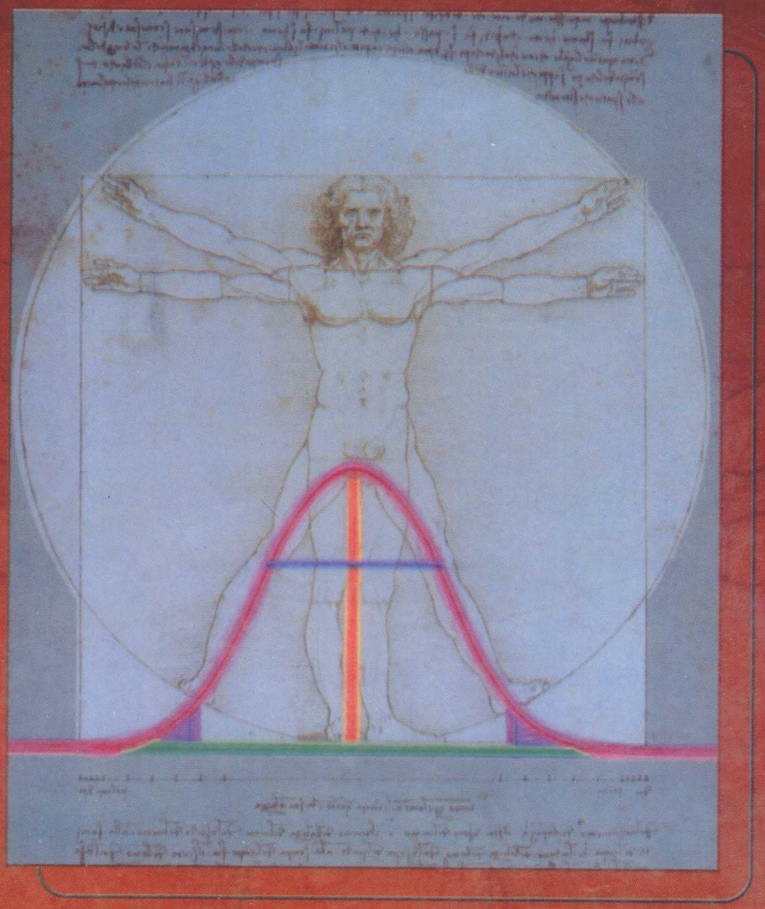

生物统计学基础

Biostatistics: The Bare Essentials

第 3 版

原　著　Geoffrey R. Norman
David L. Streiner
主　译　凌　莉

人民卫生出版社

生物統計学基礎

Biostatistics: The Bare Essentials

[redacted]

[redacted]

生物统计学基础

BIOSTATISTICS: The Bare Essentials

第 3 版

原　著　Geoffrey R. Norman
David L. Streiner

主　译　凌　莉

副主译　王　彤　于　浩　宇传华

译　者（以姓氏笔画为序）

于　浩（南京医科大学）
于晓洁（山西医科大学）
马昭君（南京医科大学）
王　彤（山西医科大学）
文　芸（武汉大学）
方　龙（武汉大学）
闫丽娜（山西医科大学）
宇传华（武汉大学）
苏　萌（中山大学）
何　书（南京医科大学）
沃红梅（南京医科大学）
张汝阳（南京医科大学）
陈　君（中山大学）
陈　昂（中山大学）
陈　雯（中山大学）
陈建平（南京医科大学）
林嵩艺（武汉大学）
易洪刚（南京医科大学）
周敏林（南京医科大学）
赵　杨（南京医科大学）
祝慧萍（武汉大学）
索瑞鑫（山西医科大学）
凌　莉（中山大学）
彭　荣（中山大学）
葛爱华（南京医科大学）
戴启刚（南京医科大学）
魏永越（南京医科大学）

人民卫生出版社

Translation of **Biostatistics: The Bare Essentials** 3e by Geoffrey R. Norman, David L. Streiner

The original English language work has been published by BC Decker Inc. Hamilton, Ontario, Canada

Now distributed and published by
People's Medical Publishing House-USA, Ltd.
2 Enterprise Drive, Suite 509
Shelton, CT 06484, USA
Tel: (203) 402-0646
E-mail: info@pmph-usa.com

Translated into Chinese by People's Medical Publishing House

Beijing, China

图书在版编目（CIP）数据

生物统计学基础/凌莉主译. —北京：人民卫生出版社，2010.1
ISBN 978-7-117-12359-4

Ⅰ.生… Ⅱ.凌… Ⅲ.生物统计 Ⅳ.Q-332

中国版本图书馆 CIP 数据核字（2009）第 200703 号

生物统计学基础

主　　译：凌　莉
出版发行：人民卫生出版社（中继线 010-67616688）
地　　址：北京市丰台区方庄芳群园 3 区 3 号楼
邮　　编：100078
E - mail：pmph @ pmph.com
购书热线：010-67605754　010-65264830
印　　刷：北京智力达印刷有限公司
经　　销：新华书店
开　　本：889×1194　1/16　**印张**：24　**插页**：4
字　　数：685 千字
版　　次：2010 年 1 月第 1 版　2010 年 1 月第 1 版第 1 次印刷
标准书号：ISBN 978-7-117-12359-4/R·12360
定　　价：70.00 元

