Statistica & Applicazioni - Vol. XVIII, n. 2, 2020, pp. 125-136

ON EVALUATION OF SAMPLE SIZE TO TEST HYPOTHESIS ON BENFORD'S DISTRIBUTION

Janusz L. Wywiał*

SUMMARY

Benford's distribution,also known as first-digit law, is used to detect several kinds of frauds on the basis of data. Several statistical tests are used to verify that real-life sources of data have Benford's distribution. In this paper we consider in detail the well-known chi-square test and the Kolmogorov test of goodness of fit. The considerations are focused on determining the sample size that provides the assumed significance level as well as the power of the test. The necessary sample size is evaluated on the basis of simulation analysis under reasonable formulated alternative distributions to Benford's distribution and under assumed significance levels and powers of the statistical tests.

Keywords: Auditing, Chi-Square Test, Kolmogorov Test, Power, Fraud Detection.
DOI: 10.26350/999999_000038
ISSN: 18246672 (print) 2283-6659 (digital)

1. INTRODUCTION

Testing the hypothesis on googdnees of fit of the data distribution with the Benford distribution usually considered a preliminary step before the actual more complex digital analysis of the data or eventually including financial audit.Benford's law refers to the probability distribution of digits and it is used-for the detection of, among other, tax fraud, environmental law compliance and campaign finance, and for determining the reliability of survey data. For instance Carslaw (1988) considered Benford's law in the field of accounting. Özera and Babacanb (2013) used the chi-square test to verify the hypothesis that the distribution of Turkish banks' financial data is consistent with the Benford distribution. In general, this allowed an overall assessment of the quality of banking data. Diekmann (2007) used Benford's law to analyse the relibility of assessments values of e.g. regression coefficients estimated in sociological scientific studies. He showed that the second significant digits of the data do not follow the Benford distribution. Another examples of Benford's law applications are considered e.g. by Cho and Gaines (2007), Giles (2007), Hales, Sridharan, Radhakrishnan, Chakravorty, and Siha, 2008; Hales, Chakravorty and Sridharan, 2009), Judge and Schechter (2009), Rauch, Göttche and Brähler (2011).

* Department of Statistics, Econometrics and Mathematics - University of Economics in Katowice - street: 1 Maja 50, 40 - 287 KATOWICE, Poland (e-mail: wywial@ue.katowice.pl).

Moreover let us note that Pinkham (1961) has shown that Benford's law is invariant with the change of scale. Larger review of Benford's law properties as well as literature was shown by Nigrini (2012).

In practice there are considered several statistical tests for the hypotheses on Benford distribution. But they let us control only only significant level of the tests. Therefore the main purpose of this paper is evaluating the sample size under the assumed significance level and the power (probability that the error of the second type do not occur). That is why we have to specify alternative distributions to the Benford's one. These distributions are shown in the Table 1.

Let the values $d = 1, ..., 9$ of the random variable D be the first possible digits. The probability function of D is as follows (see Benford (1938) and Newcomb (1881)):

$$P(D = d) = p_{B,d} = log_{10}\left(1 + \frac{1}{d}\right). \tag{1}$$

The tested hypotheses are as follows:

$$H_0 : P(D = d) = p_{B,d}, \quad H_1 : P(D = d) = p_{i,d}, \quad \text{for all} \quad d = 1, ..., 9 \tag{2}$$

where $p_{i,d} \neq p_{B,d}$ for at least one d, is a fixed probability function of an alternative distribution to Benford's distribution, $i = 1, 2....$ Let us underline that in practice it is difficult to propose any concrete competitive distribution to Benford's one. Therefore the alternative hypothesis is usually complex and it is specified as follows:

$$H_1 : P(D = d) \neq p_{i,d}, \quad \text{for at least one} \quad d = 1, ..., 9.$$

