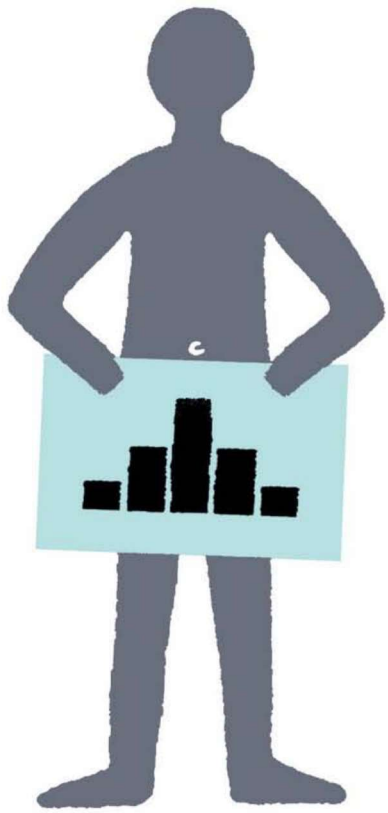
赤裸裸的统计学

除去大数据的枯燥外衣，呈现真实的数字之美

[美] 查尔斯·惠伦（Charles Wheelan）◎著
曹槟◎译



统计数字很容易说谎，
但没有它们，你就无法在大数据时代找到真相、预测未来！

Naked Statistics

Stripping the Dread from the Data

中信出版社·CHINACITICPRESS

赤裸裸的统计学

除去大数据的枯燥外衣，呈现真实的数字之美

[美] 查尔斯·惠伦（Charles Wheelan）◎著
曹槟◎译

Naked Statistics

Stripping the Dread from the Data

中信出版社 · CHINACITICPRESS · 北京 ·

图书在版编目（CIP）数据

赤裸裸的统计学 /（美）惠伦著；曹槟译. —北京：中信出版社，2013.11
书名原文：Naked Statistics
ISBN 978–7–5086–4215–4
①I. 赤… II. ①惠… III. ①经济统计学－通俗读物 Ⅳ. ①F222–49
中国版本图书馆CIP数据核字（2013）第215055号

赤裸裸的统计学

著　　者：[美]查尔斯·惠伦
译　　者：曹　槟
策划推广：中信出版社（China CITIC Press）
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座　邮编　100029）
（CITIC Publishing Group）
承 印 者：北京诚信伟业印刷有限公司

开　　本：787mm×1092mm　1/16　　印　　张：19.25　　字　　数：240千字
版　　次：2013年11月第1版　　印　　次：2013年11月第1次印刷
京权图字：01–2013–1461　　广告经营许可证：京朝工商广字第8087号
书　　号：ISBN 978–7–5086–4215–4 /F·3002
定　　价：42.00元

我为什么憎恶微积分却偏爱统计学？

我天生就很排斥数学。我对数字本身没有任何好感，对那些在现实世界中毫无用处的骗人公式也没有什么好印象。我尤其不喜欢高中的微积分课，原因很简单，因为从来就没有人告诉过我学习这门课的意义是什么——有谁会在乎抛物线下方的区域代表什么？

而事实上就在高中三年级的时候，我迎来了人生中的一个重要时刻，那时我正在准备第一学期微积分课程的期末考试，虽然那几天我也算用功学习了，但总体来说还是以偷懒为主，因为几个星期前我就申请到了理想的大学，当然随之而来的是我对这门课本来就少得可怜的学习动力也消失殆尽。考试那天我盯着试卷上的题目，发现它们竟是如此陌生。这已经不是会不会答的问题了，而是根本就搞不清楚题目问的是什么。我对“裸考”其实

并不陌生，借用美国国防部前部长唐纳德·拉姆斯菲尔德的话说就是，我总是知道我有不知道的东西。但这次考试比以往的题目都难，我草草地翻了一下试卷，几乎没有会答的题。我走到教室前面，来到监考老师——我们的微积分老师卡罗·史密斯的面前，“史密斯夫人，”我说，“试卷上的很多东西我都不认识。”

