HZ BOOKS
华章科技
que®

知名统计学专家、多部畅销书作者Conrad Carlberg倾情撰写，循序渐进系统讲解Excel决策分析的各种技术、方法和实践，指导读者充分利用已有数据优化业务和投资决策，Amazon全五星评价

从基本原理、适用范围、数据构造需求和实际执行方法等方面，由浅入深介绍决策分析涉及的主要统计学方法，包括逻辑回归、单变量及多变量方差分析、判别分析、主分量分析和聚类分析，包含大量实用案例

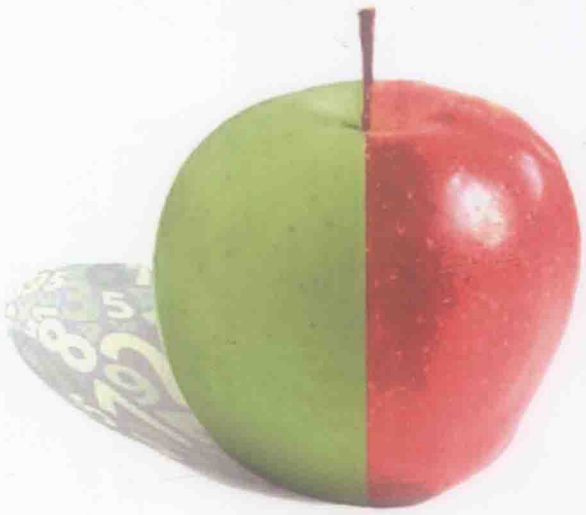
数据分析
技术丛书

Decision Analytics: Microsoft Excel

决策分析

以Excel为分析工具

（美） conrad carlberg 著
姚军 / 译



机械工业出版社
China Machine Press

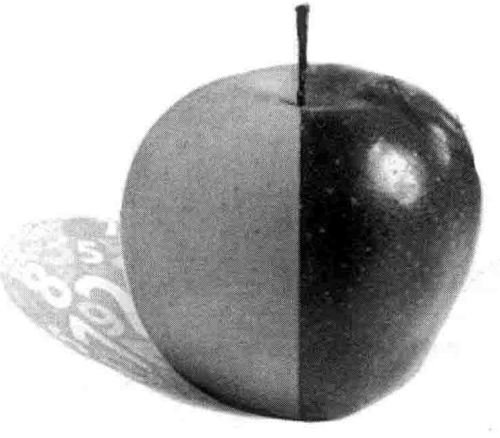
Decision Analytics: Microsoft Excel

决策分析

以Excel为分析工具

〔美〕 Conrad Carlberg 著

姚军 / 译



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

决策分析：以 Excel 为分析工具 /（美）卡尔伯格（Carlberg, C.）著；姚军译 . —北京：机械工业出版社，2014.11
（数据分析技术丛书）
书名原文：Decision Analytics: Microsoft Excel

ISBN 978-7-111-48389-2

I. E… II. ① 卡… ② 姚… III. 表处理软件 IV. TP391.13

中国版本图书馆 CIP 数据核字（2014）第 250026 号

本书版权登记号：图字：01-2013-8816

决策分析：以 Excel 为分析工具

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：关　敏　　责任校对：董纪丽
印　　刷：三河市宏图印务有限公司　　版　　次：2015 年 1 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：14.5
书　　号：ISBN 978-7-111-48389-2　　定　　价：49.00 元

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）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

The Translator's Words 译 者 序

随着大数据的兴起，当今世界真正成了数据爆炸的世界，许多企业担心的不是缺乏数据，而是如何在海量数据中挖掘出能够指导商业决策、客户分析等的运营措施，从而为企业长期目标提供支持。

决策分析的基础是统计学，这一古老的综合学科经过两千年的发展，已经形成了许多经典的理论和方法，而这些方法在进入计算机时代之后更是得到了很大的推动，原来乏味费时的计算被高速的计算机自动化代码所代替，人们也从这些应用中获得了很多成果，每一位注意该领域的人，无不为数据挖掘中的许多经典案例拍案称奇，许多高级统计软件（如SAS、SPSS）的出现，更将统计学的商业应用推向了顶峰。

