STATISTICAL LINGUISTICS

统计语言学

刘 颖 编著

清华大学出版社

STATISTICAL LINGUISTICS

统计语言学

刘 颖 编著

清华大学出版社

北 京

内 容 简 介

统计语言学是一门涉及语言学、计算机科学和数学等多门学科的交叉学科，覆盖面广。本书详细阐述语言统计知识、语言统计的R语言实现、统计结果的直观展示和统计结果的语言分析。主要介绍语言学的基本统计、参数假设检验、非参数假设检验、方差分析、文本聚类、文本分类和综合运用这些统计知识的计量风格学研究。

本书结构完整，层次分明，条理清楚。既便于教学，又便于自学。可作为中文、外语、计算机等专业高年级本科生和研究生教材，也可供从事语言统计和计量分析的研究者参考。

图书在版编目(CIP)数据

统计语言学/刘颖编著. --北京：清华大学出版社，2014
ISBN 978-7-302-37815-0

Ⅰ.①统…　Ⅱ.①刘…　Ⅲ.①统计语言学　Ⅳ.①H087

中国版本图书馆CIP数据核字(2014)第198079号

责任编辑：马庆洲
封面设计：曲晓华
责任校对：赵丽敏
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：17.25　　**字　　数**：413千字
版　　次：2014年9月第1版　　**印　　次**：2014年9月第1次印刷
印　　数：1～1500
定　　价：56.00元

产品编号：061450-01

本书由清华大学亚洲研究中心·2014年出版资助

前　言

统计语言学，研究如何利用概率论、数理统计、信息论等统计的、非离散数学的方法和计算机来对自然语言进行统计和分析。自然语言是其统计和分析的对象，概率论和数理统计等统计知识是其统计的理论基础，计算机是其可以实现统计的工具。因此，对语言进行统计不仅要有语言学方面的知识，而且还要有数学和计算机科学方面的知识。

本书分 9 章，详细阐述如何把语言学知识、数学知识和计算机知识结合起来对语言进行统计和分析。

第 1 章主要介绍统计语言学的基本概念，厘清了统计语言学、语料库语言学、计量语言学和计算语言学的区别、研究内容和应用领域，给出了统计语言学的研究步骤以及本书较为详细的研究内容。这是后面章节内容的总括。

第 2 章主要介绍了语料库的相关内容。阐述了语料库的定义、特点，根据不同标准的分类，并对国内外具有重要意义的语料库、其加工标注和应用进行了详细介绍。

第 3 章主要介绍了语言研究中的基本统计量：包括概率论和统计学的一些基本知识，方差、标准差、平均数、频率、概率，以及互信息、Dice 系数、对数似然比、N 元模型、汉字熵、Zipf 法则、Z 评分、Yule 图、Fuchs 公式以及词语的使用度和通用度等等。

第 4 章主要介绍了在语言研究中广泛使用的假设检验，根据语言研究中的总体是否为正态分布，分为参数假设检验与非参数假设检验。讨论了参数假设检验中的 U 检验、t 检验、F 检验以及χ^2检验；非参数假设检验中的χ^2检验以及秩和检验。详细地比较了不同检验使用的条件、公式和应用领域。

第 5 章主要介绍了方差分析，其主要应用于三个或三个总体以上的差异比较。讨论了单因素方差分析、无重复双因素方差分析、可重复双因素方差分析以及单因素的多重比较。

第 6 章主要介绍在语言研究中常用的一种机器学习方法——文本聚类。详细介绍了文本聚类的流程和主要算法，重点介绍了层次聚类和 k-means 聚类。

第 7 章主要介绍了语言研究中常用的另一种机器学习方法——文本分类，并且详细介绍了文本分类的过程和主要的分类模型，包括朴素贝叶斯模型、KNN 以及支持向量机等。

第 8 章介绍了在语言研究中经常使用到的一种程序语言——R 语言，其具有强大的统计分析功能和绘图功能。重点介绍了 R 的基本操作、主要绘图功能，以及本书中用于语言研究的统计方法的 R 语言实现。

