Goodness of fit for uniform distribution: Benford's law method

Determining the underlying distribution of a certain data set is an important problem within Statistics. There are a number of ways, both parametric and non-parametric, to accomplish the task. Perhaps, the most popular and widely used technique among them is chi-squared goodness of fit test.[1] However, chi-square test is highly sensitive to sample size, and a reasonably strong relationship may not appear significant if the sample size is small. Again, in large samples, the test may produce significant result, although they are not statistically significant. [4, 5] Other methods have some limitations as well.[2] In this paper, we propose a method, based on Benford's law, for checking whether a uniform a distribution underlies a certain data. Benford's law, [3] named after physicist Frank Benford and initially discovered by Newcomb, simply states that the leading significant digit is expected to be small. For instance, in a set where the law fits, most numbers begin with digit 1, accounting for over 30% of total numbers, and the frequency of numbers starting with later digits falls almost exponentially. Our method is based on the fact that values obtained by taking antilogs of a uniform random variable very significantly follow Benford's law. We proved empirically as well as with simulation that the methods works better that the prevalent methods. We tested the method on biomedical data, in which measurements concerning treating cancer patients were taken on Tissue Maximum Ratio (TMR), field size, and treatment time. In the data set, the method was found to outperform chi-squared test in detecting uniform distribution. Additionally, we also modeled treatment time with a gamma distribution.

References

1. Pearson, Karl (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling" (PDF). Philosophical Magazine. Series 5. 50: 157–175.
2. Stephens, M. A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons". Journal of the American Statistical Association. American Statistical Association. 69 (347): 730–737.
3. Frank Benford (March 1938). "The law of anomalous numbers". Proc. Am. Philos. Soc. 78 (4): 551–572. JSTOR 984802.
4. Liu, Qiang; Lee, Jason; Jordan, Michael (20 June 2016). "A Kernelized Stein Discrepancy for Goodness-of-fit Tests". *Proceedings of the 33rd International Conference on Machine Learning*. The 33rd International Conference on Machine Learning. New York, New York, USA: Proceedings of Machine Learning Research. pp. 276–284
5. https://www.mtholyoke.edu/courses/etownsle/qr/Chi%20square%20limitations.htm