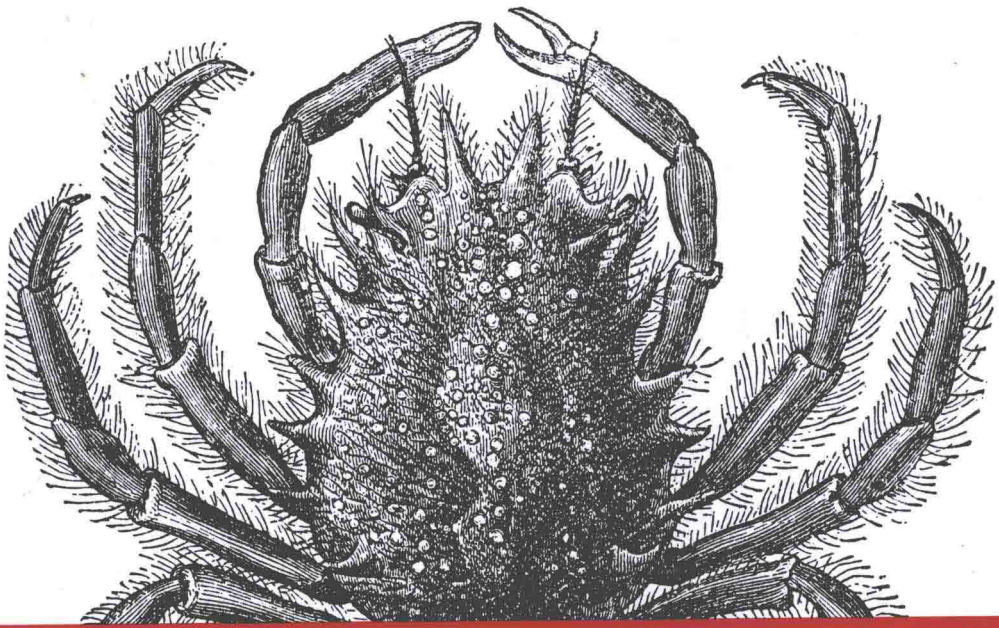
原书第2版



统计学

及其应用

Statis hell

O'REILLY®
机械工业出版社
China Machine Press

Sarah Boslaugh 著
孙怡帆 等译

原书第2版

统计学及其应用

Sarah Boslaugh 著

孙怡帆 等译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc.授权机械工业出版社出版

机械工业出版社

图书在版编目（CIP）数据

统计学及其应用（原书第2版）/（美）博斯劳（Boslaugh, S.）著；孙怡帆等译.
－北京：机械工业出版社，2016.6
书名原文：Statistics in a Nutshell, Second Edition
ISBN 978-7-111-53388-7

I. 统… II. ①博… ②孙… III. 统计学－高等学校－教材 IV. C8
中国版本图书馆CIP数据核字（2016）第066181号

北京市版权局著作权合同登记
图字：01-2013-3070号

书　　名/　统计学及其应用（原书第2版）
书　　号/　ISBN 978-7-111-53388-7
责任编辑/　和静
封面设计/　Randy Comer，张健
出版发行/　机械工业出版社
地　　址/　北京市西城区百万庄大街22号（邮政编码 100037）
印　　刷/　北京诚信伟业印刷有限公司
开　　本/　178毫米×233毫米　16开本　33.75印张
版　　次/　2016年6月第1版　2016年6月第1次印刷
定　　价/　119.00元（册）

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：(010)88379426；88361066
购书热线：(010)68326294；88379649；68995259
投稿热线：(010)88379604
读者信箱：hzit@hzbook.com

O'Reilly Media, Inc.介绍

O'Reilly Media通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自1978年开始，O'Reilly一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了Make杂志，从而成为DIY革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项O'Reilly的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar博客有口皆碑。”

——Wired

“O'Reilly凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference是聚集关键思想领袖的绝对典范。”