第 9 章讨论了计算风格学。从字符、词汇、句子、词类、短语和段落方面全面阐述计算风格学研究使用的语言特征。以莫言和余华各自六部小说为例，从字符、词汇、句子、词

类、短语和段落方面，分别运用基本统计、假设检验、文本聚类和文本分类等来对两位作者的写作风格进行系统地研究。这些特征的统计主要利用 R 语言来实现。因此，可以说，第 9 章是把全书各章节内容结合的一个范例。

本书可作为中文、外语、计算机等专业高年级的本科教材，教授时间可为 32～64 学时。如果学生掌握了语言学知识和基本的统计理论，并能用 R 语言实现本书介绍的统计模型，则对学生掌握计算机统计自然语言和分析语言打下坚实基础。

本书在写作时尽量做到通俗易懂，所有的统计都利用真实的小说语料进行了示例和分析。本书的读者如果具有一定的概率和统计知识和 R 语言编程，则能自如地利用本书的统计知识进行语言处理。若在此基础上还掌握计算机编程知识（数据库、Java 编程或 C 语言编程），则可容易地扩展本书的现有内容并进行更广泛的语言统计和分析。

本书的写作参考了许多学者的论文和著作，本书能够出版与他们所作的工作紧密相关，谨向他们表示衷心感谢。

由于本人水平和时间限制，本书难免存在疏漏和不足之处。欢迎各位读者批评指正。

刘　颖

2014 年 7 月 15 日

目　录

第 1 章　概论 …… 1

1.1　统计语言学 …… 1
1.2　统计语言学与其他学科 …… 1
1.2.1　计量语言学 …… 1
1.2.2　计算语言学 …… 2
1.2.3　语料库语言学 …… 2
1.2.4　与三个学科的联系与区别 …… 3
1.3　使用统计方法研究的语言特征 …… 3
1.4　统计语言学基本研究方法 …… 4
1.5　统计语言学研究的步骤 …… 5
1.6　统计的语言学应用 …… 6

第 2 章　语料库 …… 7

2.1　语料库的定义 …… 7
2.2　语料库的类型 …… 7
2.2.1　口语语料库与书面语语料库 …… 8
2.2.2　单语语料库、双语语料库与多语语料库 …… 8
2.2.3　通用语料库与专用语料库 …… 8
2.2.4　共时语料库与历时语料库 …… 9
2.2.5　动态语料库与静态语料库 …… 9
2.2.6　同质语料库与异质语料库 …… 10
2.2.7　生语料库与标注语料库 …… 10
2.3　国内外主要语料库 …… 10
2.3.1　国外的语料库 …… 10
2.3.2　国内的语料库 …… 26
2.4　本章小结 …… 37

第 3 章　统计在语言研究中的基本应用 …… 38

3.1　统计学的基本概念 …… 38
3.1.1　总体、个体、样本 …… 38

3.1.2 参数与统计量 …… 39
3.1.3 常量、变量 …… 39
3.1.4 实际值与观测值 …… 39
3.2 平均数 …… 39
3.2.1 简单算术平均数 …… 40
3.2.2 加权算术平均数 …… 41
3.3 方差与标准差 …… 41
3.3.1 未分组数据的方差与标准差 …… 41
3.3.2 分组数据的方差与标准差 …… 42
3.4 频度、频率、概率、条件概率、贝叶斯定理 …… 43
3.4.1 概率论中的常用概念 …… 43
3.4.2 概率 …… 44
3.4.3 独立性 …… 45
3.4.4 贝叶斯定理 …… 46
3.4.5 频度与频率 …… 47
3.5 互信息 …… 47
3.6 Z评分 …… 49
3.7 Dice系数 …… 50
3.8 Phi平方系数(Φ^2) …… 50
3.9 对数似然比 …… 51
3.10 N元模型 …… 51
3.10.1 N元语法 …… 51
3.10.2 N元语法模型 …… 52
3.11 语言学三大统计规律 …… 53
3.11.1 Zipf法则 …… 53
3.11.2 Menzerath-Altmann定律 …… 56
3.11.3 Piotrowski-Altmann定律 …… 57
3.12 熵 …… 57
3.12.1 静态平均信息熵 …… 57
3.12.2 极限熵 …… 58
3.13 Yule图 …… 59
3.14 Fuchs公式 …… 60
3.15 使用度与通用度 …… 61
3.15.1 使用度 …… 61
3.15.2 通用度 …… 62
3.16 本章小结 …… 62

