Broadview®
www.broadview.com.cn

Scammed by Statistics

揭开数据真相

从小白到数据分析达人

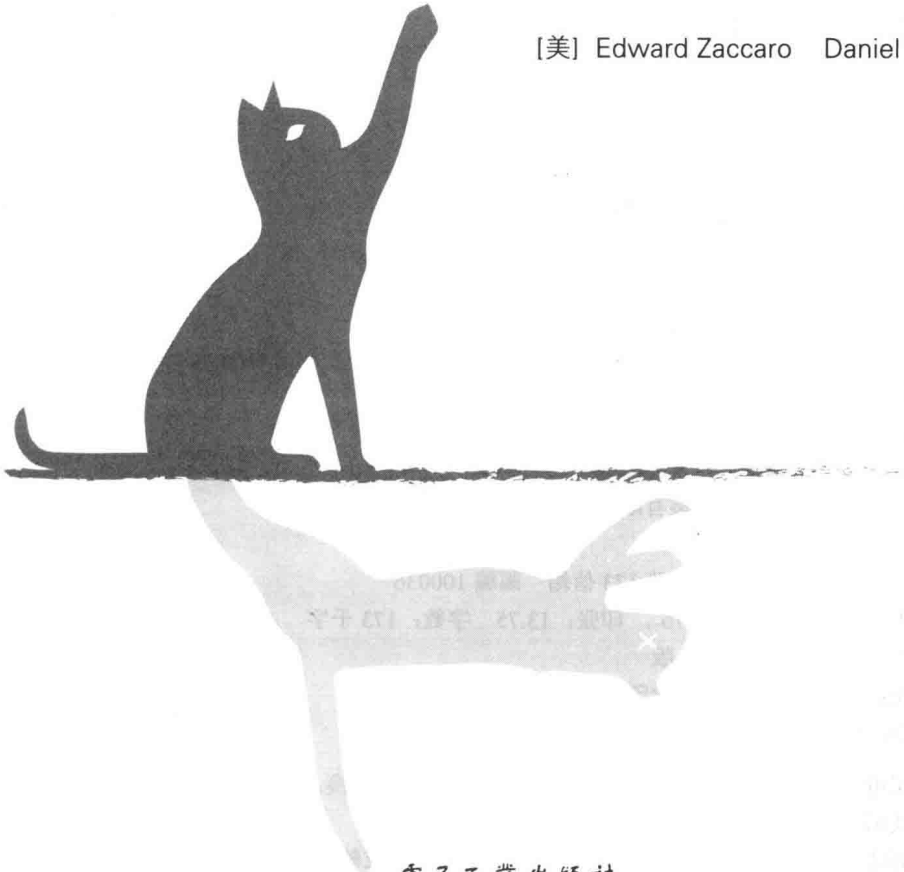
[美] Edward Zaccaro　Daniel Zaccaro　著
李芳　译

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Scammed by Statistics

揭开数据真相

从小白到数据分析达人



[美] Edward Zaccaro　Daniel Zaccaro　著
李芳　译

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

统计数据之所以强大有力，原因在于它对我们的希望、梦想和信仰无动于衷——数据让我们客观地看待事物。不过，当人们不喜欢数据告诉我们的结果时，常常对其进行操纵……因此懂得解释统计数据，了解各种歪曲、滥用数据的技术对于理解数据真相是非常必要的。

本书教给读者神圣的技术，让读者学会如何质疑“看得见”的数据，并挖出“看不见”的数据真相，还原基本的事实。

本书适合所有对数据分析感兴趣的读者。

版权贸易合同登记号　图字：01-2015-7154

图书在版编目（CIP）数据

揭开数据真相：从小白到数据分析达人 /（美）爱德华·佐卡罗（Edward Zaccaro），（美）丹尼尔·佐卡罗（Daniel Zaccaro）著；李芳译. —北京：电子工业出版社，2016.11

书名原文：Scammed by Statistics

ISBN 978-7-121-29953-7

Ⅰ. ①揭… Ⅱ. ①爱… ②丹… ③李… Ⅲ. ①数据处理－研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 228491 号

