

Python

数据分析基础

阮敬 编著

- 实用性强，强调对统计分析和数据分析知识的领悟和应用。
- 语言通俗，避免用复杂数学公式推导分析所用的基本原理。
- 通过python来实现数据分析和得到结论的全部过程。
- 配有配套案例数据，下载地址：www.zgtjcbbs.com

 中国统计出版社
China Statistics Press

Python

数据分析基础

阮敬 编著

- 实用性强，强调对统计分析和数据分析知识的领悟和应用
- 语言通俗，避免用复杂数学公式推导分析所用的基本原理
- 通过python来实现数据分析和得到结论的全部过程
- 配有配套案例数据，下载地址：www.zgtjcbbs.com

内容简介

本书通过真实案例，全面介绍 python 编程基础和数据分析工具的应用，并培养读者通过数据分析问题、解决问题以及对结果评价的能力。全书内容包括：python 基本配置和编程基础、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、时间序列分析等，将读者关注的数据分析与数据挖掘技术进行剖析。

图书在版编目 (CIP) 数据

Python 数据分析基础 / 阮敬编著. -- 北京 : 中国统计出版社, 2017.9

ISBN 978-7-5037-8320-3

I. ①P… II. ①阮… III. ①软件工具—程序设计
IV. ①TP311.561

中国版本图书馆 CIP 数据核字(2017)第 217456 号

Python 数据分析基础

作者/阮敬

责任编辑/姜洋 王立群

封面设计/李雪燕

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.zgtjcbps.com/>

印刷/河北鑫兆源印刷有限公司

经销/新华书店

开本/787mm×1092mm 1/16

字数/510 千字

印张/27

版别/2017 年 9 月第 1 版

版次/2017 年 9 月第 1 次印刷

定价/89.00 元

版权所有。未经许可，本书的任何部分不得以任何方式在世界任何地区以任何文字翻印、拷贝、仿制或转载。如有印装差错，由本社发行部调换。

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.zgtjcb.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

第 1 章 Python 编程基础	1
1.1 Python 系统配置	1
1.2 Python 基础知识	5
1.2.1 帮助	6
1.2.2 标识符	6
1.2.3 行与缩进	7
1.2.4 变量与对象	7
1.2.5 数字与表达式	9
1.2.6 运算符	10
1.2.7 字符串	11
1.2.7.1 转义字符	11
1.2.7.2 字符串格式化	12
1.2.7.3 字符串的内置方法	13
1.2.8 日期和时间	17
1.3 数据结构与序列	18
1.3.1 列表	19
1.3.1.1 列表索引和切片	19
1.3.1.2 列表操作	20
1.3.1.3 内置列表函数	20
1.3.1.4 列表方法	21
1.3.2 元组	22
1.3.3 字典	23
1.3.4 集合	24
1.3.5 推导式	26
1.4 语句与控制流	27
1.4.1 条件语句	27
1.4.2 循环语句	28
1.4.2.1 while 循环	28
1.4.2.2 for 循环	29
1.4.2.3 循环控制	30
1.5 函数	30
1.5.1 函数的参数	32
1.5.2 全局变量与局部变量	32
1.5.3 匿名函数	33

1.5.4	递归和闭包	33
1.5.5	柯里化与反柯里化	35
1.5.6	常用的内置函数	36
1.5.6.1	filter 函数	36
1.5.6.2	map 函数	36
1.5.6.3	reduce 函数	37
1.6	迭代器、生成器和装饰器	37
1.6.1	迭代器	37
1.6.2	生成器	38
1.6.3	装饰器	40
1.7	类	42
1.7.1	声明类	42
1.7.2	方法	44
1.7.2.1	实例方法	44
1.7.2.2	类方法	45
1.7.2.3	静态方法	46
1.7.3	属性	47
1.7.3.1	实例属性和类属性	47
1.7.3.2	私有属性和公有属性	48
1.7.4	继承	49
1.7.4.1	隐式继承	49
1.7.4.2	显式覆盖	50
1.7.4.3	super 继承	51
1.7.4.4	多态	52
1.7.4.5	多重继承	54
1.8	模块	54
1.9	包	55
1.10	文件 I/O	55
第 2 章	数据预处理	59
2.1	numpy 基础	59
2.1.1	向量	61
2.1.2	数组	62
2.1.2.1	数据类型与结构数组	63
2.1.2.2	索引与切片	64
2.1.2.3	数组的属性	68
2.1.2.4	数组排序	69
2.1.2.5	数组维度	70
2.1.2.6	数组组合	72
2.1.2.7	数组分拆	75