第 4 章　假设检验 …… 64

4.1　假设检验的相关概念 …… 64
4.1.1　假设检验的基本原理 …… 64
4.1.2　假设的分类 …… 64
4.1.3　检验统计量与临界值 …… 65
4.1.4　双尾检验与单尾检验 …… 65
4.1.5　假设检验的一般步骤 …… 66
4.1.6　假设检验中的两类错误 …… 66
4.2　参数假设检验 …… 66
4.2.1　正态分布 …… 67
4.2.2　U 检验 …… 73
4.2.3　t 检验 …… 75
4.2.4　χ^2 检验 …… 79
4.2.5　F 检验 …… 80
4.2.6　参数假设检验比较 …… 81
4.3　非参数假设检验 …… 83
4.3.1　χ^2 检验 …… 83
4.3.2　秩和检验 …… 88
4.3.3　非参数假设检验比较 …… 92
4.4　本章小结 …… 92

第 5 章　方差分析 …… 94

5.1　方差分析的定义及基本思想 …… 94
5.1.1　方差分析的定义 …… 94
5.1.2　方差分析的基本思想 …… 95
5.2　方差分析的基本概念和使用条件 …… 95
5.2.1　方差分析中的基本概念 …… 95
5.2.2　使用方差分析的条件 …… 97
5.3　方差分析的类型和一般步骤 …… 102
5.3.1　方差分析的类型 …… 102
5.3.2　方差分析的一般步骤 …… 102
5.4　单因素方差分析 …… 102
5.4.1　各个因素水平间的样本容量相同 …… 103
5.4.2　各个因素水平间的样本容量不完全相同 …… 107
5.4.3　方差分析中的多重比较 …… 108
5.5　双因素方差分析 …… 110
5.5.1　无重复双因素方差分析 …… 111

5.5.2 可重复双因素方差分析 …… 116
5.6 本章小结 …… 122

第 6 章 文本聚类 …… 123

6.1 文本聚类概述 …… 123
6.1.1 文本聚类定义 …… 123
6.1.2 文本聚类的流程 …… 124
6.2 文本聚类中的数据 …… 125
6.2.1 聚类分析中使用的数据结构 …… 125
6.2.2 数据归一化处理 …… 127
6.3 相似度计算 …… 130
6.3.1 文本相似度的计算 …… 130
6.3.2 特征相似度的计算 …… 132
6.4 聚类算法 …… 134
6.4.1 层次聚类 …… 134
6.4.2 划分聚类 …… 142
6.4.3 划分聚类与层次聚类的联系和区别 …… 146
6.5 文本聚类性能评价 …… 147
6.5.1 纯度 …… 147
6.5.2 归一化互信息 …… 148
6.5.3 精确度 …… 149
6.5.4 *F* 值 …… 151
6.6 本章小结 …… 152

第 7 章 文本分类 …… 153

7.1 文本分类的定义 …… 153
7.2 分类方法 …… 154
7.2.1 基于知识工程的方法 …… 154
7.2.2 基于机器学习的方法 …… 154
7.3 分类步骤与流程 …… 155
7.4 文本表示与特征选择 …… 156
7.4.1 特征项选择 …… 156
7.4.2 词袋模型 …… 156
7.4.3 向量空间模型 …… 158
7.4.4 特征筛选与权重 …… 159
7.5 向量相似度测量 …… 162
7.6 分类模型 …… 163
7.6.1 朴素贝叶斯(Naïve Bayes) …… 163

7.6.2 k-最近邻(k-Nearest Neighbor) …… 166
7.6.3 支持向量机(Support Vector Machines) …… 167
7.7 文本分类的评价 …… 169
7.7.1 准确率、召回率 …… 169
7.7.2 正确率、错误率 …… 170
7.7.3 F 值 …… 170
7.7.4 微平均和宏平均 …… 171
7.8 本章小结 …… 171

