语言统计理论与实践

YUYAN TONGJI LILUN YU SHIJIAN

■ 周世界 著

大连海事大学出版社

语言统计理论与实践

周世界　著

大连海事大学出版社

图书在版编目(CIP)数据

语言统计理论与实践 / 周世界著 . —大连 : 大连海事大学出版社,2013. 1
ISBN 978-7-5632-2829-4

Ⅰ. ①语… Ⅱ. ①周… Ⅲ. ①语言统计 Ⅳ. ①H0 - 05

中国版本图书馆 CIP 数据核字(2013)第 027768 号

大连海事大学出版社出版

地址:大连市凌海路 1 号 邮编:116026 电话:0411-84728394 传真:0411-84727996

http://www.dmupress.com E-mail:cbs@dmupress.com

大连华伟印刷有限公司印装 大连海事大学出版社发行

2013 年 2 月第 1 版 2013 年 2 月第 1 次印刷

幅面尺寸:185 mm × 260 mm 印张:10

字数:228 千 印数:1 ~ 500 册

责任编辑:陆 梅 宋彩霞 版式设计:小 月

封面设计:王 艳 责任校对:华云鹏

ISBN 978-7-5632-2829-4 定价:25.00 元

前　言

统计学是研究数据的收集、整理和分析的方法论科学。在学科分类上，统计学分为数理统计学和应用统计学两大部分。数理统计学研究统计学的基本原理，而应用统计学则是研究如何应用统计方法解决某个领域内（如语言学）的实际问题。

语言统计理论与实践，又称语言统计学，是以语言学、语言教学及语言测试等理论为指导，用统计学的原理和方法研究语言现象的一种多科性的应用统计学。语言研究使用了多种统计方法，能够更客观、更准确地描述和研究语言的各种现象，揭示语言现象的规律性，并深化对语言的认识。

本书借助SPSS统计分析软件及对案例应用的介绍，使读者能由浅入深地了解和掌握语言统计分析方法，并灵活运用到语言分析中。

本书适合于从事语言统计研究的各相关专业的读者，尤其适合于从事语言学、应用语言学、教育测量学和测试学研究的专业人员。可作为高等院校语言文学类研究生或本科生的教材使用。

本着对读者负责的态度，在本书的编写过程中，笔者对全部内容反复推敲，力求达到准确、正确、易于理解。但由于笔者水平有限，缺点和错误在所难免，祈望专家、学者和广大读者批评指正。

周世界于2011年

目 录

理论篇

实践篇

理论篇

第一章　统计学的基本概念

语言统计以语言学、语言教学及语言测试等理论为指导，用统计学的原理和方法研究语言的各种现象。它使用大量的统计方法，能更客观、更准确地描述和研究语言，揭示语言现象的规律性，并深化对语言的认识，其核心是统计学（statistics）的理论和方法。

Statistics 一词源于中世纪的拉丁词 statisticum 和意大利语 statista（statesman），后由德语的 statistik 传入到英语并最终演化为 statistics。统计学的学科性质经历了由最初的实质性学科到现代方法论学科的转变过程。

从形位学上讲，statistics 的构词形式可有两种解释。首先，statisitcs 可简单地拆分为“state + istics”两部分，称为国务学（science of state），又叫政治算术，是对国情及国力的各项数据进行比较分析，并以此为依据为社会经济的发展进行规划，属于一门实质性的学科。到了 19 世纪中期，国务学逐渐向方法论学科的方向转变，开始研究数据的收集、整理和分析，以揭示事物（如语言）总体的规律性，始称统计学。在句法上，作为“统计学”意义的 statisitcs 不受其他任何词的修饰或限制；作句子主语时，谓语用动词的单数形式，如“Statistics is an art.”，又如“Statistics is the only mathematical field required for many social sciences.”。

