从此“回归”、“聚类”、“判别”不再是熟悉的陌生人，它们将会成为你“专业”的标志
大数据再也不是不敢踏足的高深领域，迅速看懂其中各种技巧

从零开始学统计

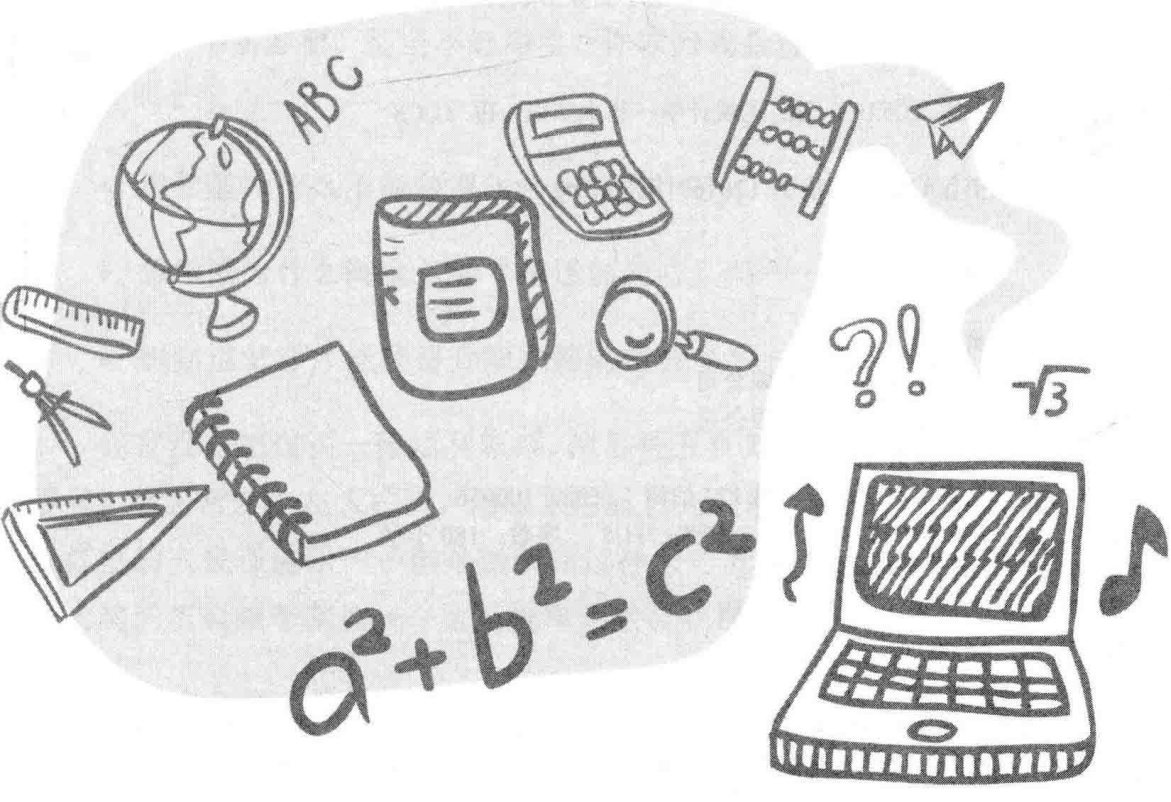
归璐　编著

用通俗易懂的文字解开“统计学”的神秘面纱，
带你走进它的世界！

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

从零开始
学统计

归璐　编著



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

内 容 简 介

大数据时代，每个人都要懂一点统计学，我们缺的不是数据，而是正确分析数据的路径，从海量数据中撷取有用信息、产生新价值，甚至用以推估未知的事物，并且已经成为个人和企业的关键竞争力。这是一本关于统计轻知识的书，作者希望借助轻松幽默的语言来激发读者对统计学的学习热情。内容从描述性统计到推断性统计，通过将生活中有趣的事件一一展开，了解统计学中的核心知识点，最后是常见疑问的答疑汇编。本书偏重于对案例和图表的引用，不会过多关注于数学推导。

