用正确的方法让你的数据有意义

no starch press

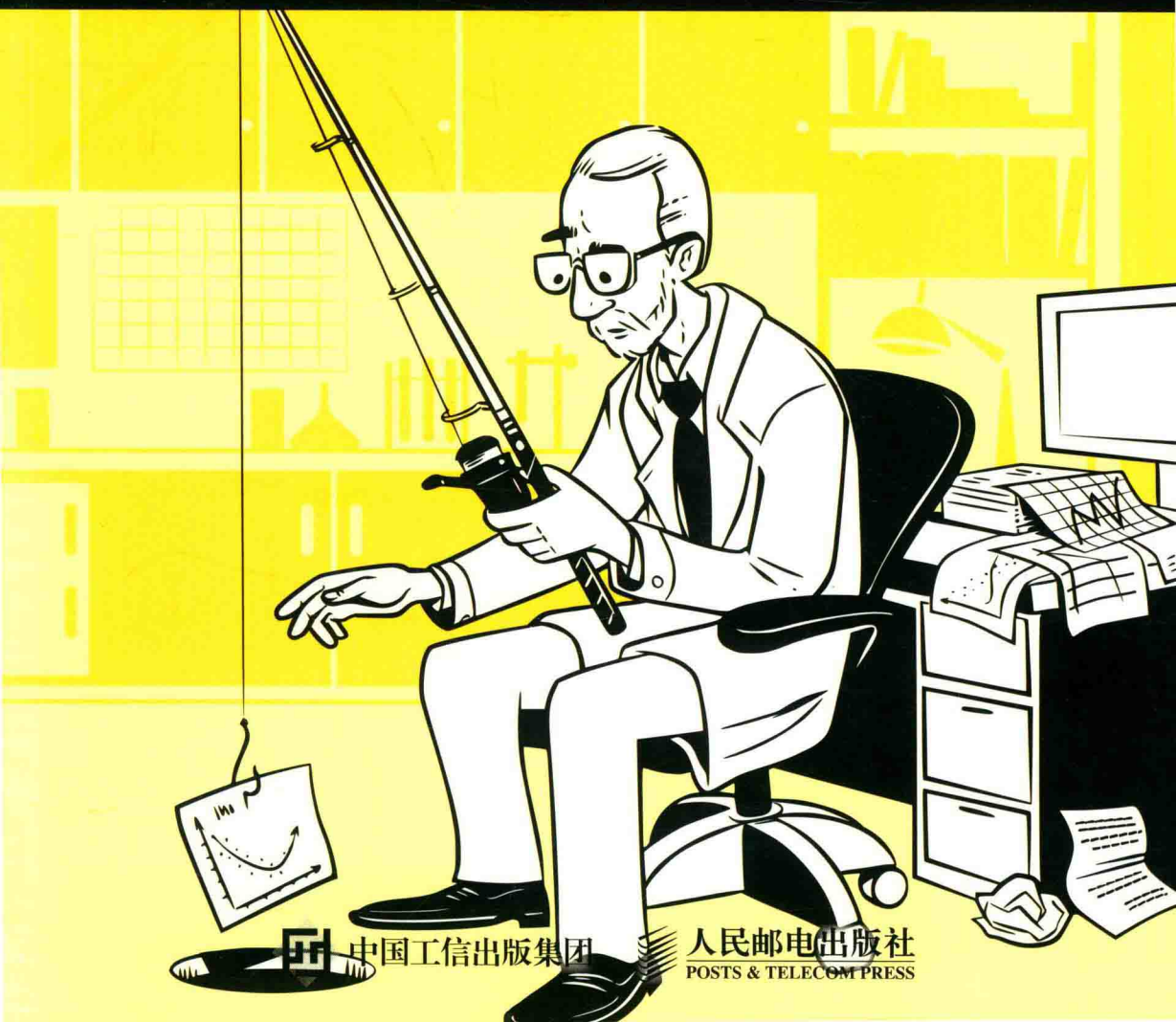
统计会犯错

如何避免数据分析中的统计陷阱

STATISTICS DONE WRONG

THE WOEFULLY COMPLETE GUIDE

[美] Alex Reinhart 著　　刘乐平 译



中国工信出版集团　人民邮电出版社
POSTS & TELECOM PRESS

统计会犯错

如何避免数据分析中的统计陷阱

STATISTICS DONE WRONG

THE WOEFULLY COMPLETE GUIDE

[美] Alex Reinhart 著　　刘乐平 译

人 民 邮 电 出 版 社

北 京

图书在版编目（CIP）数据

统计会犯错 ：如何避免数据分析中的统计陷阱 /
(美) 亚历克斯·莱因哈特 (Alex Reinhart) 著 ; 刘乐
平译. -- 北京 : 人民邮电出版社, 2016.9
ISBN 978-7-115-43374-9