译 者 序

对统计学的学习来说，最主要的是掌握统计思想，理解相关的统计原理，能够根据实际情况提出一个或几个解决问题的合适方法，并能够从中选出最优的解决方案。因此，一本理想的统计学教材，应该能够兼顾专业特点、深入浅出地阐述统计学基本原理和方法，同时在轻快风趣的讲述中激发读者的学习兴趣，培养统计思维，并辅之案例分析，对实际使用中容易发生的错误加以提醒，切实提高读者应用统计方法分析解决实际问题的能力。《生物统计学基础》(第3版)就是这样一本优秀的统计学教材。本书回避了纯数学式的描述，以十分贴近生活的例子，语言轻松活泼且通俗易懂，加之以“基于问题”的模式深入浅出地阐述了统计学的基本概念和原理，并在大多数章节的最后都编写了电脑操作的步骤，以帮你们进行SPSS(社会科学统计软件包)的操作，同时在书后还附上了相关的参考文献，帮助你进一步深入学习。本书适用人群广泛，不仅可以作为教师和研究生的参考教材，同时对临床工作者、管理人员、心理学家、社会工作者等在实际工作中应用统计学也有很大的帮助。

《生物统计学基础》(第1版)自1996年出版后，其轻松幽默的风格受到了广大读者的一致好评，销量达1万余册。很多原本“痛恨”统计学或对统计学一知半解的读者都认为这是他们读过的最好的统计书。随着统计学方法和计算机技术的不断发展，并在读者的强烈要求下，本书分别于2000年和2007年进行了再版，增加了许多新的统计学方法的介绍，以及SPSS的操作步骤。本书的两位作者都具有相当高的学术水平，同时教学经验丰富。其中，Geoff R. Norman是加拿大临床专长认知维度研究会的主席、McMaster大学教育研究与发展规划学院的副院长，同时还是临床流行病与生物统计学系的教授。David L. Streiner是多伦多Baycrest中心Kunin-Lunenfeld应用研究所的高级研究员、Toronto大学精神病学系教授和McMaster大学临床流行病与生物统计学系的荣誉教授。

在本书的翻译过程中，我们始终秉着尊重原著、服务读者的态度，力求保持原著的“原汁原味”，对于其中一些由于文化差异导致的可能会造成您理解困难和不习惯的地方，我们都以“译者注”的形式进行了注释，希望能够帮助您更好地理解本书的内容。

几经修改，本书终于能和广大读者见面。本书第1-6章以及自我测验部分由凌莉教授、苏萌、陈雯、陈君、陈昂、彭荣和李欣负责翻译；第7-12章由王彤教授、于晓洁、索瑞鑫、闫丽娜负责翻译；第13-20章由宇传华教授、祝慧萍、文芸、方龙、林嵩艺、夏欣以及董国营负责翻译；第21-29章由于浩教授、赵杨、马昭君、张汝阳、陈建平、戴启刚、魏永越、周敏林、葛爱华、何书、沃红梅以及易洪刚负责翻译。在此，要衷心感谢王彤、宇传华和于浩三位教授对本书翻译工作做出的巨大贡献和陈雯同学在校稿过程中付出的巨大努力，以及其他译者的辛勤劳动。由于译者水平有限，书中难免有疏漏和不当之处，恳请读者批评指正。

译 者

2009年11月

第 3 版前言

我们又回来了。在第 2 版的前言里，我们曾经提到过统计学知识的半衰期差不多和一头大象的寿命一样长。如今，统计学貌似已经发展到了一个“初步老龄化”的阶段，但事实是统计学发展的脚步并没有因此而停滞不前。7 年前旧版本上描述的方法在我们眼前一闪而过，如层次线性模型，现在几乎在所有的心理学杂志上都能找得到，并且开始被生物医学杂志上发表的文章所使用。另外，当我们第一次编写这本书时，我们想过它可能会被一些导师作为推荐阅读的资料，被人们当做一种用于帮助解释那些“真正的”、“成熟”的统计学课本的背景知识。我们收到了许多读者的电子邮件，数量之多让我们感到惊奇和高兴，这些读者在邮件中说这本书是他们收藏在书架上唯一的一本统计学书籍，或者至少是唯一一本让他们用心去看的统计学书籍。

以上两种原因促成了本书第 3 版的出版。对于第一类问题，我们在“改变的度量”这一章节中加入了增长曲线分析这样的主题，并令其本身的章节内容更符合最新的层次线性模型。为了使本书更像一本课本，我们主要增加了两部分内容：增加了一个新的关于等效性和非劣效性检验的章节，另外由于本书大多数章节最后一部分的内容都是运用 SPSS 进行试验，因此这一版本中增加了如何使用电脑数据程序 SPSS 这一内容。其他包含在各个章节内的改变没那么明显。例如，随着杂志数目的不断增加，仅报告一个试验有意义是不够的；现在杂志社要求作者为统计检验补充置信区间和一些效应大小的预测值，因此我们在需要的地方增加了这些内容。我们在回归和方差分析的章节里增加了更多的内容。简而言之，就是使一本优秀的书籍变得更优秀。

