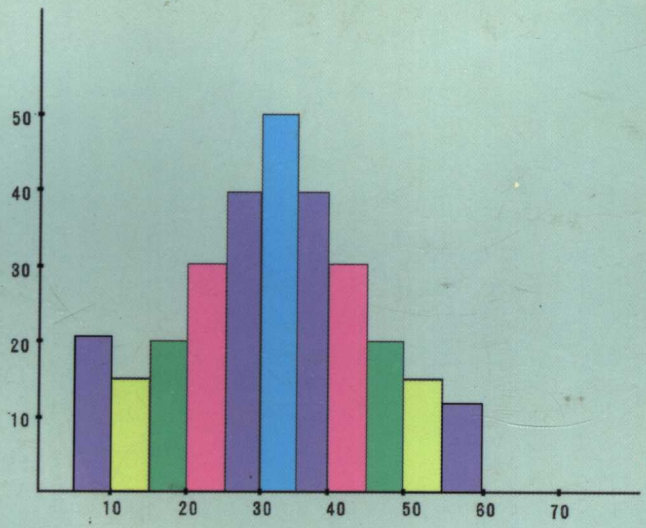
# 语言研究中的统计方法

## Statistics in Language Studies

Anthony Woods
Paul Fletcher
Arthur Hughes
著

陈小荷 徐 娟
熊文新 高建忠
译




北京语言文化大学出版社

# 语言研究中的统计方法

# STATISTICS IN LANGUAGE STUDIES

ANTHONY WOODS
PAUL FLETCHER 著
ARTHUR HUGHES

陈小荷 徐 娟
熊文新 高建忠 译

北京语言文化大学出版社

（京）新登字 157 号

**图书在版编目（CIP）数据**

语言研究中的统计方法/（英）伍兹（Woods，A.）等著；陈小荷等编译 .—北京：北京语言文化大学出版社，2000
ISBN 7-5619-0804-0
Ⅰ. 语…
Ⅱ. ①伍… ②陈…
Ⅲ. 语言统计
Ⅳ. H0-05
中国版本图书馆 CIP 数据核字（2000）第 01069 号
著作权合同登记图字 01-2000-0135 号

责任印制：汪学发
出版发行：北京语言文化大学出版社
（北京海淀区学院路 15 号　邮政编码 100083）
印　　刷：北京北林印刷厂
经　　销：全国新华书店
版　　次：2000 年 4 月第 1 版　2000 年 4 月第 1 次印刷
开　　本：850 毫米×1168 毫米　1/32　印张：11
字　　数：260 千字　印数：0001-2000
书　　号：ISBN 7-5619-0804-0/H·9109
定　　价：20.00 元

# 译者的话

由统计学家 Woods 和语言学家 Hughes, Fletcher 合著的《语言研究中的统计方法》(Statistics in language studies)是剑桥语言学系列教材之一。这本书结合语言习得、语言变异和语言测试等方面大量的研究实例,介绍了统计分析的基本概念、方法和技术。读者可以把这些技术应用到自己的研究领域中去,也可以作为一种知识基础,评价和利用统计分析文献。

书中涉及许多统计学术语,我们主要参考齐玉霞编订的《英汉数学词汇》(科学出版社,1982 年第二版)。

本书前言、第 1~5 章由高建忠翻译,第 6~9 章由熊文新翻译,第 10~12 章以及附录由陈小荷翻译,第 13~15 章以及英汉对照术语表(根据原书索引表)由徐娟翻译。最后由陈小荷、徐娟统稿。

由于我们知识水平有限,错误在所难免,敬请读者指正。

# 前　言

本书缘起于语言学家(Hughes,Fletcher)跟统计学家(Woods)关于第一语言、第二语言学习和测试的具体问题探讨的一些初期接触。这些接触使我们越来越意识到统计学跟语言学和应用语言学的其他领域的相关性,意识到在这些领域工作的人如同社会科学的其他研究者一样,有责任使他们的数据经得起统计学的同样细致的检查。语言学专业的学生越来越频繁地使用雷丁(Reading)大学应用统计学系提供的指导服务。这一点很快就变得显而易见:如果说统计学家跟语言学专业学生的对话曾是非常有用的话,那么从学生这方面来说,他们还需要掌握统计学的基本概念。于是第二步就是为语言学专业的学生开设统计学课程(Woods)。本书大致就是这门课程的教材。那些希望用统计学来处理数据的语言学专业的学生究竟需要知道些什么?本书反映了我们几位作者对于这个问题的共同看法。

