应用语言学译丛

语言研究中的统计学

——R软件应用入门

〔德〕斯蒂芬·托马斯·格莱斯 著

商務印書館
The Commercial Press
创于1897

应用语言学译丛

语言研究中的统计学

——R软件应用入门

〔德〕斯蒂芬·托马斯·格莱斯 著

韦爱云 译

李德高 审校

商務印書館 The Commercial Press 创于1897

2018年·北京

图书在版编目(CIP)数据

语言研究中的统计学:R 软件应用入门/(德)斯蒂芬·托马斯·格莱斯著;韦爱云译.—北京:商务印书馆,2018
(应用语言学译丛)
ISBN 978-7-100-16178-7

Ⅰ.①语… Ⅱ.①斯… ②韦… Ⅲ.①语言统计—统计分析—应用软件 Ⅳ.①H0-05

中国版本图书馆 CIP 数据核字(2018)第 111763 号

应用语言学译丛
语言研究中的统计学
——R 软件应用入门
〔德〕斯蒂芬·托马斯·格莱斯 著
韦爱云 译
李德高 审校

商 务 印 书 馆 出 版
(北京王府井大街 36 号 邮政编码 100710)
商 务 印 书 馆 发 行
北京市艺辉印刷有限公司印刷
ISBN 978-7-100-16178-7

2018 年 9 月第 1 版 开本 787×960 1/16
2018 年 9 月北京第 1 次印刷 印张 23½
定价:62.00 元

Stefan Th. Gries

STATISTICS FOR LINGUISTICS WITH R

A PRACTICAL INTRODUCTION

2nd revised edition

《应用语言学译丛》

编辑出版委员会

To Pat, the most supportive Department Chair I could have ever wished for.

——献给 Pat，最支持我的系主任

中文版序言

在过去15年左右的时间里，语言学这门学科似乎比以往任何时候都更多地采用定量统计方法。除了像心理语言学、计算语言学，可能还可以算上社会语言学等这些一直比较倾向使用定量统计方法的语言学分支学科外，越来越多的分支学科以及理论框架在进行问题研究和数据处理时也都开始使用统计方法了。但是，尽管有了上述发展，目前这方面的入门书籍仍显短缺，多数语言学研究者在研究生阶段都未能接受定量研究方法的相关训练，从而难以跟上学科的发展。正是在这种状况下，Baayen（2008）、Johnson（2009）、Gries（2009，2013）和 Levshina（2015）面世了。

感谢您使用这本书。德古意特出版社联系我，打算出版此书第二版，我欣然着手准备了。与此同时，这本书也被翻译成其他语言出版——目前朝鲜语译本已经出版，巴西葡萄牙语译版也即将完成。当刘海涛教授提出希望由商务印书馆出版韦爱云女士翻译的汉语译版时，我兴奋万分。现在汉译本也面世了。

我很自豪地告诉大家，在过去几年中我收到过此书的诸多好评，其中有些评价来自包括中国内地和香港在内的全球各地举办的工作坊培训班的学者们，有些是通过电子邮件形式传递的。因此，我希望您也能像其他版本的读者一样从中受益。

在此我要重申一些内容——虽然书中已经提及，但其重要性再怎么强调也不过分：这是一本实用操作的入门书，从第二章开始，您需要自己动手操作而不仅仅是阅读。如果想要从中受益，则需要从头到尾熟练操作，运行从随附网页上下载的示例代码，好好使用书中包含的数据集，同时结合您自己的样本数据，尝试操练应用每一段代码。