策划编辑：刘　皎

责任编辑：王　静

印　　刷：三河市鑫金马印装有限公司

装　　订：三河市鑫金马印装有限公司

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱　邮编 100036

开　　本：720×1000　1/16　印张：13.75　字数：173 千字

版　　次：2016 年 11 月第 1 版

印　　次：2016 年 11 月第 1 次印刷

定　　价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819　faq@phei.com.cn。

本书献给我的父亲——

卢克·N·扎卡罗（Luck N. Zaccaro，1924—1977），

他是一位数学家、批判性思维者、哲学家。他春风化雨，

在孩子们心中种下对生活的爱、理性和责任。

“有一天，统计思维将和读写能力一样成为高效公民的必备技能。”

——H·G·威尔斯（H.G.Wells）

引 言

“数学是宇宙与人类交流及吐露真相的方式。”

——伽利略

“数学比其他任何在人类社会中传承的知识结构更为强大有力。”

——笛卡儿

“数字是最高级别的知识。它就是知识本身。”

——柏拉图

统计之所以如此美妙、如此强大有力，原因在于它对我们的希望、梦想和信仰无动于衷——统计让我们客观地看待事物。可惜，统计数据常常被当作裁判，当我们不喜欢统计数据告诉我们的结果时，我们可以与之辩论，对其操纵。下面的实例提醒我们，忽视统计数据传递的信息极为危险。

1999 年，一家大型制药企业生产的一种名为万络（Vioxx）的轰动一时的止痛药物进入最后实验阶段。万络能止痛，却不像阿司匹林那样会引发胃肠道并发症，它前途无量，不仅有可能帮助成千上万的人，而且能为制药厂赚取数十亿元的真金白银。

制药公司明白，必须小心对待万络的最终实验——尤其必须要小心选择万络的竞争药物。经过深思熟虑，制药公司决定选择萘普生（Aleve）作为实验竞争药物（因为这种药物对心脏病是否有防护作用还未知）。

9 个月以后，经过对临床数据进行分析，得出了惊人的结果！服用万络的实验组发生心脏病的次数是服用萘普生的实验组发生心脏病的次数的 4 倍。统计结果提供的信息非常清楚——万络是心脏病发作的重大原因，这一点可能性很大。

可惜，解释统计数据的人往往做不到或不愿意客观地审视统计数据，他们很容易受到才能、意愿和贪婪的影响。因此，研究结论未指出万络导致心脏病的发病风险提高 400%，而是指出萘普生导致心血管疾病的发病风险降低 80%。这个解释让人难以置信，因为，前面已经提到，萘普生不像阿司匹林，它对心脏的保护作用尚未可知。实际上，如果萘普生确实能将心脏病的发病风险降低 80%，那么它的效果将达到阿司匹林的 2～3 倍！

尽管临床实验清楚地表明万络存在危险，万络还是得到美国食品和药品管理局（FDA）的批准，随后被数百万人选用。4 年以后，万络从市场上被撤下，然而这时它引发的心脏病以及死亡人数已经令人胆寒。FDA 估计万络引发了 88 000 至 139 000 例心脏病——其中 30%～40%致命。[1]

在万络/萘普生研究中得到的统计值显示出清晰的信息，但这些信息遭到忽视，造成千上万人死去。

“数学是宇宙与人类交流及吐露真相的方式。”

——伽利略

“数学比其他任何在人类社会中传承的知识结构更为强大有力。”

——笛卡儿

“数字是最高级别的知识。它就是知识本身。”

——柏拉图

这些话是伽利略、笛卡儿、柏拉图对数学的力量的真知灼见。统计的力量在我们的社会中已经作用了数百年；使用得当时，这种力量有可能拯救数百万人的性命。可惜，“统计警告”被歪曲、操纵、最小化的例子不胜枚举。这种知识和道德上的失败所造成的结果是——数百万人丧失本来不必丧失的生命。

我们对于下列问题的统计警告实在反应太慢：

- 烟草
- 石棉
- 苯
- 万络
- 胃药
- 铅
- 赖式综合征/阿司匹林关系
- 酒精

由于滥用、操纵统计数据造成的惨剧不应该致使我们相信——统计永远会被操纵，永远无用，永远不可信。每一例不恰当使用统计的意外事件总是对应着上百例公正、合理使用统计的实例——这给社会带来极大好处。下列 5 个实例向我们展示了统计的有利用途：

- 有一个统计模型帮助人们在 18 个月里防止了 100 000 多例由于医院过错导致的死亡。
- 奥克兰运动家队聘用队员的薪水差不多是业界最低的，却依靠统计成为最佳棒球队之一。
- 事实证明，一个数学公式比一群专业品酒师能更准确地预测出葡萄酒的质量。
- 一个统计模型比一群全国著名的法律专家能更准确地预测出最高法院的投票结果。
- 统计被用于帮助急诊室医生做出更好的判断。

