“十三五”全国统计规划教材

（第2版）

Python 数据分析基础

阮 敬 编著

中国统计出版社
China Statistics Press

“十三五”全国统计规划教材

（第2版）

Python 数据分析基础

阮 敬 编著

中国统计出版社
China Statistics Press

内容简介

本书通过真实案例，全面介绍python3编程基础及其数据分析工具的应用，培养读者通过数据提出问题、分析问题、解决问题以及对分析结果评价的能力。全书内容包括：python3基本配置和编程基础、面向对象编程及并行处理、数据预处理、数据描述与可视化、统计推断、相关分析、关联分析、回归分析、主成分和因子分析、聚类、判别与分类、列联分析、对应分析、定性数据分析、时间序列分析、神经网络与深度学习等，将数据分析工作中的基本理论、方法和应用进行深入剖析。

图书在版编目(CIP)数据

Python数据分析基础 / 阮敬编著. —— 2版. —— 北京：中国统计出版社，2018.8

“十三五”全国统计规划教材

ISBN 978-7-5037-8614-3

Ⅰ. ①P… Ⅱ. ①阮… Ⅲ. ①软件工具-程序设计-高等学校-教材 Ⅳ. ①TP311.561

中国版本图书馆CIP数据核字(2018)第190456号

Python数据分析基础(第2版)

作　　者/阮　敬
责任编辑/姜　洋
封面设计/张　冰
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲6号　邮政编码/100073
电　　话/邮购(010)63376909　书店(010)68783171
网　　址/http://www.zgtjcbs.com
印　　刷/河北鑫兆源印刷有限公司
经　　销/新华书店
开　　本/787×1092mm　1/16
字　　数/550千字
印　　张/31
版　　别/2018年8月第2版
版　　次/2018年8月第1次印刷
定　　价/128.00元

国家统计局
全国统计教材编审委员会第七届委员会

主任委员： 宋跃征

副主任委员： 叶植材　许亦频　赵彦云　邱　东　徐勇勇　肖红叶
耿　直

常务委员（按姓氏笔划排序）：
万东华　叶植材　许亦频　李金昌　杨映霜　肖红叶
邱　东　宋跃征　陈　峰　周　勇　赵彦云　耿　直
徐勇勇　徐　辉　郭建华　程维虎　曾五一

学术委员（按姓氏笔划排序）：
方积乾　冯士雍　刘　扬　杨　灿　肖红叶　吴喜之
何书元　汪荣明　金勇进　郑京平　赵彦云　柯惠新
贺　铿　耿　直　徐一帆　徐勇勇　蒋　萍　曾五一

专业委员（按姓氏笔划排序）：
万崇华　马　骏　王汉生　王兆军　王志电　王　彤
王学钦　王振龙　王　震　尹建鑫　石玉峰　石　磊
史代敏　冯兴东　朱启贵　朱建平　朱　胜　向书坚
刘玉秀　刘立丰　刘立新　米子川　苏为华　杜金柱
李　元　李金昌　李　勇　李晓松　李　萍　李朝鲜
杨仲山　杨　军　杨汭华　杨贵军　吴海山　吴德胜
余华银　宋旭光　张　波　张宝学　陈　峰　林华珍
罗良清　周　勇　郑　明　房祥忠　郝元涛　胡太忠
洪永淼　夏结来　徐国祥　郭建华　唐年胜　程维虎
傅德印　虞文武　薛付忠

出版说明

全国统计教材编审委员会成立于1988年，是国家统计局领导下的全国统计教材建设工作的最高指导机构和咨询机构。自编审委员会成立以来，分别制定并实施了“七五”至“十三五”全国统计教材建设规划，共组织编写和出版了“七五”至“十二五”六轮“全国统计教材编审委员会规划教材”，这些规划教材被全国各院校师生广泛使用，对中国的统计和教育事业作出了积极贡献。自本轮规划教材起，“全国统计教材编审委员会规划教材”更名为“全国统计规划教材”，将以全新的面貌和更积极的精神，继续服务全国院校师生。

《国家教育事业发展“十三五”规划》指出，要实行产学研用协同育人，探索通识教育和专业教育相结合的人才培养方式，推动高校针对不同层次、不同类型人才培养的特点，改进专业培养方案，构建科学的课程体系和学习支持体系。强化课程研发、教材编写、教学成果推广，及时将最新科研成果、企业先进技术等转化为教学内容。加快培养能够解决一线实际问题、宽口径的高层次复合型人才。提高应用型、技术技能型和复合型人才培养比重。

《“十三五”时期统计改革发展规划纲要》指出，“十三五”时期，统计改革发展的总体目标是：形成依靠创新驱动、坚持依法治统、更加公开透明的统计工作格局，逐步实现统计调查的科学规范，统计管理的严谨高效，统计服务的普惠优质，统计运作效率、数据质量和服务水平明显提升，建立适应全面建成小康社会要求的现代统计调查体系，保障统计数据真实准确完整及时，为实现统计现代化打下坚实基础。

