

实用统计分析方法

蒋庆琅 原著[美]
方积乾 等译



北京医科大学 联合出版社
中国协和医科大学

实用统计分析方法

[美] 蒋庆琅 原著
方积乾 等译

一九九九年十月十日

北京医科大学 联合出版社
中国协和医科大学

(京) 新登字 147 号

SHIYONG TONGJI FENXI FANGFA

图书在版编目 (CIP) 数据

实用统计分析方法/蒋庆琅原著; 方积乾等译. —北京: 北京医科大学、中国协和医科大学联合出版社, 1998. 10

ISBN 7-81034-873-6

I 实… II. ①蒋… ②方… III. 统计资料-分析-方法
IV. C812

中国版本图书馆 CIP 数据核字 (98) 第 12751 号

北京医科大学 联合出版社出版发行
中国协和医科大学

(100083 北京学院路 38 号 北京医科大学院内)

责任编辑: 张 侠 暴海燕

责任印制: 郭桂兰

北京东晓印刷厂印刷 新华书店经销

* * *

开本: 787×1092 1/16 印张: 22.25 字数: 565 千字

1998 年 12 月第 1 版 1998 年 12 月北京第 1 次印刷 印数: 1—3000 册

定价: 45.30 元

献 给

我的父亲和母亲

蒋庆琅

1960.10

翻译组成员

方积乾、柳 青、骆福添、周文清、万崇华、
秦学军（以下姓氏按笔划为序）

宁方芹、史明丽、刘 颜、陈曙光、
吴少敏、林爱华、周凤琼、凌 莉、
崔太亮、游志颖、傅承主、黎 黎

译者序言

本书英文原版的作者是著名华裔美籍生物统计学家蒋庆琅教授。他在美国加州大学伯克利分校生物统计系工作了四十余年，除研究与讲授生命现象的随机过程和寿命表的统计理论外，还曾亲自讲授应用统计学课程并编写了全套讲义。我在伯克利分校攻读博士学位时，曾担任过这门课的助教，亲身体验过这门课深受大批研究生欢迎的情景。经多年积累，这门课的讲义充实、丰满、生动活泼，英国牛津大学出版社早就希望正式出版。蒋先生在退休之后，以其一丝不苟的科研精神认真整理和修改了这门课的讲义，并亲自输入电脑。1994年夏，我再访母校时，有幸获得手稿一套，并连续数日随蒋先生上班，当面聆听其编写原则，细节处理等经验之谈。蒋先生身居海外，心系国内科教事业。遵先生之嘱，我返国之后立即着手组织我的年轻同事和研究生们将此书翻译成中文。其实我完全可以独自翻译，既快又方便。然而，考虑到这也是年轻人学习专业英语的极好机会，宁可让他们每人翻译一章，由我对照原文逐字逐句修改润色。他们付出的劳动自不必说，我的苦心则有稿面上密密麻麻的红字为证。

在反复推敲译文的过程中，原作的清新与精美常令我叹为观止。概念解释之生动而准确、实例列举之有趣而深刻以及习题设置之多采而翔实，其他同类课本实难媲美。我们将此中译本奉献给国内广大师生，建议用作大学应用统计课程的教本、实际工作者的自学读物以及教师和学生们的参考书。

在这里译者们特别感谢本书的特邀责任编辑——北医大生物数学与生物统计教研室张侠副教授。为保证出书质量，他付出了辛勤的劳动。

尽管我们以“慢工出细活”自慰，由于人多手杂和我本人能力所限，中译本不免有疏漏和错误，敬请广大读者和我的导师斧正。

方积乾

1997年5月1日

于广州中山医科大学

前言

本教科书为应用统计学课程而写。它与我在加利福尼亚大学伯克利分校生物统计学专业讲授多年的一门课程有关。这门课为生物学和卫生科学的研究生而开设，同时也吸收了校内其它系的学生。由于这门课有如此广泛的吸引力，我决定在这门课提纲的基础上写一本教科书。为了使本书为更多的读者服务，我对提纲作了必要的修订，加进了如方差分析中的混合模型和随机效应模型、实验设计中区组有效性的估计等内容，并用较多的例子来解释。本书包含三部分内容。前四章是概论；第二部分包括第五章和第六章为统计推断；第三部分为余下的十一章，讲述特定的分析方法，目的是使读者较好地理解统计学和统计分析方法。

前四章是为那些初学统计方法的人而准备的，已具备一些统计学知识的读者，也可用作复习。第一章的描述性统计简单直观。概率对于理解和恰当应用统计方法是必备的认识，第二章简要而适当地重复了这个重要的主题。我总是用一些简单的例子来阐明新概念和新术语的意义，全书均贯穿这种做法。

在第三章和第四章，介绍随机变量和概率分布。随机变量是统计学的基本元素，我将解释其意义并给出例子。第四章的二项分布、超几何分布和泊松分布有指导性作用，并有其实际的应用。近年来，在生存分析和失效分析中指数分布令人瞩目。在第四章第九节读者将遇到所有概率分布之“母”——正态分布。