也许正因为有了这些很好的软件，普通的商业分析人员越来越少去探究决策分析方法的基础理论和方法，而更多地借助软件中的现成功能。由于高级统计软件价格不菲，它们注定也只能是大型企业的宠儿，对于中小规模企业的分析人员，以及对高级软件分析结果进行一些简单处理的分析人员来说，深入了解这些理论和方法，并且用更容易得到、更经济的解决方案去完成分析，便成了一项迫在眉睫的任务。

本书的目的正是为广大分析人员提供这样的方案。作为量化分析方面的资深人士，作者用几乎所有 PC 上都拥有的通用办公软件 Excel，完成了许多在大型软件上才能完成的分析工作。从本书的例子中就可以看出，Excel 可以得出与大型软件相同的分析结果，只要编写一些代码，它也具备相当的自动化能力，从本书英文版出版商网站上可以下载的本书配套 Excel 工作簿中，读者可以得到许多分析所需的通用代码，并根据自己的需求进行改编，构造自己的小型分析系统。

更重要的是，本书系统地从原理、适用范围、数据构造需求和实际执行方法多个角度，介绍决策分析中的主要统计学方法，包括逻辑回归、单变量及多变量方差分析、判别分析、主分量分析和聚类分析。本书中介绍的例子，许多都是统计学发展历史上的经典实例。细细读来，读者就会发现，数据挖掘中的许多神奇发现其实可能就在我们的周围，只要你掌

握了基本原理和方法，加上细心的观察、分析和数据收集，下一个创造奇迹的也许就是你。

翔实的解说和实例，使本书成为一本不可多得的决策分析入门教程。在此我们诚挚地向有志于此的读者推荐本书，希望它能成为你成功路上的帮手。书中涉及大量统计学知识，由于译者水平所限，错误在所难免，希望广大读者见谅，并多多提出意见。

本书的翻译工作主要由姚军完成，徐锋、陈绍继、郑端、吴兰陟、施游、林起浪、陈志勇、刘建林、宁懿等人也为本书的翻译工作做出了贡献。

译　者

Preface 前　言

首先要告诉读者的是：本书所用的Excel工作簿可以从本书英文版出版商的网站quepublishing.com/title/9780789751683上下载。下载链接有时候难以和常规文本区分，但是链接和工作簿都在那里。

从第2章开始，每章都有自己的Excel工作簿，每章中的每幅插图都是一个单独的工作表。还有几个附加的工作簿用于执行聚类分析、判别函数分析和其他没有自己的工作表函数的分析过程。

好吧，我们来做个澄清：

本书不是关于获取、存储和分割所谓“大数据”的书籍。本书讲述的是关于如何了解数字的含义——它们到底是“大数据”，还是“小啤酒”。

我们都碰到过这种情况：有30个变量需要处理，每个都可能很重要，各自针对有趣现象的不同侧面，可能是12个月的生存率，或者投资盈利的可能性，或者了解新雇员的业绩。不管你需要处理的是200个还是200 000个记录，真正的问题是如何处理这30个变量。如何组合或者抛弃它们，以做出关于药物效果，是否提供资金，以及雇用哪位应征者的正确决策。

本书内容

本书的主题是：寻找你所掌握的变量的最佳组合，以便尽可能做出明智的决策。

这是使用定量分类技术的一种实践，此类技术有如下几种。

判别函数分析有悠久的光荣历史。它的用途很广泛，范围从根据法律记录辨别19世纪政治家所属党派，到根据扣减金额和调整金额，标记可能的不实1040表格[⊖]。第5章和第6章带你经历这种分析，探索所涉及的数据简化技术。它们能够让你看到，在工作表和图表

⊖ 1040表格（form 1040）是美国个人收入联邦税申报表。——译者注

的环境中，判别函数分析如何起作用。

因为判别分析依赖多变量方法处理连续变量，所以我加入了第 7 章。第 7 章能够帮助你了解特征值和特征向量等概念，因为它们与相关矩阵有关——同样，也是在熟悉的 Excel 工作表和图表环境中。

你还可以下载一个工作簿，其中包含了运行完整的判别函数分析并输出显著性测定、函数系数、典型相关和其他功能的 VBA 代码，在正文中将对此进行解释，并在该章的工作簿中进一步展示。

