本福特定律

1938 年，物理学家本福特发表了一篇学术论文，声称在一个数据集中，数字的首位数越小，其出现频率越高。换句话说，在一个数据集中，以 1 开头的数字比以 2 开头的数字多，以 2 开头的数字又比以 3 开头的数字多，依此类推。这个模式与人们通常认为的所有首位数从 1 到 9 出现的频率大致相同的直觉不符。

更有趣的是, 在论文中，本福特甚至指出，一个数据集中任意数字以数字 d 开头的概率等于 $\log_{10}(1+1/d)$，其中 $d \in \{1, \ldots, 9\}$ （$\log_{10}$ 表示以 10 为底的对数）。因此，根据本福特的定律，在一个数据集中的任意数字以 1 开头的概率约为 0.30，而以 9 开头的概率则降到了不到 0.05。下图展示了这些概率。这一论断被称为“本福特定律”。

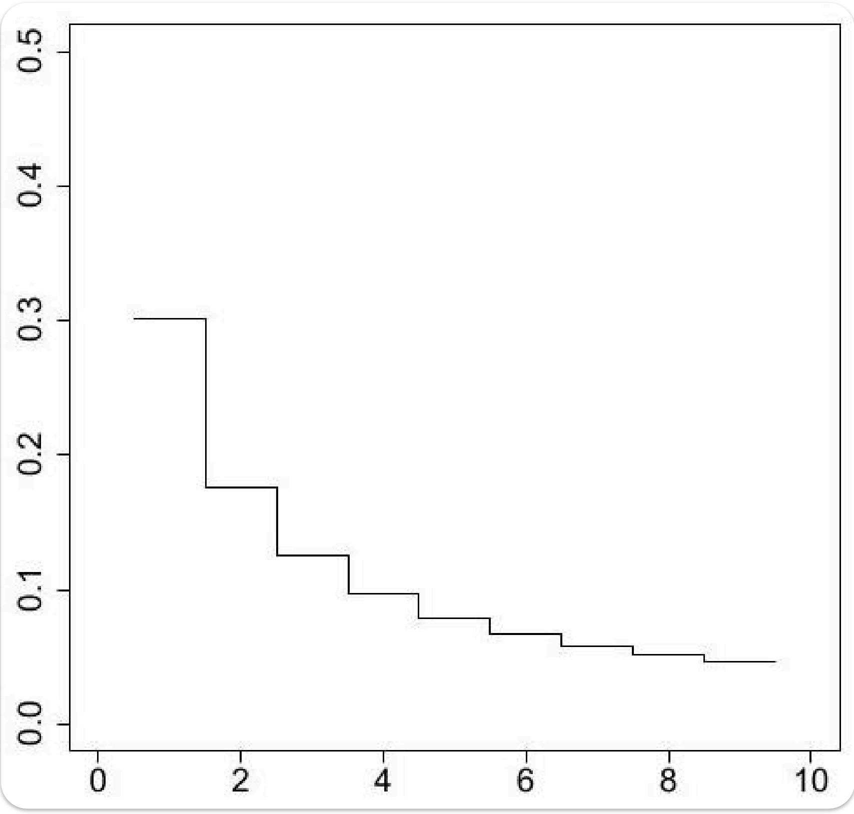


图: 根据本福特定律，不同首位数的概率。

本福特并不是第一个发现这一规律的人。早在五十多年前，即 1881 年，美国天文学家纽科姆就发表了一篇包含相同发现的学术论文。纽科姆注意到，带有对数表的书籍的前几页比后几页更脏，磨损也更严重。由于书籍的前几页包含的是首位数较低的数字，而后几页则包含首位数较高的数字，纽科姆推断，低首位数的对数比高首位数的对数被查询得更频繁。

让我们找个数据集试验一下到底是不是如此。

我们通过使用 CIA《世界概况》（2006 年 2 月）中的数据，这个数据集包含所有国家人口数量，下图展示了一个以这些国家人口数据为基础的首位数的直方图（是一个经过标准化的总面积为 1 的直方图），以及本福特频率。首位数的频率似乎很好地遵循了“本福特定律”。

