全国高等学校外语教师丛书·科研方法系列

Statistical Methods in Language Research with R

基于R的语言学统计方法

王家钺　编著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

全国高

科研方法系列

Statistical Methods in Language Research with R

基于R的语言学统计方法

王家钺　编著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS
北京 BEIJING

图书在版编目（CIP）数据

基于R的语言学统计方法 / 王家钺编著. -- 北京 : 外语教学与研究出版社，2019.9
（全国高等学校外语教师丛书. 科研方法系列）
ISBN 978-7-5213-1192-1

Ⅰ. ①基… Ⅱ. ①王… Ⅲ. ①语言统计－统计方法 Ⅳ. ①H0-05

中国版本图书馆CIP数据核字（2019）第212901号

出 版 人　徐建忠
项目负责　段长城
责任编辑　解碧琰
责任校对　段长城
封面设计　覃一彪　彩奇风
版式设计　吴德胜
出版发行　外语教学与研究出版社
社　　址　北京市西三环北路19号（100089）
网　　址　http://www.fltrp.com
印　　刷　北京虎彩文化传播有限公司
开　　本　650×980　1/16
印　　张　16.5
版　　次　2019年10月第1版　2019年10月第1次印刷
书　　号　ISBN 978-7-5213-1192-1
定　　价　61.90元

购书咨询：（010）88819926　电子邮箱：club@fltrp.com
外研书店：https://waiyants.tmall.com
凡印刷、装订质量问题，请联系我社印制部
联系电话：（010）61207896　电子邮箱：zhijian@fltrp.com
凡侵权、盗版书籍线索，请联系我社法律事务部
举报电话：（010）88817519　电子邮箱：banquan@fltrp.com
物料号：311920001

记载人类文明
沟通世界文化
www.fltrp.com

总　序

“全国高等学校外语教师丛书”是外语教学与研究出版社高等英语教育出版分社近期精心策划、隆重推出的系列丛书，包含理论指导、科研方法和教学研究三个子系列。本套丛书既包括学界专家精心挑选的国外引进著作，又有特邀国内学者执笔完成的“命题作文”。作为开放的系列丛书，该丛书还将根据外语教学与科研的发展不断增加新的专题，以便教师研修与提高。

编者有幸参与了这套系列丛书的策划工作。在策划过程中，我们分析了高校英语教师面临的困难与挑战，考察了一线教师的需求，最终确立这套丛书选题的指导思想为：想外语教师所想，急外语教师所急，顺应广大教师的发展需求；确立这套丛书的写作特色为：突出科学性、可读性和操作性，做到举重若轻，条理清晰，例证丰富，深入浅出。

第一个子系列是“理论指导”。该系列力图为教师提供某学科或某领域的研究概貌，期盼读者能用较短的时间了解某领域的核心知识点与前沿研究课题。以《二语习得重点问题研究》一书为例，该书不求面面俱到，只求抓住二语习得研究领域中的热点、要点和富有争议的问题，动态展开叙述。每一章的写作以不同意见的争辩为出发点，对取向相左的理论、实证研究结果差异进行分析、梳理和评述，最后介绍或者展望国内外的最新发展趋势。全书阐述清晰，深入浅出，易读易懂。再比如《认知语言学与二语教学》一书，全书分为理论篇、教学篇与研究篇三个部分。理论篇阐述认知语言学视角下的语言观、教学观与学习观，以及与二语教学相关的认知语言学中的主要概念与理论；教学篇选用认知语言学领域比较成熟的理论，探讨应用到中国英语教学实践的可能性；研究篇包括国内外将认知语言学理论应用到教学实践中的研究综述、研究方法介绍以及对未来研究的展望。