有一样东西还没有改变（我们希望），那就是我们相信在同样的句子里能把统计学词汇变得有趣、不敬，但是不会引起矛盾。我们仍然将一个学生因为在浴缸里读这本书时笑得很开心摔倒了，而不得不再去买一本来看作为对我们最高的赞赏。所以说，学习和欢乐一路相随。

DLS

GRN

第 2 版前言

我们非常高兴能收到对于《生物统计学基础》第 1 版的积极评价。许多人发电子邮件告诉我们，他们第一次真正理解了什么是统计学和学习统计学的乐趣，这完全出乎我们的意料之外，这所带来的喜悦不亚于打开普通邮件拿到版税支票。我们曾经就是否应该写第 2 版争论了好一段时间。我们的犹豫归咎于两方面的考虑。第一，如果医学知识的半衰期是 5 年的话，那么统计学知识的半衰期肯定要比一头大象的寿命长。毕竟，我们在统计学上仍然还在使用 Galton(1911 年离世)所提出的相关系数，以及 Ronald Fisher 在 20 世纪 90 年代初期提出统计学的基本核心知识。第二，我们还需要处理其他的日常生活事情，譬如吃饭、睡觉和陪伴家人等等。

那是什么让我们决定出版第 2 版的呢？首先，统计学已经改变了。通径分析和结构方程模型已经盛行了差不多 1/4 个世纪，但是根据最近的关于能在手提电脑里轻松进行上述分析的程序介绍，它们的使用在过去几年里得到了激增。在没有了解某个领域任何一个实例的情况下，去阅读这个领域(例如心理学)的期刊几乎是不可能的。这样的情况同样发生在其他高度应用电脑的分析技巧上，如 Logistic 回归和多变量方差分析。所以，我们在第 2 版中增加了一些关于这些主题的章节。现在，在将近半个世纪的争论后，我们开始在测量变异的最佳方法上取得共识，而这方面的知识也足以构成一个独立的章节。

编写第 2 版使得我们可以改正我们以及其他人在这几年中发现的错误和误解。然而，我们也有可能在编写新的章节时犯下新的错误，因此，请读者们擦亮眼睛，欢迎给我们来信指正。最后，我们想借此机会感谢三个人，他们一直以来都非常热心地帮我们指正错误和阅读一些新章节的初稿，这三个人分别是：来自于 Villanova 大学的 Bill Marks，来自于 St. Louis 大学的 Kathleen Wyrwich 和来自于 Universidad de la Frontera 大学的 Jose Luis Saiz。

DLS

GRN

第 1 版前言

准备好庆祝了吗？你是否终于能够面对这些年自己不愿意承认自己忽视了统计的事实，能够面对那些在科学会议上发生的非常令人困窘的事件，以及那些关于你在接受医药公司接待时对别人提出的类似“协方差分析”的谈话只能承认自己感到困惑的即时评论？你准备好去认识自己的条件和处理自己的问题了吗？去面对你是一个害怕暴露自己有数字恐惧的人吧！

现在你已经公开承认了你的秘密，我们帮你解决你的问题。首先，有必要让你明白的是所有的统计学家并不是生来就一样的，所以所有的统计学书籍也是不同的。一个家居装修的例子可能可以帮助你了解这个事实。有三种人参与了家居装修。首先是建筑师，他们从理论的角度设计房屋，他们只在意理念和美学，这样的房子除了皮肤科医生之外没人能够买得起。其次是做家庭装修的木匠，他们都很专业并且技术熟练，他们使用例如板、窗台、椽子、护套、R28 等专业术语，这些术语从实际的角度来描述事物。

最后，还有自己动手的人（DIY 者），即那些对家居装修一无所知，但仍莽撞地进行尝试并且自己能对装修中的问题进行补充的人。虽然把钉子钉成一个 2×4 的形状，或者是对地基、墙壁、天花板、管道和电线进行其他装修并不困难。但令熟练的 DIY 者失望的是，关于自己动手做的书籍都是由建筑师或木匠编写的，而并不是由真正优秀的 DIY 者创作的，对他们而言这些书都没有作用。因此，你要么可以从一个造价 200 000 美金的浴室中得到一些关于美学的启示，要么得到的是一本从头到尾只讲“如何换保险丝”的 DIY 书籍。

不幸的是，同样的情况也发生在统计学中。像建筑师一样，有一些统计学博士为统计学理论做出了贡献，他们在《*Biometrika*》期刊上发表文章或撰写专著，但这些文章只能被他这个群体里的其他人读懂。另外还有一些木匠一样的统计学家，他们是为数最多的一类统计学家。这些人通常有统计学博士学位，但他们实际上并没有为统计学的学科基础作出贡献，只是在做统计。他们不经常在统计期刊上发表文章，那些不是他们的主要工作。最后还有 DIY 式的统计学家，这类人是像我们一样先学习其他诸如心理学或教育学等学科，再学习统计学的人。随着现代统计软件包和电脑的出现，几乎任何人都可以成为 DIY 式的统计学家，你当然也可以。请注意，和其他许多统计书籍不一样的是，在这本书中我们假设你今后不会真正做统计。除了在学统计学课程的学生，没有人会做 20 年的方差分析。如果上帝真想要人们做统计，那么他就不会发明电脑了。