本书跟其他为语言学家而写的导论性统计学教科书有两点主要区别。第一,加重了概率与统计推断在书中的分量。为了阐明样本跟总体的关系,我们比较详细地讨论了概率、建立统计模型以及(用正态分布作为统计模型的例子)从样本估计得到总体值估计的问题。尽管这几章(第4章至第8章)刚开始读时会比较困难,但是我们特别建议,读者如果使用本书后面章节中的技术时希望完全理解自己正在做什么,就不要放弃这几章。

第二点区别是我们所介绍的统计方法的范围。从第13章的后半部分开始,考察了一些跟语言学数据有关的多元分析技术。多元回归、聚类分析、判别函数分析、主成分分析和因子分析近年来已经

应用于许多语言学问题并且报道了结果。一本以语言学专业的学生为对象的教科书,其主要目标之一就是要使他们有评价研究文献的能力。因此,仅凭这一理由就足以把多元分析方法包括进来。当然还有另一个理由,那就是使学生或研究者了解这些方法,以便在他们自己的工作中发现这些方法的潜在的应用,并且掌握这些必要的知识,跟统计学家进行有效率的讨论。

也许有必要强调,除了算术基础之外,理解本书中出现的计算并不需要别的数学知识。我们已经说过,掌握这些概念可能需要花些力气,另外,也需要花些时间来熟悉某些符号。主要是出于这个原因,我们提供了习题:做完这些习题会使概念变得清晰,符号用得熟练。其他专门的数学知识就不需要了。

我们从同事和学生方面得到的意见和数据非常之多,不能一一列举,在此对他们表示感谢。我们特别要感谢 Lynne Rogers 十分细致和耐心地录入原稿。

ANTHONY WOODS
PAUL FLETCHER
ARTHUR HUGHES
1985 年 7 月于雷丁大学

# 目　录

# 第 1 章　语言学家为什么需要统计学

语言学家也许会怀疑统计学对他们有什么帮助。目前在语言学界居支配地位的理论框架是生成语法，对于句子合法性自有一套语料评判标准。这些判断通常是根据语言学家自己的主观认识形成的，是“非此即彼”的，跟同一言语社团中理想的本族语者的语言能力相关。对这些语料似乎无须进行数值的定量分析并由此作出推断。这里看来没有统计学的地位。

尽管生成语法在过去的 25 年中对语言学知识作出了重大贡献，但它并不是语言学研究的惟一论题。语言学中也有其他一些领域，要求对所观察到的数据进行统计处理。本书将详细地考察语言学一些领域的研究，并希望能表明统计学在这些领域中的必要性。在这篇简短的引论里，我们举一些研究实例来说明我们面临的主要问题。

我们想通过这本书表明，统计学使得我们能够分析复杂的数值型数据，如果需要的话，还可以从中作出推断。实际上，有时可以把统计学区分为描写统计学和推论统计学两种。分析和推论的必要性在于，数据值中存在变异（就是说，测到的值不一样）。如果不存在变异，就不需要统计学了。

设想一位语音学家对说英语者的清浊对立方式感兴趣，研究的第一步是测量词首塞音的嗓音起始时间（Voice Onset Time，VOT），即从塞音除阻到嗓音开始所经历的时间。第一组数据包括 20 个人分别对 10 个以/p/开头的单词重复 10 遍发音的结果。如果不管是在词之间还是在发音人之间 VOT 都没有差别，统计学在这里就没有什么必要，只要记录一个 VOT 值就行。事实上这些值几乎没有相同的。这一组发音人发出的 VOT 值也许都是有区别的，比方说当他们在发同一个词的第一个音时。此外，一个发音人对于不同的词，甚至

同一个词的几次重复，其 VOT 值也可能会有所不同。因此这位语音学家可以得到多达 2000 个值。统计学的第一个贡献是为分析这些结果提供有意义的和易于理解的方法。通用的方法是用一个“典型”值来代表所有 VOT 值，并且给出所有 VOT 值围绕这个值上下浮动的范围（均值与标准差——见第 3 章）。这样一来，就把大量的数值精简为两个。