TABLE 1. - *Benford's distribution and alternative distribution*

d	p_B	p_1	p_2	p_3	e_4	e_5	e_6	e_7
1	0.3010	$\frac{1}{9}$	0.272	$\frac{9}{45}$	-0.04	-0.020	-0.01	-0.005
2	0.1761	$\frac{1}{9}$	0.204	$\frac{8}{45}$	-0.03	-0.015	-0.01	-0.005
3	0.1249	$\frac{1}{9}$	0.152	$\frac{7}{45}$	-0.02	-0.010	-0.01	-0.005
4	0.0969	$\frac{1}{9}$	0.110	$\frac{6}{45}$	-0.01	-0.005	-0.01	-0.005
5	0.0792	$\frac{1}{9}$	0.085	$\frac{5}{45}$	0.00	0.000	0.00	0.000
6	0.0670	$\frac{1}{9}$	0.064	$\frac{4}{45}$	0.01	0.005	0.01	0.005
7	0.0580	$\frac{1}{9}$	0.048	$\frac{3}{45}$	0.02	0.010	0.01	0.005
8	0.0512	$\frac{1}{9}$	0.036	$\frac{2}{45}$	0.03	0.015	0.01	0.005
9	0.0458	$\frac{1}{9}$	0.027	$\frac{1}{45}$	0.04	0.020	0.01	0.005
	δ	0.402	0.031	0.09	0.30	0.019	0.01	0.003
	Δ	0.268	0.049	0.101	0.10	0.050	0.04	0.020

In Table 1, the probabilities of Benford's distribution and uniform distribution are denoted by p_B and p_1, respectively. The plots of the probabilities p_2 and p_3 are

similar to the exponential and triangular distributions, respectively. Finally, the probabilities $p_i = p_B + e_i$, $i = 4, 5, 6, 7$ represent specific deviations from Benford's distribution.

2. TESTS OF GOOGNEES OF FIT

The test statistic of the chi-square test is as follows (see e.g. Cramér (1946) or Santner and Duffy (1989)):

$$U_n = n \sum_{d=1}^{9} \frac{(w_d - p_{B,d})^2}{p_{B,d}} \tag{3}$$

where $w_d = \frac{n_d}{n}$. Under a sufficiently large sample size the statistic U_n has non-central chi-square distribution with 8 degrees of freedom with the non-centrality parameter $n\delta$ where:

$$\delta = \sum_{d=1}^{9} \frac{(p_d - p_{B,d})^2}{p_{B,d}}. \tag{4}$$

The parameter δ can be treated as the distance between Benford's distribution and the alternative distribution, see Table 1. The significance level and the power of the test are defined by $\alpha = P(U_n \geq u_\alpha | H_0)$ and $\beta = P(U_n \geq u_\alpha | H_1)$, respectively.

Several statistician (see e.g. Bergsen (1938) or Zenga (1978)) argues that the chi-square test of goodnees of fit has tendency to rejection of the tested hypothesis H_0 under large sample size. The properties of this test let us conclude that the probabilities α and β are close to 0 and 1, respectively. We could suppose that from practical point of view we can expect that some small departures from Benford's distribution could be admissible. Maybe in this case we should to consider rather the hypothesis defined as follows:

$$H_0 : \; p'_B \leq p_{B.d} \leq p''_B \;\text{ for }\; d = 1, ..., 9 \;\text{ and }\; p'_B < p_{B.d} < p''_B \;\text{ for at least one }\; d.$$

In general we can say that this hypothesis specifies the ''approximate'' Benford's distribution. The alternative hypothesis could be specified as the simple negation of H_0. These hypotheses are complex because each of them defines sets of distribution functions. This problem becomes more complex and needs studies in a separate paper.