本书主要针对未曾学习过统计学或初学统计学并对此有兴趣的读者，以及希望通过学习相关知识补充数据分析技能的在职人士。

图书在版编目（CIP）数据

从零开始学统计 / 归璐编著. —北京：电子工业出版社，2017.1
ISBN 978-7-121-30165-0

Ⅰ. ①从… Ⅱ. ①归… Ⅲ. ①统计学－基本知识 Ⅳ. ①C8

中国版本图书馆 CIP 数据核字（2016）第 254842 号

责任编辑：黄爱萍
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
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：11.5　字数：180 千字
版　　次：2017 年 1 月第 1 版
印　　次：2017 年 1 月第 1 次印刷
定　　价：45.00 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。

学统计的理由

Hi，亲！很高兴遇见你，虽然你看不到我，我也无法目睹你的容颜，但当你翻开这本书的时候，我们就已经通过文字这个载体见面了！

我猜你应该是被本书的标题吸引才会翻开它的吧？那么聪明的你应该知道，这是一本关于统计学的图书。统计学是一门有趣而实用的学科，它将会成为你生活、工作中的好帮手（别告诉我你不炒股、不玩微博、不买彩票，甚至不逛淘宝，你以为我会告诉你这些都和统计有关吗？）。

- 想知道为什么不能赌博吗——学统计吧！
- 想知道为什么淘宝总能“猜透你的心”吗——学统计吧！
- 想知道怎样才能获得升职加薪的捷径吗——学统计吧！

你有没有想过买一张福利彩票，然后被五百万元大奖砸中？我就想过，那通常发生在大白天，两眼呆滞且目光涣散，幻想自己抱着一堆红色的人民币傻乐……但是当我回过神来之后，我就清楚地意识到中大奖的机会微乎其微——这是概率论教会我的。

你也许会想：这是我小时候就懂的道理，你还要读了概率论才知道？

要知道，概率论诞生于赌博游戏。一两次的小赢，甚至接连几次

都赢是有可能发生的，这属于概率的正常波动。其实，如果在完全公平的情况下，输赢概率应该各为50%。但为什么总感觉赌的时间越长，越容易输呢？这是因为我们忽视了一个重要的因素，那就是输赢各半的前提是可以进行**无限多次**的赌博，但事实上我们根本不可能有那么多的资金和精力。要知道，得出抛硬币正反面出现概率各为50%的结论，是建立在上万次试验结果之上的。所以，你若知道**概率还蕴含积分**的数学思想，就不难理解为何“十赌九输”了。

你有没有想过，“万能”的淘宝为何总能在你搜索宝贝的时候顺便推送一些名为“猜你喜欢”的产品，而且这些推送有时还能被你成功加入购物车？其中就用到了推荐算法。推荐算法不仅涉及文本挖掘技术，而且与统计学中频率的计算和关联性知识有紧密联系。

在我们的日常工作中，如果你从事的是销售、财务工作，或者你是某项目的策划者，当领导询问你对即将上架的产品，或者要削减某项开支，或者某项目的推广方案的看法时，你该如何回答？

如果你对自己所做的工作有过翔实的数据采集，例如，对需要销售的产品做过统计，就可以得出一系列图表来证明该产品在某个时间段或针对某些特殊人群有明显的销量提升（这通常涉及方差分析）；再如，你对公司的财务数据做了详细的台账记录，则可以清楚地知道缩减哪些开支既不影响生产销售又可以提高营业利润（这时可以运用相关分析）；又如，你使用定量方法将推广方案的定性数据量化，通过分析得出最佳方案。试着使用数据来说话，慢慢培养统计思维，你会发现，你的工作将会事半功倍。

生命和统计息息相关

如果上述例子无法给你学习统计的充分理由，那么，当数据和生命联系在一起时，会是怎样的呢？

手术中，麻醉师的用药剂量与病人的个体情况有着严格的匹配要求；新药物上市之前，必须经过无数次试验检验；用药说明书上的剂量指导，更是建立在海量试验检验基础之上的。其中就涉及抽样调查、假设检验和实验设计等多种统计学的理论知识。