第二个子系列是“科研方法”。该系列介绍了多种研究方法，通常是一本书介绍一种方法，例如问卷调查、个案研究、行动研究、有声思维、语料库研

究、微变化研究和启动研究等。也有的书涉及多种方法，综合描述量化研究或者质化研究，例如：《应用语言学中的质性研究与分析》、《应用语言学中的量化研究与分析》和《第二语言研究中的数据收集方法》等。凡入选本系列丛书的著作人，无论是国外著者还是国内著者，均有高度的读者意识，乐于为一线教师开展教学科研服务，力求做到帮助读者“排忧解难”。例如，澳大利亚安妮·伯恩斯（Anne Burns）教授撰写的《英语教学中的行动研究方法》一书，从一线教师的视角，讨论行动研究的各个环节，每章均有“反思时刻”、“行动时刻”等新颖形式设计。同时，全书运用了丰富例证来解释理论概念，便于读者理解、思考和消化所读内容。凡是应邀撰写研究方法系列的中国著作人均有博士学位，并对自己阐述的研究方法有着丰富的实践经验。他们有的运用了书中的研究方法完成了硕士、博士论文，有的采用书中的研究方法从事过重大科研项目。以秦晓晴教授撰写的《外语教学问卷调查法》一书为例，该书著者将系统性与实用性有机结合，根据实施问卷调查法的流程，系统地介绍了问卷调查研究中问题的提出、问卷项目设计、问卷试测、问卷实施、问卷整理及数据准备、问卷评价以及问卷数据汇总及统计分析方法选择等环节。书中各个环节的描述都配有易于理解的研究实例。

第三个子系列是“教学研究”。该系列与前两个系列相比，有两点显著不同：第一，本系列侧重同步培养教师的教学能力与教学研究能力；第二，本系列所有著作的撰稿人主要为中国学者。有些著者虽然目前在海外工作和生活，但他们出国前曾在国内高校任教，也经常回国参与国内的教学与研究工作。本系列包括《英语听力教学与研究》、《英语写作教学与研究》、《英语阅读教学与研究》、《英语口语教学与研究》、《翻译教学与研究》等。以《英语听力教学与研究》一书为例，著者王艳副教授拥有十多年的听力教学经验，同时听力教学研究又是她博士论文的选题领域。《英语听力教学与研究》一书，浓缩了她多年来听力教学与听力教学研究的宝贵经验。全书分为两部分：教学篇与研究篇。教学篇中涉及了听力教学的各个重要环节以及学生在听力学习中可能碰到的困难与应对的办法，所选用的案例均来自著者课堂教学的真实活动。研究篇中既有著者的听力教学研究案例，也有著者从国内外文献中筛选出的符合中国国情的听力教学研究案例，综合在一起加以分析阐述。

教育大计，教师为本。“全国高等学校外语教师丛书”内容全面，出版及时，必将成为高校教师提升自我教学能力、研究能力与合作能力的良师益友。编者相信本套丛书的出版对高校外语教师个人专业能力的提高，对教师队伍整体素质的提高，必将起到积极的推动作用。

文秋芳

北京外国语大学中国外语与教育研究中心

2011 年 7 月 3 日

前　言

统计方法是语言研究的重要基础。经几代学者的努力，统计方法在我国外语界已得到很大普及，很多外语院系在研究生阶段开设了与统计相关的课程。近二十年的数种优秀教材，如桂诗春、宁春岩（1997）的《语言学方法论》，杨端和、李强（1998）的《语言统计学》，李绍山（1999）的《语言研究中的统计学》，韩宝成（2000）的《外语教学科研中的统计方法》，马广惠（2003）的《外国语言学及应用语言学统计方法》，秦晓晴（2003）的《外语教学研究中的定量数据分析》，秦晓晴、毕劲（2015）的《外语教学定量研究方法及数据分析》等，为在外语界普及统计方法起到了巨大的推动作用。本书拟在已有研究成果的基础上继续努力，为从事语言学研究的高校外语教师、研究生和科研人员提供相关参考。

本书的特点是基于 R 而不是 SPSS。R 作为统计软件相当年轻，但已风靡国际，是国外很多大学公用计算机上预装的统计工具。它是开源软件，对个人完全免费，但统计功能一点也不少，而且有许多独特的优势，比如制图功能非常强大，这对撰写论文、编写教材和专著等无疑是个好消息。随着国际学术交流日益频繁，越来越多的学者将 R 带回国内，个别外语院系开始基于 R 开设统计课程；近年有些硕士学位论文也提到或用到了 R。随着学术研究跨学科、跨专业的趋势日渐明显，外语界应该对 R 投入足够的注意。这不只是为了多一种选择，更重要的是 R 提供的巨大灵活性使之有更大的发展潜力，非常有助于激发研究者的创造性。如今外语界掌握 R 的人士越来越多，相信不久以后 R 将成为国内外语界的主要统计工具。

