漫谈现代统计“四大天王”：卡尔·皮尔逊

李殊勤

> 这是一个最好的时代，也是一个最坏的时代；
> 这是一个智慧的年代，这是一个愚蠢的年代；
> 这是一个信任的时期，这是一个怀疑的时期；
> 这是一个光明的季节，这是一个黑暗的季节；
> 这是希望之春，这是失望之冬；
> 人们面前应有尽有，人们面前一无所有；
> 人们正踏上天堂之路，人们正走向地狱之门。
>
> 狄更斯

统计，是数学作用于现实生活中的一场思想革命，它正持续的进行着，我们每个人亲历其中。但人们谈起它，也往往有着如狄更斯这样复杂的情愫：当普罗大众可以在不经意间谈论“风险”“概率”“相关”这些概念的时候，它早已悄悄地改变了人们关于科学、关于世界的底层信念；借着今天大数据、人工智能的春风，它必将如火如荼地蔓延开去，日新月异地改变我们的生活。而另一方面，统计可能是最不严谨的数学子学科，像“建立在沙土上的摩天大厦”，很多本源的理论问题至今并没有得到令人满意的解答，也导致人们在工作生活中越来越广泛地使用统计思想和模型的同时，产生了越来越多的怀疑和忧虑。

这场革命从何谈起呢？又将何去何从？

KARL PEARSON

为回答这个问题，我们有必要一起回溯这场革命的源头。本文围绕现代统计四位开山立派的人物——卡尔·皮尔逊、费希尔、埃贡·皮尔逊和内曼之间不得不说的“爱恨情仇”展开，带大家重温现代统计发源之时，风起云涌、天才辈出、群星璀璨的黄金年代，在领略绝顶高手的思想交锋之中，探求统计理论形成和发展的脉络，体验统计学背后哲学思想之美。

当然，叙事完全是个人视角，评论完全是个人趣味，思考完全是个人拙见，欢迎读者存疑与讨论、批评与指正。

本文是《漫谈现代统计计“四大天王”》系列随笔的第一篇，讲述现代统计奠基人卡尔·皮尔逊的精彩人生传奇以及由他开启的现代统计的发源思想。

1 世界的本质是随机的吗?

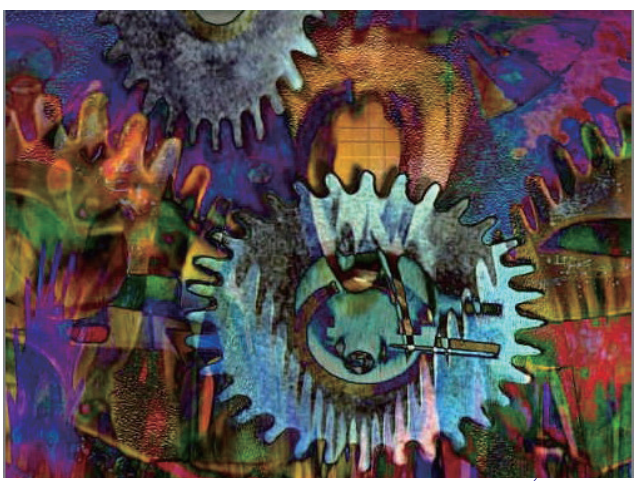
谈统计，我们却不得不从这样一个哲学问题开始，因为它涉及整个学科存在的合理性。

如果我们拿这个问题去问任何一个统计学家，他 / 她一定会回答：是！——否则，还要统计干什么呢？但要大众文化接受这样的观念却并不容易。

一者，自 19 世纪以来，以牛顿力学为代表的科学，应用于现实世界，在各行各业取得了巨大的成就，也让一种“决定论”的世界观深入人心——世界的本质就像一个大时钟运行着，于是，我们只需要少量的数学公式，不仅可以描述现实，还能预测未来。

二者，“随机”在人们日常的理解中就是“未知、复杂、毫无规律”的同义语。比如，讲故事的时候说“海盗把宝藏随机埋在了海岛上”——基本就是说，你绝无可能找到宝藏了（假设根本没有藏宝图，海盗都是打死不说）。那么，就算世界上还有很多未解之谜，也不能说本质是“随机”的吧。

所以，这个问题还真是“烧脑”啊！不过，这也是为什么统计学有意思的原因：统计是一场科学的革命。——萨尔斯伯格博士



“钟表”般运行着的宇宙

但如果我们能注意到下面两个事实，事情理解起来，可能就没有那么困难了。

第一个事实非常简单：这个“钟表的世界”也有点太不精确了吧？且这种不精确也太普遍又太显而易见了：回忆一下我们中学做过所有定量的实验（物理、化学等等），你大概从来没有一次测得的结果能恰好等同理论值。老师会告诉你，那是实验的“误差”造成的。通常，写上几页厚厚的误差分析能帮你拿个高分：大意是，如果观测和计量更精确，误差就会减小，直至消失为0。顺便说一句，把实验观测值和理论值的差值作为“误差函数”来处理这个发明源自于大数学家拉普拉斯。他对这些“随机”的、无关紧要的误差函数做过深入的研究，给出了首个概率分布。