进行判别分析的最佳方法是利用多变量方差分析（MANOVA）。你将会看到，MANOVA 能够帮助你确定执行判别函数分析是否有意义——因变量（非独立变量）之间是否相关，以及区分不同的人和行为分组的能力，以支持进一步分析。因此，第 4 章讨论 MANOVA，你可以下载一个单独的工作簿，运行多个因变量的单因素 MANOVA。

如果你有很久没思考过 ANOVA 或者 MANOVA 的问题了，可能应该通读第 3 章。作为 MANOVA 的背景知识，在工作表的环境中了解 ANOVA 管理变量的能力是很有帮助的。

除了判别函数分析之外，对人或者市场行为（或者政治家、室内植物）进行分类的另一种方法是逻辑回归。这是一种实用的方法，它避免了判别分析可能犯的一些错误。例如，逻辑回归不会像判别分析那样，做出关于数据分布方式的所有假设。所以，如果担心数据违背了那些假设（老实说，即使这些假设不成立，你的分析也不一定无效），往往可以使用逻辑回归来代替，作为决策分析的基础。

另一方面，那些假设给判别分析带来了统计能力——成功和可靠地区分不同对象组的能力。在其他情况相同时，判别分析对分类的指导比逻辑回归更敏感。

在我的前一本书《 Predictive Analytics: Microsoft Excel 》中为逻辑回归保留了两章。在本书第 2 章中我对此进行了介绍，更多的是一种复习，而非完整的讨论。

第 8 章和第 9 章介绍了其他决策分析方法。在逻辑回归和判别分析中，你知道分组的情况。你有一个或大或小的数据样板，观测值包括所属组（幸存与否、盈利与否、输赢）和你希望用来帮助你做出好的决策的变量（人口统计学数据、财务数据、购买历史）。

但是在聚类分析中，你不知道自己的分组。例如，你有一组人口统计学变量，希望知道如何用它们对人们进行分类。 你对数据集实施聚类分析的某一变种，希望它聚合样本中的人，使得同一个群集中的人在人口统计学上的差异较小，而不同群集的人之间差别相对大。

Leland Wilkinson 在 1986 年对这种决策分析方法做出了一种恰当的描述，他写道，“粗略地说，这种方法就像一种单向的差异分析，其中的组别未知，最大的 F 值通过重新安排每个组的成员来求得。”（参见 SYSTAT 手册“Cluster”节的第 1 页）。

为什么使用 Excel

感谢大家购买我写的书。但是，我还是一位顾问，我希望客户理解我对他们交给我的数字做了什么。我认为这是 20 年之后，我仍然工作在这个行业的主要原因之一。

我不喜欢交给客户一堆 R 输出，不管是采用字面形式还是电子形式。我这么说并不是对 R 语言有什么意见。尽管文档难以理解，结果就像 Fortran 语言输出的，它仍是很好的统计程序。我经常使用 R 来进行 Excel 中所完成工作的基准测试。

SAS、SPSS、Stata 和类似的软件包在文档上比 R 好得多，分析结果也采用更直接的方式输出。但是，它们太贵了。而且，和 R 一样，使用它们需要付出相当多的精力进行研究，才能学会用户界面的正确用法和正确地处理命令语法。

相比之下，我的大部分客户都很适应熟悉的 Excel 环境，并且欣赏在 Excel 图表中简单地查看一组数字的能力。当然，也很难在公司或者教学环境中找到一台没有安装和运行某种 Excel 版本的 Windows 电脑。

但是 Excel 能够处理分析（特别是决策分析）所需的复杂数据归纳方法吗？显然，我相信它可以。Excel 确实是一个通用的数字分析软件包，从一开始就不是为了提供特殊的统计功能而设计的。它们没有 WILKS() 工作表函数。

Excel 提供了一个 MDETERM() 工作表函数，如果将它指向一个组内矩阵和一个总体矩阵，就可以得到自己的 Wilks Lambda 值。假定你是一位分析的新手，或者企业的高级管理人员，想要知道为什么有人认为 Wilks Lambda 值说明应该回避某项业务。我主张，在任何一种情况下，你更应该了解的是：为什么它能够告诉你一些情况，而不仅仅是知道这个数值是什么意思。