以上的例子说明了目前存在的两个问题。

第一，现在做数据统计确实是比做水管工容易，但不幸的是，统计学上错误也被更好地隐藏起来，在统计学上没有类似漏水管这样明显的错误。此外，虽然期刊的编辑们都希望在统计学界有类似工程监理或建筑准则的东西存在，但那只是他们的希望而已。

第二，大多数自学类的统计书籍是由男商人编写的。他们拥有很多东西，也可能觉得有点心虚，所以他们不在《*Biometrica*》上发表文章。因此，他们犯下了两个基本错误。首先，他们不得不用这个策略的神秘性来迷惑你，并且十分聪明地暗示你他们一定要征服统计学这个领域。

为了达到这个目的，他们在书中大肆地使用专业术语，运用无数的衍生词，引入代数学以使这些书看起来像是科学书籍。更重要的是，他们写的书呆板、正式、完全不具有可读性，并且将这种呆板的、正式的、没有可读性的文字视为他们书籍可靠性的基础。至今为止，这类统计书是所有统计书籍中最常见的。

然而，他们还有第二个战略。在意识到实际上头脑清醒的人都不会把自己的血汗钱用来购买上述这类书时，一些木匠式的统计学家已开始出

版小而薄的统计书，这些书语言活泼，他们真诚地希望能够深入浅出地为读者解释统计这个领域并且赚到大钱。唯一的问题是，他们通常认为目前常用的那些真正的统计方法，对于平常的自学者来说理解起来太复杂。因此，这些书籍由始至终都在讲 20 世纪末一些流行的统计方法。有一个用来为这些书籍辩解的论点说："在仔细研究过生物医学文献和像因子分析这样很少有人使用的当代的有影响力的统计方法后，我们决定只是教一些经常出现的方法"。说这句话的时候，他们或许忘记了辩论有循环论证的特性。

我们有些消息要告诉你。事实上，当代统计学并不是都那么复杂，在电脑几乎解决了一切繁琐的工作的今天，统计分析比过去少了很多痛苦。当然，如果和生理学或者物理学比较，那它是没有痛苦的。但是，统计书籍的作者写书的愿望必须是真诚的，同时他要努力尝试让读者理解书中的内容。让我们回到刚刚最后那一个 DIY 的例子。熟练的 DIY 者真正参加了两种活动。对于日常的琐事，他们希望他们能够聘请专业人士，并有信心能够在他们做得好或不好的时候察觉出来。也就是说，他们知道自己没法完成所有的工作，但自己肯定一眼就能够辨认出那些粗制滥造。对于剩下的工作，他们就可以决定由自己来完成。对于生物医学研究人员面临的统计来说，上述两个途径都是可行的。一方面，在检查别人做的分析时，能够辨别别人做的好与坏是必需的，尽管人们可以选择不由自己来做这个工作。另一方面，由于许多使用起来灵活又简单的当代统计软件包的出现，几乎任何人都可以参与到统计中来。

我们应统计学消费者的要求写了第一本书——《快速学习统计学》(*PDQ statistics*)(Norman and Streiner, 1986)。我们发现，用很少的数据资料和证据就可以从理论上去解释大多数的当代统计学问题。但是，相比辨别别人做的好与坏时，当你自己实际操作时将会需要更多的知识与技巧，比如安装管道、电线或者是做统计。这也是我们编写这本书的意图。如果你从来没有打算要做统计，那节省几块钱买一本《快速学习统计学》就行。但是，如果你是真正地在做研究，抑或是《快速学习统计学》或其他一些介绍性的书籍已经激起了你对统计学的兴趣，那就向售货员付款，买下这本书吧。

以下是一些对于这本书格式的评论。通过对本书内容的细读发现它的格式和其他传统格式的统计学书籍是一样的。我们用"基于问题"的模式来写这本书，不仅仅因为我们毕业于实行"基于问题"学习模式的医学院，也是因为这样能够使我们的书更有时代感，更畅销(我们从未说过我们是利他主义)。但是，我们认为，这将贬低"基于问题"学习模式的内涵。这本书是一种资源，而不是一个课程。无论如何，我们将鼓励读者在遇到统计问题时看这本书，从而学一学这种"基于问题"的学习模式。不过"基于问题"的学习模式对读者的背景并没有限制——那些仍然狼吞虎咽地啃着 Harrison 和 Merck 手册的医学院学生，无论他们来自哪所医学院校，这本书都是适用的。我们认为按照传统的统计学书籍的顺序可以更好地从概念上解释基础知识。

当然，不是所有的章节都是同样的格式。大多数章节会从一个引例开始。通常这些例子是我们用丰富的想象力虚构的，我们希望它们是有趣的。有时，我们也用到真实的数据，因为有时候真实世界就如同虚构的世界一样的奇特。虽然许多对于统计书籍的评论都对使用真实例子的作者大加赞扬，而批评了其他的作者，但鉴于以下几个原因，我们仍坚持我们的观点：(1)本书旨在写给所有的卫生专业人员阅读，我们不想浪费你和我们的时间，给其他非专业人员解释其中的复杂关系；(2)真实的世界存在很多无法控制的影响因素，要找到能够简单地阐明教学点的真实的例子是很难或者几乎不可能的；(3)我们相信，能够找到好的心理学证据，证明令人难忘的("奇特的")例子对于学习和记忆概念是有帮助的。