——CRN

“一本O'Reilly的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照Yogi Berra的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去Tim似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

译者序

很高兴有机会可以翻译 Sarah Boslaugh的这本著作。说起统计这个词，绝大多数人都不会觉得陌生，但如果要描绘一下统计在你心中的样子，恐怕就有多种多样的答案了。许多人对统计的第一直觉就是常在新闻联播和门户网站中出现的各种数据，例如GDP、房价、PM2.5。一些接触过统计，甚至曾上过统计课程的人则会对统计中各种分析数据的方法印象深刻，在他们看来，统计更像一个工具箱。而对统计有过较为深入思考或者从事统计相关领域的专业人士则倾向于视统计为使用数学方法描述、分析数据并以此做出决策的研究领域，本书作者就是其中一员。

正是基于这一观点，本书将重点放在使用尤其是理解统计上，而非其中涉及的繁琐数学推导。事实上，正如本书作者在前言中所讲，“本书更希望告诉读者的是统计思维而非如何做统计”。由此，对于那些希望更多了解统计本身而非方法细节的人，本书是一个很好的选择。本书另一大特点是加入若干章来介绍统计在商业、医学、教育等几个具体领域中的应用，这将非常有利于统计初学者更为全面地了解统计。此外，本书作者基于个人从事统计教学和研究的多年经验为初学者如何用统计与他人沟通，以及如何正确评价他人提供的统计信息给出了很多中肯的建议。这些内容对从事统计的教学工作者具有极强的指导和借鉴意义。

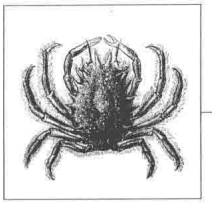
感谢本书的所有译者，包括孙怡帆（第1、3、4、16章以及全书审稿和校对工作）、颜娅婷（第2、5~9、17~20章）和扈瑞鹏（第10~15章以及所有附录）。

由于译者水平有限，书中难免不妥之处，请读者不吝赐教。

译者

2015年11月

目录



前言

从反响来看，本书第1版是非常成功的，但所有的书籍都有提升空间，所以我感谢能有机会对本书第1版进行修改。我写这本书的初衷依然没有改变：这是一本写给那些对统计感兴趣并且想要理解统计的人的书，而不是一本告诉你如何使用统计软件或者钻研统计公式背后的数学理论的书。这本书和O'Reilly轻松入门系列的其他书有些不一样——从某种程度上来说，本书既不是统计学手册，也不是统计学入门教材，而是介于两者之间。

尽管统计学不断地渗透到其他领域，但有一点始终没变：那就是在晚会上告诉别人我是一个统计学家仍然成事不足败事有余。因此，这似乎告诉我人们有多讨厌大学里必修的统计学，这也促使他们甚至引用马克·吐温的一个老掉牙的笑话“世界上有三种谎言：谎言、该死的谎言和统计”。

就个人而言，我发现统计是迷人的，我喜欢在这个领域工作。我也喜欢教授统计，而且还相信我可以热情地跟别人交流统计。但这通常是一场艰苦的战斗；因为许多人似乎相信统计就是一些扭曲事实，误导他人的技巧或者手法。但也有人持相反的意见，他们相信统计是可以帮助他们思考问题的一系列神奇方法。

那么，到底什么是统计？

在你深入学习和使用统计的技术性细节之前，先暂停一会儿，思考“统计”这个词所有可能的含义。不要担心你无法立刻理解“统计”的所有含义，在阅读这本书的时候，你就会清楚它们的意思。