相比起我对史密斯夫人的“喜爱”，她对我的“不喜爱”要更甚。是的，现在我承认作为学生会主席的我，有时会动用手中有限的权力来安排一些全校性的集会，这样史密斯夫人的微积分课就被迫取消了。我和朋友们也曾以“一位神秘的仰慕者”的名义派人在课堂上给她送花，然后看她尴尬地环顾四周，而我们则在教室后面得意地窃笑。是的，在我得知自己被大学录取之后，我就真的再也没有做过任何作业了。

所以，当我走到史密斯夫人的面前，告诉她那些题目看上去很陌生的时候，她并没有流露出一丝的同情。“查尔斯，”她大声说——表面上是对我说，但她的脸却朝着全班同学，以确保教室里的每一个人都能听到——“如果你用功了，这些题目看上去就会熟悉得多。”这一点确实很有说服力，所以我只得溜回座位。几分钟以后，我们班这门课的“尖子生”布莱恩·阿尔贝特尔走到教室前面，和史密斯夫人耳语了几句，史密斯夫人也轻声地回了几句，之后，一件十分离奇的事情发生了。“同学们，请注意一下，”史密斯夫人宣布，“我误把下学期的试题发给你们了。”当时考试已经进行了一段时间，所以这次考试不得不取消择日重考。我当时的欣喜之情无以言表。

在我之后的人生中，我娶了一位漂亮的妻子，育有 3 个健康的孩子。我出版了几本书，游览过泰姬陵和吴哥窟这样的名胜。但是，那天微积分老师得到“因果报应”的一幕，依旧是我人生中最难忘的 5 个时刻之一。（事实上，在之后的补考中我差点儿没及格，但这一点儿都没有使这一美妙的人生经历褪色丝毫。）

微积分考试的小插曲极大地说明了我和数学之间的关系，但这并不是事实的全部。有趣的是，尽管物理课也需要进行像微积分课那样令人厌烦的演算，但我在高中时却十分喜欢物理课。这又是为什么？因为物理课有一个明确的目的。我清楚地记得在世界职业棒球大赛期间，我们的物理老师教我们如何运用加速度的基本公式来预测一个本垒打能打多远。这简直酷毙了——这个公式在生活中也有很多重要的应用。

上大学之后，我彻底沉醉于概率学之中，因为它同样为我在洞察现实生活中的一些有趣场景提供了解释。回想过往，我意识到让我痛恨微积分课的不是数学，而是从来就没有人想到要告诉我数学的意义是什么。如果你没有被“高雅”的公式本身所吸引——反正我是一点儿都不觉得有什么“高雅”的——那么，你面对的只会是繁冗而机械的公式，至少我的老师当初就是这样把它们教给我的。

也正是因为这一点，我与统计学结了缘（本书所指的统计学包括概率学在内）。我爱统计学。生活中的一切一切，从脱氧核糖核酸（DNA）检测到买彩票的白痴行为，统计学通通都能做出解释。统计学能帮助我们识别诱发某些疾病的因素，比如说癌症和心脏病；统计学还能帮助我们在标准化考试中甄别作弊行为；统计学甚至能帮助你在电视游戏节目中获胜。在我的孩童时代有一档非常出名的节目，叫作《让我们作个交易》，由当时极受欢迎的蒙提·霍尔主持。在每天节目快要结束时，胜出的选手和蒙提都会站在 3 扇大门的前面，蒙提·霍尔会告诉观众和选手，在其中一扇大门的门后会有一项大奖，如一辆小轿车，而另外两扇门的门后则各站着一头山羊。玩法很简单：选手选择一扇门，然后就会得到这扇门后面的奖品。

当选手和蒙提·霍尔站在这 3 扇门的前面时，这位选手中大奖的概率为 1/3。但是，这档节目却有其微妙之处，这让统计学家们欣喜万分（却也使其他人困惑不已）。在选手选择了其中一扇门之后，蒙提·霍尔会先打开剩下的两扇门中的一扇，而打开的这扇门后面站着的永远是一头山羊。举个例子来说，假设选手选择了 1 号

门，那么蒙提会先打开 3 号门，它的后面站着一头山羊，此时 1 号门和 2 号门依然紧闭。如果大奖就在 1 号门后面，则选手获胜；如果大奖在 2 号门后面，则选手失败。但节目进行到这里的时候，会变得更加有戏剧性：蒙提会转向选手，问其是否更改之前的决定（在这个例子中就是把 1 号门改为 2 号门）。需要注意的是，此时剩下的两扇门依然是关着的，而选手得到的唯一的新信息，就是他之前没选的那两扇门中，有一扇门的后面经证实是一头山羊。

