彩票，想说爱你不容易

1998 年 8 月，电脑销售传统型中国体育彩票在江苏的第一位百万大奖得主诞生了。在此后的短短几年的时间内，彩票已造就了 200 余位百万富翁，一夜暴富的神话在社会上掀起了一股彩票狂潮。

由于中奖与否的不确定性，所以各种报章上竞相开设“彩经”专栏，各种“预测”文章纷至沓来，概率统计成了时髦的话题。那么，你手中的一注彩票中大奖的概率究竟有多少呢？

以 200072 期大扩容前的江苏体育彩票为例，它的对奖号码共有 7 位数字，前 6 位每位各有 0～9 十个数字可供选择，第 7 位为特别号，共有 0～4 五个数字可供选择，只有当这 7 位数字与中奖号码完全相同时才能赢得最高金额可达 500 万元的特等奖。

由于中奖号码的产生共有

$$10^6 \times 5 = 5\ 000\ 000\ (\text{种})$$

可能性，故一注体育彩票中大奖的概率只有五百万分之一，而目前这一概率已随着体育彩票特别号的大扩容已降到了一千万分之一。

这是一个什么样的概率？2000 年上半年，南京市区因各类交通事故而死亡的人数达 128 人，若每日在南京市区活动的人口以 280 万人计，则一个人一天之中在南京街头死于交通事故的概率就有四百万分之一。因此，说中大奖比出门遇上交通事故而致死的可能性还要低，一点也不过分。

从 2000 年上半年开始发行的江苏风采电脑福利彩票，其中奖规则与体育彩票略有不同，彩民可从 1 到 35 中任意选取 7 个数字

来构成对奖号码,这 7 个数字不能重复，只有当这 7个数字与中号码完全相同时（不考虑先后次序）才能最后赢得大奖。

由于中奖号码的产生共有

$$C_{35}^{7}=\frac{35!}{7!\ 28!}=\frac{35\times34\times33\times32\times31\times30\times29}{7\times6\times5\times4\times3\times2\times1}=6\ 724\ 520(种)$$

可能性，故一注福利彩票中大奖的概率不足六百万分之一。

为了达到吸引彩民的目的，福利彩票最初的发行宣传工作在中奖号码不排序上做足了文章，但这丝毫无助于改变其中奖率偏低的事实。有人曾在报章上发表文章，抓住福利彩票最初期大奖不断的事实，用中奖注数与投注总数之比来说明福利彩的中奖率高于体育彩票，闹了以频率代替概率的笑话，这种说法着后面几期大奖的不断轮空也就烟消云散了。

历史上的第一个概率论问题

实际上，概率论的起源与博彩有着密不可分的关系。意大利数学家和赌徒卡丹诺（Girolamo Cardano, 1501—1576）在 1564 年写成的《机遇博弈》一书中就已提出了所谓“胜率”的概念，尽管它与概率还不是一回事。

历史上的第一个得到系统研究的概率论问题，是在 1654 年 7 月到 10 月间由法国商人贡博和梅雷提出的赌徒分赌金问题。

梅雷的基本问题是：甲、乙两人以掷硬币赌输赢，掷出正面甲得一点，掷出反面则乙得一点，积满三点者赢得全部赌注。现甲已得两点，乙只得一点，赌局意外中止，两人应怎样分配赌本才合理？

梅雷将这个问题交给了当时法国著名的数学家帕斯卡（Blaise Pascal, 1623—1662），帕斯卡与另一位法国大数学家费尔马（Pierre de Fermat, 1601—1665）通信讨论了这一问题，后者以“费尔马大定理”而著称于世。两人在不同的模型下运用组合理论对这一问题给出了各自的答案。