第 8 章 R 语言简介 …… 172

8.1 R 语言的帮助文件 …… 172
8.1.1 R 的基本知识在线帮助 …… 173
8.1.2 R 程序中的关键字符及函数的在线帮助 …… 173
8.2 R 程序包 …… 175
8.2.1 程序包的安装 …… 175
8.2.2 程序包的载入 …… 176
8.3 R 语言的数据结构及基本函数 …… 176
8.3.1 R 语言的对象类型 …… 176
8.3.2 R 语言的对象的建立 …… 177
8.3.3 数值型向量的常用统计函数 …… 184
8.4 数据的读取和存储 …… 185
8.4.1 数据的读取 …… 185
8.4.2 数据的存储 …… 188
8.5 R 的基本绘图 …… 189
8.5.1 饼图(Pie Plot) …… 190
8.5.2 条形图(Barplot) …… 191
8.5.3 直方图(Hist) …… 193
8.5.4 折线图(Matplot) …… 194
8.5.5 箱线图(Boxplot) …… 196
8.5.6 散点图(Scatter Diagram) …… 198
8.5.7 散点图矩阵(Scatterplot Matrices) …… 199
8.6 假设检验 …… 200
8.6.1 参数假设检验 …… 200
8.6.2 非参数假设检验 …… 204
8.7 方差分析 …… 209
8.7.1 方差齐性检验 …… 209
8.7.2 单因素方差分析 …… 210
8.7.3 双因素方差分析 …… 214

8.8 本章小结 …… 216

第 9 章 计算风格学研究 …… 218

9.1 计算风格学研究使用的语言特征 …… 218
9.1.1 字符方面 …… 219
9.1.2 词汇方面 …… 219
9.1.3 句子方面 …… 220
9.1.4 词类方面 …… 222
9.1.5 短语和语法结构方面 …… 222
9.1.6 段落方面 …… 222
9.2 计算风格学研究中常使用的方法 …… 223
9.3 莫言与余华小说计算风格学研究 …… 223
9.3.1 基于频率的风格分析 …… 223
9.3.2 假设检验的文本风格分析 …… 227
9.3.3 基于文本聚类的风格分析 …… 231
9.3.4 基于文本分类的风格分析 …… 233
9.3.5 小结 …… 234
9.4 本章小结 …… 235

附录 常用的统计数表 …… 236

附表 1 标准正态分布函数数值表 …… 237
附表 2 正态性检验统计量 W 的系数$a_i(n)$的值 …… 238
附表 3 正态性检验统计量 W 的 α 分位数W_α表 …… 240
附表 4 正态性检验统计量 Y 的 α 分位数Y_α表 …… 241
附表 5 t 检验临界值表 …… 242
附表 6 χ^2检验临界值表 …… 243
附表 7 F 检验临界值表 …… 244
附表 8 Wilcoxon 秩和检验临界值表 …… 249
附表 9 统计量 H 的分位数$H_{1-\alpha}(r,f)$表 …… 252
附表 10 多重比较$q_{1-\alpha}(r,f)$表 …… 253

参考文献 …… 256

第1章 概 论

对语言进行统计研究，是语言学研究的一种重要的方法。传统的语言学研究，往往以个人内省为主要依据——语感，从而对语言的词汇、句子、语法等进行考察，进而归纳出纷繁复杂的言语现象背后的语言规律。然而，语言研究的科学化，一直是现代语言学家的主要努力方向和重要目标。对语言的科学化研究，即是要求通过一些实验方法和数学模型来对语言进行考察，并且得出的结论是可验证的，而且对于语言规律和发展具有较强预测能力(刘海涛 2012)。

统计——通过实验方法和数学模型对于研究对象进行考察从而得到可验证的结论，对于现代自然科学的大发展起到了极为重要的作用。统计已广泛应用于语言学、文学、历史研究、心理学、社会学、人类学、教育学、生态学、考古学、农业、动物学、审计学、牙医学、工程、流行病学、金融、遗传学、地理学及工业等各个领域中。

统计是一种极为有用的研究手段和方法，因此将统计方法引入语言学研究并发展，是实现语言研究科学化的重要途径和手段。

1.1 统计语言学

使用概率论、数理统计等统计学的方法来对语言进行研究，称为统计语言学(冯志伟 2012)。

统计学可以帮助语言学者对语言学研究中的数据进行分析和推断。通过统计数据来揭示反应语言内在的和固有的一些规律性。

自然语言是其统计和分析的对象，概率论和数理统计等统计知识是其统计的理论基础，计算机是其可以实现统计的工具。因此，对语言进行统计不仅要有语言学方面的知识，而且还要有数学和计算机科学方面的知识。

