Journal of Mathematical Sciences: Advances and Applications
Volume 41, 2016, Pages 57-73
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/jmsaa_7100121718

BENFORD'S LAW AND HOSMER-LEMESHOW TEST

ZORAN JASAK

NLB Banka d.d.
Sarajevo
Bosnia and Herzegovina
e-mail: zoran.jasak@nlb.ba

Abstract

Benford's law is logarithmic law for distribution of leading digits. It's named by Frank Albert Benford [2] who formulated mathematical model. Before him, the same observation was made by Simon Newcomb. This law has changed usual preassumption of equal probability of each digit on each position in number. Testing procedure by Hosmer-Lemeshow test for Benford's law is presented. Such test can be, particularly, used to detect anomalies in samples of two or more partitions.

1. Introduction

In article "Note on the frequency of use of the different digits in natural numbers" [1], Simon Newcomb asserted that the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones. The first significant figure is oftener 1 than any other digit, and the frequency diminishes up to 9. Newcomb did not give mathematical explanation of this observation, just relative frequencies which were verified later [1].

2010 Mathematics Subject Classification: 62E10, 62Q99.
Keywords and phrases: Benford's law, Hosmer-Lemeshow test, data partitions, decils.
Received September 18, 2016

© 2016 Scientific Advances Publishers

The same phenomenon was re-discovered by Benford (1938) [2] who gave the mathematical formulation

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

In next table (Table 1), probabilities for first leading digits are presented.

Table 1. Probabilities of first leading digits

Digits	Probabilities
1	0.30103
2	0.17609
3	0.12494
4	0.09691
5	0.07918
6	0.06695
7	0.05799
8	0.05115
9	0.04576

This law is extended to groups of leading and non-leading digits.

Practical problem is how to test conformity to this law. In this paper, test deduced from Hosmer-Lemeshow test is proposed.

2. Hosmer-Lemeshow Test

2.1. Introduction

Hosmer-Lemeshow test is proposed as a tool to asses fit of the logistic regression model ([3], p. 147-156) in case when population is divided on two disjunctive subpopulations, partitions.

Goodness-of-fit statistic $\widehat{C}$ is obtained by calculating the Pearson chi-square statistic form $g \times 2$ table of observed and estimated expected frequencies. A formula defining the calculation of $\widehat{C}$ is:

$$\widehat{C} = \sum_{k=1}^{g} \frac{(O_k - n'_k \overline{\pi}_k)^2}{n'_k \overline{\pi}_k (1 - \overline{\pi}_k)}.$$

Here g is number of groups, n'_k is total number of subjects in the k-th group, c_k denotes the number of covariate patterns in the k-th decile,

$$O_k = \sum_{j=1}^{c_k} y_j$$

is the number of responses among the c_k covariate pattern and

$$\overline{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \widehat{\pi}_j}{n'_k}$$

is the average estimated probability.

Main preassumptions for this test are:

• Sample is divided on two separate subpopulations corresponding to cases of presence and absence of some property.

• Probabilities for covariance pattern, unique combination of values of predictor variables, are π_k and $1 - \pi_k$ for presence and absence of some property, respectively; their sum is 1 for k-th decile.

• Estimate of expected frequencies are $m_j \widehat{\pi}_j$ and $m_j (1 - \widehat{\pi}_j)$, respectively, for the cell corresponding to $y = 1$ and $y = 0$ rows.

• Sum of observed and expected frequencies for k-th decile are the same

$$O_{1k} + O_{0k} = E_{1k} + E_{0k} = N_g.$$

Here O_{1g}, E_{1g}, O_{0g}, E_{0g}, N_g denote sample $Y = 1$ values, expected $Y = 1$ values, sample $Y = 0$ values, expected $Y = 0$ values, number of observations in group g, respectively.

Central problem of this test is how to make groups of values. Hosmer and Lemeshow ([3], p. 148) proposed two strategies. With the first method, percentiles of risk, use of $g = 10$ groups result in the first group containing the $n_1' = n/10$ subjects having the smallest estimated probabilities and the last group containing the $n_{10}' = n/10$ subjects having the largest estimated probabilities. With the second method, use of $g = 10$ groups results in cutpoints defined at the values $k/10$, $k = 1, 2, \ldots, 9$ and the groups contain all subjects whose estimated probability between adjacent cutpoints.