虽然我们已经努力减少了公式的数量，但本书中的公式仍然要比《快速学习统计学》一书多。原因很简单，公式就是统计学的语言。如果我们尝试将公式全部删除，那么这本书就会以繁复的散文形式结束，并且会损失掉一些信息。虽然保留了这些公式，但是我们会继续努力尝试解释我们所强调的概念，绝不会扔下一个公式给你就不管了。

这本书另外还有一些特别之处。我们在本书中沿用了《快速学习统计学》一书中 C. R. A. P 辨析的观点，它可以帮助你发现其他人(和你自己)使用的方法中的错误。在大多数章节的最后我们都编写了电脑操作的注解，以帮你们进行 SPSS(社会科学统计软件包)的操作，SPSS 是目前普

遍使用且功能强大的数据程序之一。最后，我们承认，许多临床研究人员可以靠他们大部分的技术获得资助，因此他们能雇用其他人去做数据统计。同时，想依靠小样本的计算结果从大多数的联邦、州或省的政府机构争取到资金是不可能的。这意味着许多生物医学研究者所做的唯一的分析，是在他们的基金申请中关于样本量大小的计算。因为认识到这一客观的事实，本书的每个章节在有必要的时候都有一部分关于样本量大小计算的内容，学习过这些内容后，你就可以更容易获得基金的资助。

在排版的问题上，你会发现这本书每页的边缘都留有很宽的空白。这不是出版商的错误，也不是对造纸行业的救济。相反，它有两个用途：(1)我们能用边缘的空白部分写些提示，引申一些相关的有趣的信息，或者是一些我们认为的小幽默；(2)如果你不喜欢我们的注解，你可以用它来写自己的笔记。

最后，在风格的问题上。你们或许已经察觉到我们带有些许不礼貌的语气，本书是要提供给所有不幸出现在这些章节中的人们——统计工作者、物理学家、管理人员、护士、理疗医师、心理学家和社会工作者来阅读。我们这样写是因为我们认识到在某种程度上我们冒着得罪那些曾经受到那些获得"MD"学位的人压制的卫生专业人员的风险。但是，我们觉得如果我们完全忽略了他们，冒的风险可能会更大。我们的目的不是种族歧视、性别歧视或者其他的偏见，对待所有人我们都一视同仁。我们会尽可能努力地去平等地"侮辱"所有的专家们。

写在电脑备忘录上的内容

我们坚信我们的妈妈不是为了要我们在手工计算上浪费时间而养育我们的，这就是为什么我们会拥有电脑和统计软件包。尽管如此，学习这不可思议的电脑编码语言还是像学习统计学本身一样令人生畏。所以，在我们永不停止地探索尽可能有用的同时，我们提供了一些能满足你的要求的程序。

几年以前，选择要使用哪些程序是一件简单的事情，因为那时只有三四个程序能在家用电脑上使用，我们可以使用全部的程序，同时还能显得我们还很博学。但是现在似乎每个月都会出现一个新的、"更好的"软件包，这迫使我们要做出选择。当我们写第1版时，有一些既受欢迎功能又强大的程序如SPSS、BMDP、SAS和Minitab。所以，我们很乐意去介绍如何去操作其中的三个软件。过去的几十年里事情都相应的发生了变化。SPSS已经由微软所设计出的数据操作系统软件所完成——它耗尽了他们整个早餐。在你依旧买SAS和Minitab的时候，SPSS买下了BMDP，但它大不如前了。真正的统计学家依旧使用SAS，但是，你将会需要另外的一个书架去收藏它所有的使用手册。现实就是，无论你从哪里向社会科学和医学科学里探索时，人们都在操作SPSS。它不是对于所有的分析来说都是最有效果的，但是它的适用范围很广，并且它很容易被设计成为垄断性的软件。因为没有任何使用手册或帮助指南比得过一个博学的朋友，而熟识SPSS的朋友比掌握其他软件的朋友更普遍，在这样的趋势下我们没有理由反对使用SPSS。因此，这次我们只是包含了关于SPSS的使用介绍。(SPSS9.0及以上版本)。

祝您好运(如果你的电脑坏了，别打电话给我们)。

致谢

我们许多学生已经看过这本书早期的草稿了，并且给我们指出了许多宝贵的意见和建议。不幸的是，他们人数太多以至于无从说起(他们中大部分人的姓名我们也忘记了)。尽管如此，我们要特别感谢Dr. Marilyn Craven，他很有耐心地帮助我们解决逻辑和英语的问题。因此，你们所发现的任何错误都应该归咎于他们(原作者用语幽默，这是一句反话，译者注)。而我们只会谦虚地接受任何对我们努力的赞扬。

还有一个重要的说明(我们希望这是最后一个了)，我们要向Brian C. Decker表示衷心地感谢，他提出了这本书的主要观点，并且从始至终都给予了我们莫大的鼓励。