而且，如果使用得当，Excel 能够帮助你了解那些情况。有时候，所需要的就只是一个内建的工作表函数。直接在工作表上进行多变量方差分析完全是可能的，不需要任何附加程序。我认为这样做一两次很有好处，因为它能够帮助你巩固对概念的理解。

但是，有时候你需要 Excel Slover 规划求解加载项（Excel 自带的一个附加程序）等工具的帮助。第 2 章告诉你如何使用规划求解加载项完成逻辑回归（顺便说一句，纯统计软件包使用相同的优化算法——只是它们掩盖了这一方法，所以你没有发现）。

还有一些处理，例如寻找大型相关矩阵的特征值，它们过于复杂且依赖循环，如果没有编写子程序，尝试起来会令人发疯。但是，可以在本书的 Excel 文件中找到用 VBA 编写的子程序，有些是开放的。

强调一下我的观点：2013 年年初，我着手帮助一家公司建立一个评估未来投资的模型。客户有将近 10 万个用于开发该模型的记录。 根据数据的特性，需要采用逻辑回归，客户以文本文件的方式提供数据，这很容易用 Excel 读取。我尝试使用 Excel 在该数据集上运行逻

辑回归，我的公式导致下溢。

然后，我使用 R，让数据通过一个逻辑回归例程（R 程序库的一部分），再次遇到了下溢的情况。在很多情况下，中间结果对于 Excel 或者 R 都太小了，无法精确处理。

现在，这不再会带来真正的问题了。我打算保留一些数据用于交叉验证，所以随机地将一半数据通过 R 的逻辑回归例程，得到结果，并用剩下的一半数据进行验证。然后，我用 Excel 确认结果。当然，客户得到了 R 和 Excel 的结果，但是我注意到，客户随后利用该模型所做的工作若采用 Excel，公式的结果会更加透明。

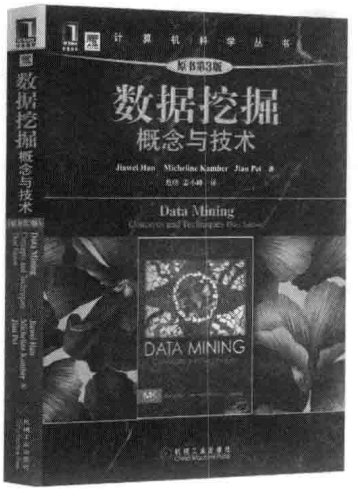
现在，这个普通的小故事不仅是件趣闻，还是一个样板。它真实发生了，是我自己使用 Excel 作为分析引擎的一次典型体验。如果你尝试本书中描述的方法，我相信你也会得出同样的结论。

说得够多了。建议你带上喜欢的饮料，打开笔记本电脑，进入第 1 章。

致谢

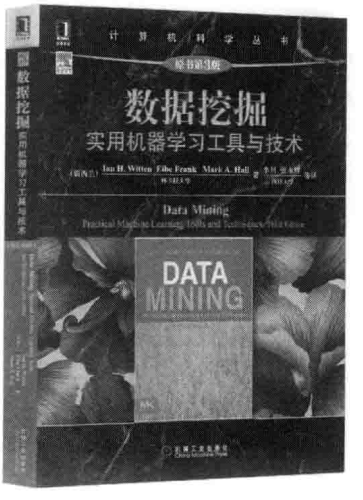
我要再次感谢 Loretta Yates，感谢她在协作过程中对本书的精妙指导，还要感谢她坚定的支持和平和的性情。感谢科罗拉多大学调查与评估方法实验室的 Michael Turner 提供很好的技术编辑，他的工作使我免于在付印的书籍中遭遇尴尬，并且在必要的时候使我回到正确的路上。对 Anne Jones，我能说什么呢？她为《Statistical Analysis:Microsoft Excel 2010》和《Predictive Analytics：Microsoft Excel》所做的封面设计和内容一样吸引读者，现在她又一次这么做了。感谢 Geneil Breeze 在编辑文稿时温和地提醒我，在初稿中有些卖弄学问了。感谢 Elaine Wiley 在百忙之中抽出时间来管理这个项目。我衷心地感谢大家！

