a mindstyle business 与思想有关

数据统治世界

如何在数据统计中挖掘商机与做出决策

[美] 冯启思 (Kaiser Fung) ◎著
曲玉彬◎译

NUMBERS RULE YOUR WORLD

- 游人如织的迪士尼是如何管理排队等候时间的?
- 高速公路为什么要修建缓行匝道?
- 服用兴奋剂的运动员在被揭穿之前是如何制造出几十次干净的药检结果的?
- 即使拥有海量的股票数据和公司财报信息,为什么大部分人还是不能在投资中所向披靡?

NUMBERS RULE YOUR WORLD

The Hidden Influence of Probability and Statistics on Everything You Do

大数据趋势必读

中国人民大学出版社
China Renmin University Press

NUMBERS RULE YOUR WORLD

The Hidden Influence of Probability and Statistics on Everything You Do

数据统治世界

[美] 冯启思 (Kaiser Fung) ◎著
曲玉彬◎译

中国人民大学出版社
·北京·

图书在版编目（CIP）数据

数据统治世界 /（美）冯启思著；曲玉彬译 . —北京：中国人民大学出版社，2013

ISBN 978-7-300-17754-0

Ⅰ . ①数… Ⅱ . ①冯… ②曲… Ⅲ . ①统计学—通俗读物 Ⅳ . ① C8-49

中国版本图书馆 CIP 数据核字 (2013) 第 154760 号

上架指导：数据决策 / 经营管理

数据统治世界

［美］冯启思 著

曲玉彬 译

Shuju Tongzhi Shijie

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	邮政编码	100080
电　　话	010-62511242（总编室）	010-62511398（质管部）	
	010-82501766（邮购部）	010-62514148（门市部）	
	010-62515195（发行公司）	010-62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京中印联印务有限公司		
规　　格	170 mm × 230 mm 16 开本	版　　次	2013年9月第1版
印　　张	16.75 插页1	印　　次	2013年9月第1次印刷
字　　数	188 000	定　　价	49.90 元

NUMBERS RULE
YOUR WORLD
The Hidden Influence of
Probability and Statistics on
Everything You Do

前言

让数据自己说话

这本书的主题并非只是谈论“该死的谎言与统计学”。达雷尔·赫夫（Darrell Huff）、约翰·艾伦·包洛斯（John Allen Paulos）、爱德华·塔夫特（Ed Tufte）、霍华德·维纳（Howard Wainer）等人，在这个常谈常新的话题上早就写下了不少垂范之作。的确，从精于操控的政治家到粗心大意的分析员，从经济学爱好者到强买强卖的广告商，我们可以举出无数的例子，来说明当数字被滥用时会引发多少问题。择优选择、过度简化、故意混淆，这几种滥用数字的花招想必我们都领教过了。在这本书中，我们将沿着不同的方向，从正面的立场来思考：当事情顺利进行的时候会出现什么结果，也就是说，当数字没撒谎时会怎么样。我对这个问题很感兴趣。

知道得越多，越不了解真相

伯尼·麦道夫（Bernie Madoff）是纽约一家投资公司的资产管理人。到2008年他认罪伏法时为止，一个存在了30年的由富人参加的投资俱乐部，被

他弄得一贫如洗。直到此时，人们才识破麦道夫的真面目——一个惊天巨骗。那么，从麦道夫欺诈案中，我们能学到些什么呢？安然公司高层拿假账混淆视听，使数千名雇员的退休储蓄金随着公司的破产而顷刻化为乌有。那么，安然高层的欺诈案又能带给我们一些什么样的思考呢？或许我们该搞清楚：为什么大量的财务数据、财务报表及办公存档几乎未能给调查员、审计员和监管机构提供任何线索，找到何人与欺诈有关呢？

我们从万络（Vioxx）事件中又能学到点儿什么呢？美国食品药品管理局（FDA）承认，该药在首次获准入市到后续的五年中，引起了上万起心脏病病例。或许我们该弄清楚：为何虽有大量现成的健康医疗资讯以及大规模的、复杂精妙的临床试验，也未能使万络的发明人默克公司、医生及病人对这种新药的致命副作用重视起来呢？

我们还应当问一下：尽管能够弄到海量的股票数据和公司报告，为何我们中的大部分人却未能在股票市场中大发横财呢？虽然严格核算每罐、每袋食品的营养指标，为何我们中的大部分人却未能成功瘦身呢？尽管在信息技术方面投入了大量资金，为何班机延误和交通拥堵却越发糟糕呢？虽然有对顾客购物行为的详细记录，为何当我们致电其服务中心时，他们却几乎给不出什么有用的提示呢？尽管在大范围的临床试验中未发现有抗癌作用，可为何 β - 胡萝卜素和维他命药片在药房却是这般抢手呢？

