Most Numbers Start with 1

Can We Validate Water Use Data with Benford's Law?

Robert B. Sowby, Ph.D., P.E., ENV SP | Hansen, Allen & Luce, Inc. | rob.sowby@hansenallenluce.com
AWWA Intermountain Section Annual Conference | Midway, UT | October 2018

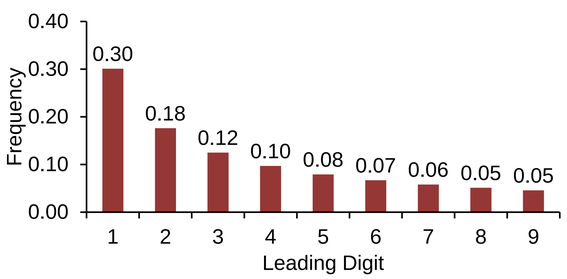
Frank Benford (1883–1948), American Physicist

INTRODUCTION

Accurate water use information is needed to plan the infrastructure, policies, and programs necessary to maintain public water services. However, most data are self-reported, so it is difficult to assess their accuracy.

Benford's Law, first observed more than a century ago, may help. Benford's Law gives the expected frequency of leading digits in numerical data. One might expect these digits to be uniformly distributed, on average 1/9 or 11.1% of the time, but in many datasets, 1 is the leading digit about 30% of the time, with digits 2 through 9 following with decreasing frequency. Benford's Law has been used to detect accounting fraud and to validate observations in the physical sciences.

Do water use data follow Benford's Law? If so, how might it help validate historical observations and modeling results?



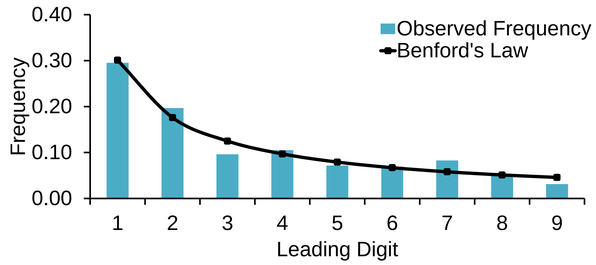
Frequency of Leading Digits according to Benford's Law

METHODS

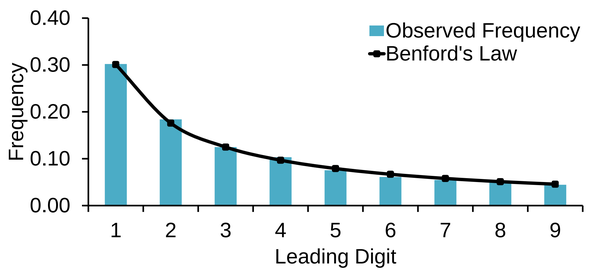
Data for this analysis originated from the Utah Division of Water Resources, the U.S. Geological Survey, and a research dataset by others. The data describe annual potable water use at system and county levels. The data were grouped into three scenarios and the Kolmogorov–Smirnov test was applied to compare the observed and expected distributions of their leading digits.

RESULTS

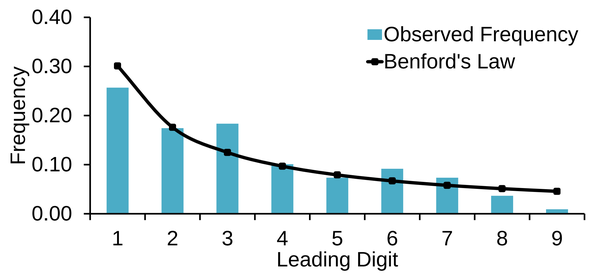
All three samples were found to conform to Benford's Law with statistical significance. The figures below compare the observations to Benford's Law.



Annual Potable Water Use of 447 Utah Water Systems



Annual Potable Water Use of 3,124 U.S. Counties



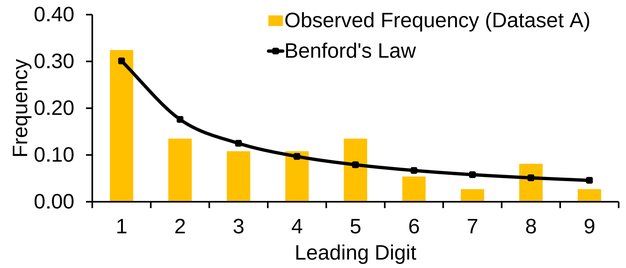
Annual Potable Water Use of 109 U.S. Water Systems

APPLICATIONS

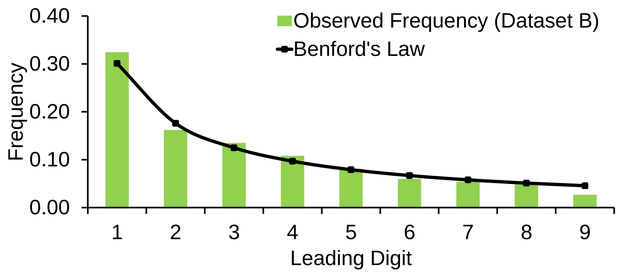
Performing a Benford analysis, as done here, could supplement other data validation methods. While Benford analysis alone will not identify specific problems, it will, even with a simple graphical comparison, indicate general conformance or nonconformance. Nonconformance could indicate incomplete data, errors, or abnormal water use.

Another potential application is checking the naturalness of modeling results. Since real water use data follow Benford's Law, then numerical models should too. Water demand projections might therefore undergo Benford analysis to consider their reasonableness.

In one example application shown below, two sets of data on the same water uses were compared. The one known to be more accurate (Dataset B) more closely matched Benford's Law.



Dataset A: Self-Reported with Limited Quality Control



Dataset B: Engineers' Data with Enhanced Quality Control

CONCLUSION

Using three separate datasets and a statistical test, the author concludes that water use data follow Benford's Law. Benford analysis may find applications alongside other data validation and analysis methods in the water industry.

Additional applications are recommended, especially to unaggregated data in larger samples with greater statistical power, in order to test the limits and identify exceptions.

HANSEN ALLEN & LUCE inc ENGINEERS