帕斯卡认为：若再掷一次，甲胜，则甲应得全部赌注；甲负，则甲应与乙平分赌注。而这两种情况发生的可能性都是 $\frac{1}{2}$，故甲应得

$$1\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}=\frac{3}{4}$$

的赌注。

费尔马认为，要结束比赛，最多再掷两次即可，有 4 种等可能情形，即甲连胜、甲先胜后负、甲先负后胜、甲连负。前 3 种情况都是甲胜，故甲应得

$$1\times\frac{3}{4}+0\times\frac{1}{4}=\frac{3}{4}$$

的赌注。

1655 年，荷兰物理学家、数学家惠更斯（Christian Huygens，1629—1695）到巴黎访问期间听说了这件事，对概率论产生了浓厚的兴趣，决定自己也来研究这一问题。1657 年，他发表了《论机会游戏中的计算》一书，尽管它比卡丹诺的遗著《机遇博弈》晚成书近 1 个世纪，但由于后者直到 1663 年才得以面世，故《论机会游戏中的计算》当之无愧地成为了历史上第一部概率论著作。

惠更斯在书中除解决了许多有趣的实际问题外，还在帕斯卡和费尔马的计算结果的基础上引进了离散型随机变量的均值——数学期望的概念，即一个人如果赢得总数 a_i 的概率是 $p_i, i=1, 2, \cdots, n$，则他可望赢得的总数为

$$\sum_{i=1}^{n} a_i p_i = a_1 p_1 + a_2 p_2 + \cdots + a_n p_n。$$

跳舞的小人

除了博彩，概率论的方法在很多方面的使用都有着悠久的历史。英国侦探小说家柯南道尔（Arthur Conan Doyle, 1859—1930）在小说《归来记》中，就曾讲述了大侦探福尔摩斯所遇到的这样一个情场奇案。

马场村庄园的丘比特先生在花园里发现了一张纸条，上面画着一群跳舞的小人。她的妻子埃尔茜看到这封密码信后惊恐万分，当场昏死过去。

福尔摩斯拿到这张纸条后，首先发现第 4、第 6、第 9 和最后 1 个位置上所画的小人是完全一样的，除了手中多出的一面小旗。从小旗的分布来看，它应该起着分隔单词的作用，故福尔摩斯断定，这 4 个小人应代表英文字母中最常见的字母 E.

“可是，现在最难的问题来了。”福尔摩斯在事后分析案情时说，“因为除了 E 以外，字母按出现次数排列的顺序大致为 T, A, O, I, N, S, H, R, D, L，但是 T, A, O, I 出现的次数几乎不相上下。要是把每一种组合都试一遍，那会是一项无止境的工作。所以，我只好等来了新材料再说。”

好在随后神秘的小人不断出现，福尔摩斯在其中一组不带小旗的 5 个小人中发现第 2 和第 4 个都是 E。这个单词可能是 sever（切断），也可能是 lever（杠杆），还可能是 never（决不），而用 never 来回答一项请求的可能性极大，故福尔摩斯找到了字母 N, V, R.

接着，他又将一个数次出现的两头是 E、中间是 3 个别的字母的单词确定为埃尔茜的名字 ELSIE，并在以‘埃尔茜’结束的带有恳求语气的一句话中，断定名字前面的一个以 E 结尾的 4 个字母的单词为 come(来)。这样，第一张纸条就变成了：

□M □ERE □□E SL□NE.

作为在纸条上出现了三次的第一个小人，就只能是字母 A. 而第二个单词的开头也只能是 H。至此，这群跳舞的小人就彻底撩开了它的面纱：

AM HERE. ABE SLANE. （我已到达。阿贝·斯兰尼。）

福尔摩斯对这起牵涉到埃尔茜婚前秘密的事一直都采取袖手旁观的态度，直到他看到阿贝·斯兰尼最后画的一行小人：

ELSIE, PREPARE TO MEET GOD. （埃尔茜，准备见上帝吧。）

频率与概率

100多年前，柯南道尔就注意到了每个英文字母出现的频数是不一样的,其排列的顺序与现代学者研究的结果相差无几(如 H 与 R,D 与 L 的先后次序略有出入）。下表就是由 G. Dewey 在统计了 438 023 个字母的文字材料后得到的一份英文字母频率表，发表于 1970 年。