Ⅰ. ①统… Ⅱ. ①亚… ②刘… Ⅲ. ①统计分析－手
册 Ⅳ. ①C813-62

中国版本图书馆CIP数据核字(2016)第202042号

版权声明

◆ 著　　[美] Alex Reinhart
译　　刘乐平
责任编辑　王峰松
责任印制　焦志炜

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
大厂聚鑫印刷有限责任公司印刷

◆ 开本：720×960　1/16
印张：13.25
字数：155 千字　　2016 年 9 月第 1 版
印数：1－3 000 册　　2016 年 9 月河北第 1 次印刷

著作权合同登记号　图字：01-2015-4186 号

定价：39.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315

内 容 提 要

面对充满不确定性的未知世界，人们在科学研究中需要大量使用统计分析方法。但是，如何正确使用统计分析方法充满玄机，即使对那些最优秀和最聪明的人也是如此。读完此书你会惊讶地发现，许多科学家使用的统计方法中其实隐藏着许多谬误和陷阱。

《统计会犯错》这本书简明扼要地指出了现代科学研究中常见的统计谬误，诸如 p 值与基础概率谬误、统计显著性和模型误用等。从这本书中，你将理解什么是统计谬误及其产生的原因，了解如何检查科学研究中隐藏的统计谬误，你还将学会如何正确地使用统计方法，如何在科学研究中避免这些统计谬误。

对本书的赞誉

“一本值得珍藏的小书……令人惊奇，统计门外汉入门必读。”

——阿尔伯托·凯若（Alberto Cairo），迈阿密大学计算科学中心可视化项目主任

“如果你正在分析数据，发现了一些规律，但不知道是否正确，请参考这本书。”

——邱南森（Nathan Yau），加利福尼亚大学洛杉矶分校（UCLA）统计学博士，数据可视化网站 *Flowing Data* 创始人

“一个令人愉快的和翔实的指南……全面而清晰。”

——约翰·沃森（John A. Wass），科学计算（*Scientific Computing*）网站

“我一定会把这本书推荐给那些对医学统计有兴趣的人和不喜欢统计学的医生。”

——卡缇·本斯（Dr. Catey Bunce）博士，穆尔菲尔兹（Moorfields）眼科医院首席统计学家

“我很喜欢这本书，并计划与我的许多学生分享。”

——妮科尔·拉齐维尔（Dr.Nicole Radziwill）博士，詹姆斯·麦迪逊大学（James Madison University）副教授

“我希望每个医生都能读到这本书。”

——埃里克·拉莫特（Dr.Eric LaMotte）博士，华盛顿大学医学院

“一本大胆的书，一本迷人的书……真正令人愉悦，并将永远改变你对统计的看法。”

——本·罗斯韦尔（Ben Rothke），信息安全专家

“一个写得很好的、有趣的、有用的指南，包含了今日统计实践中最常见的问题。”

——民间统计学家（*Civil Statistician*）网站

“任何研究人员都应该把这本书当作一个有价值的指南，来验证研究结论的正确性。”

——桑德拉·亨利-斯托克（Sandra Henry-Stocker），信息技术专家

“任何数据科学图书馆都应必备的重要读物。此外，简练的写作风格会让你的兴趣大增，而且可以成为你未来项目的创意源泉，极力推荐。”

——洞察大数据（insideBIGDATA）网站

关 于 作 者

亚历克斯·莱因哈特（Alex Reinhart），卡耐基梅隆大学（Carnegie Mellon University）统计学教师和博士生。他从德克萨斯大学奥斯汀分校（University of Texas at Austin）获得物理系学士学位，并应用物理学和统计学研发定位放射性设备。