法国大数学家 拉普拉斯
（1749–1827）

于是我们就有了第二个事实：随机，其实是有规律的，我们可以用概率分布——精确的数学公式来描述它。也因此，有人会把拉普拉斯作为统计思想的开创者。不过，统计界更普遍地把这一荣誉归属于卡尔·皮尔逊，为什么呢？这又要说回到技术背后思考问题的哲学。

回忆一下，你做实验的时候，有没有过一丝怀疑：无论怎样加强测量精度，“误差”有可能是根本不能消除的？在“决定论”根深蒂固的情况下，很难这样去怀疑。通常出现这样的情况，你的老师会微笑着告诉你：呵呵，那一定是你的实验做错了。

其实，单从“误差”这个名字本身，我们就知道，其思考哲学一定还是在“决定论”框架下——我们绝不会把“误差”作为被观测量的一部分或某种自带属性去理解，而是实验中应当尽量消灭的东西。当然，我们不能苛求前人（如拉普拉斯），因为关于这个怀疑的发现也需要我们不断提高实验手段和测量精度后才能做出——现代的事实是，随着我们实验技术的提高，测量到的误差没有像预计中的减少，甚至还增大了，且永不消除。这怎么解释呢？

那么回过头来，想想我们本节的标题：是不是就有种恍然大悟的感觉？让我们比“误差”随机走得更远一点：有没有可能，被观测量本身就是随机的呢？也就是说，我们做实验能观测到的其实应该是一个“分布”。那么，所谓的“误差”其实既不“误”也不”差”，只是被观测量的随机本质的反映。所以不管我们怎么提高观测精度，当然都不可能消除这种随机性，即所谓“误差”；且随着精度提高，随机性被观察得更清楚，所谓“误差变大”也就顺理成章了。

这是卡尔·皮尔逊做出的回答，也是我们今天统计学革命之所以合理的哲学基础。

啊哈！这是多么划时代的观念！你是不是已经开始好奇卡尔·皮尔逊是何方神圣了。或者你非常惊讶，他怎么能想有这样的天才的想法呢？其实，这个想法也绝非无中生有，横空出世。源自哪里？我们有必要先说说皮尔逊的老师高尔顿爵士在优生学上的发现。

2 思想缘起高尔顿：回归与相关

弗朗西斯·高尔顿是个典型的“维多利亚时代的天才”——多是独立而富有的贵族，以科学研究为乐，常以全才或广博著称，在多个领域都颇有建树。他还有个更为著名的表哥——查尔斯·达尔文——《物种起源》的作者。高尔顿非常崇拜他的表哥，并终生致力于为进化论找实证。他的一项早期工作就是去收集社会名流大家的家谱，整理那些公认的聪明的父子的数据。但鉴于当时还没有智商测量的工具，高尔顿很快意识到这个工作太过困难，于是就改为收集更容易测量的家庭成员的身高数据，试图发现一个公式，能通过父母的身高预测孩子的身高。于是他和助手做了大量的统计图表。

“维多利亚时代的天才”高尔顿爵士（1822–1911）

在这个过程中，他发现了一个他称之为“均值回归”的现象：

“如果父亲非常高，孩子往往比父亲矮；如果父亲非常矮，孩子往往比父亲高。似乎有种神秘力量让人类的身高远离极端，朝着所有人的平均靠拢。均值回归现象不仅仅适用于人类身高，几乎所有观测都面临均值回归的困扰。”

他还做了思想实验，如果高个子父亲生出的儿子更高，矮个子的父亲生出的儿子更矮，这样的规律代代保持，用不了几代，人类就要出现越来越高和越来越矮的人。但这种现象实际没有发生，平均来说，人类身高基本稳定。所以只有非常高的父亲后代平均身高比他矮，而非常矮的父亲后代平均身高比他高，才会保持这样的结果。正是均值回归维持了物种的稳定，确保了一个物种代与代之间的“相似性”。他发现了描述这种关系的一个数学度量，称之为“相关性”。

“回归”“相关”这些理念是不是与我们之前讨论过的“随机”“分布”已经高度一致了？虽然这些观念最早是由高尔顿提出的，但最终将该思想完整地以数学公式形式清晰地表达出来、且继续发扬光大的人是卡尔·皮尔逊。