国内已经出版的 R 教程大部分都是面向理工科读者，与语言研究相关的仅有 2018 年商务印书馆翻译出版的 Stefan Th. Gries 的著名教程《语言研究中的统计学——R 软件应用入门》（*Statistics for Linguistics with R: A Practical Introduction*），因此编者认为编写一本基于 R 的语言学统计教程非常有必要。

当今我国外语界除少数几个子领域外，多数并不重视统计，统计方法在研究生培养方案中所占的比例极小；多数人只求入门，偏爱“快餐”式的教程，对详读大部头教程没有兴趣，所以本教程力求简洁明了。不过本书也有自足性，可以作为高校应用语言学领域的统计方法入门教材。本书并不试图说服已经熟悉 SPSS 的读者再学习 R，因为只要熟练掌握一种统计软件就够用了。

本书分三篇。上篇前三章是对 R 的介绍，包括 R 的基本知识、安装与设置以及基本操作方式。第四章介绍如何用 R 做随机抽样，第五章介绍用 R 制图的主要方法，这些是 R 相对于 SPSS 的重要特色。中篇讲解主要的统计方法，包括描述统计的原理与方法、推断统计的重要概念，并介绍常用的推断统计方法以及如何用 R 实现。为帮助读者更好地理解内容，每章前面用尽量简短的篇幅介绍基本概念与原理，后面还附有一些练习题。下篇主要介绍统计在探索性语言研究中的应用，尤其在第十四章介绍了一些比较复杂的方法，这些已经成为量化语言研究的重要手段。当然由于可能的应用方式是无限的，本章只介绍其中几种。第十五章讲解认知语言学界已经用得非常多的搭配构式分析（collostructional analysis）方法。虽然该章的主体内容并不是 R，但这种分析的计算部分主要是用 R 实现的，能体现出 R 的优势；另外编者认为这种方法有很大的发展潜力，为帮助读者更好地了解搭配构式分析，专门用一章的篇幅做基本介绍。最后一章是本书的结语，指出了外语界统计分析应用中很少被注意到的几个问题，意在促使读者在使用统计方法的同时保持清醒，关注方法本身的合理性。书后有若干个附录，但不提供各类教程中常见的统计表，这是因为统计类软件的计算结果中都已经给出了相应的统计量和 p 值，查统计表其实是早期手工计算的需要。不过如果读者确实需要查各种关键值，本书的附录二也介绍了怎样用 R 命令来查表。另外由于 R 的最大优势之一是允许用户自己编写脚本，我们鼓励读者学习一些基本的 R 编程方法，附录二和附录三中有一些简单演示。其他几个附录介绍与语言研究有关的一些包和脚本，意在提示 R 的强大性和实用性。

欢迎本书的读者向编者（arthur0421@163.com）发邮件索要本书提到的数据表、脚本等电子文档，包括习题所用的数据以及除公式外所有图的制图脚本。

编者认为，硕士研究生阶段学习的最重要的内容是科研方法，而不在于记住多少理论和术语。虽然在阅读文献过程中也可以学到一些方法，但对初学者来说，文献中的研究方法和思维方法一般过于复杂，要真正领悟很不容易，必须有显式的研究方法指导。任何领域的研究方法都是多元的，只读一两本教程远远不够，必须将方法论与大量的研究实践相结合，在不断的探索中学习研究方法，持续不断地加深认识。研究方法从来不是自足的，再好的研究方法也必须是建立在正确的逻辑思维基础上才有意义。学习研究方法的目的不是给自己添加一两件装备，而是通过学习来培养正确的科学思维习惯。方法主要是工具，不能代替研究者的头脑。在这个意义上，任何一种方法论教程的作用都是有限的。