围绕新时代中国特色社会主义教育事业和统计事业新特点，全国统计教材编审委员会将组织编写和出版适应新时代特色、高质量、高水平的优秀统计规划教材，以培养出应用型、复合型、高素质、创新型的统计人才。

2015年9月，经李克强总理签批，国务院印发了《促进大数据发展行动纲要》系统部署大数据发展工作，我国各项工作进入大数据时代，拉开了统计教育和统计教材建设的大数据新时代。因此，在完成以往传统统计专业规划教材的编写和出版外，本轮规划教材要把编写大数据内容统计规划教材作为重点工作，以培养新一代适应大数据时代需要的统计人才。

为了适应新时代对统计人才的需要，组织编写出版高质量、高水平教材，本轮规划教材在组织编写和出版中，将坚持以下原则：

1. 坚持质量第一的原则。本轮规划教材将从内容编写、装帧设计、排版印刷等各环节把好质量关，组织编写和出版高质量的统计规划教材。

2. 坚持高水平原则。本轮规划教材将在作者选定、选题编写内容确定、编辑加工等环节上严格把关，确保规划教材在专业内容和写作水平等各方面，保证高水平高标准，坚决杜绝在低水平上重复编写。

3. 坚持创新的原则。无论是对以往规划教材进行修订改版，还是组织编写新编教材，本轮规划教材将把统计工作、统计科研、统计教学以及教学方法、方式的新内容融合在教材中，从规划教材的内容和传播方式上，实行创新。

4. 坚持多层次、多样性规划的原则。本轮规划教材将组织编写出版专科类、本科类、研究生和职业教育类等不同层次的统计教材，并可以考虑根据需要组织编写社会培训类教材；对于同一门课程，鼓励教师编写若干不同风格和适应不同专业培养对象的教材。

5. 坚持教材编写与教材研讨并重的原则。本轮规划教材将注重帮助院校师生学习和使用这些教材，使他们对教材中一些重要概念进一步理解，使教材内容的安排与学生的认知规律相符，发挥教材对统计教学的指导作用，进一步加强统计教材研讨工作，对教材进行分课程的研讨，以促进统计教材的向前发展。

6. 坚持创品牌、出精品、育经典的原则。本轮规划教材将继续修订改版已经出版的优秀规划教材，使它们成为精品，乃至经典，与此同时，将有意识地培养优秀的新作者和新内容规划教材，为以后培养新的精品教材打下基础，把“全国统计规划教材”打造成国内具有巨大影响力的统计教材品牌。

7. 坚持向国际优秀统计教材学习和看齐的原则。不论是修订改版教材还是新编教材，本轮规划教材将坚持与国际接轨，积极吸收国内外统计科学的新成果和统计教学改革的新成就，把这些优秀内容融进去。

8. 坚持积极利用新的教学方式和教学科技成果的原则。本轮规划教材将积极利用数据和互联网发展成果，适应院校教学方式、教学方法以及教材编写方式的重大变化，立体发展纸介质和利用数据、互联网传播方式的统计规划教材内容，适应新时代发展需要。

总之，全国统计教材编审委员会将不忘初心，牢记使命，积极组织各院校统计专家学者参与编写和评审本轮规划教材，虚心听取读者的积极建议，努力组织编写出版好本轮规划教材，使本轮规划教材能够在以往的基础上，百尺竿头，更进一步，为我国的统计和教育事业作出更大贡献。

国家统计局

全国统计教材编审委员会

第2版前言

为适应数据科学与大数据技术领域的飞速发展，《Python数据分析基础》第2版经过将近1年时间的广泛教学实践和市场检验，终于面世并顺利入选“‘十三五’全国统计规划教材”。在保留第1版全部优点和特色的基础上，第2版做了许多优化、改进和创新，具体内容如下：

1. 全书基于python 3.6.4对全部内容进行了更新。

2. 将第1版的编程基础部分根据教学难度和教学要求调整为两个章节。即，第1章强调编程基础，第2章强调编程的高级技能，并补充了类特性、异常捕获与容错处理、并行计算等编程的进阶内容。

3. 增加了“神经网络与深度学习”章节。深度学习是当前数据科学、人工智能领域较为热门的研究内容，第2版增加了对神经网络和深度学习基本思想、基本框架以及基本步骤的介绍，以及如何利用python提供的tensorflow框架工具进行解决实际问题的案例，帮助读者理解深度学习的理论基础和基本算法。

4. 可读和易用性进一步提高。本书第1版在去年9月份正式出版之后，被全国几十所高等院校采纳为基础课、专业课和选修课的教材。经过多次与授课教师和学生的沟通交流及意见反馈，第2版针对教学过程中的突出问题进行了仔细斟酌和调整，力争使得本书内容更加生动、深入浅出和言简意赅。

5. 第2版除了继续提供书中案例数据（请访问中国统计出版社官方网站www.zgtjcbs.com下载），同时还提供python编程基础和编程进阶章节的课程课件PPT，请教师读者联系作者索取。