The next test statistic is as follows:

$$Z_n = \sqrt{n} \sup_d |F_n(d) - F_B(d)| \tag{5}$$

where in our case $F_B(d) = \sum_{k<d} p_{B,k}$ is Benford's distribution function and $F_n(d) = \sum_{k<d} w_k$ is the sample distribution function. Kolmogorov (1933) derived the limit distribution of the statistic Z_n, but for cases when the hypothetical distribution function deals with the continuous random variables. In our case, $F_B(d)$ is the

distribution of the discrete random variable D. Hence, in our case the critical value of the test can not be determined by means of the limit distribution of the statistic Z_n. Due to this, in the next chapter the critical values of the test based on the statistic Z_n are evaluated on the basis of computer simulation experiments. This is known as Monte-Carlo test because the distribution of the test statistic is derived based on a computer simulation (see e.g. Hall and Titterington (1989) or Joenssen (2013)). The distance between Benford's distribution and the assumed alternative distribution, denoted by $F(d)$, is defined as follows:

$$\Delta = \sup_d |F(d) - F_B(d)| \tag{6}$$

where $F(d) = \sum_{k<d} P(D = d)$.The last two rows of Table 1 let us conclude that the distance (in the sense of both parameters δ and Δ) between the uniform distribution, denoted by p_1, and Benford's distribution is the largest. Benford's distribution and the distribution denoted by p_7 are the closest to each other.

3. EVALUATION OF THE SAMPLE SIZE

The minimum sample sizes for chi-square and Kolmogorov goodness of fit procedures are evaluated under the assumed significance level and the assumed test power. For the chi-square test, it is possible to derive the sample sizes on the basis of the limit distribution of the statistic U_n or on the basis of its simulated distribution.

TABLE 2. - *Sample sizes evaluated on the basis of the limit distribution of the chi-square test statistic*

α	β	p_1	p_2	p_3	p_4	p_5	p_6	p_7
0.10	0.90	40	520	190	220	860	1600	6390
0.05	0.90	50	620	220	260	1010	1900	7570
0.10	0.95	50	630	220	260	1040	1940	7760
0.05	0.95	60	740	260	310	1210	2260	9920
0.01	0.95	80	950	330	390	1550	2910	11620
0.05	0.99	80	980	340	410	1610	3010	14950
0.010	0.990	100	1220	430	500	2000	3740	14950
0.005	0.990	100	1310	460	540	2150	4030	16090
0.010	0.995	100	1320	460	550	2550	4070	16260
0.005	0.995	110	1420	500	590	2330	4350	17440
0.001	0.995	130	1630	570	670	2680	5010	20020
0.005	0.999	130	1660	580	690	2730	5100	20400
0.001	0.999	150	1880	660	780	3090	5790	23170

The first method of evaluating the sample size is as follows (see e.g. by Santner and Duffy (1989)). The distribution U_n, given by (3), is well approximated by the chi-square distribution when the sample size is large and $E(nw_d) \geq 5$, $d = 1, ..., 9$ (see e.g. Cochran (1952)). Hence, in our case, $E(nw_9) = n0.0458 \geq 5$ when $n \geq 110$. Therefore, under the assumption that the sample size is at least equal to 110, the algorithm is as follows. Firstly, under the assumed significance level the critical value of the chi-square test is determined. Next, the non-centrality coefficient $n\delta$ is calculated under the assumed alternative distribution specified by hypothesis H_1 (given by (2)), which lets us evaluate the power of the test on the basis of the non-central chi-square distribution. When the power is not less than the assumed level, the algorithm is stopped, otherwise the sample size is increased and the algorithm is repeated. Table 2 shows the results of determining the sample sizes for several alternative distributions and pairs (α, β) where α is the significance level of the test and β is its power.