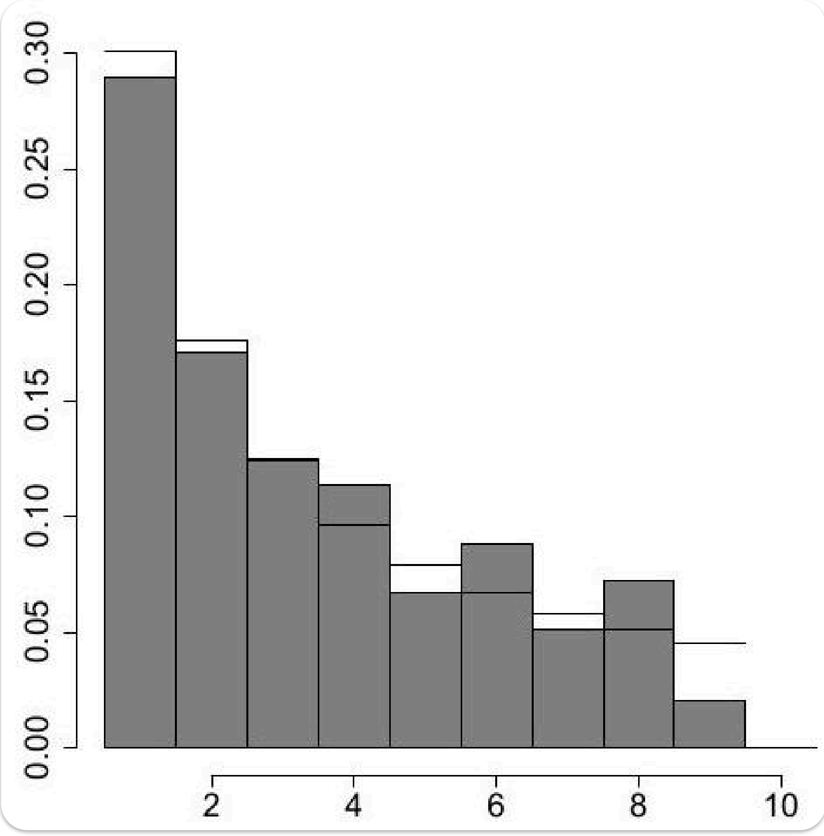


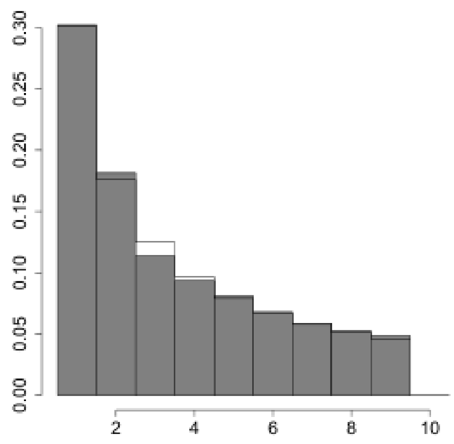
图: 世界各国人口数据集中首位数从 1 到 9 的直方图。图中直方图以外的阶梯函数表示基于本福特定律的预期频率。

许多数据集已经被用于验证本福特定律的有效性，从实验室中测量的物理量到地理信息（如河流长度和首都的人口数量），从企业会计数据到货币转换系数。在几乎所有情况下，本福特定律都大致成立。当然，并非每个数据集都适用。纯随机数（例如反复掷骰子的结果）或受限制的数字（如人口年龄或电话簿中的电话号码）并不符合本福特定律。

财务报表中的数字，例如一些大公司的账目，通常大致符合本福特定律。因此，这一定律可以用于核查账目以及调查欺诈和不一致情况。一个实施欺诈的员工，若试图掩盖其行为，通常会以某种方式捏造或篡改数字，使得首位数的分布趋于均匀。如果该员工经常篡改或捏造数字，那么他的行为将改变首位数的分布，使其偏离本福特定律所预测的分布。例如，如果账目中 9% 的数字以 9 开头，那么几乎可以肯定会对这些账目进行调查，因为根据本福特定律，只有 4.6% 的数字应该以 9 开头。

然而，也并非那么绝对, 偏离本福特定律并不意味着就一定存在欺诈。在某些情况下，人们可能更喜欢以 9 开头的数字；例如，价格为 99 欧元的产品通常比价格为 100 欧元的产品卖得更好。这种情况下出现首位 9 是相当合理的.

只有结构性(作用于整体结构的)欺诈可以通过本福特定律检测到。如果有一笔大额资金转入私人账户，而只关注本福特定律的偏差，那么这种单笔转账不会被察觉, 因为这不是结构性欺诈和偏差。下图展示了一家大公司的 150 万条账目数据中的首位数字的直方图（面积为 1），以及根据本福特定律预期的频率。这些账目的数字似乎很好地遵循了本福特定律。



尽管对本福特定律进行了大量研究，但为什么某些数据集符合这一定律，而其他数据集却不符合，仍然没有完全清楚的解释。

符合本福特定律的一个例子是指数增长的情况。让我们更详细地研究这种情况。因为只对数字的首位感兴趣，所以我们将一个数字 z 写成 $z = x \times 10^n$，其中 $1 \leq x < 10$ 且 $n \in \mathbb{Z}$。这种表示法适用于所有正数。我们称 x 为对应于 $z = x \times 10^n$ 的标准化观测值。数字 z 的首位等于 x 的首位(个位)。

