United States

Naval Research Laboratory

Benford's law:

applications in CYBER & Soft-Biometrics (Machines and Human Behavior)

Stephen Russell

stephen.russell@nrl.navy.mil

October 24, 2013

Conference on Applied Statistics in Defense

Overview

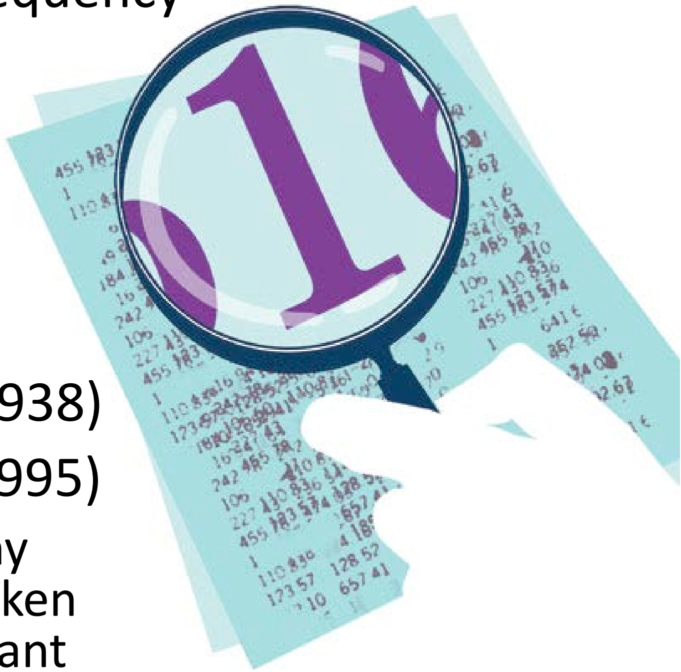
- Cyber, soft-biometrics, machines, and human behavior
- Background on Benford's Law
  - Limitations of Benford's Law
  - Identifying the right metric(s)
- Applying Benford's Law to web behavior analytics
- Applying Benford's Law to E-Mail spam filtering
- Potential applications and future work

Cyber and Soft-Biometrics: Machines & Human Behavior

- Given the diverse patterns of behavioral activity that are embedded in temporally-variable sensor data, *it is a challenge to isolate human-specific behavioral observations*
  - Impediments to solving this problem:
    - Modeling the machine or the behavior?
    - Temporal scalability
    - Sensor noise
    - Volume of data
- A fast, scalable, simple technique that does not depend on data introspection may be useful in addressing this problem

Benford’s Law: Background

- Simon Newcomb was the first to note, in an 1881 *American Journal of Mathematics* article, that not all possible first digits appear with equal frequency in large sets of “natural numbers”
  - “The law of probability of the occurrence of numbers is such that all mantissae of their logarithms are equally probable”
- Frank Benford, “re-discovered” Newcomb’s hypothesis and demonstrated its utility in a variety of empirical applications (Benford 1938)
- A proof of the law was developed by Hill (1995)
  - If distributions are selected at random (in any "unbiased" way) and random samples are taken from each of these distributions, the significant digits of the combines sample will converge to the logarithmic (Benford) distribution