TABLE 3. - *Sample sizes evaluated on the basis of the simulated distribution of the chi-square test statistic*

α	β	p_1	p_2	p_3	p_4	p_5	p_6	p_7
0.10	0.90	40	500	170	240	880	1580	6390
0.05	0.90	50	610	210	280	1060	1920	7620
0.10	0.95	50	610	210	290	1100	1960	7810
0.05	0.95	60	710	240	340	1280	2240	9970
0.01	0.95	70	920	320	430	1630	2970	11700
0.05	0.99	80	940	320	460	1720	3100	12160
0.010	0.990	100	1140	390	560	2120	3710	15080
0.005	0.990	100	1310	460	540	2150	4030	16090
0.010	0.995	100	1670	440	620	2340	4200	16410
0.005	0.995	110	1300	460	660	2420	4280	16830
0.001	0.995	130	1630	570	670	2680	5010	19980
0.005	0.999	130	1660	580	690	2730	5100	20400
0.001	0.999	150	1660	600	890	3200	5520	21590

In Table 3 there are sample sizes evaluated on the basis of the Monte-Carlo version of the chi-square test. The algorithm for evaluating sample size is similar to that explained above, except that the critical value of the test statistic is determined through at computer simulation experiment. This is equal to a quantile of order $1 - \alpha$ determined from a set of simulated $M = 100000$ values of the test statistic under the assumption that hypothesis H_0 is true. Those test statistic values are calculated on the

basis of independent samples (each of an assumed size) generated from Benford's distribution. More precisely, the sample frequencies w_d are generated from multinomial distribution under the assumed sample n and probabilities $p_{B,d}$, $d = 1, ..., 9$.

Next, the M values of the chi-square test statistic are generated under the assumption that the alternative hypothesis H_1 is true. The test statistic values are calculated on the basis of an independently generated sample from the multinomial distribution for the sample size n and the probabilities $p_{i,d} > 0$, $d = 1, ..., 9$, $\sum_{p=1}^{9} p_{i,d} = 1$, $i = 1, 2...$ The assessed power of the test is equal to the frequency of the test statistic values greater than the critical value previously estimated by means of the algorithm explained above. If the power is less than the assumed level β, the sample size is increased to the level $n + 10$ and the algorithm is repeated. When the power is not less than β, the sample size n is treated as optimal, which means that n is the minimal sample size of the test for significance level α and power β.

The above algorithms explained are implemented by means of a computer procedure written in R program. The samples of size n are replicated independently $M = 100000$ times. The results of the simulation procedures are shown in Tables 3 and 4.

TABLE 4. - *Sample sizes evaluated on the basis of the simulated distribution of the Kolmogorov test*

α	β	p_1	p_2	p_3	p_4	p_5	p_6	p_7
0.10	0.90	30	630	190	200	750	1300	5070
0.05	0.90	30	810	230	230	910	1570	6130
0.10	0.95	40	780	230	240	920	1610	6310
0.05	0.95	40	970	270	290	1110	1920	7490
0.01	0.95	60	1420	380	530	940	2570	10100
0.05	0.99	60	1250	380	420	1540	2610	10140
0.010	0.990	70	1730	490	530	2020	3380	13250
0.005	0.990	80	1940	540	580	2240	3710	14220
0.010	0.995	80	1890	550	590	2210	3680	14360
0.005	0.995	90	2120	580	630	2430	4020	15660
0.001	0.995	110	2580	720	760	2890	4660	14420
0.005	0.999	110	2420	680	770	2920	4760	18540
0.001	0.999	120	2980	800	890	3330	5530	21550

Analysis of Tables 2 and 3 let us say that the sample size evaluated based on limit distribution of the chi-square test statistic is significantly different from the appropriate sample sizes calculated based on the simulated distribution of the test sta-

tistic only in the case of the alternative distribution p_7 or for $\alpha = 0.001$ and $\beta = 0.999$. Hence, evaluating sample size using the limit distribution of the chi-square test is preferable because evaluating the critical value or p-value of the test is easier and more convenient than in the case of the simulation version of this test. The analysis of the tables let us say that this approach leads to very close to results obtained based on the simulation procedure.

Analysis of Tables 3 and 4 leads to the conclusion that the Kolmogorov test needs a smaller sample size than the chi-square test statistic but in the case of the alternative probability distributions denoted by p_6 or p_7. These distributions are closest to Benford's distribution in the sense of the distance coefficients δ and Δ. The departure of any probability distribution from the Benford's is when $\delta > 0$ or $\Delta > 0$. In general the analysis of Table 2 and 3 let us conclude that under the fixed α and β the sample size increases when coefficients $\delta > 0$ and $\Delta > 0$ decreases.

