

Implementing Digital Data Analysis Using SAS®

Thomas J. Winn, Jr.

Texas State Auditor's Office, Austin, Texas

[SAS is a registered trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.]

Abstract

Benford's Law is an empirically-verified mathematical formula which describes the expected distribution of digits in a collection of everyday numbers. Since most people aren't familiar with these theoretical frequencies, it often is possible to compare an actual distribution of digits with the Benford probabilities, in order to determine if the data were contrived (invented, fraudulent, or rounded) somehow. The paper describes the SAS code that was written by the author for use at the Texas State Auditor's Office.

Introduction

Benford's Law says that the probability of the first digit being a "d", $P(D1=d) = \log_{10}(1 + (1/d))$, where $d=1,2,3,\dots,9$. This phenomenon concerning everyday numbers was first observed empirically by astronomer Simon Newcomb in 1881, and afterwards by physicist Frank Benford in 1938. Benford referred to it as "the law of anomalous numbers". This mathematical property concerning first digits can readily be extended to provide probabilities for second digits, first and second digit combinations, etc.

For Benford's Law to apply to a set of data, the data must have the following characteristics:

- (1) the numbers should represent the sizes of similar phenomena,
- (2) there should be no built-in maximum or minimum values,
- (3) the data should not consist of assigned numbers, that is, numbers that are used for naming purposes, but without quantitative meanings (for example, telephone numbers or social security numbers),
- (4) there should be more small items than large items. (Moreover, the mean would be greater than the median, and the skewness value would be positive.)

Interestingly, in 1961, Roger Pinkham proved that the distribution of first digits is scale invariant under multiplication, so that multiplication of all of the numbers in a Benford set by a nonzero constant would result in a new set of numbers, which also would follow Benford's Law.

In 1992, Mark Nigrini's doctoral dissertation observed that many types of accounting data, including sales figures and expenditure claims, follow Benford's Law. He also showed how deviations from the expected distribution might be detected using statistical tests. Nigrini's work is the beginning of using Benford's Law for fraud detection in auditing. It should be pointed out that Mark Nigrini sells DATAS, an application software package that he wrote for performing digital data analysis. Nigrini's book and his DATAS package set forth certain other statistical methods, which are not implemented in the program described in this paper. These are the number duplication test, the number

frequency factor test, the relative size factor test, and the distortion factor model.

Early in 2001, the Texas State Auditor's Office became interested in implementing this approach, and in developing its own computer programs for auditing purposes. I developed the SAS programs that are used at the State Auditor's Office, independently of the DATAS programs. However, I acknowledge that I closely followed the procedures that were outlined in Nigrini's book. This paper describes the primary program, the SAS program which performs the most fundamental digital data analysis tests. Besides the program herein described, I also have written other digital data analysis programs that are used at the S.A.O.

Explanation of the Program

The program begins by reading in the data which are to be analyzed for consistency with Benford's Law. Only positive data values are selected. If any negative data values exist, these would be analyzed separately. Various combinations of digits (first, second, first two, first three, and last two digits, as well as the decimal fractional components) are stripped from the data values, using various SAS character-string manipulation functions. Then PROC FREQ is used to compute the frequencies and proportions for each digit combination, and PROC SUMMARY is used to count the total number of observations that went into each frequency table.

Using Benford's Law and its logical extensions, expected probabilities are computed for all of the first, second, first-two, and first-three digital combinations. Theoretically, each of the last-two-digital combinations and decimal fractional components would be equally likely.

Next, the actual and expected frequency distributions for the digital combinations are combined and compared. For each digital combination under consideration, the differences between the actual and expected proportions are calculated. And z-statistics are computed, as a measure of the statistical significance of each of the differences.

The z-statistic is computed using a formula that is found in Nigrini's book (on pages 42 and 73). It is as follows:
$$zstat = \frac{(\text{abs}(p - \text{benfp}) - (1/(2 * \text{num})))}{\sqrt{(\text{benfp} * (1 - \text{benfp}) / \text{num})}}$$

The $1/(2 * \text{num})$ term in the numerator is a continuity correction term and is only used when it is smaller than the absolute difference which appears as the first term in the numerator. Then, 95% upper and lower tolerance limits for the expected proportions are calculated. The upper and lower bounds used in the tolerance intervals are based on the z-statistic. Nigrini (p. 43) asserts that, for values of the z-statistic which are greater than 1.96, the difference between the actual and the expected proportions will be significant at the 5

percent level (and the z-statistic cutoff at the 1 percent significance level would be 2.57).

Just before printing the reports, the mean absolute difference between the actual and expected proportions is computed. The mean absolute difference is an unweighted average of all of the absolute differences for each digit-combination under consideration. Nigrini says that this is the best goodness-of-fit measure for use in an auditing context (page 79). It has the advantage of being easy to understand, and it does not vary according to the size of the data set. Another of my SAS programs compares the mean absolute differences associated with the values of a BY-variable which is used for classification purposes, listing the mean absolute differences in descending order.