字母	频率	字母	频率	字母	频率
E	0.126 8	L	0.039 4	P	0.018 6
T	0.097 8	D	0.038 9	B	0.015 6
A	0.078 8	U	0.028 0	V	0.010 2
O	0.077 6	C	0.026 8	K	0.006 0
I	0.070 7	F	0.025 6	X	0.001 6
N	0.070 6	M	0.024 4	J	0.001 0
S	0.063 4	W	0.021 4	Q	0.000 9
R	0.059 4	Y	0.020 2	Z	0.000 6
H	0.057 3	G	0.018 7		

从表中可以看出，前 6 个字母出现的频率就占到了全部字母的 52.2%，其中仅头 3 个字母 E,T,A 就超过了 30%。看来，无论“过去”和“现在”；“吃”(eat 或 ate)都是最重要的。

当然，在一篇具体文章中，每个字母出现的频数会有所不同，排序结果也会出现某些差异，这就是频率的随机波动性。所以福尔摩斯对于他所得到的第一个样本，采取了审慎的态度。但是，随

着样本数的不断增大，频率会逐渐稳定于某个常数，这个常数就是我们所关心的概率。

早在 3 个世纪前，瑞士数学家伯努利(Jakob Bernoulli, 1654—1705)就在依概率收敛的意义上,证明了 n 次独立重复试验中事件 A 发生的频率,当 n 趋向于无穷大时，趋向于事件 A 发生的概率 $P(A)$，这就是概率论中的第一个极限定理——伯努利大数定律。

虽然当那本包含了这一定律的著作《猜度术》于 1713 年出版时,伯努利已经去世了 8 年，但他从此奠定了概率论作为一门独立的数学分支的基础。因此，在实际问题中，我们可以用频率来作为概率的估计值,当然这只有在 n 充分大(即试验次数充分多)的时候才是有效的。

汉字的用字频率

至于汉字，到目前为止，还没有一个权威的频率统计表。20世纪70年代前后,曾有不少人对《毛泽东选集》1～4卷中出现的汉字进行过手工统计,其中有一份资料显示“毛选”四卷中共用字66万个,其中不重复的单字2 981个,出现次数在2 300次以上的有50个,占总字数的39.0%.

下面即为这50个汉字的出现频率表。

单字	频率	单字	频率	单字	频率
的	0.052 6	们	0.009 3	为	0.005 2
是	0.018 2	了	0.009 0	就	0.005 2
一	0.016 3	有	0.009 0	以	0.005 0
国	0.014 2	地	0.008 4	产	0.004 9
民	0.011 8	党	0.007 0	于	0.004 7
不	0.011 8	个	0.007 0	对	0.004 7
和	0.011 2	我	0.006 8	日	0.004 5
在	0.010 9	要	0.006 1	命	0.004 4
中	0.010 9	大	0.005 9	动	0.004 4
人	0.010 6	义	0.005 8	革	0.004 3
这	0.010 1	军	0.005 6	反	0.004 3
战	0.009 7	争	0.005 4	方	0.004 2
主	0.009 4	政	0.005 3	上	0.004 2

续表

单字	频率	单字	频率	单字	频率
作	0.004 1	而	0.003 9	来	0.003 6
时	0.004 0	之	0.003 8	同	0.003 5
能	0.004 0	他	0.003 6	力	0.003 5
阶	0.004 0	会	0.003 6		

汉字号称有 8 万多个,但常用字却并不多;“毛选”中只用了不足 3 000 个字,恐怕是很多人没有想到的。从表中可以看出,前 14 个单字出现的次数就占了总字数的 20.7%,其中排名第一的“的”字更是独领风骚,平均每 19 个字中就出现一次。

当然,由于时代背景和语言习惯的不同;“毛选”中一些单字出现的频率偏高(如“战”“争”等),但“毛选”作为现代汉语的范文,这一统计结果还是有一定的参考价值的。

古典概型

在概率论的初创时期，所用的模型大多是等可能概型，即所谓古典概型。其特点是试验中所有可能的结果可以划分为一些基本事件，而每个基本事件发生的概率都相同。

若记 n 为基本事件的总数，k 为事件 A 中所包含的基本事件数，则事件 A 发生的概率为

$$P(A)=\frac{k}{n}。$$

在等可能概型的计算中，人们会发现这样的情况，那就是用两种不同的观点来进行演算的时候往往会出人意料地得到两个不同的结果，这常常令初学者感到苦恼。下面的弹子球游戏就是一个典型的例子。

一个男孩有一个弹子球，一个女孩有两个弹子球。他们向远处的同一个目标把球弹出，球离目标最近者胜。假定男孩和女孩的弹球技巧完全相同，测量也完全精确而足以定出胜负，那么女孩胜出的概率是多少呢？

第一种观点：女孩有两次机会弹球，而男孩只有一次，男孩和女孩的水平又是相当的，故女孩赢球的概率为 $\frac{2}{3}$.