以后我们还要谈到这位语音学家得出的数据，但现在让我们再看另外一个例子。在这个例子中，一位心理语言学家感兴趣的是外语学习能力的本质。作为研究的内容之一，先对 100 个被试进行语言能力测试，经过一段时间的语言教学后又进行了该语言的学习效果测试。这位心理语言学家希望了解的情况之一是两次测试成绩的相关形式。观察这两组得分会看出一些问题：比如，某人在前一次学能测试中成绩出众，那么效果测试的成绩也可能很好。但是这位心理语言学家无法通过逐一对比每个人前后两个得分而得到这 200 个分数所负载的信息。尽管那位语音学家采取的分析方法会有所帮助，但这些方法并不会让这位心理学家得出两组得分的关系。然而一个直接的统计方法会帮我们把这个关系的强度用一个值表示出来（相关——见第 10 章）。统计学又一次把繁复的数据减少到可以把握的数目。

这是数据分析中大幅度简化数据的两个例子。这两个例子都考察了一组被试的语言表现。当然，语言研究者的最终兴趣并不在于**样本**本身的表现。他们通常希望归纳出更大群体的语言表现。那位手头有 20 个发音人样本的语音学家，希望能说出所有讲某种英语方言、甚至所有讲英语的人在发音上的某些共性。类似地，研究者是对所有以/p/开头的单词感兴趣，而不仅仅是对样本中的那些单词感兴趣。这里就有统计学的用武之地了。如果样本符合一定的条件，研究者可以用一些技术来估计，一定容量的样本的“典型”分数跟他们所要作出归纳的群体的典型分数的近似程度。（见 §4.4、§5.5）

让我们再来看那位语音学家的例子。当他收集以/p/开头的发音数据时，他同时要求这 10 位被试对 20 个以/b/开头的单词进行发音。他把这些数据减少为两个分析结果：这组发音的平均 VOT 和标

准差。我们暂时只考虑前一个值，即**均值**，或者说典型值。这位语音学家继而会发现/p/和/b/两组数据的典型 VOT 值不一样，/p/组的典型 VOT 值大一些。问题就出在这儿，到底均值上的这种差异能说成是代表将被归纳的更大群体的真实情况呢，还是出于偶然产生了这种结果？（试想，如果测量是精确的，几乎可以肯定样本值会有所不同。）统计技术会让这位语音学家对样本中的差异能否真正反映更大群体的可能性作出判断。在我们给出的这个（关于 VOT 的）例子中，语音学文献已经很明白地确定存在这种差异（例如，见 Fry 1979:135－7），但我们应该清楚，还有很多类似的问题需要作类似的统计学处理。

以上两例说明了统计学能够而且应该对语言学研究中的数据分析和推断有所帮助。可以运用统计学的研究领域相当广泛——应用语言学、语言习得、语言变异和语言学本身。此时与其简短地概述上述每一个领域，还不如详细考察一下一个关于习得研究的例子，以便更好地理解研究者所面临和统计学文献所涉及的问题：研究者要解决的问题是什么，采用什么方法来测定语言行为，适合于这些测定方法的统计处理是什么，结果的可靠性如何。

我们再回到 VOT 问题，并且跟语言习得联系起来，列出一些有关文献。在儿童习得英语清、浊辅音发音区别的过程中，他们要经历哪些阶段？（Macken，Barton 1980a 中进行过这一讨论，其他语言中关于清浊对比习得问题的研究，读者可参阅：西班牙语，Macken、Barton，1980b；法语，Allen，1985；葡萄牙语，Viana，1985。）调查首先从观察儿童早期发音的转写资料开始，发现他们发塞音时常常不区分/p/、/b/，都读做/b/，这或许至少是发音转写时听觉上的印象。有没有这种可能，儿童已经是在区分/p/－/b/了，而转写发音资料的成人却听不出来？VOT 就是用来区分词首清浊音的一条重要的感知线索。英语中浊塞音有一个“短延时”的 VOT 范围（其中唇音、舌尖音范围是 0 至 30 毫秒，软腭音是 0 至 40 毫秒），清塞音则有一个“长延时”的 VOT 范围（60 至 100 毫秒）。说英语的人会把 VOT 值小于 30 毫秒的塞音（这是对唇音、舌尖音而言，对软腭音而言则是小于 50 毫秒）当