1.2 统计语言学与其他学科

1.2.1 计量语言学

计量语言学，是指使用计量的方法，对于真实语言交际活动中产生的各种语言现象、语言结构、结构属性以及其相互之间的关系进行研究，从而找出语言现象背后深层的数学

规律，描绘出语言现象的数学面貌，并对其进行原因分析和理论阐释的一门学科。计量的方法，通常包括，概率论、统计学、微积分、随机过程等数学定量方法。

计量语言学以真实语料为基础，用计量的方法研究语言的结构和发展规律，其目的在于探索语言的数学面貌并发现隐藏在语言现象中的内在的数学规律。

计量语言学的根本任务是从真实文本中抽象出来的数量规律来描述与理解语言的各组成成分的关系。其发现的语言规律对于精确地描写与解释相应的语言现象以及构建现代科学意义上的语言学理论都具有极为重要的作用和意义。

计量语言学研究的主要内容有：语言成分统计规律、字频、汉字熵、词频与词序的秩的关系、词语的频率分布、词的音节构成及分布规律、英语动词的义项分布规律、名词分布规律、类符形符比、词语的使用度和通用度、文本中词长数量的分布、描述语言结构在语言系统和语言使用中的分布定律、描述不同的语言结构（及其属性）间相互关系的函数定律、描述语言演化规律的演化定律、基于短语结构树库的计量研究、基于依存树库的计量研究、概率配价模型、基于词汇系统的协同语言学模型和基于句法系统的协同语言学模型等等（刘海涛 2012）。

计量语言学广泛应用于音位学、形态学、句法学、词汇学、语义及语用学、地理语言学及方言学、类型学与语言的历时研究等领域。

1.2.2 计算语言学

计算语言学，也称自然语言处理或自然语言理解，它是研究如何利用计算机来分析、处理和理解自然语言。

计算语言学主要研究语音的分析和生成、分析型语言的分词、印欧语系语言的形态分析、词性标注、句法分析、语义分析、篇章分析等。

计算语言学主要用于语音的自动识别和自动生成、自动文摘、自动校对、信息自动检索、人机对话、自动问答、自动分类、信息自动抽取、机器翻译等领域。

计算语言学是利用计算机来分析和处理自然语言，其目的在于建立自然语言的应用系统，其研究方法有统计和规则两种方法。

1.2.3 语料库语言学

语料库语言学是利用计算机强大的检索、统计和处理语料的能力，从大规模的语料库中检索符合研究问题的实例，对其进行统计。在大量实例和统计数据的基础上，对研究问题进行定性分析，从功能上对其进行语言学解释。利用计算机和语料库可以对语言各个层面的特征（单个特征或多个特征）进行分析和研究。

语料库语言学的主要研究内容是研究机器可读的自然语言文本在词汇、句法、语义、语篇各个层面的采集、存储、检索、统计和分析。

语料库语言学主要应用在词典编纂、语言（社会语言或方言等）调查、语域变异、历时语言研究、语言习得、作家作品风格分析和教育语言学等领域的研究中。

语料库语言学对语料库既有定量的统计，又有定性的功能解释，对语言的描写更加全面。

1.2.4　与三个学科的联系与区别

统计语言学、计量语言学、语料库语言学和计算语言学的研究都是涉及语言学、数学、统计学以及计算机科学等多个学科和领域，是典型的文理工交叉学科，具有鲜明的跨学科研究性质。这些学科的研究对象都是自然语言组成的大规模语料库，研究工具都是利用计算机的软硬件，研究的理论基础是数学的概率统计知识和语言学的语音、词汇、句法、语义、语篇和语用知识。几个学科有些研究内容可能存在交叉，不能完全割裂开来。

