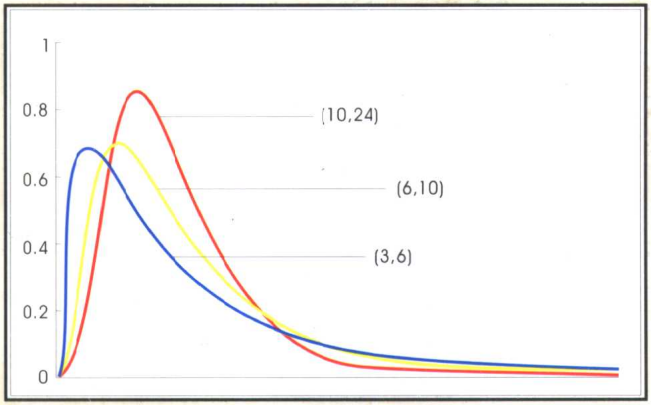
HANYU YANJIU
JIBEN SHULI TONGJI FANGFA

汉语研究基本数理统计方法

张卫国【著】
ZHANG WEIGUO ZHU



中国书籍出版社
ZHONGGUO SHUJI CHUBANSHE

HANYU YANJIU
JIBEN SHULI TONGJI FANGFA

汉语研究基本数理统计方法

张卫国◉【著】
ZHANG WEIGUO ZHU

中国书籍出版社
CHINA BOOK PRESS

图书在版编目(CIP)数据

汉语研究基本数理统计方法/张卫国著. －北京:中国书籍出版社,2002.1

ISBN 7－5068－0805－6

Ⅰ.汉… Ⅱ.张… Ⅲ.数理统计－应用－汉语－研究 Ⅳ.H1

中国版本图书馆 CIP 数据字(2002)第 024900 号

书　　名/汉语研究基本数理统计方法
书　　号/ISBN 7－5068－0805－6/H·271
责任编辑/张　瑞
责任印制/王大军　刘颖丽
封面设计/恳垦工作室
出版发行/中国书籍出版社
地　　址/北京市丰台区太平桥西里 38 号(邮编:100073)
电　　话/(010)63455164(总编室)　(010)63454858(发行部)
经　　销/全国新华书店
印　　刷/北京京海印刷厂
开　　本/850 毫米×1168 毫米　1/32　6.25 印张　168 千字
版　　次/2002 年 1 月第 1 版　2002 年 1 月第 1 次印刷
印　　数/0001－1000 册
定　　价/20.00 元(册)

序

王力先生 1980 年 10 月在武汉中国语言学会成立大会上的讲话《我对语言科学研究工作的意见》中讲道“现在培养的人才都有点瘸腿，社会科学方面的逻辑学、统计学、社会学，自然科学方面的数学、物理学等等也懂得不多，不合现代语言学的要求。我这一辈子吃亏就吃亏在不懂得数理化上。”王力先生是我国语言学界的泰斗，道德文章永为后人景仰，他对我国语言学人才的评价十分中肯，他勇于面对自己的缺陷表明他很谦逊，但是说的也是实话。知之为知之，不知为不知，这正是老一辈语言学家的学者风范。

我们这一代学问远不如老一辈，但是缺陷却一点也没有改进，绝大多数人不懂得数理化，特别是数学基础差，所以学社会语言学和计算语言学，涉及到调查统计的地方真的难于上青天。七十年代末我开始学习社会语言学，不得不学一点统计，由于数学基础太差，一见公式就头疼，就看不下去，所以学得很苦很苦。不得已只能处处从简。结果我关于女国音的调查文章发表后，儿子从美国来信，批评我没有做验证的数学计算。我只得说实话，让他代替我补做，我做不了。学计算语言学就更苦了，儿子给我寄来了最新的著作，我读了三遍也读不下去，以失败而告终，更谈不上掌握和使用了，而儿子读一遍就懂，因为他数学好。可是年逾古稀，从头学数学是不可能了。因此我希望年轻同志及早花一点时间学一学必要的数学知识，至少要改变文科学生那种天马行空，无拘束无根据地驰骋想象的习惯，逐步养成重数据，一步一步严格推理的科学习惯，至于具体的公式倒可以查，不一定一一死记硬背。大概正由于大家的数理化基础比较差，所以生成语法在我国就难成气候，社会调查统计也不为人理解和重视。不少人对社会调查持高度怀疑态度，甚至嗤之以鼻，就是因为不了解科