其次，statistics 一词可进一步拆分为“state-ist-ic-s”四部分：state 是词根；-ist 表示“人，研究……的人”；-ic 既可作形容词后缀，又可作名词后缀；-s 是复数名词的标志。-ic 作形容词后缀时，statistic 等同于 statistical，表示“与统计有关的”，如，statistic（al） tables/data/figures 统计表（资料、数字）；而作名词后缀时，statistic 表示统计过程中任何一项具有代表性的统计量（值），其复数形式 statistics 表示“多种统计、统计数字、统计资料、统计方法、统计量等”。在句法上，复数形式的 statistics 具有普通名词的特性，可受其他词的修饰或限制，如 descriptive statistics 描述性统计（法）、inferential statistics 推断性统计（法）；作句子主语时谓语用复数形式，如“These statistics are misleading.”，又如“The statistics from the Census for apportionment are available.”。

一、总体和样本

总体（population），即统计总体的简称，指依据研究目的将大量具有某种共同性质的个别语言现象的观察值（observation）汇总在一起而形成的集合体，是该语言现象观察值的集合，具有同质性和差异性共存的特征。

总体由个体构成。个体组成的总体存在一定程度的共性，即所有个体在共同性因素上均发挥作用，因而总体具有同质性（homogeneity）；同一总体内的个体之间又存在着差异，称为变异（variation）。例如，名词具有人称、数、格的共同属性，并可充当句子的主语、宾语等共同句法功能；又存在着普通、专有名词之分，单、复数之别，具体、抽象之差异。总体没有同质性不能构成统计意义上的总体，没有变异性（variability）则无须统计学的存在。统计学的任务是在变异的背景下探究总体的同质性。

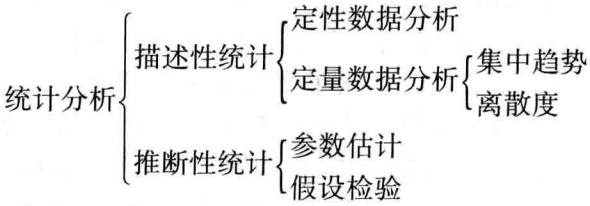
总体分为有限总体(finite population)和无限总体(infinite population)两类。例如,莎士比亚作品无论数量多少,所包含的单词是有限的,所有单词构成一个有限总体;英语单词虽然在一定时期内保持数量的相对稳定,但它们可无限地重复使用,从而产生一个无限总体。

当有限总体的容量适中时,我们可分析整个总体。但对于较大的有限总体或无限总体,一个不漏地观察其中的每一个体通常是不可能的,有时也没有必要。科学的办法是从总体中抽取一部分具有代表性的个体(individual)构成一个样本(sample),并进行深入细致地观察和测量。

统计量(statistics)属于样本指标,而总体指标称作参数(parameter),统计量和参数统称为尺度(measure)。利用统计学的原理和方法,透过样本统计量对总体参数进行描述或推断,是统计学的核心内容。

二、描述性统计和推断性统计

按其目的和功能,语言统计一般分为描述性统计(descriptive statistics)和推断性统计(inferential statistics)两种方法。



描述性统计是语言统计的出发点,主要研究如何把调查到的语言数据经过一系列的观察和计算,用统计表、统计图、统计量等方法,对数据的数量特征及其分布规律进行测定和描述,全面、客观地反映语言数据的全貌,为进一步统计分析和推论提供可靠的依据。

推断性统计以描述性统计为基础,通过样本统计量推断总体参数,即通过对语言样本观察值的统计分析,推断该组样本数据所代表的语言总体特征。既可用于对总体参数的估计,也可用作对总体分布特征的假设检验。统计推断的意义在于:由于各种主客观原因,当统计研究的范围大于实际可能获取数据范围时,必须用统计去推断。它不对问题做出绝对肯定的结论,只是在一定可靠程度保证下,做出能满足研究精度的弹性结论。因此,统计学的推断结论从来不会完全肯定或完全否定。