不久前，“雾霾致癌吗”这个话题异常火爆。关于这个命题的真伪，在此不做评述，但众所周知，吸烟是有害健康的，吸烟致癌也被大家广为接受。但你知不知道，“吸烟是否是引起肺癌的原因”这个论题曾经在统计学界掀起了轩然大波？当时，费希尔（统计学界的泰斗级人物）极力反对这个观点，其实，在证明吸烟与肺癌关系的过程中，更值得讨论的是对于试验的设计和流行病医学里的因果关系的论证。直到目前，仍然没有一种有效的方法能够证明统计学和哲学双层面的因果关系。但随着统计学的飞速发展，医学统计逐渐流行起来，并发展成为一门热门学科。

生活中的每一部分都和统计密切相关

当一门学科发展到可以通过量化数据来解密人体科学的时候，还能说它不值得去学习了解吗？比如，在大数据时代，如果你不会两个统计名词，怎能充分利用大数据的价值？从事金融行业的不会数据分析，不能跑代码，怎么体现你的专业素养？如果没听说过什么是Hadoop/R/SAS，你怎么做合格的程序员？还有机器学习、词频分析、文本挖掘、数据挖掘……所有这些都离不开统计理论的支撑。所以，

如果你想走在时代的前沿，就抓紧时间学统计吧！

当然，即使有千万个学习统计的理由，但总有一个理由会让你拒绝学习，那就是数学！你不热爱数学，所以拒绝学习和数字有关的学科。但是，这并不能成为你不学习统计的理由，因为统计和数学并不相同。我认为，统计学就是“高冷”数学和深奥哲学的平衡点。

其实，我天生对数学也没有兴趣，丝毫看不出那些积分符号优美在何处。但是这并不能阻碍我对统计学的热爱。诚然，统计理论是完全建立在数学基础上的，数理统计对数学的要求很高，但是统计学里还有一个分支叫应用统计，本书就是为了应用而生的。

本书不会有繁冗的数学公式推导，不过在有些时候，为了说清楚问题，数学公式和定理是不可或缺的。水平有限，力争通过通俗易懂的语言让大家明白统计是怎么回事，以及统计可以用来做些什么。

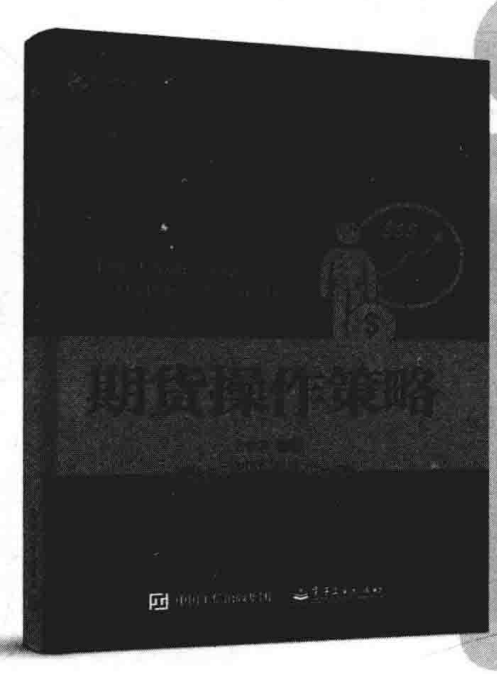
你不用惧怕巨大的计算量，这些都可以通过软件来完成。喜欢编程并想深入研究理论知识的，可以使用 Stata、SAS、R；想要快速解决问题的，可以使用 SPSS；甚至可以使用 Excel 完成绝大多数统计分析工作。