2.1.2.8	ufunc 运算	76
2.1.3	矩阵	81
2.1.4	文件读写	81
2.2	pandas 基础	82
2.2.1	pandas 的数据结构	83
2.2.1.1	Series	83
2.2.1.2	DataFrame	87
2.2.2	pandas 的数据操作	96
2.2.2.1	排序	96
2.2.2.2	排名	98
2.2.2.3	运算	100
2.2.2.4	函数应用与映射	101
2.2.2.5	分组	102
2.2.2.6	合并	103
2.2.2.7	分类数据	106
2.2.2.8	时间序列	107
2.2.2.9	缺失值处理	116
第 3 章	数据描述	122
3.1	统计量	122
3.1.1	集中趋势	122
3.1.1.1	均值	123
3.1.1.2	中位数	124
3.1.1.3	分位数	125
3.1.1.4	众数	125
3.1.2	离散程度	126
3.1.2.1	极差	126
3.1.2.2	四分位差	127
3.1.2.3	方差和标准差	127
3.1.2.4	协方差	128
3.1.2.5	变异系数	128
3.1.3	分布形状	128
3.1.3.1	偏度	129
3.1.3.2	峰度	129
3.2	统计表	130
3.2.1	统计表的基本要素	130
3.2.2	统计表的编制	131
第 4 章	统计图形与可视化	135
4.1	matplotlib 基本绘图	135
4.1.1	函数绘图	135

4.1.2	图形基本设置	140
4.1.2.1	创建图例	140
4.1.2.2	刻度设置	141
4.1.2.3	图像注解	142
4.1.2.4	图像大小	143
4.1.2.5	创建子图	144
4.1.2.6	其他绘图函数	145
4.1.3	面向对象绘图	146
4.1.4	绘图样式	148
4.2	pandas 基本绘图	148
4.3	基本统计图形	150
4.3.1	折线图	150
4.3.2	面积图	153
4.3.3	直方图	153
4.3.4	条形图	155
4.3.5	龙卷风图	158
4.3.6	饼图	159
4.3.7	阶梯图	160
4.3.8	盒须图	161
4.3.9	小提琴图	163
4.3.10	散点图	164
4.3.11	气泡图	166
4.3.12	六边形箱图	167
4.3.13	雷达坐标图	168
4.3.14	轮廓图	169
4.3.15	调和曲线图	169
4.3.16	等高线图	170
4.3.17	极坐标图	170
4.3.18	词云图	171
4.3.19	数据地图	174
4.4	其他绘图工具	176
第 5 章	简单统计推断	178
5.1	常用数据分析工具库	178
5.1.1	scipy	178
5.1.2	statsmodels	179
5.1.3	sklearn	180
5.2	简单统计推断的基本原理	180
5.2.1	数据分布	180
5.2.1.1	总体分布	181