关 于 译 者*

刘乐平，中国人民大学统计学系博士毕业，现为天津财经大学统计学、金融学教授，博士生导师，大数据统计研究中心主任。

---

* 本书的翻译由天津财经大学刘乐平和研究生高磊、毕莎莎、董婵、申亚飞共同合作完成。

“首要原则是你不能欺骗自己，但一叶障目，自欺欺人却又屡见不鲜。”

——费曼（Richard P. Feynman）

“当你要求统计学家对一个已完成的统计实验做事后重复检验时，他们的回答常常是：‘抱歉，试验已无法重复了’。”

——费希尔（R.A.Fisher）

自　序

几年前，我是德克萨斯大学奥斯汀分校的一名物理专业的大学生。在一门研讨课上，每个学生都要选择一个主题做25分钟的陈述演讲。

我告诉布兰特·艾弗森（Brent Iverson）博士，我选了关于阴谋论的主题，但他不满意这个选题，他说这太宽泛了，一个引人入胜的演讲需要重点和细节。我琢磨着放在我面前的主题建议列表。他问："科学欺诈和滥用这个主题如何？"我接受了他的建议。

我不明白与阴谋论相比，科学的欺诈和滥用为什么是一个较窄

的主题。但没有关系，经过几次粗略的研究，我感到对于科学欺诈的兴趣至少自己还能接受，与科学家承担的所有责任相比，这大多不是他们有意而为的。

我没有资格讨论统计，虽然如此，我还是挖掘出好几十篇报告科学家经常犯的大量的统计错误的研究文章。通过对这些文章的阅读和概括，我设计了一个令艾弗森博士满意的演讲报告。我决定未来当一个科学家（目前自认为是统计爱好者），我应该选学些统计课程。

两年时间里，在学习了两门统计课程后，我考入卡内基梅隆大学，成为统计学研究生。不过，我仍然着迷研究“因统计而错”的科学方法。

统计产生的错误可能导致更严重的结果，因为他们常常冠以科学的名义，而一些科学家接受的是非正规的统计教育。本书不是正式的统计教科书。一些读者通常会跳过第一章，但我建议至少浏览一下，以熟悉我的注解风格。

我的目标不仅仅是教你常见的统计错误名称和提供笑料。我将尽可能不用详细的数学推导，解释为什么统计谬误是陷阱，还将告诉你这些陷阱是如此的无所不在。有些深度的问题会导致阅读困难，但我认为这个深度是有价值的，对于科学领域的每一个人来说，都必须加深对统计方法的基本理解。

对于那些日常工作是做统计分析的人，大多数章节结尾的“提示”是解释你可能使用的统计技术，以避免通常易犯的错误。但这不是教科书，所以我不会教你如何使用这些技术的任何细节。

我只是希望让你意识到最常见的问题，这样你就可以选择最恰当的统计技术。

如果我激起了你对这个话题的好奇心，这里包括了一个广泛的、综合的文献书目，和每一个统计谬误的引用参考。在这本指南中我省略了大量的数学内容，取而代之的是容易理解的概念，但如果你喜欢一个更缜密的推导过程，我鼓励你去读所对应的原文。

在你读这本书之前，我必须提醒你。每当我们想了解几乎没有人做的事情的时候，这就吸引着我们找到每一个证明它的机会。统计犯的错误可能奇迹般地成为纽约时报最好的卖点，我希望看到 Paul Graham 所说的“中间立场”来回应大众媒体上的任何科学新闻，与其花时间了解科学新闻中的奇谈怪论，还不如去反驳脱离实际的统计学家所做的研究，对他们的统计设计提出批评*。

这已经发生在大多数讨论科学新闻的网站上，它会无休止地干扰我，查阅这本书用来证明它。这类新闻占第一位的评论总是类似于“他们没有控制这个变量”和“样本量太小”等的议论，而且 10 个当中至少有 9 个批评家从不读科学论文去关注他们的不满。

这是不明智的。对统计的一点点了解并不是一个拒绝所有现代科学的理由。一篇研究论文的统计方法，可以通过研究设计、测量技术、成本的限制和目标之外的上下文细节加以判别。运用你的统计知识，以更好地了解研究的优点、局限性和潜在的偏误，

