从抛硬币试验看概率论的基本内容及统计方法

作者 张洋 | 发布于 2012-11-20

概率 数理统计 数学

一般说到概率，就喜欢拿抛硬币做例子。大多数时候，会简单认为硬币正背面的概率各为二分之一，其实事情远没有这么简单。这篇文章会以抛硬币试验为例子并贯穿全文，引出一系列概率论和数理统计的基本内容。这篇文章会涉及的有古典概型、公理化概率、二项分布、正态分布、最大似然估计和假设检验等一系列内容。主要目的是以抛硬币试验为例说明现代数学观点下的概率是什么样子以及以概率论为基础的一些基本数理统计方法。

概率的存在性

好吧，首先我们要回答一个基本问题就是概率为什么是存在的。其实这不是个数学问题，而是哲学问题（貌似一般存在不存在啥的都是哲学问题）。之所以要先讨论这个问题，是因为任何数学活动都是在一定哲学观点前提下进行的，如果不明确哲学前提，数学活动就无法进行了（例如如果在你的哲学观点下概率根本不存在，那还讨论啥概率论啊）。

概率的存在是在一定哲学观点前提下的，我不想用哲学术语拽文，简单来说，就是你首先得承认事物是客观存在的，并可以通过大量的观察和实践被抽象总结。举个例子，我们经常会讨论“身高”，为什么我们都认为身高是存在的？因为我们经过长期的观察实践发现一个人身体的高度在短期内不会出现大幅度的变动，因此我们可以用一个有单位的数字来描述一个人的身体在一段不算长的时间内相对稳定的高度。这就是“身高”作为被普遍承认存在的哲学前提。

与此相似，人们在长期的生活中，发现世界上有一些事情的结果是无法预料的，例如抛硬币得到正面还是背面，但是，后来有些人发现，虽然单次的结果不可预料，但是如果我不断抛，抛很多次，正面结果占全部抛硬币次数的比率是趋于稳定的，而且次数越多越接近某个固定的数值。换句话说，抛硬币这件事，单次结果不可预料，但是多次试验的结果却在总体上是有规律可循的（术语叫统计规律）。

下面是历史上一些著名的抛硬币试验的数据记录：

试验者	试验次数	正面次数	正面占比
德摩根	4092	2048	50.05%
蒲丰	4040	2048	50.69%
费勒	10000	4979	49.79%
皮尔逊	24000	12012	50.05%
罗曼洛夫斯基	80640	39699	49.23%

可以看到，虽然这些试验在不同时间、不同地点由不同的人完成，但是冥冥中似乎有一股力量将正面的占比固定在50%附近。

后来，人们发现还有很多其它不可预测的事情都与抛硬币类似，例如掷骰子、买六合彩等等，甚至渐渐发现不只这些简单的事情，人类社会方方面面从简单到复杂的很多不可预测的事情宏观上看都具有统计规律。于是人们推测，在某些条件下的一些不可预测事件，都是有统计规律的，或者直观说很多不可预测结果的试验在多次进行后总体上看结果会趋近于一些常数（这个现象后来被严格定义为大数定律，成为概率论最基础的定理之一，下文会提到）。这种可观测现象，成为概率存在的哲学基础，而这些常数就是概率在朴素观点下的定义。

概率模型

在认识到上述事实后，人们希望将这种规律加以利用（人类文明的发展不就是发现和利用规律么，呵呵），但是想要利用就首先要对概率进行严格的形式化定义，也就是要建立数学模型。比较知名的数学模型有古典概型、几何概率模型和公理化概率，本文将会讨论古典概型和公理化概率。

古典概型

古典概型是人类对概率和统计规律最早的建模尝试，表达了朴素的数学原则下人们对概率的认识。在表述古典概型之前，需要先定义一些概念。

首先是随机试验。

如果一个同时试验满足下面三条原则，则这个试验称为随机试验：

1、可在相同条件下（相对来说）重复进行。

2、可能出现的结果不止一个，但事先明确知道所有可能的结果（可以是无限个，例如所有自然数，但必须事先明确知道结果的取值范围）。

3、事先无法预测在一次试验中哪一个结果会出现。

显然上面的抛硬币试验是一个随机试验。

然后需要定义样本空间和样本点。一个随机试验的样本空间是这个试验所有可能结果组成的集合，而其中每个元素是一个样本点。例如，抛硬币试验中，样本空间为$\{F, B\}$，其中F表示正面，B表示背面，而F、B就是两个样本点。