总而言之，学习统计学和通过使用 R 做统计分析，从某种程度上就像学习一门语言——一些语言学家或语言学教师目前可能还不太能接受这个不太完美的类比：(1)R 函数有点儿像生词：需要您在实际的数据环境和情境中使用、操练并学习；(2)R 函数又有些像动词：首先，这类函数需要参数，这些参数像词序一样按一定的顺序排列；其次，它们根据情境发生变化，就像动词根据不同情境使用不同的屈折/派生语素一样。因此，要熟练掌握使用这些函数，您需要花时间动手操作。不能指望通过简单阅读相关资料就精通一门新语言，就像我们不能期望只在几个星期里不定期地操练一下就能轻易熟练掌握一种新技能一样，所以，我们也不要期望仅用几个星期就能把自己培养成统计分析专家。但是，如果认真考虑我前面关于动手操作的提议，相信经过一段时间，您将培养起使用数据的能力，做许多意想不到的事情，比如，当你打开数据，能发现那些肉眼无法看到的模型(这些模型即便有再好的直觉也不可能察觉到)。这时您就会发现这一切都是值得的。学习 R 的知识可以帮助您领会和理解语言数据，甚至还能从一些新的视角体味科学和科学论断的旨趣，因此，给自己一个机会坚持下去，充分利用这本书。在此祝您一切安好，研究进展顺利。

斯蒂芬·托马斯·格莱斯(Stefan Th. Gries)

2017 年 7 月

前　言

本书是 Gries (2009b)第一版的修订和扩展。和第一版相比,本书主要有四个方面的变化。一是对第 5 章的全面调整。几十期用 R 进行语言学统计的培训班之后,我意识到,对初学者来说,回归建模部分最难的内容是对建模过程逻辑的理解、对数据结果的诠释以及如何使建模形象化且富有启迪性。因此,所有关于回归建模的内容都是重新编写的。为了便于理解,还增加了许多新的内容。

二是更新了第 1 章和第 4 章的相关内容。新版本第 1 章中的单侧和双侧检验概念采用了更好的讨论方式;第 4 章围绕一组问题,介绍了在具体研究中借助视觉化工具选择统计检验的方法。

三是针对如何编程和如何编写函数的问题,增加了一些自己和读者都能使用的超小函数。

四是新版本不仅修正了读者反馈的一些错误,还进行了许多微调。我非常感谢他们花了许多时间找出这些错误,同时也希望这一版没有添加新的错误。有些微调是显性的,有些则"隐藏"在代码中,因此,只有在使用这本书和书中的代码时才能感觉到这些变化。同时,本书的所有代码都放在一个文件里。这样一来,处理代码和查阅函数就非常方便了。

我希望读者会受益于这本书,受益于这一版本中的诸多变化和改进。我一如既往地衷心感谢德古意特出版社的团队,他们很早就支持我出版第二版的想法,所以现在第二版和大家见面了。另外,我还要感谢 R 的核心开发团队以及对 R 的缺陷进行修补和配置服务包的许多贡献者,同时感谢 R. Harald Baayen,是他让我们第一次接触到 R;如果不是他,我真不敢想象我的研究会是怎样的……

目　录

第 1 章　实证研究中的一些基本原则

如果能评价自己所说的事情，并且能用数字把你的评价表述出来，则说明你对这些事情已略知一二；但如果你不能对它进行评价，或不能用数字进行表述，则说明你对它的理解还是模糊的。这样的理解也许只是初步涉猎知识，几乎还没有进入科学的阶段。

——威廉·托马斯，洛德·凯尔文

（<http://hum.uchicago.edu/~jagoldsm/Webpage/index.html>）

1. 引言

这是一本统计学的入门书。类似的书籍已经很多，为什么我们还要再写一本呢？正如第一版所述，这本书与其他相关书籍有很大差异，主要表现在以下几个方面：

——本书是专门为语言学家写的。已经有很多关于心理学、经济学、生物学等统计学入门书，但是，像本书一样解释与语言学问题相关的统计学概念和方法，并且是专门为语言学家编写的入门书却很少；

——本书介绍了如何运用大多数统计学方法，其中有“手动”操作的，也有用统计软件操作的。这既不需要数学专业知识，也不需要花费很多时间去理解复杂的方程，而很多其他的入门书要求读者把大量时间花在数学基础上，这对刚入门的人来说是很困难的；还有些书没有解释数学基础而直接引入一些设计完善的软件，忽略了一个设计精湛的图形界面背后的统计检验逻辑；