统计语言学、计量语言学、语料库语言学和计算语言学都可以对语言学的语音、词汇、句法和语义等层面进行统计和研究。但它们的研究方法和研究目标存在不同：

统计语言学和计量语言学都是利用统计方法来实现对语言成分的统计，计量语言学以发现语言成分或语言成分间的数学规律为目标。而统计语言学以所统计的语言特征在统计学上显著和不显著为目标。语料库语言学对大规模语料库进行词汇、句法和语义等统计，依据统计数据和实例上下文对所研究的对象进行语言学层面定性的分析，是定量分析和定性分析的结合，以研究语言的结构和运用为目标。计算语言学以语言结构的理解与生成为研究目标，以统计和规则为基本研究方法。1990 年以来，统计方法在计算语言学领域占据主流，但计算语言学的统计模型——隐马尔科夫模型、最大熵模型、条件随机场模型等和实现算法更复杂，利用这些模型进行语音识别与合成、汉语的切分、词性标注、句法消歧、语义消歧或机器翻译等时，利用的都是动态规划算法，来实现对语言的理解和自动处理(刘海涛 2012)。

本书所介绍的统计语言学包含了语言与统计的研究内容、计量语言学的一些研究内容以及计算语言学中的文本分类和文本聚类等。

1.3　使用统计方法研究的语言特征

随着语料库加工深度的增加和计算机软硬件的发展，目前，从语音、词汇、句子、短语、段落等层面都可以对自然语言进行统计。

1. 语音层面

可以统计的特征有：声母、韵母、句尾韵、声调以及上述几个特征的结合。

2. 字符特征

目前，可以用来统计的字符特征有：字母、汉字、大小写字母、数字、标点符号、词首字母、词尾字母、词首字、词尾字、空格等。

3. 词汇特征

可以用来统计的词汇特征有：特定词的分布、词频、虚词、高频词、一次词比例、句首词、句尾词、段首词、段尾词、词长、平均词长、词汇丰富度、词长离散度、词汇总量、词类比例、特定类词汇数量比例(比如方言词、新造词、语气词、专业术语、国际性词语或外来词语、古语词、成语、惯用语等)、关联词语的种类及其分布等。

4. 句子特征

可以用来统计的句子特征有：句长、平均句长、句长离散度、不同句型的比例、句式的

分类统计及其数量关系(比如主动句、被动句、把字句、倒装句等)、句子结构(比如主谓结构、方位结构、动宾结构等)、用句总量等。

5. 段落特征

统计段落长度,可相对划分为长段落与短段落,并求出段落的平均长度。

统计段落的起始句、开端词、终结句和末端词,以分析分段特点。

段落的离散度可看出不同段落与平均段落离散的程度。

6. 短语特征

统计的最简单短语特征就是 N 元语法(N 元字符串、N 元词串、N 元词类串等),还可以统计名词短语、动词短语和形容词短语等分布。

除此之外,还可以对语料库进行语用特征和篇章特征等的统计。

统计不同层面的语言特征,对语料库加工的深度要求不一样。对字符层面的统计、不需要对语料库进行加工和标注。对词类进行统计,需要对语料库进行词性标注。对汉语进行词频统计,需要对语料库进行汉语分词。统计名词短语的分布,就需要对语料库进行名词短语的标注。随着语料库加工深度的增加,对语言进行的统计特征会越来越多。

1.4 统计语言学基本研究方法

1. 统计描述法

研究如何取得反映客观语言现象的统计数据,通过图表的形式对所统计的数据进行加工、处理和显示,进一步通过分析和综合得出反映语言客观现象的规律性的数量特征。

实现描述的统计量主要有:频率、概率、表示数据集中的平均数、表示数据分散的标准差和方差、互信息、Z 评分、Dice 系数、Phi 平方系数、对数似然比、N 元语法、信息熵、极限熵、词语的实用度和通用度和 Yule 图等。互信息、Z 评分、Dice 系数、Phi 平方系数和对数似然比可用来描述两个字之间或两个词之间结合的紧密程度。熵是对语言符号不确定性的度量,表示该语言每一个字符所包含的平均信息量的大小。极限熵是将充分考虑上下文关系的情况下达到的最小条件信息量。Yule 图是用来考察文本中词汇丰富度的大小的一个度量。使用度是用于衡量词语常用性的重要指标。通用度是衡量词语在不同领域的使用情况的重要指标。