当人们说到统计的时候，他们通常指的是以下几点：

1. 数值数据，比如失业率、每年因蜜蜂叮咬而死亡的人数，或者与1906年相比，纽约2006年的人口数。
2. 描述样本数据的数字，这是相对参数（用来描述总体的数字）来说的。例如，广告公司可能比较关注《体育画报》订阅人群的平均年龄。为了回答这个问题，可以从订阅者中抽取一个随机样本，计算样本均值（这是一个统计量），然后用这个均值来估计订阅者总体的年龄均值（这是一个总体参数）。
3. 分析数据时所使用的一些特定方法以及这些方法的结果，比如t统计量或者卡方统计量。
4. 开发和使用数学方法来描述数据并且据此作出决策的研究领域。

第1条定义的统计不是这本书主要的关注点。如果你只是简单地想找到关于失业、健康或者任何由政府或其他组织定期发布的关于其他主题的统计数据，你最好去咨询图书管理员或者该领域的专家。但如果你想知道如何解释这些数字（例如，理解为什么用均值来表述平均值通常会产生误导，或者理解粗死亡率和标准化死亡率之间的差异），那么本书肯定可以帮到你。

第2种定义所提及的概念会在第3章中进行讨论，这些概念涉及推断统计，但这些概念也会贯穿整本书。虽然从某种程度上说，这是一个术语的问题（*统计量*是用来描述*样本*的数，而*参数*是用来描述总体的数），但却强调了统计实践的一个基本思想。推断统计的基本思想就是利用通过研究样本得来的信息对总体做出推断，推断统计是本书主要的关注点（正如它是大部分统计书籍的关注点一样）。

第3种定义也是本书大部分章节的基础。学习统计的过程在某种程度上来说就是学习某些特定统计方法的过程，这些统计方法包括如何计算和解释这些统计量、在特定情况下如何选择一个合适的统计量等。事实上，大部分刚开始学习统计的学生都会赞同这个定义；对他们来说，学习统计意味着学会如何运用一系列统计方法。这不是学习统计的有效方法，因为它是不完整的；学会运用统计方法是统计实践所必需的一部分，但这并不是统计实践的全部。此外，由于计算机软件使得人们更加容易进行统计分析而不管人们有没有数学背景，因此理解和解释统计的需要已经远远超过了学习如何计算的需要。

第4种定义和我所想的最为接近，因为我选择统计作为我的专业。如果你是一个中学生或者大专生，那么你可能知道统计的这种定义，因为现在许多大学和大专院校要么有一个独立的统计学院，要么在数学学院下设有专门的统计学专业。讲授统计学的高中学校也越来越多，并且在美国，把统计学当做大学先修课程的院系也越来越多。

统计学不仅仅是大学水平的一个专业学科。很多大学院系要求学生不仅要修完本专业课程，还要选修一门或者多门统计学课程。此外，我们还应该知道现代统计中许多重要

的技术都是由在其他领域工作，并且把学习和使用统计作为他们工作一部分的人所开发的。Stephen Raudenbush是分层线性模型的先驱，他在哈佛大学研究政治分析和评价，Edward Tufte是统计图形方面的世界专家，他刚开始的职业是一名政治科学家，他在耶鲁大学写了关于美国民权运动的博士论文。

因为统计在许多领域都有应用，并且各个层次（从管理者到一线工作人员）对统计学的应用越来越多，对于那些离开学校多年的人来说，获得基本的统计学知识已经成为必须要做的事情。然而，这些人接触到的常常是美国大学本科课程的教材，而这些教材都太专业，太专注于计算并且太过昂贵了。

最后，不能把统计完全留给统计学家，因为统计需要参与到现代公民的生活中，尤其是在理解你所读到的报纸、看到的电视和听到的无线广播的时候。统计知识是避免误导性言论扩散或者虚假言论扩散（不管这些虚假言论是政治家、广告商还是社会改革者发布的）的最好方法，而这些言论在我们日常新闻中所占的比例似乎越来越高。这就是Darryl Huff 1954年的经典著作《统计数字会撒谎》为何仍然在发行的原因：统计很容易被滥用，大约几十年来，常见的统计方法一直存在失真的现象，面对那些利用统计数字来撒谎的人，最好的方法就是武装自己，这样你就能够发现这些谎言，并且阻止说谎者散布谎言。