5.2.1.2	样本分布	181
5.2.1.3	抽样分布	181
5.2.2	参数估计	183
5.2.2.1	点估计	184
5.2.2.2	区间估计	184
5.2.3	假设检验	185
5.2.3.1	假设检验的基本思想	185
5.2.3.2	假设检验基本步骤	186
5.2.3.3	假设检验中总体的几种不同情况	187
5.3	单总体参数的估计及假设检验	189
5.3.1	单总体的参数估计	189
5.3.1.1	单总体均值的参数估计	189
5.3.1.2	单总体方差、标准差的参数估计	190
5.3.1.3	单总体比例的参数估计	191
5.3.2	单总体参数的假设检验	191
5.3.2.1	总体均值的假设检验	191
5.3.2.2	总体比例的假设检验	194
5.4	两总体参数的假设检验	194
5.4.1	独立样本的假设检验	195
5.4.1.1	独立样本均值之差的假设检验	195
5.4.1.2	独立样本比例之差的假设检验	197
5.4.2	成对样本的假设检验	198
第 6 章	方差分析	201
6.1	方差分析的基本原理	201
6.2	一元方差分析	205
6.2.1	一元单因素方差分析	205
6.2.1.1	方差同质性检验	206
6.2.1.2	方差来源分解及检验过程	206
6.2.1.3	多重比较检验	207
6.2.1.4	方差分析模型的参数估计和预测	208
6.2.1.5	方差分析模型的预测	210
6.2.2	一元多因素方差分析	210
6.2.2.1	只考虑主效应的多因素方差分析	211
6.2.2.2	存在交互效应的多因素方差分析	215
6.3	协方差分析	217
第 7 章	非参数检验	220
7.1	非参数检验的基本问题	220
7.2	单样本非参数检验	221

7.2.1	中位数(均值)的检验	221
7.2.2	分布的检验	223
7.2.3	游程检验	224
7.3	两个样本的非参数检验	225
7.3.1	独立样本中位数比较的 Wilcoxon 秩和检验	225
7.3.2	独立样本的分布检验	227
7.3.3	成对(匹配)样本中位数的检验	228
7.3.4	两样本的游程检验	228
7.4	多个样本的非参数检验	229
7.4.1	多个样本的分布检验	229
7.4.2	独立样本位置的检验	230
第 8 章	相关分析与关联分析	233
8.1	相关分析	233
8.1.1	函数关系与相关关系	233
8.1.2	简单相关分析	234
8.1.2.1	用图形描述相关关系	234
8.1.2.2	用相关系数测度相关关系	235
8.1.2.3	相关系数的显著性检验	236
8.1.3	偏相关分析	238
8.1.4	点二列相关分析	239
8.1.5	非参数相关分析	240
8.1.5.1	Spearman 相关系数	240
8.1.5.2	Kendall tau-b 系数	241
8.1.5.3	Hoefding's D 系数	241
8.2	关联分析	243
8.2.1	基本概念与数据预处理	243
8.2.2	Apriori 算法	245
8.2.3	FP-growth 算法	249
第 9 章	回归分析	251
9.1	线性回归	251
9.1.1	回归分析的基本原理	251
9.1.1.1	参数估计的普通最小二乘法	253
9.1.1.2	回归方程的检验及模型预测	254
9.1.2	一元线性回归	255
9.1.3	多元线性回归	262
9.1.4	含有定性自变量的线性回归	266
9.2	非线性回归	270
9.2.1	可线性化的非线性分析	270

9.2.2	非线性回归模型	273
9.3	多项式回归	276
9.4	分位数回归	279
第 10 章	离散因变量模型	285
10.1	线性概率模型	285
10.2	二元选择模型	287
10.2.1	线性概率模型的缺陷与改进	287
10.2.2	二元选择模型的基本原理	287
10.2.2.1	模型构建和参数估计过程	288
10.2.2.2	模型检验	289
10.2.3	BINARY PROBIT 模型	289
10.2.4	BINARY LOGIT 模型	293
10.3	多重选择模型	295
10.4	计数模型	298
第 11 章	主成分与因子分析	301
11.1	数据降维	301
11.1.1	数据降维的基本问题	302
11.1.2	数据降维的基本原理	302
11.2	主成分分析	303
11.2.1	主成分分析的基本概念与原理	303
11.2.2	主成分分析的基本步骤和过程	304
11.3	因子分析	313
11.3.1	因子分析的基本原理	313
11.3.1.1	因子分析模型	313
11.3.1.2	因子旋转	314
11.3.1.3	因子得分	314
11.3.2	因子分析的基本步骤和过程	315
第 12 章	列联分析与对应分析	326
12.1	列联分析	326
12.1.1	列联表	326
12.1.2	列联表的分布	329
12.1.3	χ^2 分布与 χ^2 检验	330
12.1.4	χ^2 分布的期望值准则	331
12.2	对应分析	332
12.2.1	对应分析的基本思想	332
12.2.2	对应分析的步骤和过程	333
12.2.2.1	概率矩阵 P	333
12.2.2.2	数据点坐标	333