三、随机现象、频次、频率、概率

语言现象纷繁复杂、多种多样。有一类现象,在一定条件下必然发生。如,语言交际必然使用词、词组或句子,这种现象称为确定性现象或必然现象。另外一类现象,如语言交际中的单词或句型,事先无法做出判断,称为不确定性现象。人们经过长期实践并深入研究发现这类不确定性现象在大量重复实验或观察下,其结果呈现出某种规律性。这种在个别现象中呈现不确定性,而在大量重复的样本中又具有统计规律的现象称为随机现象(random event)。

在任何一个语料总体中,某一特定单词的实际出现次数称为频次(frequency),该特定单词与该语料总体中所有单词数之比称作频率(relative frequency),即语料库语言学中的类符/形符比(TTR)。表 1-1 给出了 LOB 语料库中频次/频率居前的 10 个单词:

表 1-1　LOB 语料库频次/频率居前的 10 个单词

类符	频次	频率(TTR)
the	68 216	6.73%
of	35 733	3.53%
and	27 813	2.74%
to	26 742	2.64%
a	23 023	2.27%
in	21 174	2.09%
that	11 153	1.10%
is	10 968	1.08%
was	10 486	1.03%
it	9 924	0.98%

频率具有随机波动性。在容量相同的不同语料库,同一单词出现的频次不尽相同:语料样本较小时,频率随机波动幅度较大;随着样本容量逐渐增大,频率呈现出相对的稳定性。

在不同容量的数个语料总体中对某一单词进行 N 次抽样,得出该单词在不同语料总体中的所有频率。其所有频率总和与抽样次数 N 之比(即频率的平均数)称作该类符的概率(probability)。概率描述随机语言现象的可能性大小,用 P 表示。随机现象的概率在 0 与 1 之间,即 $0 \leqslant P \leqslant 1$。$P$ 越接近于 1,表明某现象发生的可能性越大;P 越接近于 0,表明某现象发生的可能性越小。严格地说,$P=1$,表示现象必然发生;$P=0$,表示现象不可能发生。习惯上,将 $P \leqslant 0.05$,称为小概率事件,表示发生的可能性很小,构成推断性统计中的显著性水平($\alpha=0.05$)。

四、变量及其测量水平

变量(variable)是相对于常量而言的,指可变动的量。它可以是语言学理论知识,如音位、形位、单词、短语、句子、段落或语篇;也可能是语言习得及学习策略;或语言测试成绩,等等。

不同类型的变量需要运用不同的统计方法分析,识别变量的不同类型对于语言统计有着非常重要的作用。从根本上讲,变量分为自变量(independent variable)和因变量(dependent variable)两种。例如,表 1-2 是某次英语期末考试成绩的汇总表,共有 11 列(姓名、性别、听力、阅读、词汇与结构、完型填空、英译汉、写作、总分、考评、百分比),因此有 11 个变量。第一项变量(姓名)具有决定性作用,称作自变量;其他变量值随自变量的变化而变化,称作因变量。