那么，这位选手是否应该更改之前的选择？

答案是肯定的。为什么呢？本书之后的内容会做出解释。

统计学的悖论就在于，从棒球比赛的击球成功率到美国总统大选的民意调查，它几乎无处不在，但是这个学科本身却因为乏味无趣和难以理解而“臭名昭著”。许多统计学方面的书籍和课程也都过多地充斥着数学和术语。相信我，技术细节十分重要（也十分有趣），但是如果你不知道它们的出发点是什么，那么摆在你面前的将会是一堆天书般的符号。如果连你自己都不相信学习统计学是一件有意义的事情，那么你或许根本不会去关心所谓的出发点。本书中的每一章都旨在回答我向高中微积分老师提出的那个基本问题：学习统计学的意义是什么？

这是一本有关直觉的书。书中很少出现计算、公式和图表；当用到它们的时候，我保证它们都存在一个清晰和富有启发性的目的。与此同时，书中常常会出现很多例子，目的就是让你相信，学习统计学是很有必要的。统计学真的可以非常有趣，而且其中绝大部分的内容也没有那么难。

在学习过史密斯夫人讲授的微积分课程后不久，我就萌发了写这本书的想法。那段“不堪回首”的经历就发生在我读研究生期间，那时我学的是经济学与公共政策专业。在开始学习这门课之前，我和班上的大部分同学都毫无意外地被指派到了一个“数学营”进行集训，为接下来的“数学轰炸”作准备。在 3 周的集训时间里，

我们整天待在一间没有窗户的地下室里学数学——真的一点儿都不夸张。

就在其中的某一天，我离顿悟仅有毫厘之差。那时，负责集训的老师正在费劲地教我们在某些情况下能够从一个无穷级数求得一个有限数。请不要跳过这一段内容，因为这一概念马上就会清晰起来（现在，你可以想象我在那个没有窗户的教室里是什么感受了吧）。无穷级数指的是一个可以无限地写下去的数字组合，如1+1/2+1/4+1/8……最后的省略号表示这个算式还将无限地继续下去。

到了这一步，我们基本上已经开始感到困惑了。老师试图通过一些我早已遗忘的定理向我们证明，一个无穷尽的算式依然可以通过求和得到一个（大概）确定的数值。尽管有很多令人信服的数学证明，但班上的威尔同学却死活不能接受这一结论（老实讲，我自己对此也心存疑惑）。无限的东西经过叠加怎么可能得到一个有限的结果呢?

突然我灵光一现，更准确地说，是我的直觉让我想通了老师要表达的意思。我对威尔说了我的头脑里刚刚闪现出来的想法：想象自己站在离一堵墙正好两英尺（约0.6米）的地方。

现在朝墙壁的方向移动1/2的距离（即1英尺），这样你离墙壁就只剩下1英尺的距离了。

再面向墙壁的方向移动1/2的距离（即6英寸或1/2英尺），继续重复相同的动作（即移动3英寸或1/4英尺），再移动剩下距离中的1/2（即1.5英寸或1/8英尺），不断重复。

最终你将十分贴近墙壁，假设现在你离墙壁只剩下1/1 024英寸，然后你还需要朝墙壁的方向移动1/2的距离，即1/2 048英寸，但你永远都不会撞到墙壁，因为理论上你所移动的每一步都只有剩余距离的1/2。也就是说，你将无限接近墙壁但永远碰不到墙壁，如果我们统一用英尺作为计量单位，那么你所移动的距离就可

以表示为 1+1/2+1/4+1/8……

问题的核心就是：即使你正在不停地靠近墙壁，而且每一步都是剩余距离的1/2，但你所走过的总距离永远都不可能超过两英尺，也就是一开始你与墙壁之间的距离。出于计算的目的，你所走路程的总长度可以简单地估算为两英尺，但数学家会说 1+1/2+1/4+1/8……最终收敛于 2，这也是那天老师想要教给我们的。

关键在于我说服了威尔，也说服了自己。虽然我不记得这道题的数学推理论证过程，但我总是可以在网上寻找答案，而且当我找到答案的时候，我或许还能看出一点儿门道来。以我的经验来看，直觉会让数学和其他技术细节更加容易理解，但是反过来就不一定说得通了。