另一个非常重要的概念就是随机事件（简称事件）：样本空间的一个子集称为一个事件。例如，抛硬币试验有四个不同的事件：$\emptyset$，$\{F\}$，$\{B\}$，$\{F, B\}$，分别表示“既不出现正面也不出现反面”，“出现正面”，“出现反面”和“出现正面或反面”。在不考虑硬币立起来等特殊情况时，第一个事件不可能出现，但它确实是一个合乎定义的事件，叫不可能事件；而最后一个事件必然出现，叫必然事件。

有了上面概念，就可以定义古典概型了：

如果一个概率模型满足 1）样本空间是一个有限集合，2）每一个基本事件（只包含一个样本点的事件）出现的概率相同，则这是一个古典概型。例如，在上面的抛硬币试验中，再定义$\{F\}$，$\{B\}$的概率均为0.5，则就构成了一个古典概型。

古典概型简单、直观，在早期的概率研究中广泛被使用。但是这个模型太朴素太不严格了，在这种不完善的定义下，根本没有办法做严格的数学推理，而且有限样本空间和等可能性在很多现实随机试验中并不满足，甚至对等可能不同定义会导致不同结论。因此必须使用一个更严格的定义，以符合现代数学公理化推导的要求，这就是公理化概率。

公理化概率

公理化概率对概率做如下定义：

概率是事件集合到实数域的一个函数，设事件集合为E，则如若$A \in E \xrightarrow{p} P(A) \in \mathbb{R}$满足：

对于任意事件A，$P(A) \geq 0$。

对于必然事件S，$P(S) = 1$。

对于两两互斥的事件，有$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$。

公理化概率对概率做了严格的数学定义，可以较好的基于公理系统进行推导和证明。但是，概率模型只是给出了概率“是什么”（定性），没有回答“是多少”（定量）这个问题。也就是说，仅有概率模型，是不能定量回答抛硬币问题的。下面介绍对概率进行定量分析的方法。

度量与估计概率

从公理化概率的角度，我们可以这样定义抛硬币试验的概率：设N是全部抛硬币的次数，而C_F是正面向上的次数，则如下函数定义了这个概率：

$$P(A) = \begin{cases} 0 & A = \emptyset \\ \frac{C_F}{N} & A = \{F\} \\ 1 - \frac{C_F}{N} & A = \{B\} \\ 1 & A = \{F, B\} \end{cases}$$

容易验证，这个定义完全符合公理化概率的所有条件。下面就是确定N和C_F。不幸的是，显然N是无法穷尽的，因为理论上你不可能抛无数次硬币。由于不能精确度量这个概率，因此你必须通过某个可以精确度量的值去估计这个概率，而且还要从数学上证明这个估计方法是靠谱的，最好能定量给出这个估计量的可信程度。而对不可直接观测概率的一个估计度量值就是频率。

频率估计

频率是这样定义的：事件A的频率是在相同条件下重复一个实验n次，事件A发生的次数在n次实验中的占比。一种简单的估计概率的方法就是用频率当做概率的估计。

例如，我刚刚抛完十次硬币，其中六次正面，四次背面，因此根据此次实验，我估计我这枚硬币出现正面的概率为0.6。这就是频率估计。

不过你一定有疑惑，为什么可以使用频率估计概率？有上面理论依据？如何对估计的准确性做出定理的分析？下面解答这些问题。

大数定律

频率估计的理论基础是大数定律。毫不夸张的说，大数定律是整个现代概率论和统计学的最重要基石，几乎一切统计方法的正确性都依赖于大数定律的正确，因此大数定律被有些人称为概率论的首要定律。

大数定律直观来看表述了这样一种事实：在相同条件下，随着随机试验次数的增多，频率越来越接近于概率。注意大数定律陈述的是一个随着n趋向于无穷大时频率对真实概率的一种无限接近的趋势。