学的调查统计方法本身早就考虑到了调查数据的代表性和合理的误差，在抽样方法和统计过程中已经尽可能加以避免和纠正。还有人不懂什么是“随机抽样”，认为是“随便抽样”，也就是随便找几个人来调查调查，那样的结论可靠吗！这都是因为很多人根本不懂什么是调查取样、什么是数理统计。

也应该看到，我国老一辈语言学家中也有个别的人，大概是一两位吧，在数学和数理逻辑方面有很深的造诣，在中青年语言学家中也有一两位精通数学；但是，遗憾的是，由于整个语言学界懂数学、数理逻辑的人太少了，因此这样的国宝也很难充分发挥他们的作用。

这样看来，学一点数学恐怕已经是每一个语言工作者的当务之急，完全不学恐怕是要混不下去了。

根据我个人的经验，学一点统计学虽然对一个数学基础很差而年龄又偏大的文科出身的人来说，的确不容易，但是还可以忍受，只要咬紧牙关，挺一挺，花半年几个月也就可以入门，当然也就是入门，能从事语言学研究的调查统计就可以了，要求也不能过高。但是不想花力气、花时间，那是绝对不行的。

张卫国同志数学基础很好，所以学什么都很快，他在英国学了计算机和计算语言学，回国后自学数学和数理统计。他曾经跟我学过语言学，但是就数学和计算机科学而言，我连当他学生的资格都不具备。他写了一部“语言学中的数理方法”的书稿，我读了一半给“枪毙”了，就是因为数学味道太浓，说实话，我看不下去，看不懂。这一次只谈数理统计，因为我啃过一阵，所以还能读下去。至于比我年轻的同行，你们的数学都比我强，只要肯下工夫，一定能读下去并有所收获。因此我非常支持这本书的出版，并且盼望作者今后多在通俗化方面下工夫，让数学和数理逻辑这样非常有用但是面目过于狰狞让人望而生畏的学问能更为平易近人，让多数在这些方面有缺陷的同行有可能补一补课，从而使我国的语言学研究能普遍地更上一层楼。

胡明扬 2002 年 1 月于北京

前　言

《汉语研究基本数理统计方法》这本书是为了适应从事汉语研究、尤其是从事面向信息处理的汉语研究方面的工作和学习的人员在工作和学习中对数理统计的基本理论、手段、方法及使用的需要而编写的。

在语言学的研究和学习、特别是从事面向信息处理汉语研究中，作者自己越来越发现数理统计知识和方法的重要，同时也越来越感到这方面知识和技能的不足和缺乏，于是有了学习有关数理统计的冲动。《汉语研究基本数理统计方法》的最初的稿子，实际上是在作者自己学习数理统计的基本理论、手段、方法的一些体会和经验的基础上编写而成的。

这本书共 6 章，包含了汉语研究中常用到的数理统计的最主要内容，后面附有数理统计常用的数表和概率分布表。大体上说，前两章属于数理统计中描写统计的内容，后四章属于数理统计中推理统计的内容。前两章从介绍数理统计的“样本”、“总体”等基本概念开始，讨论如何科学地抽取样本，概率和概率分布，如何取得样本提供的数据，如何处理和表现样本数据及其反映出来的总体的大概数量特点。后面四章讨论在对样本数据基本处理的基础上，如何从样本数据科学的估计总体的数量特点，如何进一步对统计的结果和结论进行检验，如何进行数理统计中使用广泛、非常重要的方差分析和相关分析。