本书的目的就在于使重要的统计学概念变得更加直观和便于理解，不仅让我们这些被迫在没有窗户的教室里苦学过的人，更可以让任何对数字和数据的惊人力量感兴趣的人都爱上统计学。

刚刚我还在说统计学的核心并没有那么的直观和好理解，现在我却要提出一个貌似自相矛盾的观点：统计学可以变得非常好理解，任何人只要拥有数据和一台电脑，就可以通过简单地敲击几下键盘来完成复杂的统计流程。问题是如果数据不足，又或者统计方法错误，那么得出的结论将会谬以千里，甚至还会有潜在的危险。就比如下面的这条虚构的网上新闻快讯：工作时小憩的人更易死于癌症。假如你在上网时这个标题突然从页面弹出呈现在你眼前，你会怎么想？一项基于 3.6 万名办公室白领（多大的数据组啊！）的调查显示，那些表示会在工作期间偶尔离开办公室休息 10 分钟的员工在未来 5 年内身患癌症的概率要比那些从不离开办公室的同事高 41%。显然我们需要为此做点什么，比如在全美国范围内掀起一股抵制办公期间小憩的热潮。

或许，我们只需要对员工在休息的 10 分钟里干了什么事情作些思考。我的工

作经验告诉我，这些离开办公室休息的员工中有很多人都聚在办公楼的入口处吸烟（其他人如果要进入或走出大楼都必须一头扎进他们吞吐的“云雾”之中）。那么，我会进一步推断是香烟而非小憩引发了癌症。我举的这个例子当然十分荒谬，但现实生活中有许多统计学结论在经过解构之后，也产生了类似荒谬的效果。

统计学就像是一种高智商武器：正确地使用它能够帮助我们，但错误地使用它也会产生灾难性的后果。本书不会将你变成一个统计学专家，但会让你对这个领域保持谨慎和尊重，不至于酿成大祸。

如果这是一本统计学教科书，那么各种概念和方法都会罗列其中，而不管普通读者是否能够消化。不过，本书的创作初衷就是介绍那些与日常生活联系最为紧密的统计学概念。科学家们是如何总结癌症诱因的？民意调查是如何发挥作用的（哪些方面又会出问题）？哪些人设计了“统计陷阱”，这些人又是如何做到的？你的信用卡公司是如何根据你的消费数据，来判断你是否会错过还款期限的（别笑，它们真的做得到）？

如果你想要理解新闻中出现的数字背后的含义，并见识到“数据”的巨大力量，统计学就是你的不二法宝。最后，我还想与大家分享瑞典数学家、作家安德烈斯的一句话：用数据说谎容易，但是用数据说出真相却很难。读罢此书，我希望你们也能感同身受。

除此之外，我还有一个更加宏伟的目标，那就是让作为读者的你真正地喜欢上统计学。这是一门充满乐趣且与我们的生活息息相关的学科，关键在于如何将学习过程中涉及的技术细节与那些重要的理念剥离开来，这就是赤裸裸的统计学。

目录

Naked Statistics

引　言　我为什么憎恶微积分却偏爱统计学?

第1章　**统计学是大数据时代最炙手可热的学问**／1

基尼系数是否是衡量社会分配公平程度最完美的指标?视频网站是如何知道你喜欢的电影类型的?祈祷真的能让病人的术后康复状况改善吗?是什么导致自闭症发病率一直走高?哪些人最有可能成为恐怖分子?

第2章　**描述统计学**／19

你一直想买的一条连衣裙,商场售价为4 999元,先降价25%后再提价25%,你能算出这条连衣裙的最终售价是多少吗?

第3章　**统计数字会撒谎**／43

1950年人们的平均时薪是1美元,2012年人们的平均时薪是5美元,你觉得我们的工资水平涨了吗?

第4章　**相关性与相关系数**／69

视频网站根本不知道我是谁,但它又是怎么知道我喜欢看人物纪录片而不是电视连续剧、动作片或科幻片的?

第5章　**概率与期望值**／81

买福利彩票,去赌场豪赌、投资股票或期货,哪种方式让你跻身《福布斯》富豪排行榜的可能性更大?