表 1-2　英语期末考试成绩汇总表

姓名	性别	听力	阅读	词汇与结构	完型填空	英译汉	写作	总分	考评	百分比
薛鹏倩	女	19	28	14.5	9.5	9	10	90	优秀	3.33%
宋飞娟	女	17	28	15	10	8	10	88	优良	36.67%
孙天涛	男	18	28	14	7.5	8.5	10	86		
王娟	女	18	26	13	8.5	7.5	12	85		
王文松	男	17	26	14	8.5	8.5	11	85		
于月梦	女	17	26	13.5	10	7.5	11	85		
孟凡达	男	20	24	12	8	8	12	84		
王彤	女	19	26	13.5	7.5	6	10	82		
白书	女	17	26	13	9	8	9	82		
徐智钱	女	18	26	12.5	7.5	6	10	80		
李晓静	女	16	26	13	8	7	10	80		
徐颖	女	18	26	13	8.5	5.5	9	80		
李代双	男	18	26	13.5	8.5	6	7	79	良好	20.00%
赵倩	女	16	26	12.5	8.5	9	7	79		
赵玥	女	17	24	11.5	8	6.5	11	78		
孙梦情	女	17	24	14	8.5	6.5	8	78		
陈春	男	17	24	13.5	7.5	7	9	78		
付长慧	男	17	24	13	7.5	5.5	8	75		
高立媛	女	14	22	12	7	6	8	69	及格	23.33%
李伟达	男	14	22	13.5	8.5	4	7	69		
李小然	男	18	22	12	8	4	5	69		
周倩婧	女	15	20	11	8.5	5.5	8	68		
杨逸	男	13	18	12.5	7.5	5	8	64		
仇昊	男	14	18	11	7.5	5.5	8	64		
李竞选	男	15	20	11	7	5	5	63		
张天翔	男	14	16	12	7.5	3.5	6	59	不及格	16.67%
魏焕龙	男	12	18	11.5	7	3.5	6	58		
周庆丰	男	11	18	11	6.5	4.5	6	57		
唐新悦	女	14	16	10.5	7	3.5	5	56		
石勇	男	7	6	5	4	2	3	27		

依据对变量数据量化等级(包括分类、标示、计算等特征)的测定,表 1-2 中 11 个变量又分为四种不同的测定层次(level of measurement),从低到高依次为定名、有序、等距、比率测定层次变量。

1. 定名测定层次 (nominal level of measurement) 变量

又称无序分类变量,仅以名称方式分组或分类(labeling),不具有排序、加减等数量化处理的功能,如表 1-2 中“姓名”和“性别”变量。定名测定层次变量进一步分为:

(1)二项(binominal)分类变量,表现为数据的非此即彼特征。如,表 1-2 中“性别”变量(男,女)。又如,动词语态分为主动语态和被动语态;陈述句分为肯定陈述句和否定陈述

句,等等。

(2)多项(nominal)分类变量,表现为互不相容的多个类别的变量。如,表1-2中"姓名"变量,它把每个姓名区别开来,使之互不包含。又如,音位/p/的不同音位变体 /p^{-},p^{h},$p^{\cdot}$/;非谓语动词分为不定式、动名词、分词;状语从句分为时间、地点、原因、方式、让步、目的、结果状语从句,等等。

2. 有序测定层次(ordinal level of measurement)变量

又称有序分类变量,除具有定名变量的分类功能外,在测量层次上又上升到对观察值的排序水平,具有分类、排序双重计算功能。如,表1-2中的"考评"变量,把"总分"变量分为优秀、优良、良好、及格、不及格等高低5个级别。

3. 等距测定层次(interval level of measurement)变量

又称数值变量(numeric variable),具有分类、排序、加减数学运算功能,如表1-2中"听力"、"阅读"、"词汇与结构"、"完型填空"、"英译汉"、"写作"、"总分"等7项变量。

4. 比率测定层次(ratio level of measurement)变量

具有分类、排序、加减、乘除数学运算功能,如表1-2中的"百分比"变量。

依据变量的特性及内在关系,上述四类变量可归结为定性数据变量和定量数据变量两种类型,如表1-3所示。

表1-3　变量类型

分组	测量层次	数学运算功能
定性数据变量	定名	分类
	有序	分类、排序
定量数据变量	等距	分类、排序、加减
	比率	分类、排序、加减、乘除

定量数据变量可分为两种类型:离散性变量(discrete variable)和连续性变量(continuous variable)。离散性变量的所有可能取值可以一一列出且只能取整数值,如单词在语料库出现的频次。连续性变量的可能取值不能逐个列举出来,所有的观察值均为某一区间,例如,评价一个班级的测试成绩,它的取值区间为[0~100]。某个姓名的成绩,可能是75,可能是75.5,也可能是75.55,在这个区间内,数值可以做无数次分割。无论哪一种分割,在其内部有一个等距离的最小单位,因此这种测量水平的变量被称作等距变量。