编者从 1998 年开始学统计，在广东外语外贸大学学习期间先后问学于吴旭东教授和桂诗春教授，并从各种统计教程中受益良多。学习 R 则主要是从 2016 年在英国兰卡斯特大学访学时开始的，当时参加了 Andrew Wilson 教授的 R 软件操作研习班；R 的特性与我的专业背景和爱好有很多共同之处，使我产生了浓厚兴趣。2017 年暑假，在该校学习期间我得到了 Vaclav Brezina 博士的指导。回国后我很快尝试将 R 用在研究生教学中。本书的初稿是在教学讲稿的基础上不断扩充修改而成的。在本书编写期间，桂林理工大学博文学院的张少林教授和广西民族大学李学宁教授给了我热情的鼓励并提出了很多意见和建议，研究生们也给了我各种积极反馈；外语教学与研究出版社高英分社的段长城、金绍康以及校外审稿人都提出了许多具体而富有建设性的意见，在此一并致谢。由于个人水平有限，书中的错误在所难免，文责自负，恳请读者批评指正。

王家钺

2019 年 4 月于南宁

目　录

总　序 …… 文秋芳 vii

前　言 …… 王家钺 x

上　篇

第一章　关于 R 的基本知识 …… 2

1.1　R 是什么 …… 2

1.2　R 与 SPSS 的主要区别 …… 2

1.3　R 的学习资源 …… 5

1.4　R 的引注 …… 6

1.5　本书的形式约定 …… 6

第二章　R 和 RStudio 的安装与设置 …… 8

2.1　安装、运行和界面 …… 8

2.2　配置软件源 …… 10

2.3　本书中使用的包和功能 …… 12

2.4　设置工作目录 …… 13

第三章　R 的基本操作 …… 15

3.1　关于数据类型的基本知识 …… 15

3.2　变量类型及其转换 …… 17

3.3　变量的命名和赋值 …… 17

3.4　R 中最常用的数据结构类型 …… 19

3.4.1　向量（vector） …… 19

3.4.2　数据框（data frame） …… 21

3.4.3　列表（list） …… 23

3.4.4 矩阵（matrix）..... 24
3.5 R 中的数据文档 27
3.6 数据文档的导入 29
3.6.1 导入 CSV 数据 29
3.6.2 导入 SPSS 数据 31
3.6.3 导入 Excel 数据 31
3.7 处理缺失值 32
3.8 数据的保存 34
3.9 其他内容的保存 35

第四章 用 R 做随机抽样和生成随机数据 36
4.1 简单随机抽样 36
4.2 系统随机抽样 37
4.3 分层随机抽样 38
4.4 抽样应用实例：从 AntConc 检索结果中抽样 39
4.5 在 R 中产生随机数样本 40
4.5.1 产生正态分布的随机样本 40
4.5.2 产生非正态分布随机样本 42

第五章 R 制图基础 43
5.1 常用图形及制作方法 43
5.1.1 直方图 43
5.1.2 箱图 47
5.1.3 散点图 50
5.1.4 折线图 52
5.1.5 用 `curve( )` 画曲线 54
5.1.6 饼图 57
5.1.7 条形图 58
5.1.8 堆叠条形图 60

5.1.9 分组条形图 61
5.2 图形的编辑修饰 62
5.2.1 在图上添加直线 62
5.2.2 在图上添加线段和文本作标注 64
5.2.3 在图上添加网格线 65
5.2.4 将多个小图并列在同一大图中 66
5.2.5 将多个图重叠在同一张图中 67
5.3 图的保存 70

中 篇

第六章 描述统计 72
6.1 基本描述统计量及其计算方法 72
6.2 用 R 计算分布的正态性：峰度和偏度 74
6.3 用 R 计算标准分 79
6.3.1 Z 分数 79
6.3.2 T 分数 80

第七章 推断统计的重要概念 82
7.1 假设检验 82
7.2 零假设和备择假设 83
7.3 置信水平、alpha 和 p 值 85
7.4 置信水平与 I 类错误、II 类错误的关系 86
7.5 对总体参数的点估计和区间估计 87
7.6 标准误和置信区间 88
7.7 研究变量之间的关系 90
7.8 参数检验和非参数检验 91
7.9 效应幅度的概念 92

