Benford’s Law and irrational rotation

First Digit Frequency According to Benford’s Law

Simon Newcomb
1835-1909
1881

Numberphile video (on log tables)
https://www.youtube.com/watch?v=VRzH4xB0GdM
FRANK BENFORD

The frequency of first digits thus follows closely the logarithmic relation
\[ F_a = \log \left( \frac{a + 1}{a} \right), \]
where \( F_a \) is the frequency of the digit \( a \) in the first place of used numbers.

TABLE II

<table>
<thead>
<tr>
<th>Natural Number</th>
<th>Number Interval</th>
<th>Observed Frequency</th>
<th>Logarithm Interval</th>
<th>Observed - Computed</th>
<th>Prob. Error of Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 to 2</td>
<td>0.306</td>
<td>0.301</td>
<td>+0.005</td>
<td>±0.008</td>
</tr>
<tr>
<td>2</td>
<td>2 to 3</td>
<td>0.182</td>
<td>0.170</td>
<td>+0.009</td>
<td>±0.004</td>
</tr>
<tr>
<td>3</td>
<td>3 to 4</td>
<td>0.124</td>
<td>0.125</td>
<td>−0.003</td>
<td>±0.002</td>
</tr>
<tr>
<td>4</td>
<td>4 to 5</td>
<td>0.094</td>
<td>0.097</td>
<td>−0.003</td>
<td>±0.003</td>
</tr>
<tr>
<td>5</td>
<td>5 to 6</td>
<td>0.079</td>
<td>0.070</td>
<td>−0.001</td>
<td>±0.002</td>
</tr>
<tr>
<td>6</td>
<td>6 to 7</td>
<td>0.066</td>
<td>0.067</td>
<td>−0.003</td>
<td>±0.002</td>
</tr>
<tr>
<td>7</td>
<td>7 to 8</td>
<td>0.058</td>
<td>0.058</td>
<td>−0.002</td>
<td>±0.002</td>
</tr>
<tr>
<td>8</td>
<td>8 to 9</td>
<td>0.050</td>
<td>0.051</td>
<td>−0.002</td>
<td>±0.002</td>
</tr>
<tr>
<td>9</td>
<td>9 to 10</td>
<td>0.047</td>
<td>0.048</td>
<td>+0.001</td>
<td>±0.003</td>
</tr>
</tbody>
</table>

also true for \( 2^n \)
(or any other \( a^n, a \neq 10^k \))

Frank Benford
1883-1948
1938

Two examples of leading digit histograms. The left figure shows the leading digit distribution for 14,144 numbers taken from U.S. Federal income tax returns. The figure on the right is for numbers produced by a computer random number generator (RNG). This shows one of the long-standing mysteries of Benford’s law—why do some sets of numbers follow the law (such as tax returns), while others (such as this RNG) do not? Many have claimed that this is some sort of secret code hidden in the fabric of nature.

Key to understanding B.L. in math is through rational approximations.
Dirichlet’s Drawer / Box principle

Dirichlet proved in 1840 that for every number \( x \) and any positive \( N \) (whole), there’s a rational \( p/q \) such that:

\[
|x - \frac{p}{q}| < \frac{1}{N} \quad , \quad q \leq N
\]

**Proof Idea:** Plot fractional values of \( x \), \( x = \frac{k}{N} \) \( k = 1, 2, \ldots, N \),

so \( |x - \frac{k}{N}| \leq \frac{1}{N} \)

\( k \) chosen  \( a \geq 6 \)

\[
|ax - \frac{p}{q}| \leq \frac{1}{N}
\]

\( \text{whole} \)

**Corollary:** (proved last time)

A line \( y = ax \) intersects a “tree” of triangles \( \exists \alpha \) (for any \( \alpha \)) whatever
at some lattice point
(Dirichlet's simultaneous approximations)

**Theorem 2.** Theorem 1 also holds for several numbers \(d_1, \ldots, d_k\), i.e., for every \(N\) there's a \(q\) such that

\[
|d_i - \frac{p_i}{q}| < \frac{1}{Nq}, \quad i = 1, \ldots, k
\]

(only now \(q \leq N^k\)). *Proof:* Exercise

**Corollary (Max Dehn's Theorem on dissecting into squares).**

Theorem: An \(a \times b\) rectangle can be dissected into squares if and only if \(a/b\) is rational.

*Proof.* The *if* part is easy (if \(a/b = m/n\), then can cut into squares \(m \times m\)).

*Only if.* Assume the rectangle is cut into \(k\) squares of sides \(s_i\), \(i = 1, \ldots, k\).

Need to prove that \(a/b\) is rational.

By Dirichlet's Simultaneous Approximation Theorem, we can assume that all vertices \((x_i, y_i)\) of squares (choose \(q\) such that \(\left| x_i - \frac{p_i}{q}\right| < \frac{1}{Nq}\) and multiply all coordinates by \(q\)).