这本书适合从事汉语研究和教学的人员和研究生学习数理统计使用。数理统计是以数学为工具的科学，让不懂高等数学和仅有中学数学基础的汉语研究工作和学习的人员学习数理统计，并非易事。有鉴于此，一方面，本书紧紧抓住数理统计是汉语研究工作和学习的人员所需要的工具这一特点进行讨论，书中的说明和例子使用语言研究和教学的问题，对数理统计本身的理论和数

学推导过程不多涉及；一方面，本书内容的安排和展开也不能处处只考虑汉语研究、教学而不考虑数理统计本身的特点和系统性。所以，本书是以数理统计本身为纲、以语言研究和教学的需要为目来进行内容安排和展开讨论的。

国内出版的文科人员使用的数理统计的书不算少，可为搞语言研究和教学的人员看的却很少。在自己学习数理统计和本书稿的写作过程中，主要参考了以下著作：《数理统计学》（李茂存 周兆麟主编，天津人民出版社，1983 年 6 月第一版）、《应用数理统计学》（周复恭等编著，中国人民大学出版社，1989 年 9 月第一版）、《统计学》（钱伯海等主编，四川人民出版社，1992 年 6 月第一版）、《统计学》（中国统计出版社，王寿安主编，1994 年 2 月第一版）、*Statistics in Language Studies*（Anthony Woods, Paul Fletcher, Arthur Hughes, Cambridge University Press, 1986）、*Statistical Techniques for the Study of Language and Language Behaviour*（Toni Rietveld, Roeland van out, Mouton de Gruyter, 1993），有的从中还引用或参考了个别例子。这里，对这些书的作者致以深深的谢意。

这本书从初稿到成书，一直得到胡明扬教授的鼓励和帮助，胡先生还在百忙中抽空为这本书写了序。在此，对胡老师再次表示感谢。这本书得以出版，还得到了李行健、殷国光、张瑞等先生及中国人民大学中文系、中国书籍出版社等的帮助，借此机会，对他们一并表示感谢。

总之，本书以让汉语研究工作和学习的人员看得懂、用得上为宗旨。本书的稿子几次作为研究生教材使用，实际效果证明，本书基本上实现了这一宗旨。

由于这本书是在作者自己学习体会和经验基础上写的，书中如有错误和不妥之处，希望读者原谅，并给以指正。

作 者

目　录

第 1 章　抽样调查

1.1 调查取样和统计

1.1.1 数理统计

在语言文字研究中，语言工作者越来越重视、越来越多地采用计量方法，在定性分析的同时进行定量分析，或者通过对调查获得的资料、数据进行统计、计算、分析、归纳，得到有说服力的结论。进行这样的研究，只对数据做简单的计算解决不了问题，而需要掌握数理统计的方法。那什么是数理统计呢？

数理统计是如何进行数据收集、整理、计算、分析和推理判断的科学和工具。收集一定的数据，进行分析计算，是为了得到这些数据所代表的语言现象或事实的数量上的特点。首先，如何收集数据才能使数据足以有代表性，是基本而重要的问题。有了数据，进行分析、计算，虽然可以得到一些数量上的信息，但这只是直观的初步的东西。要得到有说服力的数量上的结论，要通过推理判断获得。数理统计的基本内容就是研究和告诉人们如何进行数据收集、整理、计算、推理的。所以，数理统计，既是一门科学，也是指导人们正确进行数据收集、整理、计算、分析和推理判断的工具。

数理统计的基础是数学。虽然数理统计的研究少不了高等数学，可是我们是把数理统计作为一个工具来学习掌握，所以，可以不涉及高等数学的内容，有了高中数学的基础就够了。

1.1.2 数理统计和调查

数据来自调查。所谓调查，是个泛称，包括调查、观察、试验、测试等。调查是进一步统计的基础，目的是得到数据。调查的方法、形式因调查的对象、使用目的不同而不同，但基本目的都是要得到有代表性的数据，所以，对不同形式调查的根本要求是一样的，即调查必须具有真实性。所谓真实性，就是真正具有代表性，体现在以下几个方面：