---

* 顺便说一句，我认为这就是人们如此热衷“阴谋论”的原因。如果你认为你了解一些其他人不知道的事件（如政府没有公开的事实真相！），你就会抓住每一次机会去炫耀你所了解的内容，然后对所有新闻，你都能找出政府弄虚作假的原因。请不要这样对待“统计错误”。

而不是去攻击一些似乎误用了一个 p 值或与你的个人信仰相矛盾的论文。此外，请记住，由不合适的统计方法得出的结论是可以被纠正的，统计学和逻辑学的错误不会使结论错误，而仅仅是不支持结论。

简而言之，请负责任地实践统计学。我希望你和我一起去质疑，从而更加完善我们所依赖的科学。

致谢

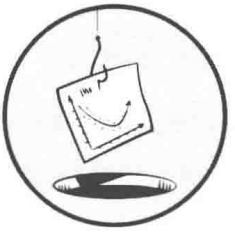
感谢 James Scott，他的统计课程开始了我的统计生涯，并为本书的写作提供了必要的统计知识背景；感谢 Raye Allen，他给 James 留的家庭作业非常有趣；感谢 Matthew Watson 和 Moriel Schottlender，他们给本书的初稿提出了有价值的反馈和建议；感谢我的父母，他们给出了反馈和意见；感谢 Dr. Brent Iverson，是他的讨论课激发了我研究统计谬误的兴趣；感谢所有的科学家和统计学家，他们不经意的错误是我写这本书的理由。

我在卡内基梅隆的朋友给了我许多好的创意，回答了我的许多问题，他们耐心地听我解释每一个新的统计谬误。Jing Lei、Valérie Ventura 和 Howard Seltman 教授给了我必要的知识。作为技术复审员，Howard 发现了几个令我尴尬的错误；如果还存在错误，它们都是我的责任，尽管我宣称它们只会出现在本书的书名里。

No Starch 的编辑为本书的初稿费了很多心血。Greg Poulos 仔细阅读了前几章，直到理解了每一个概念，他才满意。Leslie Shen 对本书最后几章进行了润色，整个团队的效率令人惊讶。

我还要感谢这本书在网络上分享时，那些给我发邮件提出建议的朋友。不分次序，感谢 Axel Boldt、Eric Franzosa、Robert O'Shea、Uri Bram、Dean Rowan、Jesse Weinstein、Peter Hozák、Chris Thorp、David Lovell、Harvey Chapman、Nathaniel Graham、Shaun Gallagher、Sara Alspaugh、Jordan Marsh、Nathan Gouwens、Arjen Noordzij、Kevin Pinto、Elizabeth Page-Gould 和 David Merfield。没有他们的评论，我的解释不可能变得如此周全。也许你也会加入这个名单。虽然我已尽力，但本书不可避免会包含一些错误或遗漏。如果你发现本书的错误，或对本书有任何疑问，或觉得我遗漏了一些重要问题，请发送邮件至 alex@refsmmat.com。本书的勘误表和更新敬请关注 http://www.statisticsdonewrong.com/。

前　言



在那本非常著名的统计读物《统计数字会撒谎》(*How to lie with statistics*) 的最后一章中，作者哈弗 (Darrell Huff) 告诉我们“任何带有医学味道的言论”或者“由科学实验室和大学发布的信息”都是值得我们相信的，虽然不是毫无条件地相信，但是肯定比“媒体”或者“政府”公布的事实可靠的多。哈弗的整本书中充满了媒体和政府利用误导性的统计信息弄虚作假的例子，但很少涉及经过专业学习的科学家所做的统计分析也可能产生误导。科学家应该追求的是对事物本质的理解，而非对付政治对手的子弹。

统计数据分析是科学的基础。随便翻开一本你喜欢的医学杂志，你就会被统计术语淹没：t 检验、p 值、比例风险模型、风险比率、逻辑回归、最小二乘拟合以及置信区间。统计学家为科学家们在复杂的数据集中发现知识和规律提供了强有力的工具，科学家们毫不怀疑欣然地接受了这些工具。