统计具有改善我们生活的能力，因此，懂得如何使用统计对我们来说是基本的技能。此外，由于操纵、欺诈和彻头彻尾的谎言常常伴随统计登场，懂得解释统计数据，对各种歪曲、滥用数据的技术有所了解也非常必要。

在阅读本书的过程中，有一些例子可能会引读者发笑，有一些则令读者愤慨。我希望，在读完本书后，读者不仅懂得如何质疑自己看见的统计数据，而且能够明白：统计学习并非像人们常说的那样枯燥、乏味。

别担心，要是别的办法都失败了，我们可以操纵数据，让它看上去能飞。

目 录

引 言 X

第 1 章 几乎不可信的各种图形 1

燕麦的降胆固醇功效 1

美化上升的犯罪率（纯属虚构） 4

哪家汽车制造公司更棒 8

条形图中的党派差异 10

在线广告衰退正式开始 12

美化 SAT 成绩 17

美国中西部加热燃料消费价格飞涨 20

交通事故死亡人数减少了吗 24

恶化房地产低迷状况 25

超大号熊猫金币 27

吊顶条形图的巧妙骗术 28

第 2 章 所比较的群体旗鼓相当吗 31

加利福尼亚州是否比伊拉克更危险 31

全球变暖和耸人听闻的飓风损失 33

某中西部城市学习成绩飙升的表象 36
租金辅助计划与犯罪率上升有关系吗 41

第 3 章　先射箭，再画靶 45

冥想实验 46
关节炎患者的天大好消息——或者相反 48
旧车换现金计划“惨败” 49
民意调查公司/智库合作关系 51
杰·雷诺居然也操纵统计数据 54

第 4 章　诚实统计的力量 56

忽视统计警告，丧失 4000 条生命 56
数学 VS 专业品酒师 58
数学对阵法律专家 60
统计——18 个月挽救 100000 条性命 62
统计——帮助急诊室医生做出更好的判断 64
统计——提高棒球队成绩？（棒球星探 VS 计算机） 65
统计的早期利用，挽救数千生命 67

第 5 章　故施迷雾 69

辛普森案 69
雷氏综合征如何导致数百例儿童死亡——这本来可以避免 73
导致年轻女子中风的厌食剂 74
烟草行业——统计操纵与故布迷阵的行家里手 75
石棉：寿险公司所知道的、石棉行业故作不知的危险 79

第 6 章 资助效应 83

钱能控制数据，钱能限制公众得知负面结果 83

制药公司刻意压制负面数据后果可能很严重
（抗抑郁药物帕罗西汀的故事） 85

钱可以影响医生，可以给医生带来偏见 87

抗抑郁剂与安慰剂——出人意料的胜出者 88

资助效应甚至会伤害新生儿重症监护室中最易受伤害的儿童 90

第 7 章 烂逻辑 93

新款雪佛兰福特汽车的惊人燃油效率：230 英里/加仑 93

为什么患糖尿病的人越来越多 95

到 2048 年，每一个美国人的体重都会超重 96

解开谜团：为什么加拿大人的预期寿命比美国高 98

夸张的广告 99

非常奇怪的逻辑 101

第 8 章 因果与相关乱象 103

恢复前囚犯的投票权将降低犯罪率 104

因果关系混淆会导致丧失生命 108

学习成绩好的关键是让家长出席家长会 112

音乐与学习成绩 113

第 9 章 要看就看全部数据 116

选举奥巴马总统搞垮了股市 116

广告商与有选择地使用数据 119

您会选择哪家宾馆 121

我该买黄金吗 123
有可能遭到操纵的合理图形 125

第 10 章　确认性偏差（所愿即所见） 128

星座效应 128
预测死亡的猫 130
分母在哪里 133
画中音乐 137
《秘密》 141
确认性偏差的负面特性 144
辅助沟通 148

第 11 章　稻草人论证术 152

医疗保健辩论策略 152
2010 年煤矿爆炸以及首席执行官的稻草人辩护术 156

第 12 章　操纵均值、中位数和众数 161

第 13 章　轶事证据 168

疾病与轶事证据 169
磁疗 171
占卜杖探测术 172
外星人奇遇 174
结论 176

第 14 章　如果你的事业缺乏统计支持，那么，创造吧 177

潜意识广告的力量 177

死亡率畸高的神经性厌食症 179
美国的 300 万名无家可归者 180
其他影响公众的错误统计 181

第 15 章 令人费解的百分数 183

被百分数愚弄的医生 183
住家孩子增长趋势 184
移民家庭的刻苦孩子 185
需求神秘下降 500% 187
我当初真应该别开始锻炼 188
了解百分数可以救人性命 191