Finally, the reports are produced. These include comparative listings of the actual vs. expected proportions for each instance, and also include the computed differences, z-statistics, and 95% tolerance intervals. Some graphical outputs also are generated, to help analysts to visualize some of the digital data analysis results.

Interpreting the Data

As with any statistical procedure, before running the digital data analysis program, it is a good idea to use standard exploratory data analysis methods to identify the general character of the data, including any outliers. The EDA plots that are produced by PROC UNIVARIATE are particularly helpful.

Above all, one should use good sense in examining digital distributions. Just because a report appears to contain some statistical results is not a good justification for the abandonment of critical thinking. A pitfall of digital data analysis is that sometimes the computed statistics can be misleading.

Data sets containing only a few hundred, or even a few thousand, records may fail to conform to Benford's Law just because of their smallness.

The z-statistic is only one clue among many regarding the identification of numbers that are fraudulent or invented. The z-statistic may not be entirely reliable, particularly with small datasets. A few apparent instances of "significant" noncompliance with Benford's Law may not be particularly material. However, if the instances of noncompliance are both large and numerous, then one might be suspicious about how they were generated.

The z-statistic suffers from an excess power problem. For large data sets, small differences may be flagged as being statistically significant even when the differences may be immaterial. If the actual proportion is reasonably close to the expected proportion, disregard the z-statistic and the associated tolerance intervals.

In the context of selecting cases for auditing, negative deviations from Benford's Law generally are not important (Nigrini's book, p. 43).

The first- and second-digit tests both are high-level tests of reasonableness. Unless there are blatant problems with the data, the first-digit test probably won't

turn up anything. The second-digit test may indicate some problematic data, but because there generally would be a large number of observations having the same second digit, by itself this test probably would not be an efficient tool for selecting specific cases for examining.

The first-two-digits test is performed to identify anomalies in the data that may not be apparent from the separate first- and second-digits tests. Very large positive deviations would indicate excessive duplications, and could be used to specify a coarse selection criteria of items for further consideration. The first-three-digits test is designed to provide a finer amount of precision than the first-two-digits test. It can be used to narrow the focus of attention in selecting cases for a careful examination.

The last-two-digits and decimal-fractional-components tests are designed to find invented numbers and rounded numbers. Rounded numbers may signal that the numbers were estimated, or were arbitrarily assigned. If the data to be tested all contain "dollars and cents" magnitudes, then the last-two-digits and the decimal-fractional-components tests would yield identical results. If excessive number duplication is suspected, then the data could be sorted and subsetted so as to identify the most frequently recurring numbers.

Conclusion

This paper presents a SAS program that performs an analysis of the distribution of certain digital combinations occurring in a batch of data values. It compares the actual distributions with the expected frequencies according to Benford's Law. Since most people might expect the distribution of digits to be uniform, this empirical distribution has been useful in detecting fraud.

My SAS code is available upon request.

Suggestions for Further Reading

- Benford, Frank (1938). "The Law of Anomalous Numbers," Proceedings of the American Philosophical Society, Vol. 78, No. 4, pp. 551-571.
- Browne, Malcolm W. (1998). "Following Benford's Law, or Looking Out for No. 1," The New York Times, August 4, 1998.
- Feller, William (1966). An Introduction to Probability Theory and Its Applications, Volume II, p. 62, John Wiley & Sons, Inc., New York.
- Matthews, Robert (1999). "The Power of One," New Scientist, July 10, 1999, pp. 26-30.
- Newcomb, Simon (1881). "Note of the Frequency of Use of the Different Digits in Natural Numbers," American Journal of Mathematics, Vol. 4, pp. 39-40.
- Nigrini, Mark J. (1992). "The Detection of Income Tax Evasion Through an Analysis of Digital Distributions," Ph.D. dissertation, University of Cincinnati.
- Nigrini, Mark J. (2000). Digital Analysis Using Benford's Law: Tests and Statistics for Auditors, Second Edition, Global Audit Publications, Vancouver B.C.
- Pinkham, Roger S. (1961). "On the Distribution of First Significant Digits," Annals of Mathematical Statistics, Vol. 32, pp. 1223-1230.

Author Information

Tom Winn
CAATS Team
Texas State Auditor's Office
P.O. Box 12067
Austin, Texas 78711-2067

Telephone: 512 / 936-9735
E-Mail: twinn@sao.state.tx.us

Appendix – Output Sample (Selected Reports)

Analysis of the Frequency Distribution of First Digits of the Positive Values in the Data