通常,为了便于数据的统计分析,一种测定层次的变量可转化为另一种类型,一般情况下只能沿着"比率→等距→有序→定名"的方向由高级向低级转化。但当样本容量达到相当规模时,低级测定层次变量也可以借用高级测定层次变量的统计方法来分析。

第二章　抽样与问卷调查

进行语言统计首先要收集被研究语言现象的原始数据或观察值。如何从纷繁复杂的语言现象中获取一定数量的具有代表性的原始数据，这一问题直接影响着统计结果的准确性、可靠性和有效性。本章主要介绍语言统计中常用的数据收集方法——抽样调查法和问卷调查法。

一、抽样调查法

如何依据语言现象的样本推断语言总体的真实特征呢？乍看起来，最理想的办法是对每一个体逐个进行观察，但这种做法一般是不现实的。在统计实践中，我们往往是从总体中抽取一定数量的个体作样本，然后根据样本统计量推断总体参数。如何从总体中抽取一定数量的具有代表性的个体呢？

抽样是从总体中抽取样本的过程，目的在于科学地挑选一定量的个体作为总体的代表，以便通过对局部的统计，准确地推断总体的特征和规律性。为了使推断统计正确可靠，对总体而言，被抽取的样本必须具有代表性，即样本容量足够大，抽取的方法恰当，样本数据精确。抽样遵循的顺序一般为确定抽样总体、选择抽样技术及确定样本量的大小。抽样方法一般分为随机抽样、分层抽样和整群抽样。

1. 随机抽样

随机抽样，又称简单随机抽样，是以完全随机的方式从总体中抽取样本的过程，适用于个体之间差异较小的总体抽样。传统的随机抽样一般采用直接抽选、抽签、摇号等手工方法完成。随着计算机的飞速发展和检索软件的不断更新，对于英语国家语料库（British National Corpus，简称 BNC 语料库）等大容量或超大容量总体，一般采用基于随机数表等方法进行随机抽样。

例如，在高达一亿词次的 BNC 语料库中，start 共出现 23 177 次。研究 start 在词性、词义、搭配、分布等方面的特点不需要检索出所有含 start 的句子，往往只抽取一定数量的句子作样本。在 http://www.natcorp.ox.ac.uk/页面中，输入 start 并点击 Go 按钮，如图 2-1 所示。

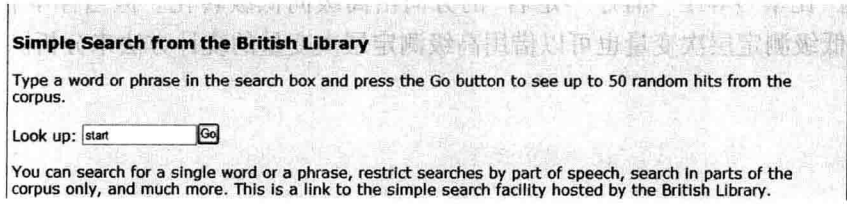


图 2-1　BNC 语料库在线检索页面

运行后，系统随机抽取并输出含有 start 的句子，如：

A3K 81 *Prices for the cottages start at £ 35,000 and for the estates at £ 85,000.*

A6V 1831 *If she can stay, if her husband is allowed to join her, she can start a new life, with people who love her...*

AAF 67 *The Canadian's carriages actually start from Toronto.*

随机抽样简单、直观，且总体中每个个体均有被抽取的机会，适合于个体差异不大且均匀分布的总体。对于总体内部个体变异程度较大的总体，不宜采用随机抽样法。

2. 分层抽样

分层抽样，又称分类抽样或类型抽样，适用于容量大、个体差异较大的总体。分层抽样是先将总体按个体的差异或特征进行分类、分层，然后在各类或各层中随机抽取一定数量的个体组成样本。分层抽样有等比抽样和不等比抽样之分：等比抽样规定在各层中的抽样比例与该层在总体中的比例相同；当总体内个体差别过大时，可采用不等比抽样。