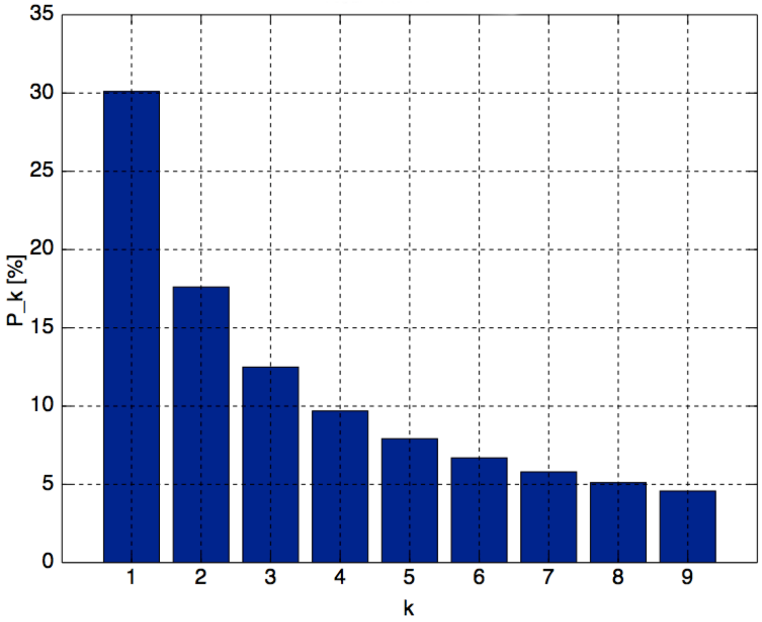
Hill, P. (1995), “A Statistical Derivation of the Significant Digit Law,” *Statistical Science*, 10, 354–363.

Benford's Law:

Localization Without Explanation

- Variances from Benford indicate bias, modification, dependence, etc.
- It has been used in a variety of other domains to identify "non natural," man-made, or fraudulent measurements/observations in data sets
  - Accounting (e.g. taxes, payroll)
  - River/Ocean/Lake volumes
  - Skyscraper heights
  - Economic and election data

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(1+\frac{1}{d}\right)$$

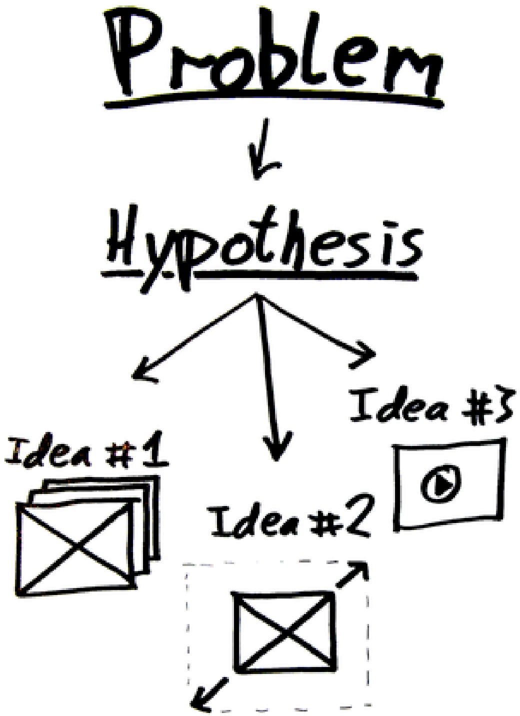


BENFORD'S LAW: With naturally occurring data, the odds of obtaining a 1 for the first significant digit of a number are much higher than the odds of obtaining any other digit

Why Benford's Law in Cyber and Soft-Biometrics?

- Exploitation of human-behavioral fit – separation of man and machine
  - Observations naturally generated by man are theoretically un-biased and those generated by machine are biased by construct
  - Assumes appropriate observation metric
    - Determining which distributions (or mixtures thereof) satisfy Benford's Law
- A test of "reasonableness of output" given a proposed model
  - I.E. "Benford-in/Benford-out" criteria
  - Fast test of data quality
- Filtering and intelligent sampling of large data
  - Minimization of heuristic-based filtering and data inspection