Email：ruanjing@msn.com

读者交流群：

阮　敬

2018年7月22日于洛杉矶

第1版前言

数据分析是科学研究中的重要环节，随着大数据时代的迅猛发展，其越来越受社会和市场的重视，是科学研究、经营管理、预测与决策等过程中必不可少的基础工作。Python是当今大数据时代下最为流行的编程工具之一，在大数据领域有着十分广泛的应用，可以实现从数据收集和数据管理到数据分析和挖掘的完整过程，其高效的编程和程序执行过程，能够完全胜任日常数据分析工作的需求。

随着数据分析作用的日益凸显，如何对现有数据进行整理、加工、处理和分析，以期得到所谓的结论，作为人们进行决策的依据进而实现数据的价值？如何利用现有数据对将来可能出现的数据结果或结论进行判断或预测？不管是针对企事业单位的管理者或决策者还是从事具体数据分析的工作人员而言，都需要进行合理数据分析流程的规划，区分数据类型，利用适合的数据分析方法，使用方便、快捷、可靠的统计软件作为工具，对特定数据进行分析与预测，从而洞察市场动向，观测人心所在，把握商机，提升竞争力。

而具有深厚数学背景的统计分析和数据分析方法往往会成为相关人员继续深入学习的门槛，甚至成为枯燥乏味的代名词，无法体验到数据分析成果带来的成效。本书就是要力求降低学习难度，通过编者积累的大量真实案例和数据，主要以文字阐述替代复杂公式推导，深入浅出剖析数据分析方法的基本原理和步骤，重点在于厘清数据分析的基本思路，合理得到恰当的分析结果。在分析过程中，本书基于python 2.7，从基础编程入手，主要通过调用python基本库和常用工具库的方式，用大量的实例来展示数据分析每一步骤的细节，带领读者走入数据分析的奇妙世界。

本书的第1章和第2章主要介绍python的基本环境、编程基础和数据预处理方面的内容，具体内容包括python数据类型及数据结构、语句与控制流、基本库、函数和面向对象编程的基础，以及数据分析最为常用的基本分析工具库numpy和pandas基础等；

第3章和第4章主要介绍利用python进行描述分析的基本过程和方法，涵盖了各种常用数据分析图形的绘制和解读以及统计量和统计表等具体内容；

第5、6、7章主要介绍利用python如何进行总体推断。在大数据时代即使数据量再大，但也离不开利用统计思想对总体特征进行推测和判断，这些具体内容包括参数估计、假设检验和非参数分析；

第8章主要介绍如何用python来分析数据之间的关系，具体涵盖了简单相关分析、非参数相关分析、偏相关分析、点二列相关分析以及数据挖掘中常用的关联分析等内容；

第9章和第10章主要介绍如何利用python来进行回归分析。回归模型可以说是大部分统计分析和数据挖掘方法的基础，本书介绍的具体内容有线性回归、非线性回归、多项式回归、分位数回归、自变量含有定性变量的回归以及因变量含有定性变量的广义线性回归分析；

第11章和第12章主要就日常数据分析中所使用的多元统计分析方法进行介绍，具体内容包括主成分分析、因子分析、列联分析以及对应分析等；

第13章和第14章主要介绍在python中进行数据挖掘所使用的聚类和分类方法。内容涵盖系统聚类、*k*-means聚类、DBSCAN聚类、距离判别和线性判别、贝叶斯判别以及数据挖掘中的*k*-近邻、决策树、支持向量机和随机森林等分类方法；

第15章主要介绍python中使用ARIMA建模进行时间序列分析的基本方法和思路。

本书以实用为主要目的，因此上述大部分的数据分析过程均会调用现有常用且公认的结果较为合理的工具库（如numpy、pandas、matplotlib、scipy、statsmodels、scikit-learn等）。对于本书提及的数据分析方法无法通过调用现成工具库实现的，本书在相应章节中使用python编制了相应的函数或类，以供读者在分析实际问题时调用和复用。读者在复用这些函数或类时，也可根据自身需要对它们进行进一步优化。

全书采用macOS Sierra操作系统下的python 2.7.13和Anaconda 4.3.1的jupyter notebook作为分析环境，希望读者参考本书的内容边做边学习。为了提高学习效果，读者应该自行把本书全部代码在python中一字一句的敲一遍并运行之，故本书不提供电子版程序代码。但为了提高学习效率，本书附送随书案例的全部数据（下载地址：www.zgtjcbs.com）。

本书由本人在原书《实用SAS统计分析教程》（中国统计出版社2013年版）基础上亲自编写完成。开源软件的显著特点大家都懂的。因此，读者可在阅读本书时对照原书进行实际操作，认真体会商业软件和开源软件分析流程和分析结果

的异同。此外，我的研究生杨磊磊和王禹提供了部分分析程序并对全书所编制的程序进行了运行验证。同时感谢中国统计出版社的支持。尽管作者已经投入了大量时间和精力来编写此书，但由于水平有限，如有不足之处，敬请专家与同行批评指正。同时也欢迎广大读者与作者积极联系，共同探讨数据分析方面的心得与体会。

Email：ruanjing@ msn.com

读者交流群：

阮　敬

2017 年 8 月 23 日

目　录