应用统计的一个重要的功能是统计推断。统计帮助科学家发现未知的世界。统计推断的两个重要分支是估计和假设检验。在第五章，将介绍基于中心极限定理和正态分布的总体均数的区间估计。第六章是关于假设检验基本概念的复习。诸如简单和复合的对立假设、两类错误、检验功效、检验有意义的水平、检验统计量和临界域等术语将充分讨论和解释，并有例子。这是进行统计推断所必备的基础知识，本书的第三部分包括特定的分析方法。无论对于特殊研究或一般性的问题，总有一个研究感兴趣的基本变量（用“ Y ”表示），同时也有其它的因素或自变量。研究的目的是确定自变量对因变量 Y 的分布的影响或估计它们联系的程度。根据 Y 是离散或连续的以及自变量是离散或连续的、随机或选定的，需要不同的统计方法来处理。这些方法总结于后面的表格中，要点细述于后。

在第七章，我们开始讨论特定的分析方法。我们介绍用来进行有关总体均数的假设检验的 Student t 统计量，并对检验的过程进行详细的描述。第八章介绍 χ^2 检验，主要用于分析分类资料，研究中的变量是离散的，观察值以频数分布的形式表达。提出了三类问题：单因素分类，两因素分类和拟合优度。在两因素分类中，有一种是关于两个随机变量（用 X 和 Y 表示）联合分布的情形；另一种是给定 X 关于 Y 的条件分布的情形。前者，检验的是 X 和 Y 间的独立性，而后者检验的是 Y 是否独立于 X 。 χ^2 检验列为表中第一组分析方法。

当研究的两个变量是连续时，第九章的回归分析可用来研究随机变量 Y 和预测变量 X 的依赖性。一般地，回归的形式是线性的，当两个变量均是随机时，第十章的相关分析提供了相互依赖性的度量，第十一章的多重回归和相关用于两个或两个以上自变量和一个因变量的情形，是第九章线性回归和第十章相关的直接扩展（见表中“回归和相关”）。

参数统计的总结——Y 是随机变化的因变量(除 χ^2 检验)

分析方法	自变量	有关的假设	章
χ^2 检验	(离散)		
(Y 是离散的)	X 是选定的	如果 Y 独立于 X	8
	X 是随机的	Y 和 X 间是独立的	
回归和相关	(连续的)		
(Y 是连续的)	X(或 X_1 和 X_2)是选定的	Y 关于 X(或 X_1 和 X_2)的回归	9,11
	X(或 X_1 和 X_2)是随机的	Y 和 X(或 X_1 和 X_2)间的相关	10,11
方差分析(ANOVA)	(离散的)		
(Y 是连续的)			
单向分组方差分析		Y_1 和 Y_2 的期望	7**
固定效应模型	处理效应是固定的	处理效应	12
随机效应模型	处理效应是随机的	处理效应的方差	12
双向分组方差分析(一个观察单位)			
固定效应模型	行和列的效应固定	行和列的效应	13
随机效应模型	行和列的效应随机	行和列的效应的方差	14
双向分组方差分析(几个观察单位)			
固定效应模型	行和列的效应及交互效应是固定的	行和列的效应及交互效应	13
随机效应模型	行和列的效应及交互效应是随机的	行和列的效应的方差及交互效应的方差	14
混合模型	行的效应是固定的,列的效应和交互效应是随机的	行和列的效应的方差及交互效应的方差	14
协方差分析(Y 是连续的)	(Z 是离散的, X 是连续的)		
	Z 和 X 都是固定的	Y_i 和 X_i 的回归	16

* * 两个处理效应的单向分组方差分析, $t^2 = F$

方差分析也研究自变量对因变量分布的影响。自变量是离散的, 而因变量是随机的和连续的, 自变量代表因素或处理, 其对于因变量的影响称为效应。据研究的因素(自变量)数目不同, 分为单向分组方差分析(单个因素), 双向分组方差分析(两个因素)等等。参见表中第三组分析方法。

第十二章为单向分组方差分析。该章的第一部分, 因素的效应是固定的, 模型称为固定效应模型, 研究的目的是估计效应并检验其是否相等; 第二部分, 因素的效应是随机的, 把它们想象为来自因素的“所有效应的总体”的一个样本, 这一模型称为随机效应模型, 统计推断的内容是相应总体中效应的方差。

第十三章的双向分组方差分析用于研究两个因素对一个因变量的效应。因素分为行和列(安排为 $r \times c$ 列联表), 它们对因变量的效应分别称为行效应和列效应。也可能存在两因素的交互效应, 当每个格子有 $n > 1$ 个观察值时可检验交互效应。本章的效应和交互效应数值是固定的, 模型称为固定效应模型。

当效应和交互效应是随机时,模型称为随机效应模型,第十四章提到这个模型。因变量 Y 的期望方差,零假设和推断的范围区别于固定效应模型。当有一个因素的效应是固定的,而另一个是随机时,便是混合模型,也在第十四章讨论。混合模型不仅是固定效应模型和随机效应模型的逻辑推论,它还有自己的特色。混合模型非常实用,有很多应用。