目　录

第1部分

数据与统计学的特征

第 1 章

基 础 知 识

在这一章中，我们会向你介绍变量的概念以及资料的不同类型：名义资料、有序资料、区间资料及定量资料。

哪些人会用到统计学？

大多数刚开始接触统计学的学生都会问这样一个问题："我们为什么要学习统计学？"暂且不说它是你攻读学位的必修课程，我们亟须解决的问题是如何通过学习统计学的方法、术语使你变得更加优秀，并获得从未有过的满足感。学习统计学是因为这个世界到处都有变化，而且有时候很难从自然变化中分辨出真实的差异。如果世界上人人都一模一样，那么我们就不需要统计学了。假设你是男性，身高 172cm，棕色的眼睛和头发，长相英俊，这样的描述也可能适合其他人。同理，如果人与人之间没有差别，而且我们知道你的期望寿命、知道某种新药是否能有效地去除你的头屑、知道下一届选举你会投哪个政党的票（假设各党派最终给你一个自由选择的机会，当然这令人怀疑），那么我们就知道其他所有人的相应信息了。

然而，情况并非如此。人与人之间存在差别。这些差别使我们难以判断一种新的治疗方法应用到某个人身上会产生怎样的疗效或者当他处于某种情况时会产生怎样的反应。我们不能对着镜子问自己"你觉得最新品牌的牙膏好不好用？"之后假设人人都会有相同的感受。

统计描述与统计推断

正是因为人与人之间存在差异，甚至同一个人在不同时间也存在差异，统计学应运而生。我们希望当你读完这本书后，会用统计量去描述一个群体的"平均"水平，并且判断这个"平均"水平是否适用于群体中的其他人，还能够判断从小部分人群中推断出的结论能否推广到全体人群中去。因此，统计学主要有两方面的作用：描述数据并且对其进行推断。

统计描述主要用于数据的展示、组织与归纳。

本章涉及了多种统计描述的方法，它们通过对数据进行组织、绘图来展示数据的特征。描述统计学还包括利用数据中的一些关键数值对数据特征进行描述。这些内容都是统计推断的基础。

统计推断让我们可以从样本数据出发去推断更大群体中个体的特征。

例如，一个皮肤病专家给 20 个因满脸粉刺而爱情受挫的青少年使用一种新型的从茄子中提取的膏状精油进行治疗，并将他们同 20 个没有接受治疗的青少年进行比较。他关心的并不只是这 40 个青少年的疗效，他还想知道的是这个疗法应用到所有患有粉刺的青少年身上是否都有效。这时，他需要从他所研究的这个较小群体入手，对更大的群体进行推断。我们将在第 6 章中对统计推

断进行更详细的阐述。以下，我们介绍一些定义。

变量

在前面的段落中，我们提到人与人在很多方面都存在着差异：性别、年龄、身高、头发及眼睛颜色、政治偏好、治疗效果以及期望寿命等方面。统计学将以上因素称为变量。

简而言之，变量就是现在正被观察或测量的东西。

变量有两种类型：自变量和因变量。实例是理解这两个概念最简单的方法，回到刚才长粉刺的青少年的例子上。我们想知道粉刺的数量是否会随青少年使用茄子精油而发生改变。这里结局（粉刺）是因变量，可以认为它因治疗而发生改变。施加的干预措施是治疗方法（茄子精油），称其为自变量。

因变量是我们所关心的结局，它因干预措施的效果而发生改变。

自变量是某种干预或是人为控制的某种措施。

听起来是不是很简单？但这样的定义太过简练，不够确切。一旦脱离实验的范畴，因变量和自变量之间的区别就变得不明显了。例如，如果我们想观察一个小孩随年龄增长其词汇量的增加状况，那么词汇的记忆量就是因变量而年龄则是自变量。从而认为词汇的记忆量取决于年龄，尽管年龄并不是某种干预，也不是人为施加的某种措施。广义上说，如果一个变量因另一变量的变化而发生改变的话，那么我们就称这个因自变量的改变而发生改变的变量为因变量。

自变量和因变量都可以取若干值中的某一特定值：对于性别而言，取值为男或女；而头发颜色则可以是棕色、黑色、金色、红色、灰色、人工染色或秃顶；像身高这类变量的取值可以从早产儿的25～40cm一直到篮球运动员和本统计书的合著者身高的200cm。

数据的类型

离散型和连续型数据

尽管我们说性别和身高都是变量，但是他们在取值的类型和数量上还是有着明显的差别。一种区分变量类型的方法是判断这些变量的取值是**离散型（discrete）**的还是**连续型（continuous）**的。

离散型的变量只能取一些有限的数值。举之前的例子，这样的变量有性别、头发和眼睛的颜色、政治偏好、一个人接受的治疗次数等等。另一个离散型变量的例子是“总数”，如一个人总的入院次数，龋齿、缺牙和填补牙齿的数目、一个家庭中孩子们的数目。由于儿童数量是离散型变量，因此人口统计学家可以明确地指出不可能有2.13个孩子的家庭。

