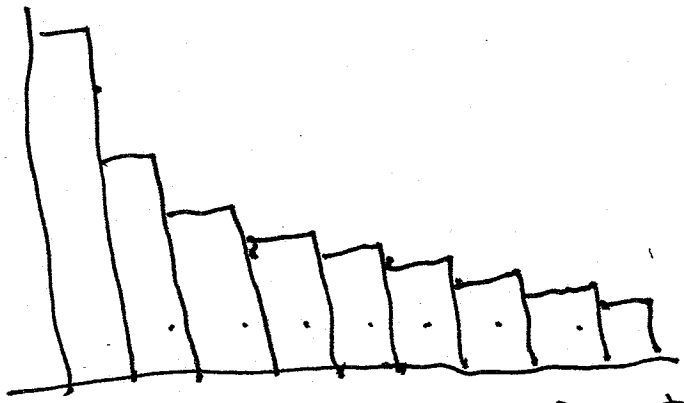
## Benford's Law or the law of anomalous digits

In certain natural collections of numbers like sizes of towns, heights of buildings, lengths of rivers, financial transactions the leading digit has a initially suprising feature.

Since the leading digit is 1, ..., 9 one would expect that each digit would occur 1/9 ≈ 11% of the time but in actuality 1 is the leading digit about 30% of the time and 9 about 5% of the time.

**DEF:** A data set satisfies Benford's law if the frequency of the occurrence of $d$ as the leading digit is

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



approximate plot

**Initial note:** If $x \in [1, 10)$ it has leading digit $d$ if $x \in [d, d+1)$ or

$$d \leq x < d+1$$

Taking $\log_{10}$

$$\log_{10} d \leq \log_{10} x < \log_{10} d+1$$

or $\log_{10} x \in [\log_{10} d, \log_{10} d+1)$

which is an interval with width or Lebesque measure

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

So if all the data is in $[1, 10)$, being Benford is assuming that the $\log_{10}$ of the data is uniformly distributed w.r.t. Lebesque measure

In general, if $x = dd_1d_2\ldots$

$= d.d_1d_2\ldots \times 10^k$ for some $k$ so

$\log_{10} x = \log d.d_1d_2\ldots \times 10^k = (\log d.d_1d_2\ldots) + k$

. So the general assumption on being Benford is that the log of the data mod 1 is uniformly distributed with respect to Lebesque.

. In certain cases one can prove data is Benford using Ergodic Theory

. Lemma: $R_\alpha : S^1 \to S^1$ $\alpha \notin \mathbb{Q}$ has the property that for any interval $J$ in $S^1$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \chi_J R_\alpha^i(x) \to \int \chi_J \, d\mu = \mu(J)$$

for all $x \in S^1$ where $\mu$ = Lebesque measure

so $\mu(J)$ is just the Length of $J$.

Proof: This is almost unique ergodicity but $\chi_J$ isn't continuous so one needs to approximate it by continuous functions.

**Theorem** The set $\{2^n\}_{n\in\mathbb{N}} = \{1, 2, 4, 8, 16, \ldots\}$ satisfies Benfords law.

**Proof** As noted above $2^n$ has leading digit $d$ if and only if

$$\log_{10}(2^n) \in [\log_{10} d, \log_{10}(d+1)) \mod 1 := J_d$$

or $n \log_{10} 2 \in J_d$

Define $R: S^1 \to S^1$ via $\theta \mapsto \theta + \log_{10} 2$ and note that $\log_{10} 2$ is irrational (!) and

$$n \log_{10} 2 = R^n(0)$$

By the Lemma

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \chi_{I_d}(R^n(0)) = \mu(I_d) = \log \frac{d+1}{d}$$

and $\sum_{k=0}^{n-1} \chi_{I_d}(R^n(0))$ is counting the average number of times (asymptotically) that $d$ occurs as the leading digit of some $2^n$. ∎

- Similar arguments show that daily balances with compound interest $(R)^n P_0$ with $\log_{10}(R)$ irrational, $\log\left((1.03)^n(20{,}000)\right)$ are Benford.

- Fraudulent checks, deposits, votes, .. are usually uniformly distributed and so don't satisfy Benfords law, so it can be used as one part of a Fraud detection suite.

We need one more proof to properly wrap up entropy:

**Lemma**: If $\rho = \{A_1, \ldots, A_n\}$, $\eta = \{B_1, \ldots, B_m\}$

Then $H(\rho \vee \eta) \leq H(\rho) + H(\eta)$

**Proof**: For simplicity we don't include terms in the sums when $\mu(B_j) = 0$ or $\mu(A_i \wedge B_j) = 0$

$$H(\rho \vee \eta) = -\sum_{i,j} \mu(A_i \wedge B_j) \log \mu(A_i \wedge B_j)$$

$$= -\sum_{i,j} \mu(A_i \wedge B_j) \log \left( \frac{\mu(A_i \wedge B_j)}{\mu(B_j)} \mu(B_j) \right)$$

$$= -\sum_{i,j} \mu(A_i \wedge B_j) \log\left(\frac{\mu(A_i \wedge B_j)}{\mu(B_j)}\right)$$

$$-\sum_{i,j} \mu(A_i \wedge B_j) \log \mu(B_j)$$

$$= -\sum_{i,j} \frac{\mu(A_i \wedge B_j)}{\mu(B_j)} \mu(B_j) \log\left(\frac{\mu(A_i \wedge B_j)}{\mu(B_j)}\right)$$

$$-\sum_{j} \mu(B_j) \log \mu(B_j)$$

using the fact that $\sum_i \mu(A_i \wedge B_j) = \mu(B_j)$ since $\rho$ is a partition

Thus $H(\rho \vee m) = -\sum_{i,j} \frac{\mu(A_i \wedge B_j)}{\mu(B_j)} \mu(B_j) \log\left(\frac{\mu(A_i \wedge B_j)}{\mu(B_j)}\right)$

$$+ H(m) \qquad (*)$$

Now fix $i$ and let $\lambda_j = \mu(B_j)$ and so $\sum_j \lambda_j = 1$

and let $\alpha_j = \dfrac{\mu(A_i \cap B_j)}{\mu(B_j)}$ so $\alpha_j \in [0,1]$ and.

$$\sum_{j=1}^{m} \lambda_j \alpha_j = \sum_j \mu(B_j) \frac{\mu(A_i \cap B_j)}{\mu(B_j)} = \sum_j \mu(A_i \cap B_j) = \mu(A_i)$$

Now recall, if $\phi(x) = -x \log x \quad x \in [0,1]$, $\phi(0) = 0 \; x=0$

Then we showed $\sum_{j=1}^{m} \lambda_j \phi(x_j) \le \phi\left(\sum_{j=1}^{m} \lambda_j x_j\right)$

Thus $-\sum_{j=1}^{m} \mu(B_j) \dfrac{\mu(A_i \cap B_j)}{\mu(B_j)} \log \dfrac{\mu(A_i \cap B_j)}{\mu(B_j)} = \sum_j \lambda_j \phi(\alpha_j)$

$$\le \phi\left(\sum_j \lambda_j \alpha_j\right) = \phi(\mu(A_i)) = -\mu(A_i) \log \mu(A_i)$$

For each $i$

Summing over $i$ yields

$$-\sum_{i,j} \mu(B_j)\,\frac{\mu(A_i \wedge B_j)}{\mu(B_j)}\log\frac{\mu(A_i \wedge B_j)}{\mu(B_j)}$$

$$\leq H(\rho)$$

Then going back to equation (*) we have

$$H(\rho \vee \eta) \leq H(\rho) + H(\eta) \quad \blacksquare$$