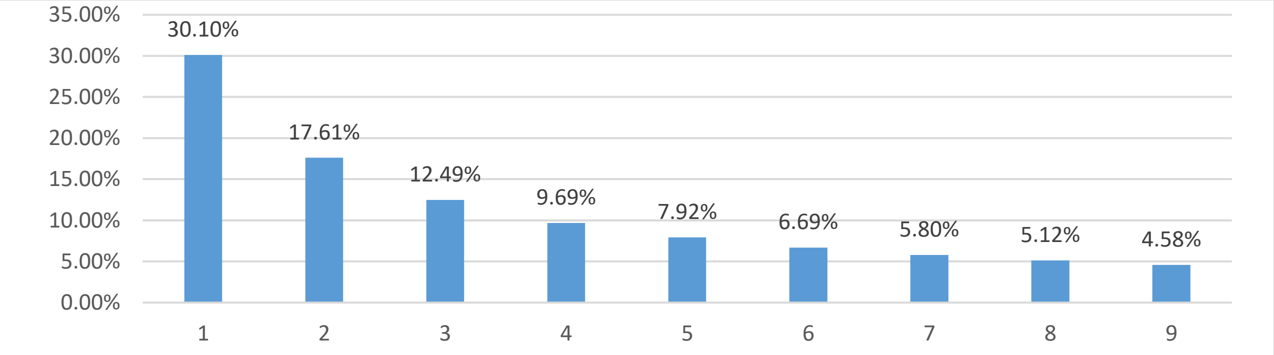
并不平均的首位数字——本福特定律

191850128 数学系 陆青阳

如果问你这样一个问题：世界上的所有数据中[1]，第一位是 1 的概率有多大？在略加思考之后，相信很多人给出的答案都是$\frac{1}{9}$，理由也很简单，因为世界上的所有数据一定是“均匀”分布的，因此 1-9 作为首位数字的概率自然应该相同。然而，本福特给出的结论可能会让你大跌眼镜——以 1 开头的数据占到总量的 30%还多！



本福特定律给出的首位数字比例

想象从 1 开始数数的过程，我们很容易发现只有当数完了 11，12，……，19 之后，才轮到 2 开头的数字，而只有当 1-8 开头的两位数都数完之后才轮到 9 开头的。因此考虑到在中间停止的情况，毫无疑问 1-9 开头的概率会依次降低，这样看来，本福特所给出的分布规律似乎也不那么难以理解。

本福特得出这一规律的过程听上去有些偶然。据说他在查看对数表时发现 1 开头的部分已经被翻烂了而 9 开头的仍然崭新，由此想到 1 开头的数是否更多。通过统计国家面积、人口等大量数据，他成功证实了这一分布规律。

首位	人口计数	人口比例	面积计数	面积比例
1	64	27.59%	70	30.17%
2	39	16.81%	43	18.53%
3	24	10.34%	26	11.21%
4	29	12.50%	27	11.64%
5	20	8.62%	15	6.47%
6	17	7.33%	17	7.33%
7	15	6.47%	13	5.60%
8	13	5.60%	8	3.45%
9	11	4.74%	12	5.17%

对全球 232 个国家和地区的人口和面积的首位数统计

本福特定律的完整形式为：

$$\text{自然产生的数据在 d 进制下,首位为 k 的数占比为} \log_d \frac{k+1}{k}$$

我们首先证明，等比数列满足这一分布规律。

证明

设进制为 d，取数列$\{a_n\} = q^n, q \in N^*, \log_d q \in R \setminus Q$,

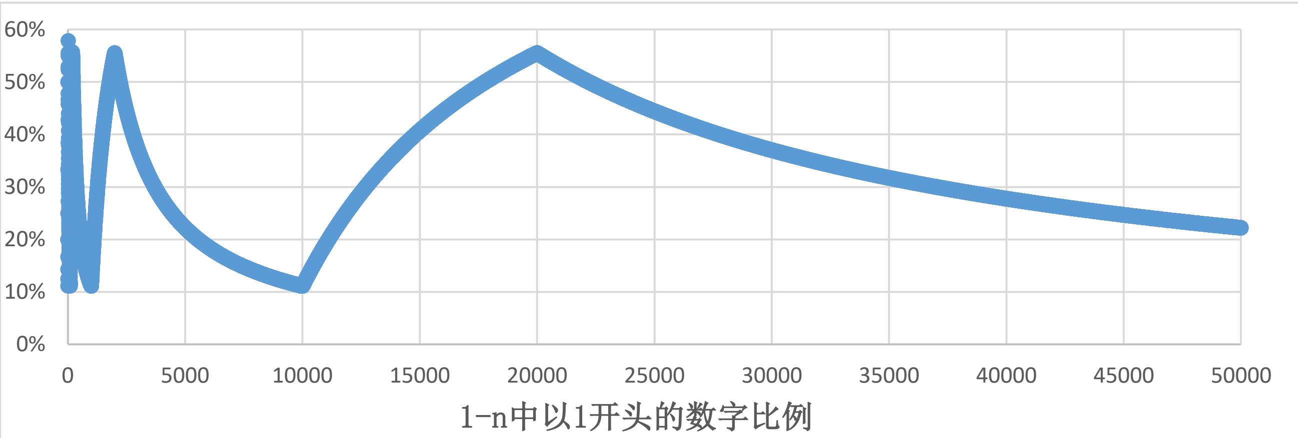
则a_i的首位为 k$\Leftrightarrow \exists t \in N, k \cdot d^t \leq a_i < (k+1) \cdot d^t \Leftrightarrow t + \log_d k \leq i \log_d q < t + \log_d (k+1)$