---

**Max Dehn**

1878-1952

1903
Now draw horizontal & vertical lines at \( x = \pi \sqrt{\frac{a}{2}} \) and \( y = \pi \sqrt{\frac{b}{2}} \) - the total length of all horizontal lines inside the rectangle is the number of horizontal segments.

\[ L = a \cdot \lceil b \rceil \]  
(Notation \( \lceil x \rceil \) is closest integer \( \geq x \))

Length of each segment

and for the vertical lines we have similarly:

\[ M = b \cdot \lceil a \rceil \]

Their number

Claim \( L = M \). Indeed, look at the lines inside the \( i \)-th square:\[
\begin{array}{|c|c|}
\hline
S_i & S_i \setminus S_i \\
\hline
\end{array}
\]
They contribute exactly \( S_i \cdot \lceil S_i \rceil \) to \( L \) and exactly the same to \( M \), so we have

\[ L = \sum_{i=1}^{n} S_i \cdot \lceil S_i \rceil = M \]

But \( L = a \cdot \lceil b \rceil \) and \( M = b \cdot \lceil a \rceil \), so

\[ a \cdot \lceil b \rceil = b \cdot \lceil a \rceil \Rightarrow \frac{a}{b} = \frac{\lceil b \rceil}{\lceil a \rceil} \text{, which!} \]
Back to Benford’s Law

How to go about proving B.L. for $2^n$?
E.g. what exactly does it mean that $2^n$ starts with digit $d$?
This means: $2^n = d \cdot 10^k + r$, $0 \leq r < 10^k$

\[ d \cdot 10^k \leq 2^n < (d+1) \cdot 10^k \]

For some $k$

How to get rid of $k$?

Take $\log_{10}$:

\[ \log_{10} d \cdot 10^k \leq \log_{10} 2^n < \log_{10} (d+1) \cdot 10^k \]

\[ k + \log_{10} d \leq n \cdot \log_{10} 2 < k + \log_{10} (d+1) \]

\[ i.e. \quad \log_{10} d \leq n \cdot \log_{10} 2 - k < \log_{10} (d+1) < 1 \]

\[ 0 \leq \left\lfloor n \cdot \log_{10} 2 \right\rfloor - k \leq \left\lfloor \log_{10} (d+1) \right\rfloor < 1 \]

So \[ k = \left\lfloor n \cdot \log_{10} 2 \right\rfloor \]

\[ \frac{1}{n} \cdot \log_{10} 2 \]

\[ \left. x \cdot \right| = x - \left\lfloor x \right\rfloor \]

So \[ \ln \cdot \log_{10} 2 \in \left[ \log_{10} d, \log_{10} (d+1) \right] \]

\[ \left. \log_{10} \frac{d}{a} \right) \]
i.e. we want to study when

\[ \log_{10} \frac{a}{x} \text{ lands inside the interval } I \]

\[ \text{of length } \frac{\log_{10} a}{x} \]

i.e. \( \frac{1}{2} \), \( \frac{1}{2} + \frac{1}{2} \), \( \frac{1}{2} + \frac{3}{2} \), ...

\[
\begin{array}{cccc}
0 & 1 & 2 & 3 \\
\hline
\end{array}
\]

\[ \text{circle of length } 1 \]

**Theorem (Kronecker's Approximation Theorem)**

For any \( \varepsilon > 0 \), irrational \( x \) and any \( \alpha \), there exists \( n \) and \( m \) such that

\[ |n\alpha - \alpha - m| < \varepsilon \]

\[ \therefore \quad |n\alpha - \alpha| < \varepsilon \]

\[ \therefore \quad \text{fractional part} \]

\[ a - \varepsilon < n\alpha < a + \varepsilon + m \]
Corollary: There exist a set \( 2^n \) begins with its fixed combination of k digits.

Proof: The only thing we need to check is that \( x = \log_{10} 2 \) is irrational. Indeed if we assume \( \log_{10} 2 = \frac{p}{q} \) is rational then \( (10^k)^{p/q} = 2 \), i.e.

\[
(10^k)^p = 2^q.
\]

\[
10^{kp} = 2^q \quad \text{false!}
\]

Proof of Kronecker’s Thm.
Let \( N \) be such that \( \frac{1}{N} < \epsilon \), then by Dirichlet there are \( a > 6 > 0 \)
Leopold Kronecker  
1823-1891  
1884  

"Die ganzen Zahlen hat der liebe Gott gemacht, alles andere ist Menschenwerk"  
"God made the natural numbers; all else is the work of man."
Final statement follows for the Equidistribution Theorem.

Theorem. Let \( \alpha \) be any irrational number and let \( \mathbb{I} \) be an interval on \([0,1)\) of length \( \delta \). Then the probability that the fractional part \( \{n\alpha\} \) for \( n = 1, 2, 3, \ldots \) belongs to \( \mathbb{I} \) is equal to \( \delta \).

This was proved in 1909 and 1910 by:

- Piers Bohl (1865-1921)
- Waclaw Sierpiński (1882-1969)
- Hermann Weyl (1895-1955)