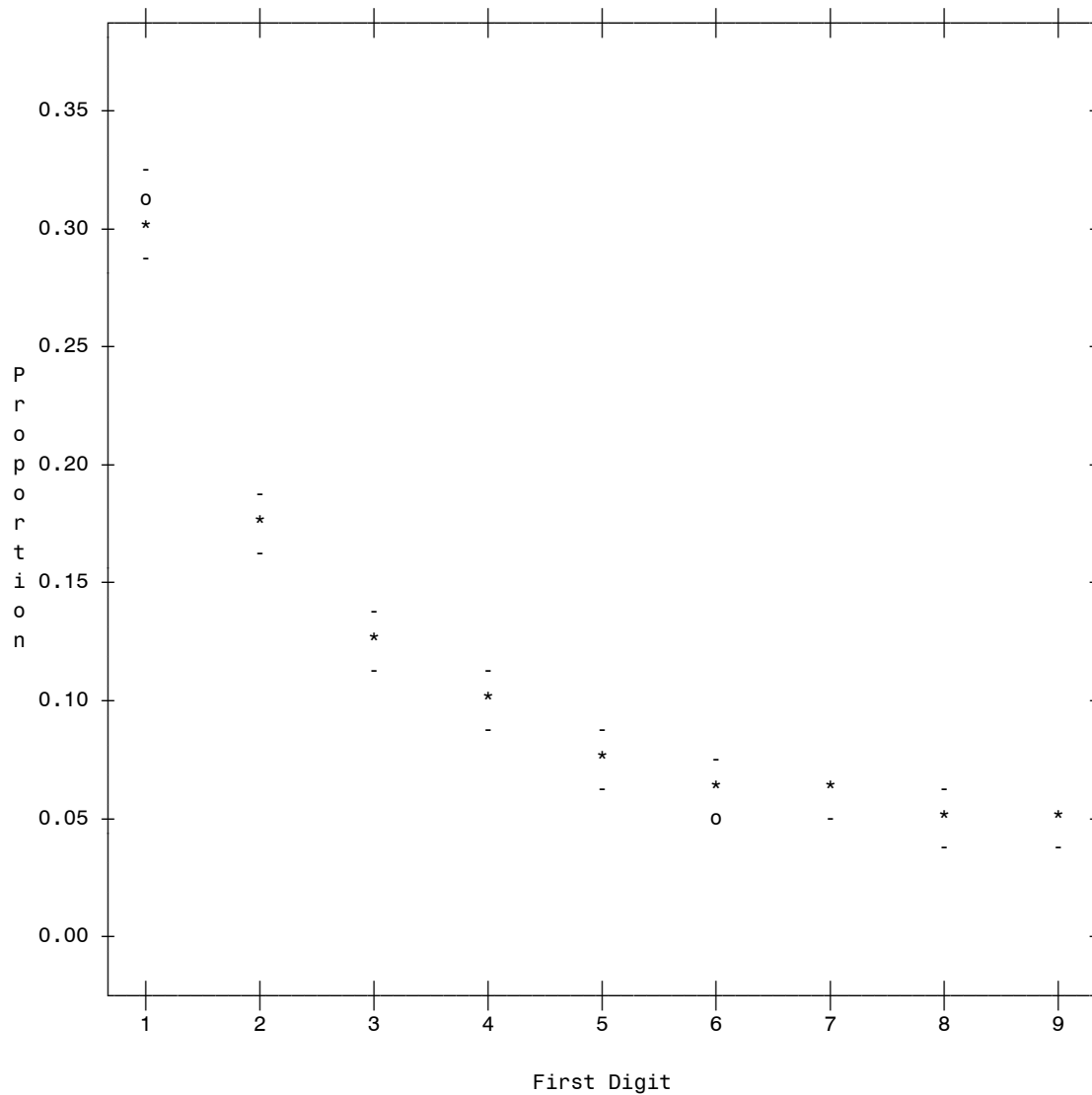
NOTE -- The Mean Absolute Difference is 0.00659

First Digit	Actual Count	Actual Proportion	Expected Proportion	Difference	Lower Bound	Upper Bound
1	774	0.30985	0.30103	0.00882	0.28284	0.31922
2	422	0.16894	0.17609	-.00716	0.16095	0.19123
3	335	0.13411	0.12494	0.00917	0.11177	0.13811
4	258	0.10328	0.09691	0.00637	0.08511	0.10871
5	211	0.08447	0.07918	0.00529	0.06839	0.08997
6	134	0.05364	0.06695	-.01330	0.05695	0.07695
7	138	0.05524	0.05799	-.00275	0.04863	0.06736
8	117	0.04684	0.05115	-.00432	0.04231	0.05999
9	109	0.04363	0.04576	-.00212	0.03736	0.05415
	=====	=====	=====			
	2498	1.00000	1.00000			

First Digit	Z-Statistic	Comment
1	0.93898	within 95% tolerance interval
2	0.91273	within 95% tolerance interval
3	1.35564	within 95% tolerance interval
4	1.04281	within 95% tolerance interval
5	0.94143	within 95% tolerance interval
6	2.62043	significantly different, at 5% level
7	0.54477	within 95% tolerance interval
8	0.93352	within 95% tolerance interval
9	0.45982	within 95% tolerance interval

Analysis of the Frequency Distribution of First Digits
of the Positive Values in the Data
(*=Expected, o=Actual, bars(-) represent 95% tolerance interval)

Plot of $\text{benfp1} \cdot d1$. Symbol used is '*'.
Plot of $\text{lower95} \cdot d1$. Symbol used is '-'.
Plot of $\text{upper95} \cdot d1$. Symbol used is '-'.
Plot of $p1 \cdot d1$. Symbol used is 'o'.



NOTE: 10 obs hidden.