下面给出大数定律的数理表述，大数定律有多重数学表述，这里取伯努利大数定律：

$$\lim_{n\to\infty} P\left\{\left|\frac{n_x}{n}-p\right|<\varepsilon\right\}=1$$

其中n_x表述在n次试验中事件x出现的次数。伯努利大数定律代表的意义是，当试验次数越来越多，频率与概率相差较大的可能性变得很小。大数定律从数学上严格证明了频率对概率的收敛性以及稳定性。这就是频率估计的理论基础。在后面关于中心极限定理的部分，还将定量给出估计的置信度（表示这个估计有多可靠）。

最大似然估计

下面给出另一种估计概率的方法，就是最大似然估计。最大似然估计是参数估计的一种方法，用于在已知概率分布的情况下对分布函数的参数进行估计。而这里分布函数的参数刚好是要估计的概率。

最大似然估计基于这样一个朴素的思想：如果已经得到一组试验数据，在概率分布已知的情况下，可以将出现这组试验数据的概率表述为分布函数参数的函数。

看到上面的话很多人肯定又晕了，我还是举个具体的例子吧（非数学严格的例子，但思想一致）。我来到一所陌生的大学门口，想知道这所大学男生多还是女生多，我蹲在校门口数了走出校门的100名同学，发现80个男生20个女生，如果我认为这所学校每个学生这段时间内出校门的概率都是差不多的，那么我会推断男生多。因为男生多的学校更大可能性产生我观察的结果。所以，最大似然估计的核心思想就是：知道了结果，但不知道结果所在总体的情况，然后计算在总体在每种可能下产生这个结果的概率，哪种情况下产生已知结果的概率最大，就认为这种情况是总体的情况。

下面正式使用这个方法估计硬币正面出现的概率。

还是上面的实验，我已经得到“抛了十次，六次正面”这个结果，下面我想知道正面向上的概率。由于这个概率是一定存在的（第一节已经说明了哈，在既定哲学观点下），而且这个概率的取值范围应该是0到1的开区间（正面背面都出现过，所以不可能是0或1）：

$$p\in(0,1)$$

由一些背景知识知道，每抛十次硬币，正面出现的次数服从二项分布：

$$C_n^k p^k(1-p)^{n-k}$$

由于已知n=10，k=6，将其带入，得到一个函数：

$$L(p)=C_{10}^6 p^6(1-p)^{10-6}$$

其中p的定义域为$p\in(0,1)$。这个函数表示的是，当出现正面的真实概率为p时，“抛十次六次正面”这个事件出现的概率。我们希望估计的p让这个函数取值最大，以下是求解过程：

因为在(0,1)区间，ln(x)是x的单调递增函数，所以最大化lnL(p)就等于最大化L(p)。这样做主要是取对数可以让连乘变成连加，方便后面求导。

由微积分知识可知：

$$\frac{dlnF(p)}{dp}=C_{10}^6\left(\frac{6}{p}+\frac{4}{p-1}\right)=C_{10}^6\left(\frac{10p-6}{p^2-p}\right)$$

让这个导数为0，解得p为0.6，这就是我们对概率的最大似然估计，与概率估计的结果一致。

显著性及假设检验

到此为止，我们已经说明了概率是存在的、建立了概率的数学模型，并能对不可直接观测的概率进行估计。但似乎还缺点什么。

大数定律只说明了理论上我们的估计是靠谱的，但是到底有多靠谱，却无法通过大数定律定量计算。这一节，我们就来解决这个问题：定量计算出估计的可靠性（术语叫显著性）。

评估显著性

还是上面我抛那十次硬币的试验。根据最优的频率估计和最大似然估计，均估计p（出现正面的概率）为0.6。但是如果有人提出异议，说我的估计可能是错的，p实际是0.5，我那个出现六次正面是因为只是偶然性的结果。这时我需要找证据反驳他，由于不能做无数次试验，我只能给出一

个较高可信度的证据，例如，我想证明至少95%的可能性出现六次正面是因为p不等于0.5，也就是说，证明如果p为0.5，则偶然出现我这个结果的可能性不超过5%（5%称作显著水平）。

中心极限定理

要评估显著性，首先要借助于中心极限定理。中心极限定理也是统计学的基石定理之一，它的一种表述是：

设随机变量$X_1, X_2, \cdots, X_n$独立同分布，且数学期望为μ，方差为$\sigma^2 \neq 0$。则其均值$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$近似服从期望为$\mu$，方差为$\sigma/n$的正态分布。等价的，$\zeta_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$近似服从标准正态分布。