推荐阅读



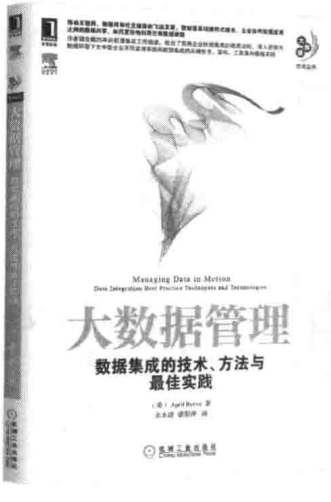
数据挖掘：概念与技术（原书第3版）

作者：Jiawei Han 等 ISBN：978-7-111-39140-1 定价：79.00元



数据挖掘：实用机器学习工具与技术（原书第3版）

作者：Ian H. Witten 等 ISBN：978-7-111-45381-9 定价：79.00元



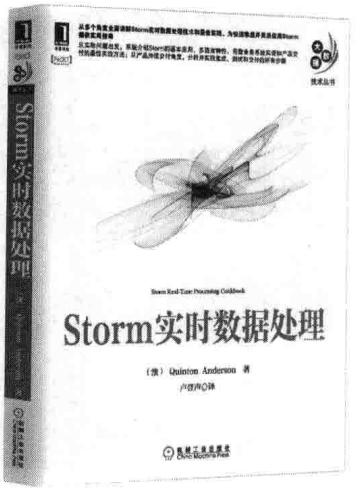
大数据管理：数据集成的技术、方法与最佳实践

作者：April Reeve ISBN：978-7-111-45905-7 定价：59.00元

大规模分布式系统架构与设计实战

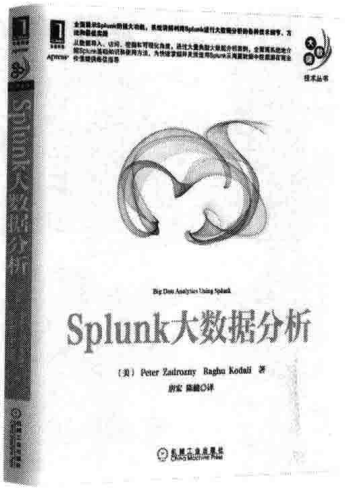
作者：彭渊 ISBN：978-7-111-45503-5 定价：59.00元

推荐阅读



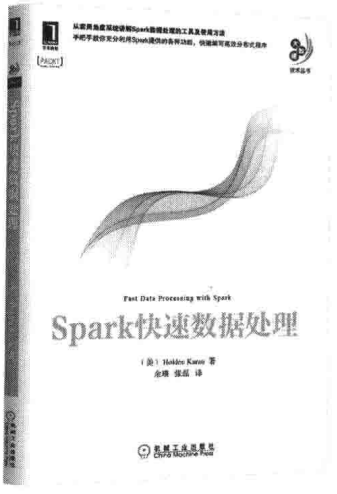
Storm实时数据处理

作者：Quinton Anderson ISBN：978-7-111-46663-5 定价：49.00元



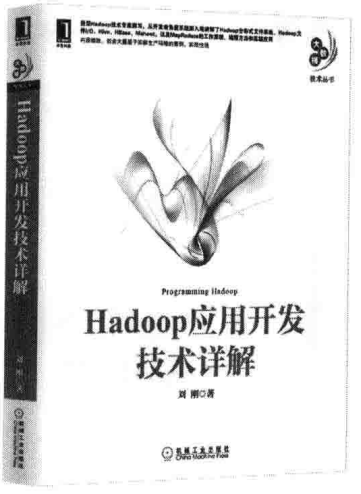
Splunk大数据分析

作者：Peter Zadrozny 等 ISBN：978-7-111-46429-7 定价：69.00元



Spark快速数据处理

作者：Holden Karau ISBN：978-7-111-46311-5 定价：29.00元



Hadoop应用开发技术详解

作者：刘刚 ISBN：978-7-111-45244-7 定价：79.00元

Contents 目　录