本书给出 R 语言中对数据直观展示和描述的几种图形方法——饼图、条形图、直方图、折线图、箱线图和散点图等。饼图用来描述同一类型具有比例关系的数值型数据。条形图适用于展示类别型变量的分布。直方图是将数据对象划分为一定数量的组。折线图是将若个数据散点连接起来形成反映数据变化趋势的图形。箱线图反映一组数据的最小值、两个四分位数、中位数以及最大值,描述连续型变量的分布。

2. 统计推断法

统计推断法包括根据统计数据得出的反应语言现象的数学规和根据样本数据对总体的均值、方差等进行的假设检验。

根据统计数据得出的单个语言特征或两个语言特征数量间满足的数学规律,主要介绍 Zipf 法则、Menzerath-Altmann 定律、Piotrowski-Altmann 定律和 Fuchs 公式。Zipf 法

则描述了词频和词的排序序号之间的反比关系。Menzerath-Altmann 定律描述了词所含音节数和音节的平均长度间的关系。Piotrowski-Altmann 定律描述了语言现象的演变规律。Fuchs 公式描述了不同语言中词的音节数目的分布规律。

假设检验是对未知的总体分布形式或总体的未知参数做出一定的假设，然后构造适合的统计量并根据样本信息进行计算，在设定的显著水平或置信度上判断假设是否成立。根据总体是否服从正态分布，可分为参数假设检验和非参数假设检验。参数假设检验主要介绍对总体均值 μ 进行检验的 U 检验、t 检验，以及对总体方差 σ^2 进行检验的 χ^2 检验和 F 检验。非参数假设检验主要介绍 χ^2 检验和秩和检验。

参数假设检验和非参数假设检验主要用于检验两个总体之间的均值或方差等是否存在显著差异，而方差分析主要用于检验更多总体均值之间差异是否显著。

3. 统计模型法

统计模型法是根据数学模型对语言中成分或文本之间关系进行推断的方法。主要介绍的数学模型是朴素贝叶斯模型、K-最近邻模型、支持向量机模型、层次聚类和划分聚类。层次聚类和划分聚类是无指导的机器学习方法，可以根据文本的一个或多个特征实现对文本的自动聚类。而朴素贝叶斯模型、K-最近邻模型、支持向量机模型是有指导的机器学习方法，根据训练语料库的文本分类来对未知的文本进行分类。

1.5　统计语言学研究的步骤

作为一门实证学科，统计语言学研究所遵循的思路和研究方法与其他实证学科基本相同 大致包括以下五个步骤(Reinhard Köhler 2005)。

如图 1-1 所示，统计语言学研究一般要经过五个步骤。

第一步，提出语言学假设。语言学假设是对自然语言现象的大胆推测。这个假设要满足一定的形式与内容，并且所提出的假设必须有相关的实证性和可验证性。对于统计语言学研究中使用很多的随机假设来说，一方面，要推翻该假设绝不能依靠个别的反例，而必须建立在足够大量的数据和进行充分的数理检验的基础上。另一方面，一个假设永远不能被认为是完全证实，即便已有的数据均支持该假设，但仍然有继续检验的必要。

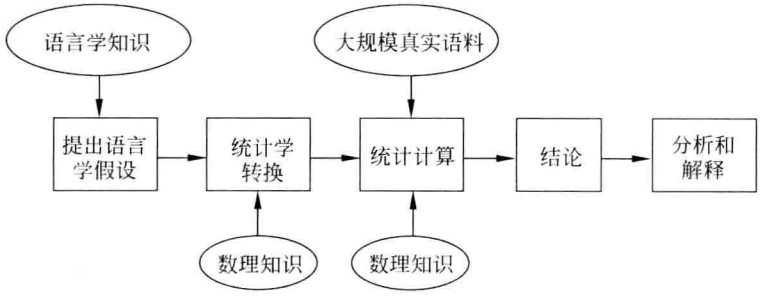

图 1-1　统计语言学研究步骤

第二步，将语言学假设转换为可以量化的特征。由于随机假设只能够通过数理方法的方法得到检验，因此，每一个假设，都必须要转换成为可以使用数理检验的形式。在转