第6章 **蒙提·霍尔悖论** / 105

在《让我们做个交易》节目中，主持人打开的3号门后面是一头羊，在剩下的1号门和2号门中必定有一扇门后面是汽车，你应该如何选择才能中大奖？

第7章 **黑天鹅事件** / 113

1%的小概率风险如何在2008年成为击垮美国华尔街的“黑天鹅”，并毁了全球金融体系。

第8章 **数据与偏见** / 131

2012年，《科学》杂志刊登了一项惊人的发现：在求偶期多次遭受雌性果蝇冷落的雄性果蝇会“借酒消愁”。那么，这些果蝇是如何一醉方休的？

第9章 **中心极限定理** / 151

一辆坐满肥胖乘客的抛锚客车停在你家附近的路上，你推断一下，它的目的地是马拉松比赛场地，还是国际香肠节展厅？

第10章 **统计推断与假设检验** / 169

垃圾邮件过滤、癌症筛查、恐怖分子追捕，我们最不能容忍哪件事情出错，又有哪件事情是可以“睁一只眼闭一只眼”的？

第11章 **民意测验与误差幅度** / 197

民调结果显示，有89%的美国人不相信政府会做正确的事，有46%的美国人认可奥巴马的工作表现。这个结果可以代表美国人的真实想法吗？

第12章 **回归分析与线性关系** / 215

你认为什么样的工作压力更容易使职场人士猝死，是“缺乏控制力和话语权”的工作，还是“权力大，责任也大”的工作？

第13章 **致命的回归错误** / 243

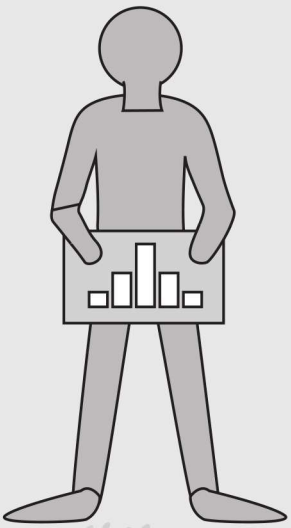
世界上3本最有声望的医学期刊上刊登的49篇学术研究论文中有1/3后来都被推翻了，所以，“尽量不要用你的回归分析研究杀人”。

第14章 **项目评估与“反现实”** / 259

哈佛大学等世界顶尖大学的毕业生进入社会后，其收入往往高于一般大学的毕业生，让他们获得高收入的究竟是常春藤大学的教育优势，还是他们本身就很出色？

结束语 统计学能够帮忙解决的5个问题 / 277

致 谢 / 293



Naked Statistics

第1章

统计学是大数据时代最炙手可热的学问

基尼系数是否是衡量社会分配公平程度最完美的指标？视频网站是如何知道你喜欢的电影类型的？祈祷真的能让病人的术后康复状况改善吗？是什么导致自闭症发病率一直走高？哪些人最有可能成为恐怖分子？

我注意到一个有趣的现象。学生们在课堂上常常抱怨统计学课程有多么难学和无关紧要；可一离开教室，他们又会在午饭时开心地讨论某位球星的击球成功率（夏天）或寒冷指数（冬天），又或者彼此成绩的平均分数（永恒的话题）。他们会指出美国职业橄榄球联盟（NFL）采用“传球效绩指数”用以将一个四分卫的场上表现浓缩为一个数字的不当之处，认为以此作为评价球员的依据略显武断，但可以通过调整其中所包含数据（完成率、平均过球码数、触地得分率、截球率等）的权重比例重新计算，以得出一个与原来不同，但同样可信的球员表现指数。但只要是看过橄榄球比赛的人都会觉得，没有比用一个单一数字来衡量四分卫的表现更加方便的了。

关于四分卫表现的这个评价指数是完美的吗？当然不是，无论是什么问题，统计学都极少提供唯一的“正确”方法。但是，这个指数是否以一种易于理解的方式提供了一些有意义的信息呢？那是肯定的，如果想快速地对某场比赛的两名四分卫的表现做出比较，那么这个指数会是一个不错的工具。我是芝加哥熊队的粉丝，在 2011 年季后赛期间，熊队与芝加哥包装工队进行了一场比赛，以后者的胜利告