Research Hypothesis

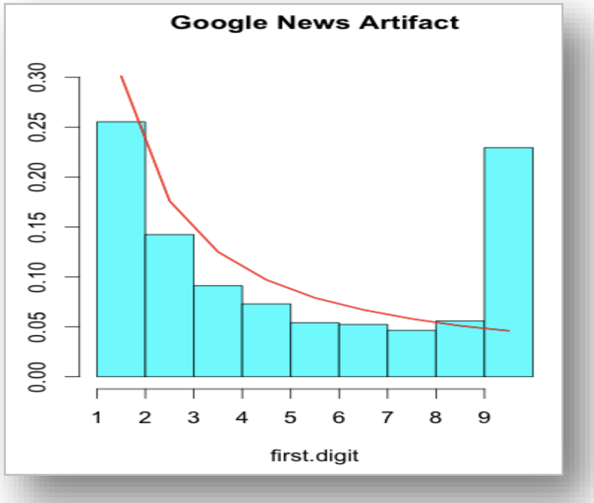


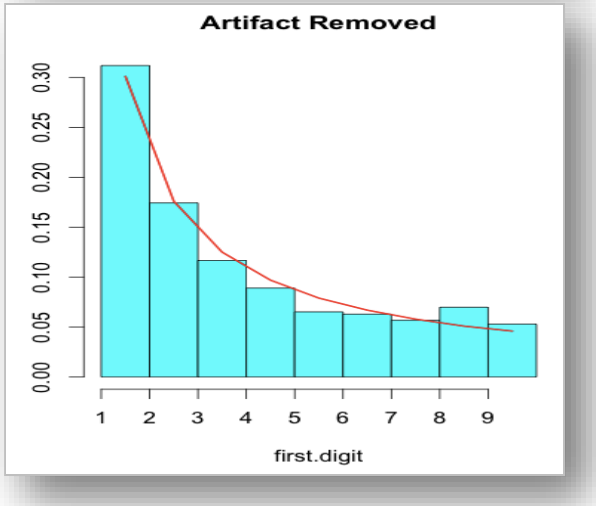
H_0: Benford analysis can aid cyber and soft-biometric applications

- The general technical approach for this research:
  - Identify appropriate metrics
  - Simple empirical evaluation
  - Evaluate predictive power, post Benford processing
  - Evaluate use of Chi Squared or Kuiper Test to evaluate similarity to previous Benford's distributions over time

Experimentation: Web Browsing Pauses

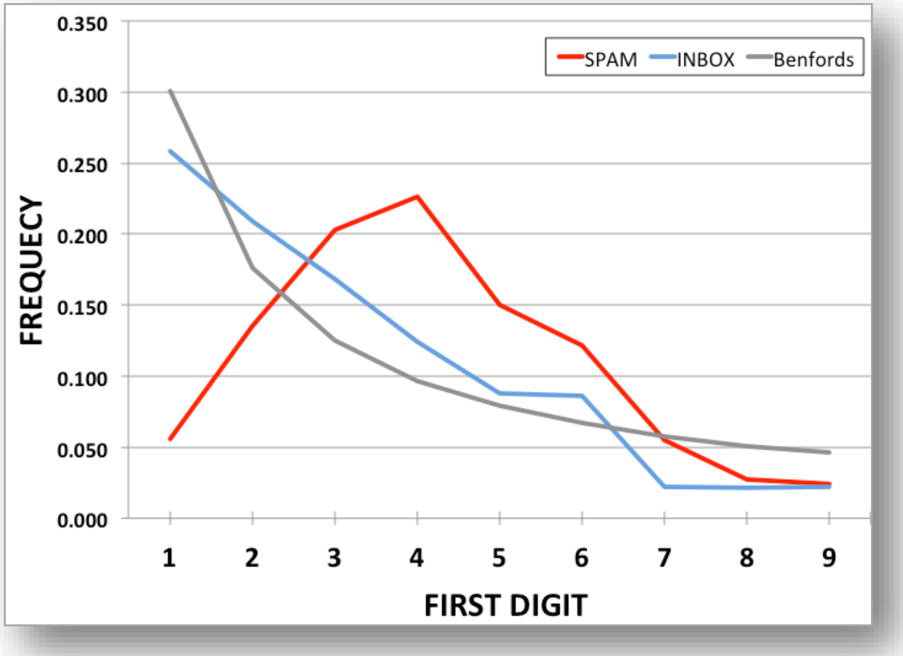
- Benford's Law applied to web browsing pauses – localizing abnormal data
  - In Web Behavior Active Authentication application data was cleansed via introspection and the use of automated heuristics
  - Benford can provide a filter to identify "where" non-human variations may exist
- **Finding**: Human click-delay in web browsing follows Benford's Law