这些例子揭示出一个令人略感不快的惊讶事实，那就是：现代人对测量的迷恋没能使我们变得更具有洞察力、明辨力。诚然，我们现在收集、保存、处理、分析的资讯要比以前多得多，但结果如何？亚里士多德说过的那句名言描述现在的情况再贴切不过了，他说："我们知道得越多，才知自己知道的越少。"

应用科学的力量

我们开始考察一些有积极意义的事例，看看那些富有进取心的人们是如何机智地利用这些新资讯来改善我们的世界的，并借助这些好消息来平复上面的陈述带给我们的挫败感。在后续的五章，你将幸会那些使明尼苏达州的高速公路保持畅通的工程师、提醒人们当心不安全食品的疾病侦探（disease detective）、替佛罗里达的居民计算他们该为自己的房子投多少飓风保险的精算师、那些致力于开发标准化考试（如 SAT）的教育专家、那些仔细检查精英运动员血液样本的实验室技师、那些声称能甄别谎言的数据挖掘师、那些涉嫌欺诈的博彩业执业人员、那些设计出奇思妙招缩短了队列的迪士尼乐园科学家、那些引发了消费信贷高潮的数学家，还有那些为人们的空中旅行提供最佳建议的研究人员。

上面的十句话像十幅速写，勾勒出了一些特殊的男人和女人，他们的工作很少有幸得到公开表彰。这种被漠视起因于他们的工作性质以及一种由来

已久的社会风尚：人们只对发明性的成就颁发奖金、授以嘉奖。这些人的工作不是发明性的，而是适应性的、提炼性的、推销性的以及需要坚持不懈的。他们的专长在于应用科学。

统计式思维

对我来说，这十幅速写最终融合为一个结论，那就是：这些杰出的科学家都仰仗于一种所谓的“统计式思维”，这种思维方式跟我们的日常思维截然不同。我把这些故事组织成五对，每对故事都与一项重要的统计学法则有关。

统计式思维到底有何独特之处?

第一，统计学家们对平均数这个流行概念不太关心；相反，对平均数的任何偏差却是情有独钟。他们反复考虑变异的程度有多大、发生的频率有多高，以及变异存在的原因是什么。在第 1 章中，研究排队问题的专家们，解释了与平均等待时间相比，我们为何更应该担心等待时间的变异性。佛罗里达州高速公路的工程师告诉我们：为什么他们解决拥堵问题时最喜爱的招数是，采取技术手段设置关卡迫使上下班的驾车人多等些时间。而迪士尼乐园的工程师们却证实说，减少等待时间最有效的工具其实并不能真的减少平均等待时间。

第二，我们不必为变异寻找一个合理的解释，尽管我们有一种与生俱来的、对任何事情进行理性诠释的欲望，但如果两件事物之间存在相关模式，统计学家同样很乐意观察它。在第 2 章中，我们追溯了疾病侦探追查污染菠菜的整个过程，又在另一个故事中撬开了产生信用分数的黑箱子。在“追查污染菠菜”这个案例中，（流行病学家）使用的是随机模型（casual models），而产生“信用分数”所使用的则是相关模型（correlation models）。我们对这两种建模方式进行了对比。令人惊讶的是，这些从业人员坦言，他们的模型不能完美地描述周围的世界，从这个意义上说，这两个模型都是“错误的”。我们接下来要看看他们是如何为自己辩护的。

第三，统计学家时常会寻找那些被错过的细微差别：统计平均数（statistical average）也许正好掩盖了各组间存在的重大差异。忽视这个差异通常预示着将来的不公平对待。分组的典型方式，比如按种族、性别或者收入，通常是有缺陷的。第 3 章介绍了保险业。为了反映海岸和内陆地区的房产在遭受飓风风险上的差异，保险公司对保险价格进行了调整。我们对这种做法所带来的混合效果进行了评价。我们也考察了标准考试的设计者为消除黑人和白人在考试表现上的悬殊差距所做出的努力以及由此所带来的后果。

第四，可对基于统计的决策进行微调，来寻找两类错误类型之间的平衡。可以想见，受动机使然，决策者们专盯着那些可能令公众蒙羞的错误，并尽量减少这类错误的发生。然而，统计学家指出，由于这种偏向，他们的决策

会加重另一种类型的错误。而这种错误通常不被注意，但后果很严重。在第4章，我们将用这个原理来解释：为何自动数据挖掘技术不能既可以识破恐怖阴谋又不会带来令人难以承受的附带性破坏；为何类固醇实验室在抓捕大多数舞弊运动员这件事上工作不力。

第五，统计学家在决定证据是否跟罪行匹配时，遵循一种叫作统计检验的特定程式。跟我们中的有些人不同，统计学家们不相信奇迹。换句话说，如果硬要拿最最巧合之事来解释那些费解之处，他们宁愿把这个案子搁置一边。在第5章，我们来看看，在加拿大人们是如何利用这个强大的工具来揭露那个规模巨大的州博彩欺诈的，以及它是如何驱散“怕飞”背后的无稽之谈的。

