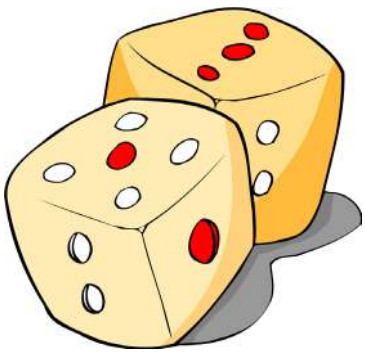


上帝掷色子
——2008 美国统计年会杂记

万精油

关于美国的年会我写过好几个。比如数学年会（题目是谁想当数学家）；羽毛球全国老年年会（题目是生命不息，拼博不止），以及美国围棋年会。但对我参加过多次的统计年会却一直没有写。一方面因为没想到好的标题，另一方面担心大家觉得统计很枯燥。最近看一篇关于量子力学的文章，提到爱因斯坦的著名论断：“上帝不掷色子”。统计学实际上就是关于掷色子的学问。根据观测到的数据来推出色子的一些性质。或者已知色子的性质算出某种情况出现的概率。用“上帝掷色子”作标题，借着爱因斯坦的名气或许可以抓一些眼球。

有了标题算是有了好的开头。每年的统计年会有意思的事情不少，写起来就比较容易了。

一英里高的城市

还是老习惯，先来一段与统计无关的轻松话题。

2008 年的年会在美国科罗拉多州的丹佛市召开。飞机刚着陆喇叭里就传出机长的迎宾词：“欢迎来到一英里高的城市”(Welcome to the Mile High City)。丹佛的海拔正好是一英里（1609 米），这也算是很巧合的事。这个高度比起其它一些高原城市来说算不了什么，比如拉萨的海拔就比这里高出一倍还多。但对于我们这些居住在平原的人来说，这个高度就有明显的效应了。

首先，天显得出奇的蓝。这种蓝天我只在云南大理看见过。回来查了一下，大理的海拔比这里还要高。另外一点就是感到氧气不足。一般走路似乎还没有什么感觉，但跑起来就明显喘不过气来。刚来的第一天开会开到很晚，已经不能出去跑步，只好到旅馆里的健身房去跑。没想到平均七分钟一英里的速度竟然坚持不下来，只好往下调。最后调到七分半钟的速度才免强跑完三英里，而且已经累得不行。第二天早上起来时的静止心跳也串到每分钟七十多下（在家时我一般都在五十以下），难怪跑不动。后来听人说一般人要好几个月才能完全适应这种情况。跑步不行就做重力训练，缺氧的情况对此不影响。实事上因为海拔高，这些铁块应该比标明的重量轻一点。或许是心理原因，在旅馆健身房几天下来，我竟然打破了我平常的重量纪录，压腿终于可以压到三倍于我的体重。

这里的人已经习惯了这种状况，跑步不受影响。我抽空去了一趟 科罗拉多斯普林斯 (Colorado Spring)， 路上看见很多跑步和骑车的人。最有意思的是有些马路上还专门给自行车留一条道（比一般的车道窄三分之一左右），在美国其它地方我还没有见过。为此对这里留下了很好的印象。

因为接近民主党大会，城市里到处挂着民主党的宣传条幅，也算是一景。

上帝掷色子

爱因斯坦“上帝不掷色子”的话针对的不是统计，而是对海森堡测不准原理所给的一个哲学断语，属于可知论与不可知论的争议范畴。统计在物理上的重要性是不可争的，它作为热力学、量子力学的理论基石之一也是众所周知的事实。爱因斯坦 1905 年发表的五篇重要文章中，除了相对论与光

电效应（因此而获诺贝尔奖）的文章外，还有一篇关于布朗运动的。这布朗运动可就是实打实的依赖于统计。统计不单是在物理这样的理论上有用，在现实中的应用更是到了无所不及的地步。政治、经济、管理、体育、制药，你想得出来的领域都或多或少地可以找到统计的应用。来开会的人除了学校的教授、研究生，相当一部分来自政府各部门（卫生部、标准局、药检局）、各大制药公司、华尔街投行等等等等。洋洋五六千人，可谓声势浩大。