——本书不仅解释了统计学的概念、检验、图形，还解释了储存和分析

数据的表格设计,以及实验设计中一些非常基础的内容;

——本书只运用开源的软件,主要运用 R,许多入门书使用 SAS,更多人使用 SPSS。这些软件有许多局限性,比如,用户必须购买昂贵的使用权,在函数的种类、数据处理能力和使用时间方面都受权限限制,师生们也许只能在校内使用这些软件,函数更新等方面完全受制于软件公司;

——本书所做的研究易于操作,操作方式也不是特别正式,本书尽量避免使用术语,软件的使用也配有非常详细的解释,配有思考题、警示语和练习,在随附的网页上还配有参考答案,并为进一步阅读提供建议。

因此,本书的目的是帮助读者进行科学的定量研究,其内容框架如下:

第 1 章介绍定量研究的基础:什么是变量,什么是假设,定量研究的结构是什么,结构蕴含什么类型的推理,如何获取实验数据以及用什么格式的文件储存数据?

第 2 章概述了编程语言和 R 的运行环境。这章的内容将应用于后面其他章节的统计图和分析,诸如如何创建、下载和使用数据来为分析做准备。

第 3 章解释了描述性统计分析的基本方法:如何描述数据,如何从这些数据中发现模型,又如何用图来描述这些发现。

第 4 章解释了推断性统计分析的基本方法:如何检验所获得的结果是否确实有意义或只是偶然发生。

第 5 章介绍了几种多因子程序,这是对几种潜在因果关系同时进行研究的程序。这一章有很多内容,但我只能介绍几个经过选择的方法,更多的是把你导向附加的参考资料。

除了章节所配备的思考题和练习之外,该书在 < http://tinyurl. com/StatForLingWithR > 上的附属网站也是个重要的资源。你必须到这些网站上下载练习、数据、参考答案及勘误表等。在 < http://groups. google. com/group/statforling-with-r > 上,你可以找到一个名为"StatForLing with R"的信息群。我建议你加入这个信息群,然后就可以做下面这些事:

——询问关于语言学统计的问题，也有希望从某个热心的群友那里获得答案；

——为附加练习的扩展和/或改善提供建议或数据；

——告诉我和其他读者你在本书中发现的缺陷，当然也从其他读者那里接收类似的信息。这意味着，如果 R 在该书中的指令或者编码与网页上的不一样，那么网页上的信息可能更可靠。

最后，必须明白你不可能通过阅读一本关于统计分析的书籍就学会统计分析。每晚关灯睡觉前 15 分钟在床上读一下这本书或者阅读任何其他关于这方面的书籍，就学会做统计分析是不可能的。确实有些书籍为了诸如市场利益之类的原因在封面上或标题中注明你可以这样学会统计分析，但实际上这是不可能的。我强烈建议，从第 2 章开始，在阅读本书的同时动手运行 R 系统，在 RStudio 里就更理想了，这样才能立刻进入所读到的 R 编码，并能试着使用从随附网站上获得编码文件的所有相关功能；通常编码文件可以提供许多重要的信息、附加的编码片段和更多使用图形解释的建议等。这在第 5 章尤其明显。有时，练习文件甚至能提供更多的建议和图示。即使你不能立刻理解具体每个编码的每个方面，但这也能为你学习本书提供帮助。

2. 语言学中的定量研究方法

如上所述，本书将讲述如何进行科学的定量研究。这类研究的目标有哪些呢？总体上有三个目标。它们是所有实证学科研究者都具备的知识体系的一部分，它们与本书的构架密切相关。

第一个目标是对某种现象的数据化描述。这意味着，研究结果必须以准确并富有启迪性的方式表述出来。以下所有统计方法将有助于实现这个目标，其中第 3 章所描述的方法尤为重要。第二个目标是对数据的解释。通常，对数据的解释离不开研究假设。多数情况下，这已经足够了。然而，有时你也许对第三个目标感兴趣，这就是预测将来会发生什么或者你什么时