至此，你应该找不到不学统计的理由了吧？

欢迎大家和我一起进入奇妙的统计学世界！

归　璐

2016 年 12 月 1 日



深度解剖期货投资，

完美呈现优秀的期货交易策略，

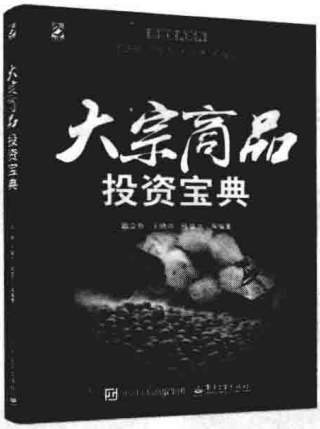
正确的期货投资理念，

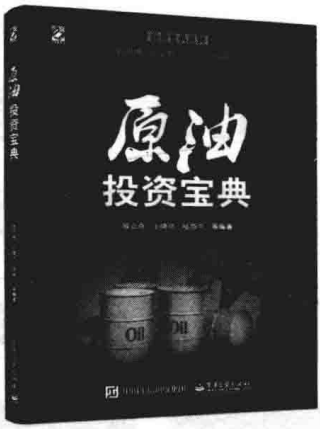
以及作者独创的金大定律！

知行合一 开卷有益 期货之惑

博弈之趣 大道至简 期货江湖

电子工业出版社好书分享

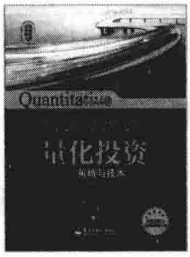




整套丛书洞见贵金属投资、原油交易、外汇投资、大宗商品交易大趋势，从技术面到消息面，从短线投资到长线投资，以完整严密的知识体系、幽默诙谐的语言，传授实用的操作技巧！

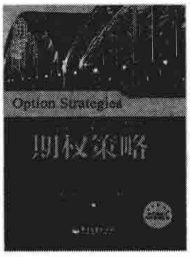
每本书都包含近百经典真实案例，让读者一眼透过现象看到本质，举一反三，活学活用！

量化投资与对冲基金丛书



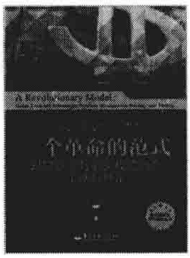
《量化投资——策略与技术（典藏版）》
丁鹏 编著
ISBN 978-7-121-24062-1
定价：118.00元

《量化投资：以MATLAB为工具》
李洋 郑志勇 编著
ISBN 978-7-121-24062-1
定价：69.00元

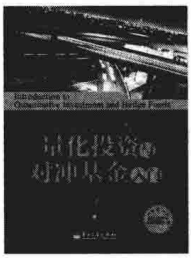


《期权策略》
[美]林万佳 卢扬洲
杨 威 黄彬金 著
ISBN 978-7-121-23617-4
定价：59.00元

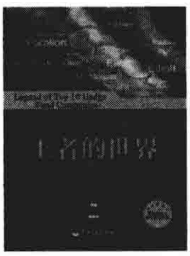
《机器学习在量化投资中的应用研究》
汤凌冰 著
ISBN 978-7-121-24494-0
定价：59.00元



《一个革命的范式——对冲基金与另类投资组合策略与理论》
对冲网阿尔法研究中心 组编
卢扬洲 等编著
ISBN 978-7-121-20189-9
定价：65.00元

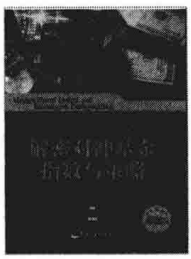


《量化投资与对冲基金入门》
丁鹏 著
ISBN 978-7-121-22292-4
定价：59.00元



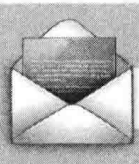
《王者的世界——全球十大对冲基金公司传奇》
对冲网阿尔法研究中心 组编
卢扬洲 黄振 等编著
ISBN 978-7-121-19090-2
定价：59.80元

《网格交易法——数学+传统智慧战胜华尔街》
林万佳 著
ISBN 978-7-121-19056-8
定价：59.80元



《解密对冲基金指数与策略》
对冲网阿尔法研究中心 组编
卢扬洲 李凌飞 等编著
ISBN 978-7-121-19088-9
定价：65.00元

《保本投资法——不跌的股票》
林万佳 著
ISBN 978-7-121-19057-5
定价：55.00元



反侵权盗版声明

目　录

第 0 章

入门阶段——带你迈入统计学的大门

0.1 我和统计学的从零开始

既然书名是《从零开始学统计》，那么本书的目录自然也从第 0 章开始。0 意味着起点，在我们开始系统地了解统计学之前，先来听我讲讲我和统计学之间的故事。

我和统计学的相识是一场美丽的意外。在选择统计学专业之前，我对统计的了解仅限于求平均数、求方差。如果说得再深奥一点，那么还能略微扯上一些概率论。对于学了统计学将来能做什么，我也是一知半解。是什么原因让我选择了这个在当时略显生僻的专业呢？原因很简单——好奇。

