Springer

Python 统计分析

[奥] 托马斯·哈斯尔万特(Thomas Haslwanter) 著　李锐 译　张志杰 审

An Introduction to Statistics with Python

With Applications in the Life Sciences

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

Python 统计分析

[奥] 托马斯·哈斯尔万特(Thomas Haslwanter) 著　李锐 译　张志杰 审

人民邮电出版社
北京

图书在版编目（CIP）数据

Python统计分析 /（奥）托马斯·哈斯尔万特（Thomas Haslwanter）著；李锐译. — 北京：人民邮电出版社，2018.12
ISBN 978-7-115-49384-2

Ⅰ. ①P… Ⅱ. ①托… ②李… Ⅲ. ①统计分析－应用软件 Ⅳ. ①C819

中国版本图书馆CIP数据核字(2018)第214278号

版权声明

◆ 著　[奥] 托马斯·哈斯尔万特(Thomas Haslwanter)
译　李　锐
审　张志杰
责任编辑　王峰松
责任印制　焦志炜

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路 11 号
邮编　100164　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京缤索印刷有限公司印刷

◆ 开本：720×960　1/16
印张：15.25
字数：295 千字　2018 年 12 月第 1 版
印数：1－3 000 册　2018 年 12 月北京第 1 次印刷

著作权合同登记号　图字：01-2017-5035 号

定价：79.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

内容提要

本书以基础的统计学知识和假设检验为重点，简明扼要地讲述了 Python 在数据分析、可视化和统计建模中的应用。本书主要包括 Python 的简单介绍、研究设计、数据管理、概率分布、不同数据类型的假设检验、广义线性模型、生存分析和贝叶斯统计学等从入门到高级的内容。

本书利用 Python 这门开源语言，不仅在直观上对数据分析和统计检验提供了很好的讲解，而且在相关数学公式的讲解上也能够做到深入浅出。本书所讲内容的可操作性很强，配套提供相关的代码和数据，方便读者动手练习。

本书适合对统计学和 Python 有兴趣的读者，特别是在实验学科中需要利用 Python 的强大功能来进行数据处理和统计分析的学生和研究人员。

作者简介

托马斯·哈斯尔万特（Thomas Haslwanter）在学术机构中有超过 10 年的教学经验，是林茨上奥地利州应用科学大学（University of Applied Sciences Upper Austria in Linz）医学工程系的教授，瑞士苏黎世联邦理工学院讲师，并曾在澳大利亚悉尼大学和德国图宾根大学担任过研究员。他在医学研究方面经验丰富，专注于眩晕症的诊断、治疗和康复。在深入使用 Matlab 软件 15 年后，他发现 Python 非常强大，并将其用于统计数据分析、声音和图像处理以及生物仿真应用。

译者简介

李锐，复旦大学公共卫生学院流行病与生物统计专业博士生，Python、R 和 Lisp 语言的爱好者，主要研究方向为统计学习和机器学习建模以及组学数据的数据挖掘。先后以第一作者身份发表学术论文 6 篇，其中 SCI 论文 4 篇。参编中文专著 2 本。

审校者简介

张志杰，复旦大学公共卫生学院副教授，多本国际医疗卫生期刊的特邀编辑，研究方向为统计建模和医学领域的统计分析方法。参与并完成国家重大科技专项、“863”计划项目、国家“十五”科技攻关课题、自然科学基金重大项目等多项国家级课题的研究，研究成果先后获 2010 年全国百篇优秀博士学位论文、2012 年上海市医学奖二等奖、上海市科技进步奖二等奖以及中华医学奖三等奖，2011 年入选复旦大学首批“卓学人才计划”，2013 年入选上海市第二批新优青人才计划。

前　言

在我自己的研究工作中，当我进行数据分析时，有两件事情经常让我深陷泥沼：（1）我了解的统计学知识不够多；（2）市面上的书籍大多是理论性的，缺少一些实践上的帮助。你手上拿着的（或者在你的平板电脑或笔记本电脑上的）这本书就是要解决这两个关键问题。这本书将为你提供足够的统计学领域知识，这样你就不会迷失，与此同时，这本书也会教你使用统计学分析所需要的工具。我认为 Python 提供的解决方案能够解决生物学家、物理学家、医学博士们在他们的工作中遇到的 90% 的问题。如果你正在攻读研究生学位，或者是一个正在分析最近实验数据的医学研究者，那么你将会在本书中找到你所需要的工具和它们的使用说明及源代码。

基于上述原因，本书将重点讲解统计学的基础知识和假设检验，并简单地介绍其他的统计学方法。我明白本书中介绍的大多数统计学检验也可以使用统计学建模的方法来完成。但是在大多数情况下，统计学建模并不是生命科学领域的期刊所使用的方法论。高级的统计分析超出了本书的范围，并且坦率地说，也超出了我对统计学的了解。

本书主要使用 Python 语言进行统计学检验和数据分析，主要是基于下面两点原因：首先，我希望这些方法能够被所有人使用。尽管市面上有一些商业的解决方案，比如 Matlab、SPSS、Minitab 等，它们提供了强大的分析工具，但是这些软件大多数只能在学术环境中合法使用[1]，而 Python 则是完全免费的[2]（“像啤酒一样免费”经常能够在 Python 社区中听到）。另一个原因是，Python 是我见过的最优美的编程语言之一，大约在 2010 年，Python 及其文档就发展得较为成熟了，这使得一个业余的编码人员也能够轻松地使用它。配合这本书一起使用，Python 和 Python 生态系统提供的优美又免费的工具包，将覆盖大多数研究者一辈子所需要了解的所有统计学分析。