Preferred strategy, by authors, ([3], p. 152) is to use deciles of risk.

2.2. Connection to Benford's law

Benford's law is known as a strong tool for detecting anomalies. There are numerous text concerning theoretical and practical issues of this law.

Usual approach in testing conformance to the Benford's law is to consider data as one sample, with no difference either any element belongs to any of two distinct subpopulations. Only criterion is leading or non-leading digit or groups of digits. There is a lot of examples in which such approach is both unpractical and has some deficiencies. This is specially case in finance and similar areas.

Suppose we want to analyse some financial data set (accounting, payments, ...) consisting of input (credit) and output (debit) transactions. It's common to merge those data into one sample and conduct statistical test. If we don't differ them in some way we can loose possible important information about anomalies on one, either credit or debit, side. We can test them separately but in this case we do not have whole context. Most of existing testing procedures, generally, do not consider such difference and treat them as they are merged in one group. For plausible investigation, it's important sometimes to make such difference for analyse them in some context, for example, detecting anomalies, money laundry etc.

One of possibilities is to use Hosmer-Lemeshow statistics, what is proposed in this paper.

According to Benford's law, we divide data in $G = 9$ groups, corresponding to leading digits $d \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Suppose data are divided in two partitions (subgroups), marked by $Y = 1$ (for credits) and $Y = 0$ (for debits), respectively.

For test we need next values:

- O_{jd} : number of observations in d-th group for partition $j \in \{0, 1\}$.

According to this is

$$N_j = \sum_{d=1}^{9} O_{jd}, \quad j \in \{0, 1\}.$$

- $O_d = O_{0d} + O_{1d}$: number of observations in d-th group. According to this is

$$N = \sum_{d=1}^{9} O_d = N_0 + N_1.$$

- b_d : Benford's theoretical probability for d-th group

$$b_d = P[D = d] = \log\left(1 + \frac{1}{d}\right).$$

- $E_d = E_{0d} + E_{1d}$: expected number of elements in the d-th group given Y, calculated by

$$E_{jd} = E(Y = j|d) = b_d \cdot N_j = N_j \cdot \log\left(1 + \frac{1}{d}\right),$$

$$d \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}, \quad j \in \{0, 1\}$$

$$\Rightarrow E_d = E_{1d} + E_{0d} = N \cdot b_d$$

$$N_j = \sum_{d=1}^{9} E_{jd} = \sum_{d=1}^{9} O_{jd}, \quad j \in \{0, 1\}.$$

Specificities for this case are:

- Probability for both partitions in group d is $b_d < 1$.

- Sum of expected number of cases in group d must not be equal to the sum of observed cases

$$E_{1d} + E_{0d} \neq O_{1d} + O_{0d}.$$

Appropriate statistic is [4]:

$$H = \sum_{g=1}^{G}\left[\frac{(O_{1g} - E_{1g})^2}{E_{1g}} + \frac{(O_{0g} - E_{0g})^2}{E_{0g}}\right].$$

After such preassumptions for $g = 9$ groups, we have

$$H_1 = \sum_{d=1}^{9}\left[\frac{(O_{1d} - E_{1d})^2}{E_{1d}} + \frac{(O_{0d} - E_{0d})^2}{E_{0d}}\right] = \sum_{d=1}^{9}\left[\frac{(O_{1d} - N_1 \cdot b_d)^2}{N_1 \cdot b_d} + \frac{(O_{0d} - N_0 \cdot b_d)^2}{N_0 \cdot b_d}\right]$$

$$= \sum_{d=1}^{9}\left[\frac{O_{1d}^2 - 2 \cdot O_{1d} \cdot N_1 \cdot b_d + N_1^2 \cdot b_d^2}{N_1 \cdot b_d} + \frac{O_{0d}^2 - 2 \cdot O_{0d} \cdot N_0 \cdot b_d + N_0^2 \cdot b_d^2}{N_1 \cdot b_d}\right]$$

$$= \sum_{d=1}^{9}\left[\frac{O_{1d}^2}{N_1 \cdot b_d} - 2 \cdot O_{1d} + N_1 \cdot b_d + \frac{O_{0d}^2}{N_0 \cdot b_d} - 2 \cdot O_{0d} + N_0 \cdot b_d\right]$$