斯坦福大学的统计教授笛阿孔尼斯掷硬币可以掷出任何给定的概率；因为他从 14 岁到 24 岁都是在各地巡回演出的职业魔术师。

大会的演讲程序表，单是题目及主讲人就列了好几十页。“线性回归”，“蒙提卡罗”，“基因矩阵”，“棒球比赛数据”，“选举加权”，五花八门的题目真是应有尽有。这也是我很喜欢来参加这个会的原因之一。总能找到有兴趣的演讲听，开会效率很高。

最近看到一本书上有一章的题目是：“上帝不掷色子，或许会玩牌”，其实还是一个意思。规律定在那里（比如万有引力、电磁场），剩下的就是按这些规律的运动。变量多了，系统就很复杂，宏观上的结果就带有很多随机性，与掷色子差不多。因为有大数定理（或者叫中心极限定理），统计总会在现实中到处派上用场。上帝的色子总是要继续掷下去的。

有偏差的样品

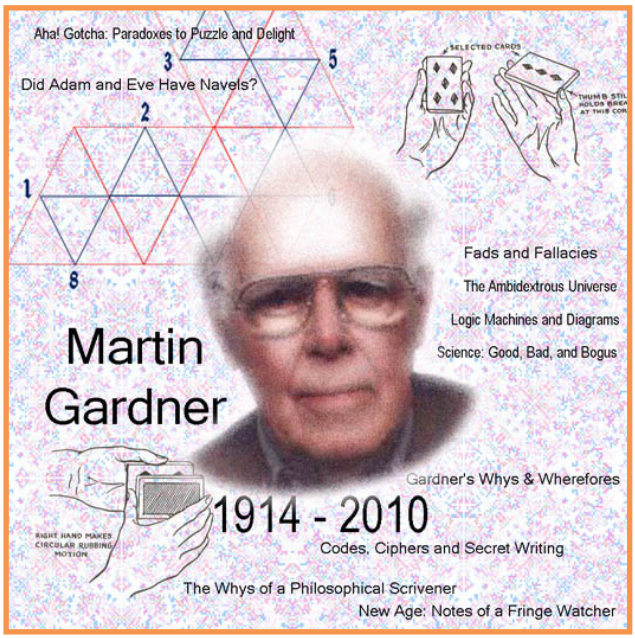
斯坦福大学的统计教授笛阿孔尼斯 (Persi Diaconis) 在课堂上给学生表演掷硬币，说是想掷头就掷头，想掷尾就掷尾，可以掷出任何给定的概率。如果用它掷出的结果做样本去估计那个硬币的性质就不会得到正确的结果，因为样本有系统误差。

笛阿孔尼斯是数学界很传奇的人物。他能准确地掷出头尾是因为他从 14 岁到 24 岁都是在各地巡回演出的职业魔术师。24 岁时他想弄清楚一些组合游戏里面的原理，就请人给他推荐一本概率书。别人给他推荐了费勒的概率数学原理，可惜他看不懂，因为他不懂微积分。为了弄懂费勒的书他决定上大学。两年就数学本科毕业。这时他已经被数学、统计这些理论东西所吸引，决定继续读研究生。而且说要读就读最好的，于是就申请哈佛。本来，凭他的成绩是进不了哈佛的，因为他第一年的微积分得了两个 D。所幸的是他有著名趣味数学专栏作家嘉德纳 (Martin Gardner) 给他写推荐信。推荐信说：“数学的东西我不是太懂，但我知道在过去十年里发明的最好的十个数学魔术中，这小子发明了其中两个。凭这点你们是不是应该多考虑一下”。几乎每个数学家都是嘉德纳的粉丝，哈佛数学教授也不例外。嘉德纳的话份量很重，笛阿孔尼斯当然就进了哈佛。事实证明嘉德纳的眼力是不错的。笛阿孔尼斯经过哈佛的熏陶终于成了数学、统计上的大家。他的研究范围很广，证明的定理当然也很多。其中一个定理在非数学界也很有名气，那就是“洗牌定理”。说的是一付 52 张的牌要洗七次才能洗匀。洗少了不匀，洗多了没必要。所以你下次打牌一定要洗七次。如果洗太少，上次有人出拖拉机，就要影响下次牌的分布。