离散型数据的取值只能是整数。

而连续型变量的取值却不尽相同。乍一看可能像身高这样取值单位离散的变量应该属于离散型变量：某人身高172cm，某个比他高一点的人是173cm，某个矮一些的人则是171cm。事实上，这是十分局限的，这种局限性是由我们的测量尺造成的。如果用一个有更精确刻度的测量尺进行测量的话，就可以精确到二分之一厘米。事实上，还可以用精确到千分之一毫米的激光去测量每一个人的身高。身高、体重、血压、血清大黄浓度、时间以及其他很多变量的取值都是连续的，而取值间的分界则是我们为了适应需要而主观制定的。这就导致了我们对测量值的认识十分机械，当精确到毫米汞柱时两个人的血压可能相同，但如果我们的测量精确到十分之一毫米汞柱时，这两人的血压可能就会显现出差异。如果数值仍然相同，那么我们可以采用更精确的测量尺度直到两人的血压值分出差别为止。

连续型数据可以取一个固定范围的任何数值。

我们可以用另外两个例子对连续型数据和离散型数据的区别加以详细说明。钢琴是一个“离散”的乐器，它只有88个琴键，人们经过长期艰苦的努力去理解Paganini（帕格尼尼）才发现原来升A调和降B调是同一个音调。然而小提琴家（对于Mason-Dixon线[美国南北分界线，译者注]以南的人来说则是“小提琴手”）却可以将这两个音调“连续”地演奏出来，并且可以在演奏中很好地区分这两个音调。类似地，廉价的电子表只能显示4位数字，并且只能精确到1分钟；而功能繁多的高档表不仅可以储存电话号码和你的银行存款余额数，还可以把时间精确到百分之一秒。物理学家的做法更妙，他们将每秒钟分成9 192 631 770次铯原子的振荡。然而，即便是这样，这仍是一种

主观的分界。只有医院中的某些高收入者能买得起一支 Patek Phillipe(百达翡丽)模拟精密计时器从而可以看见时间本来的面目:一个平滑的、不间断的过程。

你将要学到的很多统计学方法其实并不十分依赖于数据的类型是离散的还是连续的,毕竟不管怎样的数字对于这些统计方法来说只是数字而已。但有的时候,分清数据类型还是十分重要的。本书会在相应章节做详细介绍。

名义资料、有序资料、区间资料、定量资料

我们可以使用另一种方法对变量进行分类。像性别这类只能取两个值的变量,即男性和女性。某个值不比另一个值"更高"或"更好"。无论我们怎样排列这两个变量,都不会丢失任何信息。这类变量被称作**名义变量(nominal variable)**。

名义变量即各种被命名好的分类,在各种分类中没有隐含的顺序。

最简单的名义变量是 Feinstein(1977)提出的"存在"变量——即一个事物要么存在要么不存在。一个人有或者没有肝癌;某人接受或者未接受一个新的治疗方法;对于大多数生物来讲是活着还是死亡。名义变量并不一定非得是二分类的,它们可以有多种分类。我们可以将一个人的婚姻状况分为:单身/已婚/分居/丧偶/离异/同居(6种分类);可以将眼睛的颜色分为黑色/棕色/蓝色/绿色/混合色(5种分类);还可以将疾病归为几百个诊断分类中的一个。其中的关键在于你不能说棕色眼睛比蓝色的"更好"或"更坏"。分类的排序是主观决定的,改变这个顺序并不会增加或丢失任何的数据信息。

由于电脑处理数字的能力远远比处理字母要好,所以要求研究者对各种分类分别进行数字编码:例如女性可以被编码为1,而男性则编码为2;或者单身为1,已婚为2等等。在这些情况下,用于编码的数值不应超过分类的数量,同时我们也不能认为这些数字有着任何数量上的意义。同样,我们可以改变赋值,将男性编码为1而女性编码为2,我们得出的结论和之前应该是一样的(当然要在我们记得之前是如何编码的基础上)。

对一个学生的评价可以包括优秀/良好/不良3个等级。它不同于头发颜色这类变量,因为这些变量值之间存在顺序:"优秀"优于"良好",而后者又优于"不良"。而且,优秀与良好之间的差别不能单纯等价于良好与不良之间的差别。这样的差别在用字母评分时表现得更加明显;评分B+与B之间只有很小的差别,然而D-和F+之间却有很大的差别,而后者关乎你的暑假能否过得安稳。这就像赛马比赛的结果一样;第一名的马匹跑得比第二名快,第二名则跑得比第三名快。然而前两名之间的差距可能只有1秒,而第三名则可能比第一名慢上10秒。因此,字母评分和比赛名次等被称为**有序变量(ordinal variable)**。

有序变量即有序的分类,各分类之间的差异不能被认为是相等的。

在医疗领域中遇到的很多变量在本质上属于有序变量。病人通常被评价为显著好转/有些好转/没有变化/恶化/死亡;或者病情被分为紧急症/急症/一般症状。有时候也会用到数字,比如癌症的Ⅰ期到Ⅳ期。不要被数字所欺骗,这仍然是有序变量,数字(这次用罗马数字来表示,显得高级些)只不过是用来代表有序的分类而已。应用差别检验:Ⅰ阶段与Ⅱ阶段之间疾病严重程度的差异和Ⅱ阶段与Ⅲ阶段之间、Ⅲ阶段与Ⅳ阶段之间的差异是否相等?答案是否定的,因为级别间是有序的。