$$= \sum_{d=1}^{9}\left[\frac{O_{1d}^2}{N_1 \cdot b_d} + \frac{O_{0d}^2}{N_0 \cdot b_d}\right] - 2.\sum_{d=1}^{9}(O_{1d} + O_{0d}) + (N_1 + N_0) \cdot \sum_{d=1}^{9} b_d$$

$$= \sum_{d=1}^{9}\left[\frac{O_{1d}^2}{N_1 \cdot b_d} + \frac{O_{0d}^2}{N_0 \cdot b_d}\right] - 2 \cdot N + N$$

$$= \sum_{d=1}^{9}\frac{O_{1d}^2}{N_1 \cdot b_d} + \sum_{d=1}^{9}\frac{O_{0d}^2}{N_0 \cdot b_d} - N. \tag{1}$$

Another two ways to write this formula are:

$$H_1 = \left(\sum_{d=1}^{9}\frac{O_{1d}^2}{N_1 \cdot b_d} - N_1\right) + \left(\sum_{d=1}^{9}\frac{O_{0d}^2}{N_0 \cdot b_d} - N_0\right); \tag{1.a}$$

$$H_1 = \left(\sum_{d=1}^{9} \frac{O_{1d}^2}{E_1} - N_1 \right) + \left(\sum_{d=1}^{9} \frac{O_{0d}^2}{E_0} - N_0 \right). \tag{1.b}$$

This statistic has χ^2 distribution with $G - 2$ degrees of freedom. We can consider that this statistic is sum of two statistics

$$H_{11} = \sum_{d=1}^{9} \frac{O_{1d}^2}{E_1} - N_1, \quad H_{10} = \sum_{d=1}^{9} \frac{O_{0d}^2}{E_0} - N_0.$$

If we do not make difference between partitions, we have next

$$H_2 = \sum_{d=1}^{9} \frac{(O_d - E_d)^2}{E_d} = \sum_{d=1}^{9} \frac{((O_{1d} + O_{0d}) - (N_1 + N_0) \cdot b_d)^2}{(N_1 + N_0) \cdot b_d}$$

$$= \sum_{d=1}^{9} \frac{[(O_{1d} - N_1 \cdot b_d) + (O_{0d} - N_0 \cdot b_d)]^2}{(N_1 + N_0) \cdot b_d}$$

$$= \sum_{d=1}^{9} \frac{(O_{1d} - N_1 \cdot b_d)^2 + 2(O_{1d} - N_1 \cdot b_d)(O_{0d} - N_0 \cdot b_d) + (O_{0d} - N_0 \cdot b_d)^2}{(N_1 + N_0) \cdot b_d}$$

$$= \sum_{d=1}^{9} \frac{(O_{1d} - N_1 \cdot b_d)^2 + (O_{0d} - N_0 \cdot b_d)^2}{(N_1 + N_0) \cdot b_d}$$

$$+ 2 \cdot \sum_{d=1}^{9} \frac{(O_{1d} - N_1 \cdot b_d)(O_{0d} - N_0 \cdot b_d)}{(N_1 + N_0) \cdot b_d}. \tag{2}$$

This statistic has χ^2 distribution with $G - 2$ degrees of freedom, in this case is $G = 9$. Second factor on the right side can be simplified in next way:

$$S = \sum_{d=1}^{9} \frac{(O_{1d} - N_1 \cdot b_d)(O_{0d} - N_0 \cdot b_d)}{(N_1 + N_0) \cdot b_d}$$

$$= \sum_{d=1}^{9} \left[\frac{O_{1d} \cdot O_{0d} - (N_1 \cdot O_{0d} + N_0 \cdot O_{1d}) \cdot b_d + N_1 \cdot N_0 \cdot b_d^2}{(N_1 + N_0) \cdot b_d} \right]$$

$$= \sum_{d=1}^{9} \left[\frac{O_{1d} \cdot O_{0d}}{N \cdot b_d} - \frac{N_1 \cdot O_{0d}}{N} - \frac{N_0 \cdot O_{1d}}{N} + \frac{N_1 \cdot N_0 \cdot b_d}{N} \right]$$

$$= \frac{1}{N} \sum_{d=1}^{9} \frac{O_{1d} \cdot O_{0d}}{b_d} - \frac{N_1}{N} \sum_{d=1}^{9} O_{0d} - \frac{N_0}{N} \sum_{d=1}^{9} O_{1d} + \frac{N_1 \cdot N_0}{N} \sum_{d=1}^{9} b_d$$