本书的重点

市面上有许多统计学书籍，你可能很想知道为什么我再写一本统计学书籍是有必要的。最主要的原因是我还没有发现哪本书能够满足本书所提出的需求。事实上，如果允许我此刻充满诗意，那么目前的状况就像是在改写柯勒律治的诗歌《古舟子咏》中古代水手的困境，“书，书，到处都是书，却不知道该学哪本”。在这本书中，我想要强调以下几个问题：

- 需要一本专注于研究或应用背景下使用和理解统计的书，不是一堆分散的数学方法，而是对数字进行推理过程的一部分。
- 需要把测量以及数据处理这些问题的讨论纳入到统计学的介绍中。
- 需要一本不局限于某个学科领域的统计学书籍。在不同的学科领域，大部分基本的统计方法是一样的（不管数据是来自医学领域、金融领域还是刑事审判领域，*t*检验都是一样的），因此没有必要出版一大堆大同小异的书籍。
- 需要一本介绍性的统计学书籍，内容紧凑、价格便宜而且易于初学者理解，初学者既不会觉得高深莫测，也不会觉得过于简单。

因此哪些人会是本书的潜在读者呢？我重点强调以下三类人：

- 在高中、专科院校或者大学本科选修初级统计学课程的学生。
- 因为目前工作需要或者升职需要而学习统计的成年人。
- 出于求知欲而想学习统计的人。

尽管本书介绍了许多统计方法，但本书重点不是具体的统计方法，而是统计推理。你可能会说相比于做统计，这本书更关注的是如何按照统计的思维思考。这是什么意思呢？在用数据进行思考的过程中必须要做到几件事情。尤其是，我强调对数据进行思考，在这个过程中辅以统计方法。大部分章节都有一些实际应用的例子，但这些例子旨在提供一个回顾该章内容以及思考该章重要概念的机会，其目的并不是盲目地计算。

第2版对第1版中的所有例子进行了修订，大部分章节还补充了新的例子和练习。尤其增加了一些处理比例数据的例子，还增加了更多实际数据的例子，这些实际数据来源于联合国人类发展项目和行为风险因素监测系统等；所有数据集都可以免费从网上下载，因此学生可以用这些数据进行分析，也可以再现本书的分析。本版还增加了一章新内容，即第19章。增加这一章的原因是我观察到，尤其是对于出于职业目的而学习统计的人来说，交流统计信息的能力至少和进行统计计算的能力一样重要。还增加了一些新的附录，主要是为了使这本书更加完备和友好。这些附录包括常见分布的概率表、在线资源的列表、术语表和统计学符号表。

信息时代的统计

我们生活在一个信息时代——这样的说法变得越来越时髦，在信息时代，太多的事实在不断地被收集和散布，导致没有人可以与时俱进。虽然这是陈词滥调，但却是以事实为基础；作为一个社会整体，我们总是被淹没在数据中，并且问题似乎越来越多。这种情况对我们来说既有正面影响，又有负面影响。从正面角度来看，计算机技术、电子数据存储技术的发展和传播使得获取信息越来越容易，因此研究人员并不需要去图书馆或者复印这些资料。

然而，数据本身并没有任何意义。在数据变得有意义之前，需要人来组织、解释这些数据，因此，置身信息时代要求人们要熟练地理解数据，包括数据的收集方式、分析方式和解释方式。并且由于需要支持不同的结论，同样的数据常常用许多不同的方式来解释，所以即使是不从事统计工作的人，也需要理解统计是如何起作用的，以及如何识别基于数据误用的虚假结论。

本书的组织结构

本书分为四部分：介绍部分（第1~4章）是后续章节的重要基础；推断统计（第5~13章）；不同专业领域使用的特定统计方法（第14~16章）；统计工作的部分辅助性知识，即使它们看起来并不是严格的统计（第17~20章）。