笛阿孔尼斯的故事很多，可以写一本书。我们还是言归正传，谈我们的样品偏差。

样品偏差有些是人为的，比如笛阿孔尼斯掷的硬币；有的是无意识的，比如有人用佛罗里达的数据得出结论说富人死亡率高于穷人死亡率。事实上因为很多老人搬到佛罗里达去度晚年，最后死在那里。这些老人平均起来比当地人要富很多，大大影响了死亡人员的经济情况。

这次会议中听到一个有意思的样品偏差的例子。说是二次世界大战时，美国国防部有人研究战斗机应该把飞行员放在什么位置比较安全。他从所有飞回来（没有被击落）的飞



马丁·嘉德纳 (Martin Gardner; 1914-2010)，名声显赫的业余数学大师、魔术师、怀疑论者，他曾经为《科学美国人》杂志趣味数学专栏写作长达 20 多年。嘉德纳没有数学博士学位，但是他的作品能让广大普通读者和数学家为之着迷。

机上的弹孔取样做统计。发现有个位置弹孔很少，于是得出结论那个位置最安全。后来有人说：那个位置上有弹孔的飞机大概都被击落了，所以，飞回来的飞机上那个位置的弹孔最少，或许那是最不安全的位置。显然这个人的统计没有学好，或者说战争年代高人都去造原子弹去了。

这种偏差样品现实生活中也能找到很多例子。比如你如果用三鹿奶粉来测一般奶粉的成分，那就有系统偏差。另一个更切实的例子是，我经常听一些从中国回来的人说，中国人现在生活比美国好。说是他们的同学个个开好车，顿顿吃饭馆，家事有佣人。不象我们在美国下班后回家还要做家务，周末还要割草。对这些论点我不敢赞同。首先生活质量的判别有许多因素，另外，后园有草割也不见得都是坏事。但我反对的原因主要还是样品的偏差。需知这些过得好的同学都是在大城市，不能代表绝大多数农民。实际上这些同学也不能代表大城市的居民，甚至连他们的同学都不能代表。很可能的情况是这些是同学中混得最好的一小部分。你从国外回去，混得好的同学来找你，表示他们混得不比你差。而这些混得好的同学常常也是同学聚会的积极组织者。混得好不到老同学处显摆一下岂不是锦衣夜行。所以我说这些同学是带有严重偏差的样品。

博览会

数学会也好，统计会也好，与年会同时进行的都有一个博览会。就是与它有关的各个商家在这里宣传他们的产品，还有各政府部门在这里摆摊招工。最多的当然是书商，其次是各种各样的数学与统计软件。十几个篮球场那么大的大厅被这些厂家占得满满的。

每个厂家为了吸引顾客，都在自己的亭子里放一些免费小礼品，各种各样的笔、书签，鼠标垫等等。大家边看边拿，一圈走下来，差不多装半个塑料袋。有些礼品还真是很实用。比如房利美 (Fanniemae) 的笔形镙螺丝刀，体积比一支笔大不了多少，却有四种不同的螺丝头，很实用。谷歌 (Google) 的闪光胸针设计得也别致有趣。

还有些艺术家在这里卖数学艺术品。比如那个卖克莱茵瓶的就是每会必到。克莱茵瓶是二维无定向曲面。虽然怀特定理说可以把它嵌入到欧氏空间中，但那需要四维空间。要在三维里做克莱茵瓶，就必须要自相交。这自相交在什么地方交，以什么方式相交，可以产生各种各样的克莱茵瓶。这些克莱茵瓶怎么把水倒进去，倒出来都可以研究一番。我没有买过，每次看见都要想如果里面脏了怎么洗。另一个每会必到的是卖科学衫的。在 T 恤衫上印出各种科学幽默、卡通。我每次都买一两件。最喜欢的一件是：一个有曲面积分的式子，里面有椭圆函数等一长串数学符号，下面是一句问话：到底哪一步你不懂？ (Which part of this don't you understand?) 我们家的 T 恤衫除了跑步比赛发的以外， 差不多都是这些科学衫。