$$= \frac{1}{N} \sum_{d=1}^{9} \frac{O_{1d} \cdot O_{0d}}{b_d} - \frac{N_1 \cdot N_0}{N}$$

$$\Rightarrow H_2 = \sum_{d=1}^{9} \frac{(O_{1d} - N_1 \cdot b_d)^2 + (O_{0d} - N_0 \cdot b_d)^2}{(N_1 + N_0) \cdot b_d}$$

$$+ 2 \sum_{d=1}^{9} \left(\frac{O_{1d} \cdot O_{0d}}{N \cdot b_d} \right) - \frac{2 \cdot N_1 \cdot N_0}{N_1 + N_0}.$$

Last factor on right side is harmonic mean of N_1 and N_0.

Considering inequality

$$\frac{a+b}{x+y} \leq \frac{a}{x} + \frac{b}{y}, \quad a, b, x, y > 0,$$

from (2), we have

$$H_2 \leq \sum_{d=1}^{9} \left[\frac{(O_{1d} - N_1 \cdot b_d)^2}{N_1 \cdot b_d} + \frac{(O_{0d} - N_0 \cdot b_d)^2}{N_0 \cdot b_d} \right]$$

$$+ 2 \cdot \sum_{d=1}^{9} \frac{(O_{1d} - N_1 \cdot b_d)(O_{0d} - N_0 \cdot b_d)}{(N_1 + N_0) \cdot b_d},$$

or

$$H_2 \le \sum_{d=1}^{9}\left[\frac{(O_{1d} - N_1 \cdot b_d)^2}{N_1 \cdot b_d} + \frac{(O_{0d} - N_0 \cdot b_d)^2}{N_0 \cdot b_d}\right] + 2 \cdot \sum_{d=1}^{9}\left(\frac{O_{1d} \cdot O_{0d}}{N \cdot b_d}\right) - \frac{2 \cdot N_1 \cdot N_0}{N_1 + N_0}.$$

First factor on the right side is identical to H_1. This means that H_2 is more conservative than H_1, although they have the same number of degrees of freedom.

3. Numerical Examples

For demonstration, sample of financial payments is taken, consisted of 33,563 items, of which 21,244 are credits (input) and 12,319 are debits (output). In Table 2, frequencies for leading digits for all three categories are presented.

Table 2. Frequencies of leading digits in sample

Digits	All_Items	Credits	Debits
1	9,169	5,328	3,841
2	6,104	3,830	2,274
3	4,197	2,772	1,425
4	4,048	2,830	1,218
5	3,019	1,976	1,043
6	2,366	1,594	772
7	1,870	1,186	684
8	1,536	926	610
9	1,254	802	452
Total	**33,563**	**21,244**	**12,319**

Graphical presentation of relative frequencies is on next diagram.

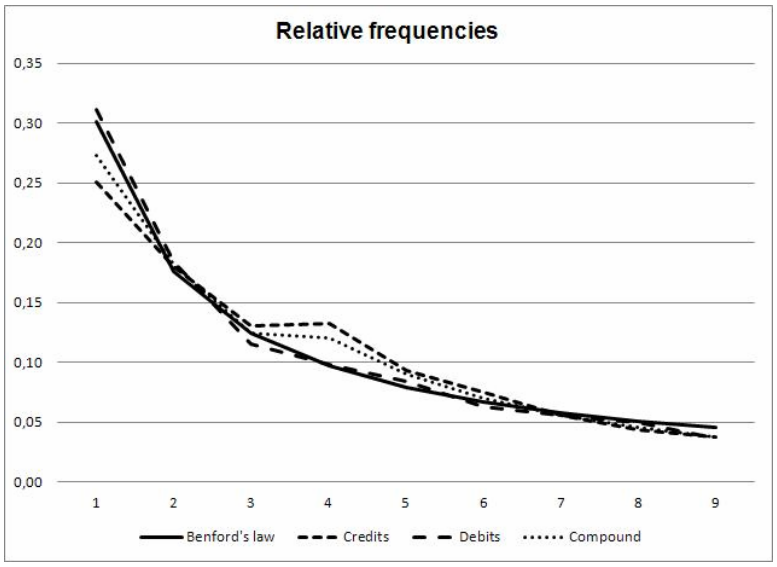

Diagram 1. Relative frequencies of leading digits for credits, debits and all transactions, compared to Benford's law.