12.2.2.3	行列变量分类降维	335
12.2.2.4	对应分析图	335
第 13 章	聚类	345
13.1	聚类的基本原理	345
13.1.1	聚类的基本原则	346
13.1.2	单一指标的系统聚类过程	347
13.1.3	多指标的系统聚类过程	349
13.2	聚类的步骤和过程	354
13.2.1	系统聚类	354
13.2.2	K-MEANS 聚类	360
13.2.3	DBSCAN 聚类	361
第 14 章	判别和分类	363
14.1	判别和分类的基本思想	363
14.2	常用判别方法和分类算法	364
14.2.1	距离判别和线性判别	364
14.2.2	贝叶斯判别	371
14.2.3	k-近邻	373
14.2.4	决策树	375
14.2.5	随机森林	380
14.2.6	支持向量机	381
第 15 章	时间序列分析	384
15.1	时间序列的基本问题	384
15.1.1	时间序列的组成部分	384
15.1.2	时间序列的平稳性	386
15.1.2.1	平稳性的含义	386
15.1.2.2	时间序列的零均值化和平稳化	387
15.1.2.3	时间序列的平稳性检验	387
15.2	ARIMA 模型的分析过程	390
15.2.1	ARIMA 模型	391
15.2.1.1	AR 模型	391
15.2.1.2	MA 模型	391
15.2.1.3	ARMA 模型	392
15.2.2	ARMA 模型的识别、估计与预测	392
15.2.2.1	模型的识别	392
15.2.2.2	模型参数估计及检验	395
15.2.2.3	模型的预测	398
附录: 各章图形		401

第1章

Python 编程基础

Python 是一种面向对象的解释型高级编程语言，其结构简单、语法和代码定义清晰明确、易于学习和维护、可移植性和可扩展性非常强。python 提供了非常完善的基础代码库（内置库），覆盖了数据结构、语句、函数、类、网络、文件、GUI、数据库、文本处理等大量内容。用 python 进行数据分析和功能开发，许多功能直接使用现成的包（packages）或模块（modules）即可，极大的提升了效率。除了内置库外，python 还有大量的第三方库，如 numpy、scipy、matplotlib、pandas、statsmodels、sklearn 等主要用于数据分析的库，提供了向量和矩阵的操作、数据可视化、统计计算、统计推断、统计分析与建模、数据挖掘和机器学习等几乎全部数据分析的功能。

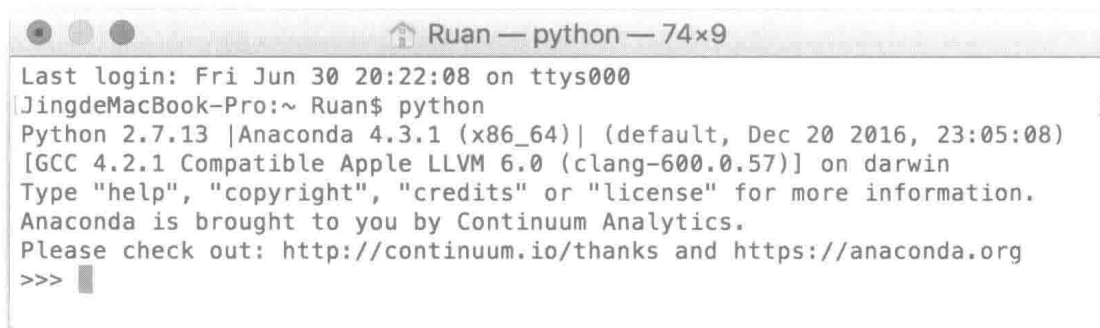
本章将就利用 python 进行统计分析和数据分析的基本功能进行详细介绍，内容包括 python 系统配置以及面向对象的编程方式，涵盖了数据结构、控制流、函数、类、模块与包等具体内容。