中心极限定理的直观意义是，随便一个服从什么的总体中，你独立随机的抽取一组样本，那么样本的均值服从正态分布，并且可以根据总体的期望和方差推导出这个均值服从的正态分布的期望和方差，然后简单变换一下就可以得到一个服从标准正态分布的随机量。由于标准正态分布的概率密度函数是已知的，那么就可以得到这个量出现的概率。

这样说貌似太抽象了，我们下面还是看这个定理的应用实例吧。

假设检验

上面说过，我要反驳的是抛硬币得到正面的实际概率是0.5，那么我就要证明如果p是0.5，则得到这组结果的概率是很小的（上面要求小于5%）。

设正面取值为1，背面取值为0。如果p是0.5，则每一次抛硬币的取值服从一个p为0.5的0-1分布。由期望及方差的定义可知，这个分布的期望和方差分别为：

$$\mu = p \times 1 + (1-p) \times 0 = 0.5 \times 1 + (1-0.5) \times 0 = 0.5$$

$$\sigma^2 = (1-\mu)^2 \times 0.5 + (0-\mu)^2 \times 0.5 = 0.25 \times 0.5 + 0.25 \times 0.5 = 0.25$$

由中心极限定理$\frac{\bar{X}-0.5}{\sqrt{0.25/10}}$近似服从标准正态分布。

而我抛的十次硬币可以看做十个独立随机抽样，它们的均值是0.6，变换后的值为$\frac{0.6-0.5}{\sqrt{0.25/10}} \approx 0.632$。

标准正态分布的概率密度公式为：

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

上面说过，我们希望显著水平是5%，所以，我需要找到x=z，使得此概率密度函数从-z到z的定积分为0.95，然后看0.632在不在[-z, z]内，如果在的话，我会认为我确实错了，至少我没有95%以上的把握说p不等于0.5，而如果0.632不再这个范围内，则我可以拍着胸脯说，我已经从理论上证明我有95%以上的把握，p不是0.5（换句话说，如果p是0.5，抛十次六次正面的可能性不足5%）。

坦白说这个z不是很好算，不过还好由于这东西特别常用，任何一本概率课本后面都可以找到标准正态分布表（或者很多工具如R语言可以直接计算分位点），下面就是我在网上找到的一个（来源http://www.mathsisfun.com/data/standard-normal-distribution-table.html）：

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890

这是一个单侧表，要保证显著水平为5%，则单侧积分上限不能低于0.475，通过查上表，可知0.475对应的z是1.96，远大于我们算出的0.632。很不幸，我在5%的显著水平下无法拒绝p=0.5的假设。同时通过上表可以看到，0.63对应的单侧概率是0.2357，也就是说，通过抛十次得到六次正面，我们只有约50%的把握说出现正面的概率不是0.5。换句话说，抛十次硬币来做频率估计是不太合适的，于是，我们需要增加试验次数。

假如，我又做了100次实验，抛出了60次正面，40次背面。那么这个试验结果可以显著的认为p不是0.5吗？用同样的方法算出 $\frac{0.6-0.5}{\sqrt{0.25/100}}=2.0$。很显然，2.0大于1.96，所以这个试验结果可以充分（超过95%的可能）说明这枚硬币正面朝上的概率确实不是0.5。通过查表可以看到，2.0的显著水平约为0.046，换句话说，这次试验结果95.4%以上表明硬币正面出现的概率不是0.5。当然，也有可能结论是错误的，因为毕竟还有4.6%的可能这是在p=0.5的情况下偶然出现的。

通过假设检验理论，可以通过增加试验次数，将犯错的概率缩小到任意小的值。

总结

这篇文章以抛硬币试验为引子引出了一系列现代数学中概率的基本模型、定理及基本的估计及显著性检验方法。写这篇文章是我无聊抛硬币时一时兴起，其中对很多东西只是给出一个轮廓，没有处处给出严格的定义和证明，不过大约说明了常用的一些统计方法及其理论基础，限于篇幅不能面面俱到，例如一个假设检验如果展开写可以单独写一篇文章。目前随着大数据概念的热炒，基于互联网的数据挖掘和机器学习也变得火热，其实很多数据挖掘和机器学习都是基于概率和统计理论的，很多方法甚至只是传统统计方法的应用。因此如果准备在这方面深入学习，不妨考虑先在概率论和数理统计方面打好基础。