但是，不少科学家并没有接受过统计教育，在科学领域中许多本科课程中也不涉及任何统计训练。

自 20 世纪 80 年代以来，学者已经揭示了无数的统计谬论，以及出现在经过同行评议的科学文献中的错误，他们发现许多科学论文，大概有一半以上，都犯过这些错误。由于统计能力不足，使得许多研究无法找到他们想要找的东西；多重比较和对 p 值误读导致了许多错误的“正确结论”；灵活的数据分析使得我们很容易找到原本不存在的相关性；不恰当的模型选择可能会使结论产生偏倚。这些错误都被同行评议人员和期刊编辑们忽视了，造成这一结果是由于他们通常并没有经过专业的统计训练，而且很少有杂志会聘请统计人员来审核投送的文章，另外，大部分文章也没有给出充足的、能够被精确评估的统计细节。

这些问题并不涉及恶意欺骗，而是由统计教育不足而造成的——一些科学家甚至指出大多数发表的研究成果可能是错误的 [1,*]。在顶级期刊中经常会出现一些要求对将要发表文章采用更高统计标准、更严格审查标准的评论文章和社论，但是只有很少

---

* 本书正文中加注的数字上标 1、2……表示参考文献的编号，具体文献信息请查阅书末的参考文献。

的科学家们响应这一呼吁，而且杂志授权标准往往被忽视。由于这些建议通常散落在一些误导性的教科书和杂志的综述中，而且对于应用型科学家们来说统计研究文章很难理解，所以大多数科学家想要提高他们的统计知识并不是那么容易的。

现代研究中复杂的方法论意味着没有经过广泛统计训练的科学家也许不能完全领会他们研究领域内发表的一些文章。例如，在医学领域中接受过标准统计入门课的医生，其所具备的统计知识只能充分理解在《新英格兰医学杂志》上刊登的 20%的学术论文[2]。大多数的医生甚至都不具备这些知识，很多医学人员并不是通过统计的必修课而是利用杂志社或者短期课程等方法非正式地学习统计[3]。我们对这些医学人员进行“医疗中常用的统计方法”测验，结果仅有不足 50%的人能够答对[4]，这证明这些非正式的方法所包含的内容并不足以让医学人员真正学会统计知识。即使是经过研究训练的医学院的教员其得分也小于 75%的正确率。

情况如此糟糕，即使是从事上述统计知识调查的作者也缺乏构建调查问券所需的统计知识——我刚才引述的数字是有误导性的，因为在上述对医疗人员进行的调查中包括一道定义 p 值的选择题，但是在这道题中却给出 4 个不正确的定义作为选项[5]。我们可以为这个作者找些借口，因为即使很多统计入门的课本中也没能正确地定义 p 值这一基本的统计概念。

当科学研究的设计者不注重对统计人员的雇佣时，他们可能会迷失在工作中，在不会得到答案的研究上花费数千美元。正如心理学家 Paul Meehl 所抱怨的那样。

> 我们野心勃勃的研究员——在逻辑科学的知识体系下的毫无畏惧并且满心喜悦的依赖于“精确”的现代统计假设检验，已经著作等身或被提升为教授。就他对心理学整体来说，他几乎什么贡献也没做——更直白地说，他是一个对多个领域均有所涉猎，却没有得出什么真正科学成果的多产科学家[6]。

对大多数的科学家来说，由于很多科学领域对 p 值的误解而指控他们不能孕育知识也许是不公平的。但是这些错误确实对现实世界有很大影响。医学临床试验指导我们的卫生保健方向，并且决定某些新强力处方药的安全性；犯罪学家评估不同的策略来减少犯罪和骚乱；流行病学家试图延缓新疾病的蔓延；营销人员和业务经理们试图找到销售产品的最好方式。这一切都归结到统计，但是统计知识却不能被正确使用。

任何人都曾抱怨过医生没有在你能够理解的范围内告诉你什么是好的或者什么是不好的。现在，我们对一些声称某些食物、饮食或运动可能会损害我们健康的新闻不屑一顾，因为几个月后的另一项研究可能会得到完全相反的结果。正如一位杰出的流行病学专家所说的那样：“我们正在变成社会所讨厌的那类人，人们不再重视我们，而一旦人们把我们当回事，我们可能会无意中做出弊大于利的事[7]。”我们的直觉是正确的：在一些科学领域，最初的结论可能与之后的相悖。过早发布令人兴奋的结论，往往比发布有充分证据支持且仔细核对过的结论有更大的压力。

尽管如此，我们不要过早地下结论。一些统计误差可能只是由于资金不足造成的。让我们看看 20 世纪 70 年代中期在美国发起