设 D 为随机变量，表示某个数据集中任意（随机）数字 $Z = X \times 10^n$ 的首位数字。假设 X 服从 ab^Y 分布，其中 $a, b > 0$，且 Y 在区间 $[0, 1/\log_{10} b]$ 上均匀分布。那么

$$
\begin{aligned}
\mathrm{P}(D=k) &= \mathrm{P}(k \leq X < k+1) \\
&= \mathrm{P}\left(k \leq ab^Y < k+1\right) \\
&= \mathrm{P}\left(\log_{10}(k/a) \leq Y \log_{10} b < \log_{10}((k+1)/a)\right) \\
&= \log_{10}(k+1) - \log_{10} a - (\log_{10} k - \log_{10} a) \\
&= \log_{10}(1+1/k),
\end{aligned}
$$

其中第 4个等号是由 $Y \log_{10} b$ 的分布得出的，即在区间 $[0, 1]$ 上的均匀分布。
因此，首位数字 D 等于 k 的概率正是本福特定律所预测的概率。如果 $b = 10$，则 $\log_{10} b = 1$，并且假设 Y 在 $[0, 1]$ 上均匀分布。

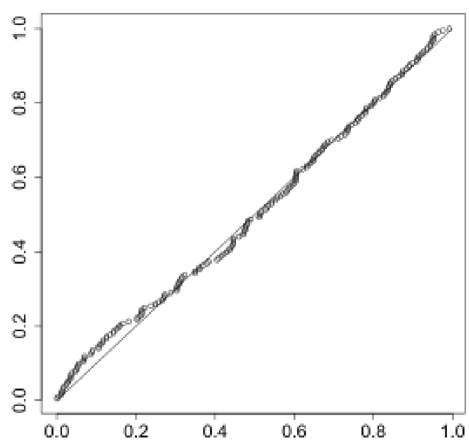


图: 展示了中的标准化人口数量的 $\log_{10}$ 的顺序统计量与 $[0, 1]$ 区间上的均匀分布分位数的 QQ 图。对于这个数据集，假设显然成立。

假设 X 服从 ab^Y 分布，其中 $a,b>0$ 且 Y 在区间 $[0,1/\log_{10}b]$ 上均匀分布，这个假设看起来不太直观，因此似乎不太现实。

然而，以下示例说明这个例子还是挺常见的。

假设一家公司市值为 d 百万欧元，每年以 $x\%$ 的速度增长。经过 t 年后，公司市值增长到 $d(1+x/100)^t$ 百万欧元。经过 $t=1/\log_{10}(1+x/100)$ 年后，公司市值增长了十倍。此时，市值的首位数字等于 $t=0$ 时的首位数字。由于这个时间跨度与初始金额 d 无关，因此时间可以任意选择，而我们只关心首位数字，所以只需要考虑 t 在区间 $[0,1/\log_{10}(1+x/100)]$ 内的取值。

设 T 为在区间 $[0,1/\log_{10}(1+x/100)]$ 上均匀分布的随机变量。对于市值为 d 的任意公司，其在时间 T 时的市值为 $Z=d(1+x/100)^T=(d/10^n)(1+x/100)^T10^n$，其中 $n\in\mathbb{N}$ 使得 $(d/10^n)(1+x/100)^T\in[1,10)$，且概率为 1。现在我们回到了前面的情况，其中 $Y=T, b=1+x/100$，$a=d/10^n$。一家公司在时间 0 的市值以首位数字 k 开头的概率等于本福特定律所给出的概率 $\log_{10}(1+1/k)$。

另一个得出相同结论的例子基于这样一个假设：某家公司市值以首位数字 k 开头的概率与该公司市值以首位数字 k 开头的时间跨度成正比。设 t_k 为公司市值从 k 百万欧元增长到 $k+1$ 百万欧元所需的时间跨度（以年为单位）；则有 $k(1+x)^{t_k}=k+1$，即 $t_k=\log_{10}(1+1/k)/\log_{10}(1+x/100)$。

因此，市值从首位数字 k（百万欧元）增长到首位数字 $k+1$（百万欧元）所需的时间跨度与根据本福特定律得出的首位数字为 k 的概率成正比。当然，这与选择的“百万欧元”单位无关。我们再次得出结论，在我们的假设下，所有公司的市值以首位数字 k 开头的比例大约为 $\log_{10}(1+1/k)$，正如本福特定律所预测的那样。