对我来说当然主要是转书铺。这里买书可以比书店便宜百分之二十。与工作有关的书可以报账，便不便宜也无所谓。但自己买书百分之二十还是比较可观的。有时还会有意外惊喜。上次买一本趣味数学书，正遇到作者 (Peter Winkler) 在那签名。 我对趣味数学有很大的兴趣，正好借机与他聊了半天，收获很大。

对数学软件我也很有兴趣。我并不是要买这些软件，而是对他们的一些设计或相关的东西有兴趣。有一次我走到软件公司 Mathematica 的亭子面前。亭子里一个工作人员过来与我打招呼。我随便瞟了一眼他衣服上别的名片，眼睛突然发亮。

我：哇，你就是大名鼎鼎的 Eric。

E：大名鼎鼎不敢当，我就是 Eric。

我：你的数学世界（MathWorld）给我太多的帮助，我真应

埃里克•韦斯坦因（Eric W. Weisstein），1969 年出生于美国，是数学方面网上百科全书数学世界的创始人。主要编辑 MathWorld、ScienceWorld 等著名网站。

该谢谢你。

E：很高兴它能对你有帮助。

我：你知不知道你的数学世界是我浏览器上的第三个常用地址。

E：让我猜一猜，第一个肯定是谷歌（Google），第二个大概是维基（WIKI）。

我：全说对了。

E：很荣幸能排到第三，我一个人的能力也不能与它们竞争。

我：难道数学世界都是你一人之力吗？

E：以前都是我一个人，后来有些人帮忙。不过 95％以上都是我自己搞的。

我：厉害厉害。谢谢。

埃里克•韦斯坦因 (Eric Weisstein) 是加州理工学院的物理博士。从高中开始就收集数学公式及相关信息。后来把它放到网上，一直发展成现在的数学世界。数学世界是数学方面的网上百科全书，相当于维基，在数学界享有盛誉。不过它比维基早很多，而且运作方式也不一样。加入 Wolfram Research 公司以后，数学世界已经扩展成科学世界 (www.scienceworld.com)，其中包括数学世界、物理世界、生物世界等等，建议大家去看一看。韦斯坦因现在是国家电子图书馆的活跃人物之一，也算是牛人。与他聊天收获很多。后来我们又聊了一些数学软件的设计，Mathematica 与 Matlab 的比较，非常有趣。临走时我给他们提了一些建议，没想到回来以后收到他们发展部门的邮件说你的建议非常好，我们正在考虑采用。

每次开这种会，我都要在博览会 (EXPO) 里呆好几个小时。收获虽赶不上听学术报告，但也算相当重要的一部分。

高维问题

虽然说是讲故事，但统计会杂记总免不了要讲一些理论性的东西。还是挑一样现在比较热门的东西来讲一讲。

传统的统计一般是三五个参数，几十上百个样本，用这些样本来估计那几个参数或者建分类模型。现在差不多倒过来了。经常出现十来个样本，几万个变量的情况。比如常见的基因矩阵数据 (MicroArray)，十几个矩阵数据，几万个“基因”都是变量。学过数学的都知道，一般情况下，如果变量比方程多，可以有无数多个解。通过传统方法用这些数据建模型，几乎可以得到任何你想要的结果。实事上现在有不少人就是这样做的，把原始数据做这样或那样的变换然后用来建分类模型。这样做出来的结果，按范剑青的话说“与随机猜测同样糟糕”。