The first thing we can note is that frequencies for input and output transactions are considerably different in comparison to frequencies if they are taken together, if we make no difference of category. Frequencies of digit 4 are notably bigger in input and output transactions but it's not so visible on whole sample level.

First step is first digit test. Standard χ^2 test is conducted and results are in Table 3.

Table 3. Chi-square test

Digits	O_j	E_j	$(O_j - E_j)^2$	$\frac{(O_j - E_j)^2}{E_j}$
1	9,170	10,103.4697	873,233.7033	86.4291
2	6,104	5,910.1509	37,577.4628	6.3581
3	4,197	4,193.3188	13.5511	0.0032
4	4,048	3,252.5908	632,675.8486	194.5144
5	3,019	2,657.5602	130,638.75714	49.1574
6	2,366	2,246.9351	14,176.4503	6.3092
7	1,870	1,946.3837	5,834.4721	2.9976
8	1,536	1,716.8321	32,700.2523	19.0468
9	1,254	1,535.7587	79,387.9400	51.6930

Value of χ^2 statistic is 416.5090 what is significantly bigger than table value $\chi^2_{7;0.05}$ = 14.0671. As the second, Hosmer-Lemeshow test is conducted. Calculations are in Table 4.

Table 4. Calculation of H statistic

Digits	**Total**	O_1	E_1	$O_1^2 / (N_1 \cdot b_d)$	O_0	E_0	$O_0^2 / (N_0 \cdot b_d)$
1	**9,170**	5,328	6,395.0812	4,438.97161	3,842	3,708.68955	3,980.10236
2	**6,104**	3,830	3,740.8827	3,921.24029	2,274	2,169.44431	2,383.59472
3	**4,197**	2,772	2,654.1985	2,985.02987	1,425	1,539.24524	1,319.23423
4	**4,048**	2,830	2,058.7563	3,890.16414	1,218	1,193.93136	1,242.55384
5	**3,019**	1,976	1,682.1264	2,321.21440	1,043	975.51295	1,115.15588
6	**2,366**	1,594	1,422.2176	1,786.53112	772	824.78445	722.59364
7	**1,870**	1,186	1,231.9809	1,141.73521	684	714.46079	654.83790
8	**1,536**	926	1,086.6842	789.07562	610	630.19908	590.44834
9	**1,254**	802	972.07210	661.68341	452	563.73228	362.41316
Total	**33,564**	21,244	21,244.00000	21,845.64568	12,320	12,320.000	12,370.93407

Value of H statistic is 21,845.64568 + 12,370.9340 – 33,564.00000 = 652.57968. This is significantly bigger than value of standard chi-square test conducted as a first step.

Suppose, for the moment, that observed frequencies in this example in total are the same as expected and that frequencies in partitions are as in Table 5.

Table 5. Calculation of H statistic

Digits	**Total**	O_1	E_1	$O_1^2 / (N_1 \cdot b_d)$	O_0	E_0	$O_0^2 / (N_0 \cdot b_d)$
1	**10,104**	6,300	6,395.0812	6,206.33243	3,804	3,708.68955	3,901.75986
2	**5,910**	3,850	3,740.8827	3,962.30012	2,060	2,169.44431	1,956.07694
3	**4,193**	2,610	2,654.1985	2,656.53749	1,583	1,539.24524	1,627.99854
4	**3,253**	2,070	2,058.7563	2,081.30509	1,183	1,193.93136	1,172.16872
5	**2,658**	1,680	1,682.1264	1,677.87630	978	975.51295	980.49339
6	**2,247**	1,433	1,422.2176	1,443.86415	814	824.78445	803.35656
7	**1,946**	1,232	1,231.9809	1,232.01908	714	714.46079	713.53951
8	**1,717**	1,092	1,086.6842	1,097.34182	625	630.19908	619.84382
9	**1,536**	977	972.07210	981.95285	559	563.73228	554.84382
Total	**33,564**	21,244	21,244.00000	21,249.52932	12,320	12,320.000	12,329.54479

Value of H statistic is 21,249.52932 + 12,329.54479 – 33,564.00000 = 15.07411. Since critical value is $\chi^2_{0.05;7} = 14.0671$ we need to reject hypothesis that leading digits in this example follow Benford's law. On the other side, we have that

$$H_{11} = \sum_{d=1}^{9} \frac{O_{1d}^2}{E_1} - N_1 = 5.52932, \quad H_{10} = \sum_{d=1}^{9} \frac{O_{0d}^2}{E_0} - N_0 = 9.54479.$$

This means that we should not reject hypothesis if we make separate tests on partitions. At the same time, value of chi-square test for whole sample is 0.00028. This means that we should not reject hypothesis for whole sample. In the another words, we have different conclusions for the same sample, depending on either we divide sample or not. At the same way, it's possible to have nonconformity on one of sides and conformity on both sides.