1.1 Python 系统配置

Python 可应用于多系统平台，如 Linux、macOS、Windows 等，用户可以直接在其官网：<https://www.python.org>，下载最新版本的 python。用户也可以下载其他发行版本的 python 进行使用，如 anaconda、enthought canopy 等，这些发行版本的 python 已经包含一些特定的分析库，用户可以直接在它们提供的环境中使用。此外，还可以在集成开发环境 IDE（integrated development environment）中配置和使用 python，如 PyCharm、Eclipse 以及 python 自带的 IDLE 等。本书主要介绍 python 数据分析的基础知识，对于这些系统配置问题本书不予赘述，请读者自行查阅相关资料。

本书使用 macOS 10.12.6 系统原生的 python2.7.13 和自行安装的发行版本 Anaconda4.3.1 所搭建的环境。

打开 Mac 系统的应用程序→实用工具→终端，输入命令：python，即可进入 python 的交互式环境，如图 1-1 所示：



```

Last login: Fri Jun 30 20:22:08 on ttys000
JingdeMacBook-Pro:~ Ruan$ python
Python 2.7.13 |Anaconda 4.3.1 (x86_64)| (default, Dec 20 2016, 23:05:08)
[GCC 4.2.1 Compatible Apple LLVM 6.0 (clang-600.0.57)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and https://anaconda.org
>>> █

```

图 1-1 python 的交互式环境

用户可以在提示符“>>>”右边输入 python 命令或者语句。如：

```
>>> print 'Hello, world!'
Hello, world!
```

在如需在 python 中安装第三方的工具库或包 (packages)，可以在终端中使用如下命令：

```
pip install package 的名称
```

在连接有网络的状态下，系统会自动下载对应名称的 package 并将其安装在当前系统环境当中。

macOS 原生版本 python 的交互式编程环境功能较为单一，而 Ipython 可以提供一个综合的交互式编程环境即 notebook，可以提供 tab 补全、富媒体、多客户端连接 kernel、交互式并行计算等，在科学计算和数据分析领域中发挥着极为强大的作用。在终端中可以使用 pip install 安装 ipython 和 notebook：

```
pip install ipython
pip install notebook
```

安装完毕后，在 python 环境中输入如下命令：

```
ipython notebook #或者 jupyter notebook
#注：ipython notebook 现已升级为 jupyter notebook
```

python 中可以使用“#”作为注释，“#”右边的一切内容均不会执行。但是“#”只能对一行内容进行注释，如需对多行内容进行注释，可以在被注释的内容前后加上“'''”或“"""”，所有注释内容在程序运行过程中不会被执行：

```
ax1=fig.add_axes([0.1,0.6,0.2,0.3])
'''
```

这是注释内容：

```
指定子图在图像中的位置坐标，图像的左下角是原点(0,0)
图像横轴方向和纵轴方向总长度都为1，(0.5, 0.5)是图像的中点
'''
```

```
line=ax1.plot([0,1],[0,1]) #绘制一条直线
ax1.set_title('Axes1')
```

运行上述的 ipython notebook 命令后，系统便会打开默认的浏览器进入 ipython notebook (现已更新为 jupyter notebook)，如图 1-2 所示：