首先，调查要有客观性。确定了调查目标和程序后，调查时要实事求是，对收集到的材料、数据要一视同仁，不能凭主观好恶或先入为主的印象加以取舍，更不能加以歪曲或捏造。[1]

其次，调查要准确。一方面，观察要准确，另一方面，对客观材料、数据的记录要准确。

再有，调查要科学。科学性体现在很多方面，最重要的是调查对象的选择要科学，以保证收集到的材料、数据的代表性，这就要掌握正确的取样方法。

1.1.3 基本概念和术语

首先，简单介绍一下数理统计中的一些最基本概念和术语。

1.1.3.1 总体、个体、样本

要研究、描写的对象的全体组成的集合，数理统计中称为“统计总体”或简称为“总体”。构成总体的元素，称为“个体”或“单位”。

进行调查时，常常不可能、也不必要对构成一个总体的每个

[1] 为了讨论的方便和说明问题，书中的例子根据有关材料做了一定的加工，这不能看做“歪曲”。

个体进行调查，而是对从总体中选取的若干个体构成的子集进行直接调查。这个作为调查直接对象的子集，称为“样本”，确定样本的过程叫做“抽样”或“取样”。样本中个体的数目叫做“样本容量”。样本中所有个体作为一个集合，也称为“样本总体”，要记住，它只是统计总体的一个子集。

有的总体是有限集合，有的总体是无限集合。例如，调查一个学校、一个班级学生学习成绩时，学生构成的总体是有限集合。这样的总体称为“有限总体”。再如，进行一种方言调查，这种方言的语句是个无限集合。这样的总体称为“无限总体”。一般情况下，元素数目极大的有限总体，相对于很小的样本来说可看做无限总体，例如，一个大中城市的人口，几十万、数百万，可近似看做无限总体。

1.1.3.2 标志、指标、变量

一个客观对象，有各种不同的属性，数理统计中称不同的属性为“标志”，例如，一个元音，开口度、舌位、口形，可以作为不同的标志加以描写或作为分类的依据。一个标志的体现，称为“指标”。例如，半高、后、圆唇是元音/o/的开口度、舌位、 口形三个标志的指标。

指标的实际表现是一些值。有的值是数字，如，两岁幼儿说话句子的长度、学生的考试百分成绩，等等；有的值是非数字，如一个元音的指标。不管是数字的或非数字的，指标的值有实际值和调查值之分。实际值，是一个指标客观实际的值，调查值，也称观察值或实验值，是通过样本得到的值。对于观察值来说，世纪值也称期望值。调查，就是希望通过调查值了解、估计乃至确定实际值。显然，前者越是接近后者，调查统计越是成功。

一个值，如果是固定不变的（数字或属性），称为“常量”，否则称为“变量”。一个指标在样本中可能有不同的值，就是一

个变量。一个变量，在一次调查或一个样本中可能具有若干不同的值，这些值构成一个系列，称为“数列”或“变量数列”。

1.1.3.3 频度、频率、比率

一个样本中，某个指标出现的次数，称为“频度”或“频次”（token），频度与样本总频度（样本中所有个体的频度之和）的比，称为“频率”（frequency），用百分数表示，称为“比率”。例如，对长度为 n 的汉语语料进行用字统计，统计到不同的汉字 m 个，它们便是不同的标志，第 i（i＝1,2,……m）个汉字使用的次数便是第 i 个字的频度（字频）t_i，t_i 除以总频度 n（$t_1+t_2+\cdots+t_m$）的商 f_i 是第 i 个字的频率（frequency），$f_i \times 100\%$便是第 i 个字在全部样本中所占的比率（proportion）。