Let us consider testing the quality of commercial US-bank deposits from 2006 using data from the web: *http://www.census.gov/support/USACdataDownloads.html#HSD*. In detail, we test the hypothesis $H_0 : p = p_B$ under the significance level $\alpha = 0.05$ and $n = 3197$. The value of the chi-square test statistic is $u = 9.2992$. Hence, $u < u_\alpha = 15.5073$. This means that the sample distribution do not significantly differ from Benford's one.

Hypothesis H_0 could be reasonably accept provide the power of the test is assessed. In order to do this we had to consider alternative distributions. The analysis of Table 2 let us say that that e.g. under $\alpha = 0.05$ and $n = 3197$ the test could reach even the power on the level $\beta = 0.99$. Moreover this is valid for the several specified distributions, denoted b p_i, $i = 1, ..., 9$. When we take into account the alternative hypothesis $H_1 : p = p_7$, then the non-centrality parameter of the test statistic is: $\delta = 8.0651$ and the power of the test is: $\beta = 0.4804$. Hence, the sample size $n = 3197$ is not enough if we wish to reach the power on the level e.g. $\beta = 0.95$. It is possible for the larger sample size n'. Using the program shown in Appendix we can evaluate that $n' = 9030$. In this situation, it is necessary to observe more data about bank deposits.

4. CONCLUSIONS

Two popular tests have been considered in order to verify the hypothesis on Benford's probability distribution. The first one is the chi-square test and the second one is the Kolmogorov test. The minimum sample sizes were evaluated in such a way that the tests reached the assumed significance level α and the power β. The analysis leads to the following conclusions. In general in the case of alternative distributions not too close to Bemford's distribution or $\alpha > 0.001$ and $\beta < 0.999$ the necessary sample sizes can be evaluated on the basis of the limit distribution of the chi-square test statistic. Under alternative distributions very close to Benford's distribution Kolmogorov test needs a smaller sample size than the chi-square test. In many practical situations, the sample sizes of accounting data are rather large, the

significance levels taken into account are not very close to zero and the powers are not very close to one. Hence, it seems that implementing the chi-square test based on the limit distribution of the test statistic is more convenient than Kolmogorov test.

ACKNOWLEDGEMENTS

The project is supported by the grant of the National Science Centre, Poland, DEC-2012/07/B/HS4/03073.

APPENDIX

The R-program implementing procedure of the sample size evaluation in order to test hypothesis on Benford distribution under assumed the significance level and the power of the test. The following program is prepared based on the algorithm explained in Section 3.

```
# The chi-square test of goodness of fit for hypoth.
# on Benford's distrib.
# Evaluation of sample size (n) on the basis of simulated distribution
# of the test statistic under fixed significance level (a)
# and power (b);
# ls - number of replication,
# dn - increase of sample size,
# n - start sample size:
n=50; dn=5; ls=10000; k=9; a=0.1; b=0.95
#H0: p=p0; H1: p=p1
#Benford's prob. function:
p0=matrix(0,k,1);
for (i in 1:k) p0[i]=log10(1+1/i)
#uniform alternative distribution:
#p1=matrix(1/k,k,1)
#quasi-exponenntial alternative distributions with exp. value
# of Benford's distr.
EX=matrix(1:9,1,9)
#p1=as.matrix(dexp(1:9,1/EX)/sum(dexp(1:9,1/EX)))
#triangular alternative distrib.:
p1=matrix(c(9/45,8/45,7/45,6/45,5/45,4/45,3/45,2/45,1/45),k,1)
#other alternative distrib.:
p12=matrix(c(-.005,-.005,-.005,-.005,0,.005,.005,.005,.005),k,1)
#p12=matrix(c(-.01,-.01,-.01,-.01,0,.01,.01,.01,.01),k,1)
#p12=matrix(c(-.04,-.03,-.02,-.01,0,.01,.02,.03,.04),k,1)
#p12=matrix(c(-.02,-.015,-.01,-.005,0,.005,.01,.015,.02),k,1)
#p1=p0+p12
qs=matrix(0,ls,1)
ws=matrix(0,ls,1)
fitchi=function(n,p,w) {#evaluation of the test statistic
n*sum((w-p)*(w-p)/p)}
bs = 0; it = 0
while ((bs<b)&(it≤10000))
{n=n+dn; for (t in 1:ls) {ws=rmultinom(1,n,p0)/n;
qs[t]=fitchi(n,p0,ws) }
qs=sort(qs)
```

```
wk=qs[floor((1-a)*ls)+1]
bs=0
for (t in 1:ls) {ws= rmultinom (1,n,p1)/n;if(fitchi(n,p0,ws)>=wk)
bs=bs+1}
bs=bs/ls
it=it+1 }
#evaluated sample size:
n
#simulated critical value:
wk
#number of processed iteration:
it
```