以下是各章节的内容简介：

第1章，测量的基本概念

讨论统计学的基本问题，包括测量水平、操作化、代理测量、随机误差和系统误差、信度和效度以及偏倚的种类。

第2章，概率

介绍概率的基础知识，包括试验、事件、独立性、互斥性、加法定律和乘法定律、排列和组合、条件概率和贝叶斯定理。

第3章，推断统计

介绍推断统计的基本概念，包括概率分布、自变量和因变量、总体和样本、常用抽样方法、中心极限定理、假设检验、第一类错误和第二类错误、置信区间和p值以及数据变换。

第4章，描述统计和统计图

介绍集中趋势和离散程度的常用测量方法，包括均值、中位数、众数、极差、四分位距、方差和标准差，并讨论了离群点。这一章还介绍了一些常用于展示统计数据的图示法，包括频数表、条形图、饼图、帕累托图、茎叶图、箱型图、直方图、散点图和折线图。

第5章，分类数据分析

复习分类数据和定距数据的概念，介绍了$R \times C$列联表。该章涉及的统计方法包括卡方独立性检验、卡方等比例检验、卡方拟合优度检验、费希尔精确检验、McNemar检验、大样本比例检验以及分类数据与顺序数据关联性度量。

第6章，t检验

讨论t分布以及单样本t检验、两个独立样本t检验、重复观测样本t检验以及异方差t检验的理论和应用。

第7章，Pearson相关系数

介绍相关性的概念，并介绍展示两个变量不同强度相关性的图形技术，讨论皮尔逊相关系数和判定系数。

第8章，回归分析和方差分析导论

介绍与线性回归和单因素方差分析相关的广义线性模型的概念，并讨论了使用这种

方法的前提假设。该章介绍了简单回归（两个变量）、单因素方差分析和事后检验等方法。

第9章，多因素方差分析和协方差分析

讨论更加复杂的方差分析设计，包括双因素方差分析、三因素方差分析和协方差分析，还讨论交互项。

第10章，多元线性回归

将线性模型扩展到多元情况，模型中包含多个预测因子。该章讨论的内容包括预测变量间的关系、标准化和非标准化系数、哑变量、建模方法和违背线性模型假设的情况（包括非线性、自相关和异方差）。

第11章，Logistic回归、多项Logistic回归和多项式回归

把回归模型扩展到二元输出变量（Logistic回归）、分类输出变量（多项回归）以及非线性模型（多项式回归）上，并且讨论模型的过拟合问题。

第12章，因子分析、聚类分析和判别函数分析

讨论三种高级统计方法：因子分析、聚类分析和判别函数分析，并且讨论不同方法适用的问题类型。

第13章，非参数统计

讨论何时用非参数统计而不用参数统计，以及非参数统计的组间设计和组内设计，包括Wilcoxon秩和、Mann-Whitney U检验、符号检验、中位数检验、Kruskal-Wallis H检验、 Wilcoxon符号秩检验和Friedman检验。

第14章，商业和质量改进统计

介绍商业和质量提高领域常用的统计方法。该章包含的分析方法和统计方法包括指数，时间序列，大中取小、大中取大和小中取大的决策方法，风险下的决策方法，决策树和控制图。

第15章，医学和流行病学统计

介绍跟医学和流行病学相关的基本概念和统计方法。该章涉及的概念和统计方法包括比、比例和比率的定义和应用，患病率和发病率的度量，粗比率和标准比率，直接标准化和间接标准化，风险的度量，混淆，简单优势比和Mantel-Haenszel优势比，精度、势和样本量的计算。

第16章，教育和心理统计

介绍教育和心理学领域常用的概念和统计方法。该章的内容包括百分比，标准化得分，检验的构造方法，经典测量理论，混合测量的信度，内部一致性的度量（包括alpha系数）以及题目分析方法。该章还简要介绍了题目反应理论。