范剑青出国前是中国科学院应用数学所的研究生，现在在普林斯顿当教授。算是中国出来的留学生中出类拔萃的人物，照网上的流行语，算是“大牛”。他在这次会上给了一个“高维数据”的报告，讲的就是这个问题。因为是大牛做报告，听的人把大厅挤得满满的。他用实例指出没有选择地全用这些高维数据推出的结果等同于随机猜测。

另一个由高维数据带来的问题就是假正问题 (False Positive)。一般的假设检验都用 5％作为分界线。小于 5％的事件被认为是小概率事件。可是，如果对每个变量做假设检验，几万个做下来，小概率事件也几乎成了肯定事件。这就是所谓假正问题。一米八五的个子是小概率事件，但在全中国找几十万个也不会有问题。当然，假正问题变量少的时候也存在，只不过当变量多的时候，这个问题就变得更加突出。

基因矩阵数据是现在很热门的话题，大会中有很多报告

本科毕业于复旦大学的范剑青现在是普林斯顿大学统计系讲座教授，2000 年 COPSS 统计大奖获得者。

都是围绕这个问题在展开。其中很多方法涉及到很深的数据分析知识（比如非负矩阵分解），对我这种有数学背景的人正对胃口，所以这种报告我几乎都去听。这也算是我现在的工作中最接近前沿的了。

统计会上的中国人

最后还是谈点轻松话题结尾。

这个大会与数学大会一样，也搞了一个知识竞赛。数学大会的竞赛叫“谁想当数学家？”。统计大会这个竞赛叫“统计杯”。 本来想谈一下这个竞赛， 可是不论从形式到内容都比数学大会的竞赛差太多，不谈也罢。还是另选话题吧。

五六千人的大会大概有四分之一的中国人。大会花名册的最后几页（从 W 到 Z）几乎被张王赵周这些中国大性占满了（还有于俞余的统一拼法 Yu）。有些小讲座从主持人到演讲者几乎都是中国人。

我读书的时候，读数学的都去摘皇冠上的明珠，搞数论、几何之类的，统计算冷门。现在讲究实用主义，统计一下变热了。学数学的如果不改行，只有在学校当教授。学统计的却可以在学校、公司、政府部门到处找到事做。统计现在是如此的热门，以至于许多从前不搞统计的人现在也往统计上靠。这次会议上碰到十几年前的一个邻居，学经济的，也摇身一变成了统计学家。在博览会看见一个人觉得面熟，聊起来原来在羽毛球比赛时见过，现在也搞起统计来了。看到大会材料中一个什么委员会的主席名字很眼熟，后来见面才发现是与我在中文学校一起打乒乓球的家长。这阵势大有全民搞统计的味道。

前几年开会还能碰到一些过去的同学，现在很少碰到了。大部分中国人都是年轻人。与一帮中国人一起吃饭，聊天中得知其中一位今年博士毕业，他的导师是我在科学院读研究生时的同学的学生。按照金庸武侠小说的说法，他应该叫我师叔祖了。相当一部分的参会者都是这样的年轻人，我这个年龄的人越来越少了。不过我现在来开会主要是来长长知识，顺便逛一下开会的城市及周边，能不能碰见老朋友不是很重要。当然，如果碰见了老朋友就多一份惊喜。

明年的统计会在华盛顿特区，希望到时候能碰见更多的朋友。

作者后记：这是两年前写的文章，现在读起来居然完全没有过时。事实上，随着计算机技术的发展，一切东西都数字化了。需要用到统计的东西越来越多起来。去年八月纽约时报一篇讨论什么专业最有前途的文章标题是，“对今天的学生来说，就一个词：统计”。（For Today’s Graduate, Just One Word: Statistics），见 http://www.nytimes.com/2009/08/06/technology/06stats.html。此文说，未来十年，统计将会是最有前途的专业。

作者介绍：

万精油，本科毕业于四川大学数学系。中国科学院数学研究所硕士，美国马里兰大学数学博士。业余时间爱好写作。以杂文，记事为主，科普为辅，偶尔也写小说。代表作为科幻小说《墨绿》，获新语丝文学二等奖。因为兴趣广泛，起笔名为万精油。