即$\log_d k \leq \{i log_d q\} < \log_d (k+1)$, 其中$\{x\}$表示 x 的小数部分

又由$\log_d q$是无理数, 根据 Weyl 均匀分布定理, $\{i log_d q\}$在$(0,1)$中均匀分布

因此a_i的首位为 k 的概率为$\log_d (k+1) - \log_d k$, 满足本福特定律

大家一定会感到疑惑，为什么是对等比数列进行证明？难道用数列$a_n = n$不行吗？很遗憾，如果用等差数列，十进制下 1-n 中以 1 开头的数所占比例将如下图所示：



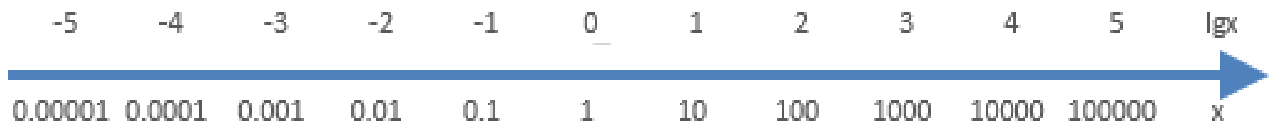
观察其两个子列$n = 10^k$以及$n = 2 \cdot 10^k$也可发现它们的极限分别为$\frac{1}{9}$和$\frac{5}{9}$, 因此极限不存在。

新的问题必然接踵而来：既然本福特通过总结海量数据得出这一规律，难道这些数据都是指数增长的吗？首先我们需要认识到，指数增长在我们的生活中广泛存在。例如存贷款中利息与本金的关系，生物繁殖过程中的数量变化等等。就拿此次新冠肺炎疫情举例，众所周知，病

毒在没有达到饱和感染前种群规模近似按照“J 型曲线”发展，也即按指数关系增长。通过对约翰霍普金斯大学给出的 187 个国家从 4 月 3 日至 16 日总感染人数进行分析，发现其首位数字分布如右表所示，可以看出，确实与本福特定律符合得非常不错。

首位	概率
1	33.64%
2	15.63%
3	10.85%
4	9.47%
5	7.96%
6	6.91%
7	6.33%
8	4.65%
9	4.57%

当然，还有一个疑问我们也不能回避，为什么国家面积、人口、物理常量这些数字也会按照等比数列分布？为了解决这一问题，我们首先从另一个角度来看本福特定律。



这是天文中常用的对数坐标轴，观察 lgx 中的任意一格，以$[0,1]$为例，我们惊奇地发现，$[0,\lg 2)$中的所有数在 x 轴上首位均为 1，$[\lg 2,\lg 3)$中的所有数在 x 轴上首位均为 2，……且在每格中都是如此。现在我们考虑在 lgx 上完全均匀分布的数据，很容易得到首位为 k 的数在每格中所占比例均为$\lg\frac{k+1}{k}$，在所有数据中所占比例自然也为$\lg\frac{k+1}{k}$，与本福特定律的描述完全相同。这一事实告诉我们，某种意义上来说人们的直觉并没有错，世界上所有的数据的确是“均匀”分布的，只不过不是在一般的数轴上，而是在对数坐标轴上！

因此，我们只要解决这个问题：为什么土地面积、人口、物理常量这些数字在对数坐标轴上是均匀的？以面积为例，首先我们需要意识到一个重要的事实：如果世界上所有面积的首位数字存在一定的分布规律，那么这一分布规律一定对任何成比例的面积放缩都是不变的。举例而言，如果所有面积以平方千米计算得到的分布比例是 A，那么以平方海里计算得到的分布比例必然也要是 A，若不然，这一结论显然是荒谬的，毕竟所有的单位都是经人为选择而创造的，而这一分布规律应当是大自然与生俱来的。有了这一结论，用反证法就可以轻松解决问题——如果所有面积在对数坐标轴上两段长度相等的区域分布密度不同，那么只要乘一个系数（即对

数坐标轴上的平移），使两块重合，便直接导出矛盾。这样一来，土地面积，人口，物理常量之类的数字满足本福特定律也就不足为奇了。

自被发现以来，本福特定律已经在多个领域得到广泛应用，例如在金融领域用于检查数据报表是否造假，在科学计算中用于生成符合真实情况的随机数，以及检验数学建模的正确性等等。当然，我们有必要强调，正如牛顿运动定律只适用于宏观低速的情况下一样，本福特定律不可避免地有一定的局限性，有它自身的适用范围，那就是指数增长，也即对数分布的数据。对于生活中那些有规律，有范围的数据，本福特定律一般是不适用的。只有做到巧用而不滥用，本福特定律才能在数据的海洋中发挥出强大的威力。

注：[1]本文中的数均用科学计数法表示，不考虑首位为 0 的情况。