第二种观点 把女孩的球记作 A 和 B，男孩的球记作 C，则有下列 4 种情况：

(1) A 球和 B 球都比 C 球更接近目标；

(2)仅 A 球比 C 球接近目标；

(3)仅 B 球比 C 球接近目标；

(4) C 球比 A 球和 B 球都接近目标。

这 4 种情况中的前 3 种都是女孩赢，所以女孩赢球的概率是 $\frac{3}{4}$。

为什么会出现这种情况？两种观点中哪一种是错误的呢？如果我们把所有可能的情形一一列出，就很容易找到答案。

按 3 个球接近目标的程度，实际上有 6 种等可能的情形，即

$$ABC、ACB、BAC、BCA、CAB、CBA,$$

而前 4 种情况均为女孩胜出，故女孩赢的概率为

$$\frac{4}{6}=\frac{2}{3}。$$

那么第二种观点又错在哪里呢？它所分析的第 1 种情况实际上包含了 ABC 和 BAC 两种情形，第 4 种情况也包含了 CAB 和 CBA 两种情形，而第 2 种和第 3 种情况却分别只包含了 ACB 与 BCA 一种情形，所以这 4 种情况发生的概率是不一样的，不能使用我们开头所讲的公式。

最有可能的性别组合

已知一个家庭中有 2 个孩子,问这 2 个孩子为一男一女的概率是多少？你可能会说，这里总共只有 3 种可能的情形：

(1)2 个男孩；

(2)2 个女孩；

(3)1 个男孩和 1 个女孩。

所以这 2 个孩子为一男一女的概率是 $\frac{1}{3}$.

如果你这样回答，你就犯了和上面弹子球游戏中同样的错误。实际上，如果给每个孩子编上号（例如按照他们的出生次序，第一个出生的为 1 号,第二个为 2 号),则会出现 4 种等可能的情形：

(1)1 号是男孩,2 号还是男孩；

(2)1 号是男孩,2 号是女孩；

(3)1 号是女孩,2 号是男孩；

(4)1 号是女孩,2 号还是女孩。

由于一男一女占了其中的 2 种，所以它发生的概率为 $\frac{1}{2}$,而不是 $\frac{1}{3}$。

若一个家庭中有 4 个孩子，如果我告诉你最有可能的性别组合是 3—1 组合(3 个同性,1 个异性)而不是 2—2 组合(2 个男孩,2 个女孩),你一定会感到惊讶。

实际上,一个家庭中有 2 个男孩、2 个女孩的概率是

$$\frac{C_4^2}{2\times 2\times 2\times 2}=\frac{3}{8}.$$

而有 3 个男孩、1 个女孩或 3 个女孩、1 个男孩的概率是

$$\frac{2\times C_4^1}{2\times 2\times 2\times 2}=\frac{1}{2},$$

而余下的 $\frac{1}{8}$ 则是 4 个孩子性别相同(4 个男孩或 4 个女孩)的概率。

同样的,在有 6 个孩子的家庭中,最有可能的性别组合是 4—2 组合而不是 3—3 组合,前者发生的概率为 $\frac{15}{32}$,后者发生的概率为 $\frac{5}{16}$,余下的依次为 5—1 组合($\frac{3}{16}$)、6—0 组合($\frac{1}{32}$)。

从上面可以看出,孩子性别相同的概率随着孩子人数的增加而急剧减少,如果你在街上遇到了一位久未谋面的老朋友,你只记得他家里至少有 2 个孩子,但是是男是女已记不清了,而你又不想显得你把他的情况忘得一干二净,你完全可以冒昧地问一句:“你儿子现在怎么样了?”,因为你至少有 $\frac{3}{4}$ 的把握。

测测你的手气

如果你是一位桥牌手，你会很关心你拿了一手什么牌。关于一手桥牌中 4 种花色的最有可能的分布，其答案也同样违反直觉。即使是有多年实战经验的桥牌手，也往往会猜想手头的 13 张牌中最有可能的花色分布是 4,3,3,3,而答案却是 4,4,3,2.