第 16 章 你的样本合理吗 195

代表性样本的重要性 195
总统大选：罗斯福与兰登 197
当研究参与人自我选择或样本有偏差，
则结果几乎总是无效的 198
双盲的重要性，随机临床实验 200
检验组大小的重要性 202

注释 205

第 1 章

几乎不可信的各种图形

“抽奖：向数学不好的人征税。”

——佚名

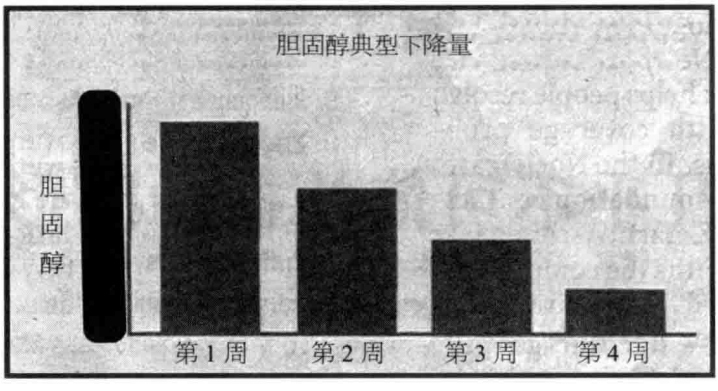
燕麦的降胆固醇功效

胆固醇是人类细胞和血液中重要的多脂肪物质，人体缺少胆固醇就会失去机能，但是，胆固醇水平过高却可能会在血管中形成脂肪堆积，进而引发心脏病。

血液中的胆固醇有两个来源：一个是食物，一个是肝脏。所以，治疗胆固醇过高往往会采用节食、锻炼和用药三管齐下。医学界目前（2010 年）认为，人体的胆固醇总量应为 200 毫克/分升。

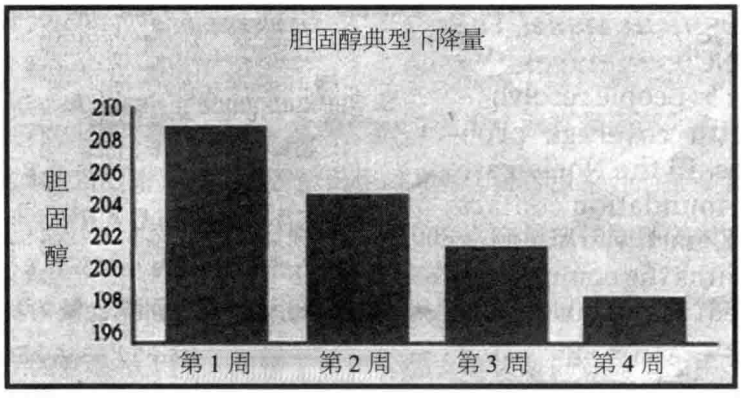
现在有一些食物以健康、无痛苦地降低人体的胆固醇含量为卖点，燕麦就是其中之一。有一家大型食品公司为了帮助人们了解燕麦的降胆固醇功能，在广告中展示了一幅条形图。下面这幅条形图的 Y 轴被遮住了，这

样你可以猜一猜在该项为期 4 周的研究中，胆固醇水平的下降幅度。



条形图的外观给我们留下了这样的印象：食用燕麦 4 周之后，人体的胆固醇水平一般会大约下降 75%，在满足以下两点假设的情况下，这样的下降比例是一个好消息：（1）你的胆固醇含量极高；（2）条形图合理地描绘了胆固醇的真实下降量。如果你的胆固醇水平为 400 点，那么，你可能会认为该水平将在 4 周内下降 100 点——只要在食物中增加燕麦就行了。

让我们揭开 *Y* 轴，看一看条形图是不是合理地描绘了燕麦的降胆固醇效果。



根据标度可以看出，真实情况是胆固醇水平只下降了 4%——并非图形