这五条原理就是统计式思维最重要的部分。读完这本书，你就可以应用这些原理来更好地做出决策了。

工作中的应用科学家

这些故事大致反映了我自己作为一名商业统计执业人员的经历。它们展现出应用科学家跟纯科学家或者说是理论科学家在工作上的某些实质的不同。

所有这些例子都包含那些以某种方式对我们的生活产生影响的决策，或是通过公共政策，或是通过商业策略，或是通过个人选择。理论科学家重在

求“新”，而实用型的工作重在求“高”，譬如“利润会爬到多高？”或者“选票数会有多高？”除了纯粹技术的标准而外，应用科学家还要考虑社会目标，就像明尼苏达州的高速公路工程师那样；或者要考虑心理学的目标，譬如迪士尼乐园的排队管理程序；或者还要考虑经济目标，比如飓风保险承保人和信贷员。

对理论科学的追求很少受到时间的限制。举一个最极端的例子，数学家安德鲁·怀尔斯（Andrew Wiles）花了七年时间周密地证明了费马大定理。这种奢侈不是为应用科学家准备的，他们必须在有限的时间内，通常是在连续几个星期或几个月内尽最大努力解决问题。外部因素，即便是绿色产品的生命周期或者酝酿中的药物发明，都会受时间约束。想想看，假如等到“大肠杆菌疫情”平息了才找到流行病的致病源，那还有何用呢？假如大量的运动员已经因服用类固醇而获得了不公平的优势，此时人造类固醇的检测方法才姗姗而来，请问这还有何意义？

理论科学中最漂亮的某些发现，产生于一组经过审慎选择的、简化过的猜想；应用科学家注意到一些令人出乎意料的细节，并进行了处理，使这些结果能够适用于真实世界。如果你读过纳西姆·塔勒布（Nassim Taleb）的著作，你会认识到钟形曲线其实就是这样一种简化，在某些情况下需要对之进行完善。另一个例子，请参看第 3 章，明显属于不同组的人本该区别对待却被混在了一起。

成功的应用科学家形成了一种本能的决策过程：他们知道主要的影响因素，掌握了自己的那套思考方式，理解自己的动机，也预见到了矛盾的来源。至关重要的是，他们重新整理了用逻辑捆扎的信息，来打动那些喜欢直觉和情感多于证据的人们。鉴于了解事情的背景对应用科学家的工作非常有价值，因此我在故事的叙述中加入了大量的相关枝节。

总结一下，应用科学对成功的量度跟理论科学截然不同。比如，谷歌就认识到了这种区别，因此出台了著名的“20%”时间政策。他们准许工程师们将每周的工作时间一分为二，一部分用于他们所选择的纯理论项目，另一部分用于应用项目。要特别强调的是后者占了 80% 的时间！

数据已经统治我们的世界。对这个事实你决不能一无所知。看看应用科学家是如何利用统计式思维来改善我们的生活的，你会惊奇地发现，在日常生活中，你也能运用数据来做决定了。

NUMBERS RULE YOUR WORLD

The Hidden Influence of Probability and Statistics on Everything You Do

目录

NUMBERS RULE YOUR WORLD

The Hidden Influence of Probability and Statistics on Everything You Do

01

关注异常值，而非平均数本身

| 解决拥堵之害 |

控制咱自己啊！若无人愿意，为何要服从？绿灯一闪，单车通行呐。

——“双子城”的上班者之歌（俳句）
“行路者”博客（Roadguy）的读者

海姆利希的嚼嚼车；精彩的电影呐，好评如潮呐，好长的队伍呐；可就坐了20秒呐。

——迪士尼之歌（俳句）
作者 无名

可怕的平均化

2008 年年初，资深记者詹姆斯·法洛斯（James Fallows）在《大西洋月刊》（*The Atlantic*）上发表了一篇令人瞠目的文章，谈及美国对中国的贸易赤字已经失去了控制。在这篇文章中，法洛斯阐述了一直以来中国人是如何支撑美国人如此高的生活标准的。

这家格调高雅的杂志很少在网络上受人追捧，这次却引起了意想不到的轰动。事情的起因是这样的，有网民读到了法洛斯的文章，他们把原标题“1.4 万亿美元之疑”（*The $1.4 Trillion Question*）弃置不用，将其改成了“平均每个美国人欠每个中国人 4 000 美元”（*Average American Owes Average Chinese $ 4 000*）。网民很欣赏这篇文章，三个月内就对其“掘”（dig）了 1 600 多次，或者说对其作出了 1 600 多次积极响应。“掘客”是个网络新名词，“掘”是唱赞歌的一种高科技方式。很显然，这则最新头条引发了热议。我们的大脑在处理 1.4 万亿美元这样的天文数字时有点儿困难，