成是浊塞音;只要 VOT 值超出这个范围,该塞音感觉上都会被当作清塞音。儿童的发音结果可能是被转写发音资料的成人归入按"长、短延时"的 VOT 定义的音位范畴了。因此,在某个发展阶段,即使儿童能用 VOT 对清浊音作出前后一贯的区分,由于 VOT 值仍然在成人的音位范畴之内,转写人出于他们的感知习惯,很有可能会忽视这种情况。如何调查这种可能性?显然这项研究会牵涉到一系列问题。

(a)我们找一组年龄适当的儿童作实验,以得出数据。像这种持续性研究,必须考虑数据是以纵向方式获取(对同一组儿童作有适当时间间隔的追踪调查),还是以横向方式获取(按年龄分组,在我们感兴趣的年龄范围里,每个年龄段为一组)。纵向方式的缺点是获取数据的时间与儿童自身发展所用的时间一样长,而横向数据则可在一小段时间内获取。然而,有了纵向数据,我们就可以有把握地追踪个体发展进程,并能将 A 时和 B 时的情况作可靠的对比,而横向数据恐怕就没这么清楚了。一旦决定了需要什么样的数据,接下来就得定出样本的容量和样本中的元素。我们对研究结果的概括能力就是取决于上述考虑。Macken 和 Barton 的研究是纵向的,他们选取了 4 个儿童,这些儿童"只说英语,没有上学的兄弟姊妹……已能说出一些以塞音开头的词,语言方面发展正常……肯合作"(1980a:42 - 3),并且,每个儿童的父母都是以英语为母语,所有儿童受到正常教育。这些对被试的各方面的规定,其理由是显而易见的;有关样本容量和结构的问题后面有讨论(详见 §4.4, §7.5)。

(b)第二个是语言学研究中很普遍的问题,就是来自每个人的数据样本的容量。此项研究中的被试是 4 名,而以/ptk/或/bdg/开头的单词数目却无比庞大。(这儿马上就有一个问题,我们是选取较少的被试并相应增加调查项目好呢,还是相反? §7.5 对此有一些讨论。)在 VOT 习得研究中,调查者同样需要对抽样频率以及 6 个词首塞音每个应取多少个词例这样一些问题做出决定。他们的抽样频率为每两周一次,每次调查的词例个数从 25 到 214 不等。(其目的在于每一个塞辅音至少有 15 个词例,刚开始时不能保证每次调查都能

成功。)

(c)一旦得到调查数据并且此前已对每次调查中每个儿童发出的每个词例都做了测量,就需要把它用一种可接受的和易于理解的形式表示出来。Macken 和 Barton 在仪器测量中,将每次调查中的每一个塞音的词例限定为 15 个。很有可能同一塞音的这 15 个词例各有不同的 VOT 值,因此,为了便于评估,需要对这些值加以归纳并且(或者)用图表显示出来。Macken 和 Barton 既用了表格形式来归纳数值,也用了图片形式来表达数据(参看第 2 章和第 3 章关于数据归纳方法的一般性讨论)。

(d)从对儿童 VOT 数据的描述性归纳中,可以就某些儿童词首塞音清浊对立发展的一个阶段得出一些有趣的结论。不要忘了,人们普遍认为唇音、齿龈音清浊塞音的感知界限是 30 毫秒。在被试 Tessa 发齿龈音的早期,词首/d/的平均值为 2.4 毫秒,而词首/t/的平均值为20.50毫秒。这两个值都在成年人浊音范围之内,因而很有可能被成年人当作浊音。但这两个值是很不一样的。从统计学角度来讲,所观察到的这两个均值的差别是显著的吗?或者从调查的角度说,是否 Tessa 已能前后一贯地用 VOT 值来区分词首的/d/和/t/,但由于这两个值处于成年人的某个音位范畴之内而不太可能被感知?有关这个问题,第 10 章做了具体的统计检验,而第 3 ~ 8 章为理解这个问题作了重要的知识准备。

(e)我们现在谈到的只是一个儿童的一个可能显著的差异。作为调查者,我们通常关注的是,基于已作分析的样本数据,将我们的发现扩展到更大的群体而不局限于参加研究的那些被试,在这方面我们能走多远?这个答案很大程度上取决于我们是如何处理上面(b)和(d)中提出的问题的,第 4 章还会对此加以讨论。

这里我们对语音例子讨论较多——并非因为它是语言学中会出现这些问题的惟一领域,而是因为 VOT 是一种容易理解的度量,可以用它来对语言研究的许多领域所共同关心的问题作出直截了当的说明。以后我们将不断地谈到这些问题,并且给出各种研究的参考