Experimentation: Email Subject Length

- 6,749 manually identified spam e-mail messages and 5,711 “real” e-mail messages analyzed
  - Taking the subject length (number of characters) illustrates that human (non-spam) subject lengths comply with Benford’s
- A spam filter was built to forward messages to SMS
  - Not perfect, e.g. password-reset mails are filtered as spam
  - Effective as “cost-saving” technique for mobile-data usage

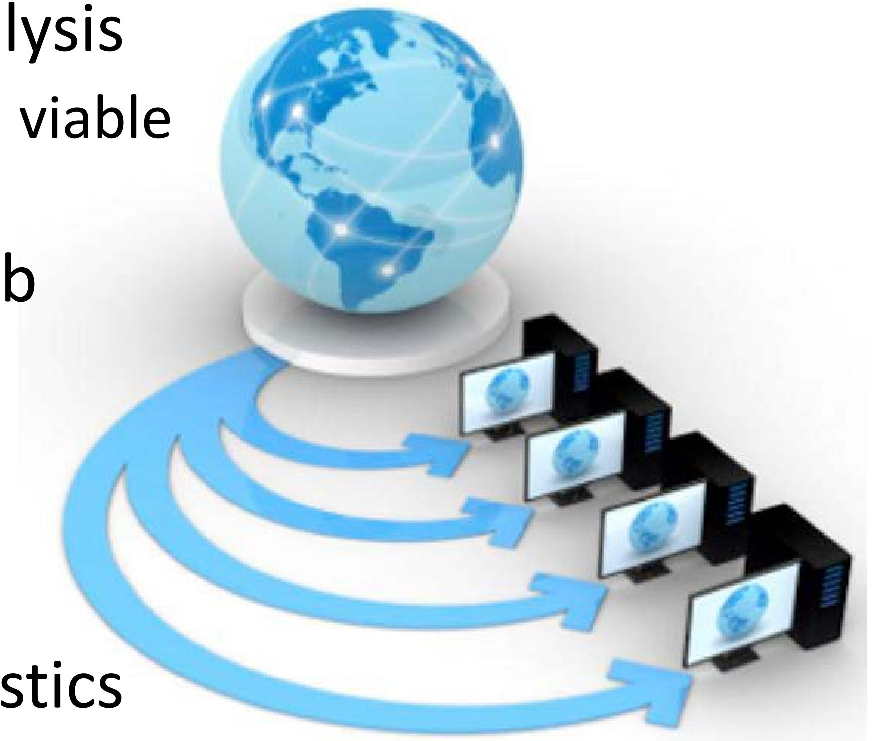


Benford's Law:
Limitations and Cautions

- Likely a good pre-processing or filtering technique and can be used as an indicator for the location of "non natural" samples
  - Does not EXPLAIN anomalies/outliers/deviations
- Useful for processing efficiency enhancement (fast, simple), thus appropriate for optimizing future/downstream processing costs
  - Less effective with small sample sizes
- Appropriate for naturally occurring observations
  - Use with improper metric can produce improper results
    - Use on contextually limited data will affect analysis, e.g. analysis of SSNs, phone numbers, time card data will not adhere to Benford's Law

Future Work

- There is an increasing need for simple tools that aid the identification of non-machine data as a first step to more detailed and costly analysis
  - Benford’s Law appears to be a viable candidate methodology
- Extend the Benford’s Law web browsing pause experiment to a larger dataset
  - ComScore data acquired and planned for use
- Evaluate temporal characteristics
- Investigate pre-processing and predictive potential



QUESTIONS?