这一类概率的计算是一件需要技巧的工作，下面我们就以这两个组合为例来具体说明一下演算过程。

从 52 张牌中选出 13 张,共有 C_{52}^{13}种选法 而这 13 张牌要形成 4,4,3,2 组合，首先要确定它们所对应的花色。

我们先从 4 种花色中任选 1 种作组合数字中的 2(例如红心),再从余下的 3 种花色中任选 1 种作 3(例如黑桃),余下的 2 种花色(方块、梅花)就是我们所要的 4,4，所以花色的选择共有 4×3种 。

接着我们再从 13 张红心中选出 2 张(共有 C_{13}^{2}种选法),13 张黑桃中选出 3 张(共有 C_{13}^{3}种选法),13 张方块、13 张梅花中各选 4 张(各有 C_{13}^{4}种选法),则 4,4,3,2 组合出现的概率就是

$$\frac{4\times3\times C_{13}^{2}C_{13}^{3}C_{13}^{4}C_{13}^{4}}{C_{52}^{13}}\approx0.215\ 5,$$

换句话说,这种组合大约每 4 到 5 圈就可以拿到一次。

至于 4,3,3,3 组合，通过类似的分析我们可以得到它出现的概率为

$$\frac{4\times C_{13}^{4}C_{13}^{3}C_{13}^{3}C_{13}^{3}}{C_{52}^{13}}\approx0.105\ 4,$$

所以它平均 9 到 10 圈才能拿到一次，其发生的可能性甚至不如

5, 3, 3, 2 组合,后者平均 6 到 7 圈就可以拿到一次。

有人常常宣称他曾经拿到过一手完满的牌（13 张牌花色完全相同），对这种天方夜谭式的故事你完全可以一笑置之，因为如果不是有人作弊，这种事情发生的概率只有

$$\frac{4}{C_{52}^{13}}=\frac{1}{158\ 753\ 389\ 900}。$$

即使你 1 秒钟就能翻出一手牌，这种情况平均也要 5 000 年才会出现一次,而 5 000 年前我们的祖先还处在刀耕火种的阶段，不识桥牌为何物也。

实际推断原理

在现实生活中，我们会遇到一些小概率事件，但它应该是在大量重复试验中才会产生的。很少有人会担心一颗预报要落在地球上的陨石会落到自己头上，即使像作为世界第一大城的南京城这样大的范围，这块陨石落在其中的概率也只有 0.000 000 127(南京明城墙所围区域面积与地球表面积之比)，除非我们面临的是一场横扫地球的流星雨。

在概率论中，我们所讨论的都是随机现象，其结果在个别试验中应呈现出不确定性，而在大量重复试验中又具有统计规律性。如果一个在理论上发生概率非常小的事件在单个试验中就出现了，我们完全有理由怀疑这样的结果是否具有随机性。

因此，在实际问题中我们认为概率很小的事件在一次试验中几乎是不可能发生的，这就是所谓实际推断原理。

下面我们来看一个例子。

某一天，你开车进了一个过去从未去过的机关停车场，发现里面共有 18 个车位,其中有 8个位置停了车,而有一连 10 个位置是空着的。这时，你可以随便找个地方把车停下吗？

由于车辆的停放太有规律，我们当然有理由怀疑它的随机性。我们先假定车辆的停放是随意的，则一连 10 个位置空着共有 9种可能的情况(从 1～10 号车位空着到 9～18 号车位空着),故这种放法出现的概率为

$$\frac{9}{C_{18}^{8}}=\frac{1}{4\ 862}\approx 0.000\ 2。$$

如此小概率的事件竟然发生了，可以肯定关于停车位置是有具体