分层抽样可以看作是在分层基础上的简单随机抽样。例如，为了调查 start 在 BNC 语料库中的具体分布情况，我们只在 newspaper 层面上进行随机抽样，以提高抽样效率，减小简单随机抽样的误差。在 http://corpus.byu.edu/bnc/页面中，做如图 2-2 的设定，进行 NEWSPAPER 层面的抽样。

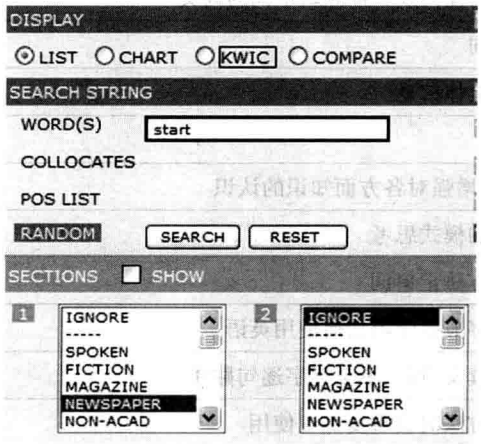


图 2-2　BNC 语料库分层次检索页面

同一个总体可从不同的角度进行分层。例如，BNC 语料库不仅有口语、小说、杂志、报纸、非学术等文体分层，还可以从地域、语域、性别、年龄等层面进行分层。对同一总体不同层面的抽样样本，既可进行独立分析，又可进行不同层面之间的显著性检验。

3. 整群抽样

按照一定标准把总体分成若干群组，从中随机抽取一定数量的群组作样本，并对群组中的所有个体进行调查的方法，称为整群抽样法。整群抽样以群为抽样单位，群的规模可大可小。群与群之间相互独立，互不包含。各群内所包含的个体数目可以相等，也可以不等，但一般不要相差太大。

整群抽样设计和组织方便，节省财力和时间。但在容量等同的情况下，整群抽样误差往往比简单随机抽样大。

二、问卷调查法

问卷调查法是根据研究课题以书面形式搜集语言数据的方法，近年来在语言研究，特别是语言教学研究方面得到了广泛的应用。

语言统计往往采用结构型问卷形式，即把问题的答案加以限制，只允许在问卷所限定的范围内进行选择。结构型问卷答案标准，方便回答，易于进行各种统计处理和分析，并有利于提高问卷的效度和回收率。结构型问卷一般分为二项式量表和里克特式量表两种类型。

1. 二项式量表

二项式量表(binominal scale)多用于表达认同性意见的事实性问卷中，用以调查受试对象的认同态度，通常用“是/否”、“同意/不同意”等表达。两种态度相互对立、互不包含，答案非此即彼，不能有其他选择。表2-1是语言统计中常用的一种二项式量表，目的在于调查和研究“阅读理解中的学习策略”问题。

表2-1　二项式量表样本一

阅读理解中的学习策略应用调查问卷

为了寻找更好的阅读方法，提高阅读效率，我们设计了本组问卷。请您实事求是地回答，谢谢合作。

1	阅读时我总是先浏览整篇文章，然后再仔细阅读	是	否
2	我利用所能找到的任何线索推测新单词的词义	是	否
3	阅读时我不查生词	是	否
4	我通过略读查找具体信息	是	否
5	阅读时我做笔记	是	否
6	我做大量阅读，以增强对各方面知识的认识	是	否
7	我尽量使用英语的模式思考	是	否
8	我会通过联想的方法记单词	是	否
9	我借助词汇表和字典等学习如何使用英语	是	否
10	我在试图理解所读文章时，不会逐字逐句翻译	是	否
11	我在学习新表达方法时会重复练习使用	是	否
12	我用英语总结文章大意	是	否
13	学习新课之前我会预习将要学习的内容	是	否
14	理解文章时我会联系到以前所学到的知识	是	否

二项式量表可转化为表2-2的格式：研究者提供若干选项，受试者从中选出认为最重要的几项。