“统计”一词起源于国情调查，最早意为国情学。首先来看看“统”字的含义。“统”字可以作三种解释：（1）充满、充盈；（2）总括，

总起来，如统一、统帅等；（3）事物的连续关系，如系统、传统等。从中可以看出，统计学的“统”更倾向于后两种解释。“计”为核算之意。那么两者相结合，表示对总体的核算和对事物连续关系的计量。结合日常生活，一些工作偏向于总体的核算，如对宏观经济数据的披露；而现如今一些职业如 Data Scientist 则需要统计学的专业背景，且更倾向于事物连续关系的挖掘。两者有一定的共性，归结起来就是统计的定义：对数据进行收集和整理，并在此基础上加以分析和科学决策。至于怎么收集和整理数据、怎么分析和决策，将在本书的后续章节详细介绍。

客观地说，数学功底好对于学习统计学大有益处，但这并不能保证你一定能够学好统计学。以笔者的经验来看，**统计学真正迷人的地方在于统计方法和统计思想**。在很多优秀的统计学著作里，通常看不到长篇大论的数学证明，有些甚至放在附录中，正文则更多地阐述数据处理方法的创新，以及建模和算法的创新。

为什么说数学好未必能学好统计学呢？首先，**数学讲究严密的逻辑演绎，而统计学则更多的是归纳推理**。比如，通常人们认为，统计结论都应该建立在数据服从正态分布的基础之上，但很多数据仅仅是近似服从。这么宽泛的条件，怎么能得到让人信服的结论？笔者试图用大数定律和中心极限定理来验证结论的可信度，但事与愿违。其中的矛盾就在于统计学往往更注重应用。在实际应用中，数据是无法达到完美的理论要求的，适当地放宽和采用近似方法反而更能够接近真相。

其次，市面上**种类繁多的统计软件**，让那些不擅长数学的人也

可以掌握统计学的知识。常用的统计软件有：龙头老大——SAS；后起之秀——R；新手福音——SPSS；擅长面板数据计量分析的 Stata/MATLAB；计量入门小能手 Eviews；数据挖掘方面也有 Clementine、Python 等。

如果你不想深入研究，只想利用统计学来解决一些非统计专业领域的难题，那么，大可不必选择高深的软件，拥有菜单操作的 SPSS 甚至 Excel 都可以满足你的统计需求。是的，只需轻轻地单击一下，结果自然呈现。但前提是你必须知道结果的含义，也知道如何选择正确的统计方法。

但如果你想要专业一些，那么还是需要学习 R、SAS 和 Python 的。R、SAS、Python 是目前比较热门的软件，通常金融类企业需要处理海量数据，SAS 使用频繁，而且较为权威；R 是免费开源的，包含各类程序包，所以现在很多分析公司也会采用 R 作为主要软件，也有很多编程爱好者喜欢研究 R，如果你的工作偏向于数据分析类，那么 SAS 和 R 可以任取其一；如果你的工作偏向于数据挖掘方向，那么可以考虑选择 Python，它的应用面非常广。

学习统计软件的过程不仅仅是为了简化运算，也不单单是为了建模。笔者之所以喜欢统计，很大一部分原因在于在学习这些软件的同时加深了对统计思想的理解。笔者通常会把数据在各类统计软件里执行一遍，看结果会有何不同；也会试着用不同的检验方法检验同样的数据，如使用参数检验和非参数检验，再来对比一下结果有何不同。尤其是在进行多元统计分析的时候，如进行聚类分析，不同的数据处理方法会带来完全不同的结果。这类小实验给笔者的统计学习带来很