4. Discussion

Main goal of this paper is to analyse possibility to use Hosmer-Lemeshow test to test conformity of sample to Benford's law. In this sense, grouping of values, by leading digits instead of decils is proposed. By this, we can have 9 groups for first leading digits, 90 groups for leading two digits etc.

Advantage of this approach is that we can detect contribution of any group in whole level of anomalies, even in case when test does not detect anomalies on whole sample level.

Another way is to divide interval $[10^{k-1}, 10^k)$, $k = 1, 2, \ldots,$ in n subintervals, where n is arbitrary chosen natural number [4]. This can be extended to digits or groups of digits on other positions.

Problems can arrise with big frequencies in some groups.

Next step is to generalize this procedure on m partitions, corresponding to values $Y = j$, $j \in \{0, 1, \ldots, m-1\}$, with G groups in each partition. Assumptions in this case are:

- O_{jd} : number of observations in d-th group for partition $j \in \{0, 1, \ldots, m-1\}$. According to this is

$$N_j = \sum_{d=1}^{G} O_{jd}, \quad j \in \{0, 1, \ldots, m-1\}.$$

- $O_d = \sum_{j=0}^{m-1} O_{jd}$: number of observations in d-th group. According to this is

$$N = \sum_{d=1}^{G} O_d = \sum_{j=0}^{m-1} N_j.$$

- b_d : Benford's theoretical probability for d-th group

$$b_d = P[D = d] = \log\left(1 + \frac{1}{d}\right).$$

- $E_d = \sum_{j=0}^{m-1} E_{jd}$: expected number of elements in the d-th group given Y, calculated by

$$E_{jd} = E(Y = j|d) = b_d \cdot N_j = N_j \cdot \log\left(1 + \frac{1}{d}\right),$$

$$d \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}, \quad j \in \{0, 1, \ldots, m-1\}$$

$$\Rightarrow E_d = \sum_{j=0}^{m-1} E_{jd} = N \cdot b_d$$

$$N_j = \sum_{d=1}^{G} E_{jd} = \sum_{d=1}^{G} O_{jd}, \quad j \in \{0, 1, \ldots, m-1\}.$$

Specificities for this case are:

- Probability for all partitions in group d is $b_d < 1$.

• Sum of expected number of cases in group d must not be equal to the sum of observed cases

$$\sum_{j=0}^{m-1} E_{jd} \neq \sum_{j=0}^{m-1} O_{jd}.$$

In this case, statistic H_1 can be interpreted as sum

$$H_1 = \sum_{j=0}^{m-1} H_{1j}, \quad H_{1j} = \sum_{d=1}^{G} \frac{O_{jd}^2}{E_j} - N_j.$$

This statistic has $G - m$ degrees of freedom. This means that $G - 1$ is the biggest number of partitions.

5. Conclusion

In this paper, use of Hosmer-Lemeshow test for goodness-of-fit for Benford's law is proposed. This means that grouping is according to leading digits is used instead of decils.

Calculations show that, in this variant, test is more sensitive to anomalies than standard χ^2 test if it's possible to divide sample in two or more partitions.

It's method is easy to implement this test in Excel or similar programs.

References

[1] Simon Newcomb, Note on the frequency of use of the different digits in natural numbers, American Journal of Mathematics 4(1/4) (1881), 39-40.

[2] Frank A. Benford, The law of anomalous numbers, Proceedings of the American Philosophical Society 78(4) (1938), 551-572.

[3] David Hosmer and Stanley Lemeshow, Applied Logistic Regression, 2nd Edition, p. 148.

[4] Zoran Jasak, Benford's law and arithmetic sequences, Journal of Mathematical Sciences: Advances and Applications 32 (2015), 1-16. ISSN 0974-5750.

[5] https://en.wikipedia.org/wiki/Hosmer-Lemeshow_test.

■