如果变量的分类之间的差距是相等的,那我们称之为**区间变量(interval variable)**。

区间变量的各变量值间的差距是相等的,但起始值是人为规定的。

我们为什么要在末尾附加上一句"起始值是人为规定的"呢,这句话有什么含义?我们加这句话的目的是给我们描述的区间变量施以限制。同时说明起始值可能没有实际意义,可以人为更改。为了说明这点,让我们用IQ测验测出的智商值和体重值进行比较。体重的起始值是有意义的,我们都知道重量的最小取值为零。我们不可能贸然决定从现在开始将所有之前称过的物品都减去10kg,并推断之前重11kg的物体现在重1kg。这不仅仅是数字上的变化那么简单,如果先前一个东西重为5kg,转化后这件东西重量变成-5kg—这显然是不可能的。

然而,智力评分则是另外一回事。我们说平均IQ是100,但这只是依照惯例而言。也许下一届IQ世界会议上,专家就可能决定从现在起,我们将IQ的平均水平升为500,即简单地将所有分值都加上400。这样的变化不会给我们带来任何收获,同样的,也没有任何损失;唯一的改变就是

我们需要调整之前对IQ平均水平的认识而已。

现在，我们来进一步了解这其中的含义。因为是等间距的，所以IQ70与IQ80之间的差别和IQ120与IQ130之间的差别是等价的。但是，IQ100并不等于IQ50的两倍。原因在于起始值的制定是主观的，并且起始值是可改变的，导致数字之间的差异是有意义的，但是它们之间的比值却没有意义。

如果起始值是有意义的，那么数值间的比值也是有意义的，这样的变量（不足为奇）就称为**定量变量(ratio variable)**。

定量变量各数值间差距相等，并且起始值有意义。

像人的身高、体重等生理特征一样，实验室的绝大多数测量数据是定量变量。一个体重100kg的人是体重50kg的人重量的2倍，就算我们将千克换算成磅，比值依然是不变的：220磅是110磅的2倍。

区间变量和定量变量的差别就先介绍到此。其实，从统计学家的角度看，这两个变量可以用相同的方法处理和分析。

我们注意到从有序变量逐步演变到定量变量，都是在上一个条件的基础上再添加一个限制条件。

变量类型	假　设
名义变量	名义分类
有序变量	在名义变量的前提上加有序分类
区间变量	在有序变量的前提上加等间距
定量变量	在区间变量的前提上加有意义的起始值

尽管名义变量、有序变量、区间变量、定量变量之间的差异在理论上很明显，但有时它们之间的界限却很模糊。例如，如前所述，智力由IQ值来评价，其平均水平为100。严格意义上，我们不敢保证IQ80与IQ100和IQ120与IQ140之间等距；也就是说，IQ最有可能是有序变量。但在实际生活中，大多数人将IQ看做是区间变量。就我们所知，目前他们并未因此关进监狱，上帝也没有惩罚他们。

尽管如此，名义变量、有序变量、区间变量、定量变量之间的差异还是要牢记在心的，因为在某种程度上这些差异限定了我们必须采取某种统计学方法去处理相应类型的数据。在以后的章节中我们会见到，某些特定的图表和“参数检验”的方法只能应用于区间资料和定量资料，而不能应用于名义资料和有序资料。相反地，如果是名义资料和有序资料，严格意义上只能用“非参数检验”来处理数据。在稍后的章节中我们再来介绍这些晦涩难懂的术语。

比例和率

到目前为止，我们讨论涉及的各种数据类型都是单一数值型——血压、课程评分或计分等。但有些时候，我们需要分析两个数字之比。尽管这是我们在小学就学过的内容，但由于某些统计学家草率用词而导致了某些概念到现在还存在混淆。作为实事求是的学者，我们要尽力消除这类错误。

比例(proportion)是比的一种，其分子是分母的一个子集。例如1/3，其含义是三个个体中的一个。百分数是比例的一种，其分母被限定为100。你可能觉得这太基础了，但为什么我们还要讲它？原因有两个。其一，我们后面会遇到许多其他比（例如，优势比），它们的分子不是分母的一部分。其二，很多统计学家经常弄错，称比例为率。

但严格地讲，**率(rate)**不只是比，还包含有时间因素。如果我们说有23%的孩子是蓝眼睛（这个数值是我们随便捏造的），那23%就是一个比例。但是，如果我们说今年每1000个人中就有1个人会患上畏光症，那么这就是率，因为我们指定了一个时间段。

好了，学习完基础知识，让我们开始学习统计学吧！

习　题

1. 在下列研究中，请指出哪些变量是因变量，哪些变量是自变量，或都不是。

a. ASA和安慰剂比较以研究是否能预防冠心病。

自变量是_____　因变量是_____

b. 低胆固醇血症与癌症的关系。

自变量是_____　因变量是_____

c. 我们知道那些禁食毒品、烟、酒和禁欲（因