_0	d_1	d_2
Kurtosis	2.13	4.60	2.71
Skewness	0.28	0.81	0.07
Jarque-Bera(Gauss)	0.41; $p\chi^2(0.41,2)$=82%	2.16; $p\chi^2(2.16,2)$=34%	0.04; $p\chi^2(0.04,2)$=98%
Jarque-Bera(Laplace)	25.1; $p\chi^2(25.1,2)=10^{-6}$	4.33; $p\chi^2(4.33,2)$=11%	18.1; $p\chi^2(18.1,2)=10^{-4}$

Table 4 shows that the largest departure between the results obtained from Chi Square statistic and from Kolmogorov-Smirnov statistic is expected to be at d_0, followed by d_2 and the lowest difference should be at d_1.

Indeed, comparing the probabilities given in Table 2 with the ones given in Table 3, largest departure is at d_0 (29% from Chi Square, Table 2; 90.8% from Kolmogorov-Smirnov, Table 3; difference: 61.4%), followed by d_2 (90.3% from Chi Square, Table 2; 94.6% from Kolmogorov-Smirnov, Table 3; difference: 4.3%) and by d_1 (98.0% from Chi Square, Table 2; 95.2% from Kolmogorov-Smirnov, Table 3; difference: 2.8%).

CONCLUSIONS

With a good confidence, given by the probability to observe the observed departures from Benford distribution, the numbers giving the total number and the number of distinct genotypes from independent runs of a genetic algorithm by using different selection and survival strategies follows Benford law.

Since (and when in general) the kurtosis of the differences between observation and the model shows closeness to the Gauss than the Laplace distribution, a more weight to the Chi Square statistic should be assigned when remarks regarding the probability of observation are made, and vice versa.

Acknowledgments. Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Carmen E. STOENOIU, and Mihaela IANCU acknowledge the financial support for conducting this study to the UEFISCSU Romania funding agency (Grant IDEAS, no. 1051/2007, 2007-2010).

REFERENCES

Barlow, J. L. and E. H. Bareiss (1985). On Roundoff Error Distributions in Floating Point and Logarithmic Arithmetic. Computing 34:325-347.

Benford, F. (1938). The law of anomalous numbers. Proceedings of the American Philosophical Society 78(4):551-572.

Berger, A. and T. P. Hill (2007). Newton's method obeys Benford's law. American Mathematical Monthly 114(7):588-601.

Christian, C. W., S. Gupta and S. M. Lin (1993), Determinants of Tax Preparer Usage - Evidence From Panel-Data. National Tax Journal 46(4):487-503.

de Marchi, S. and J. T. Hamilton (2006). Assessing the accuracy of selfreported data: An evaluation of the toxics release inventory. Journal of Risk and Uncertainty 32(1):57-76.

Diekmann, A. (2007). Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data. Journal of Applied Statistics. 34(3):321-329.

Fisher, R. A. (1920). A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error. Monthly Notices of the Royal Astronomical Society 80(S):758-770.

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. Philosophical Transactions of the Royal Society A 222:309-368.

Günnel S. and K.-H. Tödter (2008). Does Benford's Law hold in economic research and forecasting? Empirica, DOI: 10.1007/s10663-008-9084-1.

Hill, T. P. (1995). Base invariance implies Benford's Law. Proceedings of the American Mathematical Society 123(3):887-895.

Hill, T. P. (1995). Base-Invariance Implies Benford's Law. Proceedings of the American Mathematical Society 123(3):887-895.

Jäntschi, L. (2009). A genetic algorithm for structure-activity relationships: software implementation. Manuscript:
http://arxiv.org/abs/0906.4846 (abstract), http://arxiv.org/pdf/0906.4846 (PDF).

Jarque, C. M. and A. K. Bera. (1981). Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. Economics Letters 7(4):313-318.

Leemis, L. M., B. W. Schmeiser and D. L. Evans (2000). Survival distributions satisfying Benford's law. The American Statistician 54(4):236-241.

Ley, E. (1996). On the Peculiar Distribution of the U.S. Stock Indices Digits. American Statistician 50:311-313.

Newcomb, S. (1881). Note on the Frequency of the Use of Digits in Natural Numbers. American Journal of Mathematics 4:39-40.

Nigrini, M. J. and L. J. Mittermaier (1997). The use of Benford's Law as an aid in analytical procedures. Auditing-A Journal of Practice & Theory 16(2):52-67.

Pinkham, R. S. (1961). On the distribution of first significant digits. Annals of Mathematical Statistics 32(4):1223-1230.

Sandron, F. (2002). Do Populations Conform to the Law of Anomalous Numbers?. Population 57(4-5):755-762.