第八章　F 分布与方差齐性检验 94
8.1　F 分布 94
8.2　方差齐性的 F 检验 96
8.3　方差齐性的 Levene 检验 98

第九章　t 检验 101
9.1　t 分布 101
9.2　独立样本 t 检验 102
9.3　配对样本 t 检验 106
9.4　单样本 t 检验 107
9.5　习题 109

第十章　方差分析 111
10.1　方差分析的基本概念与分类 111
10.2　单因素组间方差分析 112
10.3　单因素重复测试方差分析 118
10.4　双因素组间方差分析 120
10.5　双因素重复测试方差分析 125
10.6　习题 128

第十一章　非参数检验 131
11.1　曼—惠特尼 U 检验 131
11.2　威尔柯克斯配对样本检验 133
11.3　Kruskal-Wallis 检验 134
11.4　弗里德曼秩和检验 136
11.5　习题 137

第十二章　相关与回归 139
12.1　相关和回归的基本概念 139

12.2 协方差 140
12.3 相关分析 141
12.4 皮尔逊积矩相关系数 142
12.5 斯皮尔曼等级相关系数 143
12.6 肯德尔等级相关系数 144
12.7 偏相关 145
12.8 一元线性回归 146
12.9 多元线性回归 148
12.10 logistic 回归 150
12.11 克朗巴哈系数 153
12.12 习题 157

第十三章 卡方检验 158
13.1 卡方的概念和卡方分布 158
13.2 卡方拟合优度检验 159
13.3 卡方独立性检验 160
13.4 卡方方差检验 162
13.5 卡方检验用于 2 × 2 列联表 163
13.6 Yates 连续性校正 165
13.7 Fisher 精确检验 166
13.8 McNemar 检验 168
13.9 习题 171

下 篇

第十四章 探索性研究中的统计方法 174
14.1 主成分分析与因子分析：基本概念 174
14.1.1 主成分分析 175
14.1.2 探索性因子分析 181
14.2 聚类分析 186

14.2.1 层级聚类 186
14.2.2 K-means 聚类 188

第十五章 搭配构式分析及其主要类型的实现步骤 192
15.1 搭配构式分析简介 192
15.2 搭配词位分析 193
15.3 区别性搭配词位分析 196
15.4 共变搭配词位分析 197

第十六章 关于统计方法和研究设计的思考 198
16.1 “p 值危机”带来的思考 198
16.2 “思辨加实证”式语言研究中的统计和逻辑谬误 201
16.3 “为例式研究”和“抽样的抽样” 202

习题参考答案 204

附录一 用 R 查统计表 221

附录二 自制“关键性”计算器 224

附录三 标准分的一种应用 228

附录四 pwr 包与功效分析 231

附录五 语料分析包 koRpus 的基本用法 235

附录六 网络爬虫 Rcrawler 的基本用法 241

相关文献推荐 244

上　篇

第一章　关于R的基本知识

1.1　R是什么

我国外语界大部分的统计教程都是基于SPSS的，本书则介绍另一种选择：R。R是一种统计计算与制图系统，SPSS能做的它都能做；同时它还有方便而强大的制图功能和用作基于命令行的高级计算器。但它能做的远不止于此，更像是一把统计与计算领域的高级瑞士军刀。

R是20世纪90年代初问世的开源（open-source）软件，对非商业使用完全免费，因此在学术界和教育界备受青睐。R是一个开放平台，用户可以自由修改、补充已有功能或者添加新功能，而且可以像为智能手机开发APP一样为R制作各种包（package），所有用户都可以自由下载和使用这些包。这种社区式的架构吸引了心理学、医学、生命科学、计算语言学、自然语言处理、人工智能等众多领域的专业人士为R开发了数不胜数的包，而且时常发布更新，R的功能因此不断得到拓展。自诞生以来，经无数用户和计算机高手的贡献，R在很多领域迅速普及，成为国际上最流行的统计计算与制图工具之一。

1.2　R与SPSS的主要区别

R和SPSS在统计计算功能方面不相上下，主要区别在于其他方面：

表1.1　R与SPSS的主要区别

	R	SPSS
费用	对个人完全免费	昂贵
操作界面	命令行	图形界面
语言	英文	多语种

（待续）