终。我可以通过很多种方式来描述那场比赛，包括长篇累牍的分析和令人眼花缭乱的原始数据，但这里我为大家提供了一种更加简洁的分析方法。芝加哥熊队的四分卫杰 · 卡特勒的传球效绩指数为 31.8；与此同时，格林湾队的四分卫亚伦 · 罗杰斯的传球效绩指数为 55.4。同样的，我们可以将杰 · 卡特勒与他之前跟格林湾队比赛时的表现进行对比，在那场比赛中他的传球效绩指数高达 85.6。两者相比较，我想大家就不难理解为什么熊队在常规赛时击败了包装工队，但在季后赛时却输给了包装工队。

这对于概括场上进行的比赛非常有用。传球效绩指数是否起到了简化问题的作用？是的，但这同时也反映了描述统计学的优势和劣势。仅凭一个数字，你就可以知道杰 · 卡特勒在与格林湾的那场比赛中败给了亚伦 · 罗杰斯；但你却无法从这个数字中读出运动员在比赛中的运气是好是坏；不知道他是否传出了一个漂亮的过人球却被愚蠢的队友错过了，导致这个球最终被对方截获；不知道他是否在比赛的某些关键时刻顶住压力发挥出色（因为每一次的成功发球在统计时都被同等对待，不论是决定性的三次触地还是比赛接近尾声时那些毫无意义的发球）；不知道那一场的防守是否糟糕透顶……读不出来的信息还有很多。

令人好奇的是，同样一群人，在谈论体育、天气或成绩的时候提到数据时还是兴高采烈的，但是当研究人员开始向他们解释基尼系数时，他们的手心却出汗了。基尼系数是衡量收入不均的标准经济学工具，我在之后的内容中将对其做出解释，但是现在我要说的最重要的事情是，基尼系数实质上与传球效绩指数没有多大区别，都是将一系列复杂数据浓缩成一个单一数字的便捷工具。正因如此，基尼系数也拥有描述统计学的大多数优势，如果你想比较两个国家或某个国家不同时期的收入分配情况，该系数就为你提供了一个简单易行的方式。

基尼系数用于衡量一个国家的财富（或收入）分配的公平程度，最小为 0，最

大为1。计算基尼系数可以看总资产，也可以看年收入；可以以个人为计算和比较单位，也可以以家庭为单位。所有这些数据都是紧密联系的，但不会完全相同。就像传球效绩指数一样，基尼系数只是一个用作比较的工具，其数字本身并无实质意义。在一个家庭财富均等的国家里，基尼系数为0；与此相反，如果一个国家的所有财富都集中在一个家庭里，那么这个国家的基尼系数等于1。或许你已经猜到了，一个国家的基尼系数越接近于1，那么这个国家的财富分配就越不公平。根据美国中情局提供的数据（顺便说一句，这可是一个巨大的数据收集机构），美国的基尼系数为0.45。那又怎么样？

如果将这一数字放到实际情况中，我们就可以得到许多信息。例如，瑞典的基尼系数为0.23，加拿大为0.32，中国为0.42，巴西为0.54，南非为0.65。[①]纵观这些数字，我们能够感觉到美国在收入的公平分配方面相对落后，情况比许多国家都要糟糕。我们同样可以对不同时期的收入分配的公平情况进行比较，1997年美国的基尼系数为0.41，但在接下来的10年内，基尼系数就上升到了0.45（最近一次来自美国中情局的数据是在2007年），这就客观地告诉我们在这10年的时间里，美国虽然变得更加富裕，但财富的分配也变得更加不公平。现在我们再来看一下其他国家在这一时期内基尼系数的变化情况，加拿大在过去10年中的收入分配情况基本上保持不变；瑞典经济虽然在过去20年的时间里得到了长足发展，但其基尼系数却从1992年的0.25降到了2005年的0.23，也就是说瑞典不但变得更为富裕，其社会也变得更加公平。

基尼系数是否就是社会分配公平程度最完美的衡量指标呢？绝对不是，正如传球效绩指数也不是衡量四分卫比赛表现的完美指标一样。不过，基尼系数确实以一种便捷易懂的形式为我们提供了一个重要社会现象的一些宝贵信息。

① 基尼系数有时候会乘以100得到一个整数，拿文中的例子来说，美国的基尼系数也可以是45。