For instance, when we assume that the triangular distribution is the alternative distribution and the significance level is equal to 0.1 and the power is equal to 0.95, then the above procedure provides that the sample size has to be not less than 210.

REFERENCES

Benford F. (1938). The low of anomalous data. *Proceedings of the American Philosophical Society*, **78**, 551-572.

Berkson J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, **33**, 526-536.

Carslaw C. (1988). Anomolies in income numbers: evidence of goal oriented behaviour. *The Accounting Review*, **63**, 321-327.

Cho W.K.T., Gaines B.J. (2007). Breaking the (Benford) law: Statistical fraud detection in campaign finance. *The American statistician*, **61**, 218-223.

Cochran W.G. (1952). The chi-squared test of goodnees of fit. Annals of Mathematical Statistics 23, 315-345.

Cramér H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Diekmann A. (2007). Not the first digit! Using Benford's Law to detect fraudulent scientific data. *Journal of Applied Statistics*, **34**, 321-329.

Giles D.E. (2007). Benford's law and naturally occurring prices in certain ebaY auctions. *Applied Economics Letters*, **14**, 157-161.

Özera G., Babacanb B., (2013). Benford's Law and Digital Analysis: Application on Turkish Banking Sector. *Business and Economics Research Journal*, **4**, 29-41.

Hales D.N., Sridharan V., Radhakrishnan A., Chakravorty S.S., Siha S.M. (2008). Testing the accuracy of employee-reported data: an inexpensive alternative approach to traditional methods. *European Journal of Operational Research*, **189**, 583-589.

Hales D.N., Chakravorty S.S., Sridharan V. (2009). Testing Benford's Law for improving supply chain decision-making: a field experiment. *International Journal of Production Economics*, **122**, 606-618.

Joenssen D.W. (2013). Benford Tests: Statistical Tests for Evaluating Conformity to Benford's Law. R package version 1.0.

Judge G., Schechter L. (2009). Detecting Problems in Survey Data using Benford's Law. *Journal of Human Resources*, **44**, 1-24.

Kolmogorov A.N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari* **4**, 83-91.

Newcomb S. (1881). Note on the frequency of use different digits in natural numbers, *American Journal of Mathematics*, **4**, 39-40.

Nigrini M.J. (2012). *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud detection*, John Wiley and Sons, Hoboken, New Jersey.

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

Rauch B., Göttche M., Brähler G. (2011). Fact and fiction in EU-Governmental Economic Data. *German Economic Review*, **12**, 243-255.

Santner T.J., Duffy D.E. (1989). *The Statistical Analysis of Discrete Data*. Springer-Verlag New York.

Zenga M. (1978). Suggerimenti per l'impiego del test Chi Quadrato. *Quaderni di Statistica e Matematica Applicata alle Scienze Economico-Sociali*, **1**(1), 3-16.