在取样合理的情况下，样本的容量越大，一个标志的频度便越大，但它频率不会有太大的变化，所以，频率或比率比频度更能说明统计对象的特点。

1.2 抽样的原理和方法

1.2.1 随机取样

取样调查时，要使调查值（观察值）尽量接近实际值（期望值），样本具有代表性是关键。样本和总体都是集合，它们各自的元素间具有客观存在的某种数量关系，调查统计就是用样本元素间的数量关系映射总体元素间的数量关系，两个集合间的映射关系 F 如图 1.1 所示。

映射 F：S→P 越接近同构映射，就越具有代表性。

要保证样本具有代表性，构成样本的个体必须是从总体中随机选取的，统计学中称之为随机抽样。“随机”中的“机”，是

“几率”，“概率”的意思，即出现的可能性。随机取样，就是按不同标志可能出现的机会的大小选取总体中的个体。在个体出

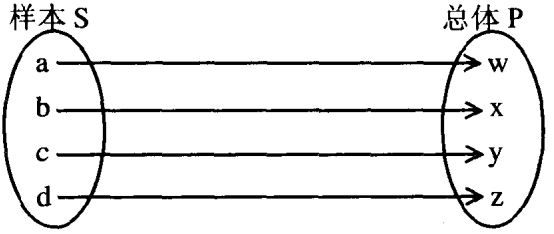


图 1.1 F：S→P

现概率（概率见 1.3）不明的情况下，保证随机取样的原则是总体中的每个个体有相同的机会被选进样本。应该注意，随机取样不是随意取样，更不是随便取样。

要保证样本的代表性，除了使每个个体有相同的中选机会外，样本的容量也十分重要。样本太小，不能保证样本的代表性，样本太大，增加了操作的难度和成本。数理统计理论研究证明，一般来说，容量不到 30 是小样本，容量在 30 以上是大样本。30 是对一个标志而言的最低数目。当然统计的标志不同，这个数目可能不同，但是对语言工作者来说，30 个已经够用，且在语言文字研究中，获得大样本一般来说并不是太难的事情。

1.2.2 抽样的方法

抽样的方法有不少，调查统计时，要根据调查统计的总体的特点，选择合适的方法，以保证抽样的随机性。

下面介绍几种常见的抽样方法。

1、经验抽样

调查人凭自己的经验、判断以及允许取样的条件，从总体中抽取若干个体作为样本。这样抽取的样本没有考虑样本的随机性，代表性很差。

2、等距抽样

把一个有限总体中的个体按某个标准排序编号，然后选取间隔固定的位置上的个体组成样本，这种抽样方法叫等距抽样，例如，一本500页的书，要取50页作为样本，可以从第一页开始，每隔10页取一页，共取50页。再如，要从全体m个学生中取n个学生作为统计样本，按学生的学号，每隔m/n个取一个学生作为样本，就是等距抽样。

等距抽样也叫系统抽样，操作较容易，但不能保证样本具有很好的代表性。

3、分组抽样

抽样前，根据统计对象的特点及统计的要求，对总体进行分组，在分组的基础上再抽取样本。因抽取样本的不同分为两种。

①整群抽样

分组后，选取其中的一组或若干组作为样本，这样的抽样方法称为整群抽样。整群抽样多用做有限总体的抽样统计。例如，要调查某学校学生方言的构成情况，可以把班级作为实际上的分组，随机地选取一个或几个班的学生作为样本。

需要时，也可以对调查对象据不同标准多层分组，然后选取其中若干组作为样本。例如，研究某位作家的语言特点，把该作家的所有作品按体裁分类，不同体裁的作品再按时期分组，然后从不同时期的作品中选取若干作为样本。这是两层分组整群抽样。

整群抽样，实施起来简单易行。但是，因为样本是总体的一个子集，子集内元素之间的差异可能不同于组与组之间的差异，所以，调查值和实际值之间可能存在差异，影响样本的代表性。

②比例抽样

把统计对象分组后，根据各组在总体里所占的比例选取一定数量的个体作为样本加以统计。由于选取样本是按比例进行的，所以也叫做按比例抽样。根据分组是一层还是多层的，又分为一