这本书是为谁写的

这本书的前提条件如下：

- 你有一些基本的编程经验。如果你之前从来没有接触过编程，你最好先从学习 Python 语言开始，在书中我提供了一些非常好的 Python 学习参考链

1　应该是免费的合法使用。——译者注。

2　free 是双关，自由软件也叫 free software。——译者注。

接给你。同时学习编程和学习统计学可能有点拔苗助长了。

- 你不必是一个统计学专家。如果你有高级的统计学分析的经验，那么借助 Python 和 Python 包的在线帮助文档，你马上就能够进行大部分的数据分析。尽管本书会让你熟悉 Python 编程，但主要还是聚焦在统计学的基本概念和假设检验上，只有在本书的最后一个部分才会涉及线性回归建模和贝叶斯统计学等内容。

本书旨在提供所有（至少大部分）你需要的统计学分析工具。我在本书中会提供足够的理论背景知识帮助你明白你正在做什么。如非必要，我不会证明任何的定理或应用数学。对于本书中提到的所有统计学检验，我都会提供一个能够正常运行的 Python 程序。总的来说，你只需要定义好你的问题，选择合适的程序，稍微修改一下程序让它符合你的需求。这样的话，就算你没有太多的 Python 编程经验，你也能快速上手。我并没有提供给你单独的 Python 包，因为我希望你能够根据你自己的需求对每个程序进行修改以适应你的设置（数据类型、自定义的绘图的标签和返回值等）。

全书共分为 3 个部分。

第一部分 简单介绍 Python：如何安装和配置 Python 运行环境，运行一些简单的 Python 程序；为了防止你犯一些常见错误，我们也提供了一些小建议。这部分也会介绍如何从不同的数据源读取数据到 Python 中，并对数据进行可视化。

第二部分 介绍统计学分析：如何进行研究设计，如何分析数据，概率分布的基本知识，概述常见的重要假设检验方法。尽管现代统计学扎根于统计学建模，但是假设检验仍然占据着生命科学领域的主导地位。对于每一个检验方法，我们都会提供一个 Python 程序来展示该检验是如何用 Python 语言完成的。

第三部分 介绍统计建模的知识并简单介绍高级统计分析的步骤。因为 logistic 回归等离散型数据检验方法使用了一种叫作“广义线性模型”的高级统计学方法，所以在这个部分中，也会包含这些内容。在本书的最后，将会展示贝叶斯统计学中的基本概念。

补充材料

随本书出版的还有大量的 Python 程序和示例数据，均可以在网上获取。这些程序包括：所有书中的程序，每章末尾示例的答案，每个检验方法的示例代码。此外还包括书中插图的绘图代码，运行代码所需要的数据，等等。

本书附带的 Python 程序和数据集可以在 Github 代码库上下载（https://github.com/thomas-haslwanter/statsintro_python），所有的材料都可以在 http://www.springer.com/de/book/9783319283159 下载。

致谢

Python 是由许多用户社区做出的贡献组成的，本书中的一些章节也基于互联网上的优秀信息（已获得作者同意在本书中引用他们的内容）。

我要特别感谢下面这些人。

- Paul E Johnson 阅读了全书书稿，并对全书的内容组织和统计细节方面提出了很有价值的反馈建议。
- Connor Johnson 写了篇博客解释了 statsmodels OLS（最小二乘法）命令的结果，这篇博客是本书“统计学模型”的基础。
- Cam Davidson Pilon 写了一本名叫《Probabilistic-Programming-and-Bayesian-Methods-for-Hackers》的开源电子书，我从中借鉴了 Challenger disaster 的例子用来阐述贝叶斯统计学。
- 多亏了 Fabian Pedregosa 的一篇关于有序 logistic 回归的博客，让我在本书中加入了相关的内容，因为我对这部分内容并不熟悉。

我还想感谢 Carolyn Mayer，他阅读了我的手稿并将一些口语化的语言润色为正式的书面语。此外我还要特别感谢我的妻子，她不仅对全书的组织结构提出了重要的建议，还提出了很多编程教学的小技巧，并且本书中和茶有关的主题都有她的支持和协助。

如果你有任何建议或者勘误，请给我的工作邮箱（thomas.haslwanter@fh-linz.at）发电子邮件。除非另有通知，如果我基于你的反馈建议对本书做出了改变，我会将你加入到贡献者名单中。如果你愿意附上错误出现位置的句子，哪怕只有一部分，也能让我更容易定位错误。页码和章节名也很好，但不如句子容易定位。非常感谢！

Thomas Haslwanter
奥地利林茨
2015 年 12 月

缩　写

ANOVA	方差分析
CDF	累积分布函数
CI	置信区间
DF/DOF	自由度
EOL	行末
GLM	广义线性模型
HTML	超文本标记语言
IDE	集成开发环境
IQR	四分位数间距
ISF	逆生存函数
KDE	核密度估计
MCMC	马尔可夫链蒙特卡洛
NAN	不是数字
OLS	普通最小二乘法
PDF	概率密度函数
PPF	百分比点函数
QQ-Plot	分位数—分位数图
ROC	受试者操作特征
RVS	随机变数样本
SD	标准差
SE/SEM	（均值的）标准误
SF	生存函数
SQL	结构化查询语言
SS	平方和
Tukey HSD	Tukey 显著差异检验

资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

配套资源

本书提供如下资源：

- 书中彩图文件。

要获得以上配套资源，请在异步社区本书页面中单击 配套资源，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，单击“提交勘误”，输入勘误信息，单击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。

扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“**异步社区**”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

“**异步图书**”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

异步社区

微信服务号

目 录