协方差分析是另一种研究自变量对因变量效应的方法。方差分析中一些自变量是离散型的,而在回归分析中自变量则是连续型的。第十六章我们考虑简单的情形,一个离散型变量 (Z) 和一个连续型变量 (X) 影响一个随机的因变量 (Y),这列为表中最后一项分析方法。

第十五章讨论试验设计,但仅包括与前面章节中方差分析有直接联系的内容,即完全随机化设计、随机区组设计、拉丁方和希腊拉丁方设计。前两种设计有适当的描述,拉丁方设计讨论得相当广泛,希腊拉丁方设计是拉丁方设计的简单扩展,仅简略地提到供读者参考。如研究目的为增加试验的有效性,拉丁方设计可形象地视为双重区组设计。但它也可用于研究三个因素、两个因素及一个区组,甚至可用于调查资料的分析而不是实验数据的分析。为了说明特定研究中设计的用处,详细描述了一种评价在拉丁方中划分区组的有效性的方法。到目前为止,提到的统计推断方法需要假定研究的随机变量服从正态分布。基于等级而不是观察值的非参数统计,不需要正态性的假定。第十七章我们介绍一系列常用的非参数方法。在前面的章节中,它们都有相应的参数方法。例如,符号检验对应于二项分布检验,Wilcoxon 两样本的秩和检验和 Mann-Whitney 检验对应于 Student t 检验。最后,Kruskal-Wallis 检验和 Friedman 检验可分别视为观察值等级的单向分组方差分析和双向分组方差分析。非参数统计虽然放在本书的最后一章,但简明扼要。在学习了方差分析和协方差分析的复杂公式后,我相信读者将乐于改变节奏,带着成就感读完本书。

虽然这是一本应用统计的书,我努力为所介绍的分析方法提供理论上的证明,必要时,我在介绍一种分析方法或检验统计量前另设一节给出定理。例如,第七章在用 t 统计量作假设检验之前给出了 Student t 分布的定理。第八章第 2 节,在介绍 χ^2 检验之前,给出了有关 χ^2 分布的定理。这样的定理不仅适合于当时的目的,也可作为后面内容的参考。例如,第九章定理 6 涉及到 χ^2 随机变量和 F 统计量的关系。在后面方差分析和协方差分析的讨论中,读者仅需认出检验统计量的分子和分母是 χ^2 随机变量,便可归结出它们之比服从 F 分布。

我从教学中得知,并非每个人都喜欢推导或证明公式或定理。一些人认为公式只有经过证明才可靠,而另一些人认为证明太讨厌。作为折衷,我把几乎所有的证明从正文中取出,放在另外一节题为“本章结果证明”中,位于每一章的末尾。读者可跳过这一节而不失课文的连续性,或可认真地学习这一节以增强理解。一些人甚至愿意自己来求解,与书中所写的核对。

读者在方差分析和协方差分析的讨论中将发现三项内容对材料的理解是相当有帮助的。第一是线性模型,它使基本假定和零假设的意义非常清楚;第二是均方的期望,期望的明确表达公式可直接引出估计和检验统计量的公式;第三是最后的方差分析表,这个表格有条理地总结了分析的要点。

用图比文字常能更清楚地解释问题,本书也如此。在准备资料时,我作了很多图以帮助读者理解,大多数达到了目的。除统计图和表之外,我还准备了与分析方法相应的练习与思考题,放在每章最后一节。这些练习与思考题和本章的材料直接有关,他们确实是本书不可分割的一部分,这门课的教师可用其作为学生作业的来源。

最后我想说,采用不同的章节,这本书既可用来作为初级课程的教科书,也可作为中级课程的教科书。主要的区别在于是否包括“本章结果证明”这一节。

1)用作初级课程,第一到第十三章,第十五和第十七章。去掉“本章结果证明”。先决条件:大学代数。

2)用作中级课程,第五到第十七章,包括“本章结果证明”。先决条件:大学代数和微积分。

在初级课程中,第十四到十六章可作为进一步学习的内容。在中级课程中,学生可自学第一到第四章,作为复习。但是作为中级课程的先决条件的微积分学,主要是指学生数学上的成熟,而不是微分和积分的技巧。一般地,那些善于领悟概念和对符号运用自如的学生应觉得本书容易阅读。

我非常感谢很多对统计分析方法的发展作出贡献的人,没有他们的工作,不可能写成这本书。我要特别提出 J. Neyman 和 E. S. Pearson,他们在统计估计和假设检验的奠基者之列, R. A. Fisher,对试验设计和方差分析的发展比其他任何人的贡献都大;还有 W. G. Cochran,对 χ^2 检验有重要的贡献。

我也要感谢我的同事 S. Selvin 和 C. Langhauser,他们为我提供了一些示例放在正文或“练习与思考题”一节中。我非常感激他们的慷慨。我还要对 C. Y. Liu 在计算机方面、M. C. White 和 J. Hicks 在编辑方面、S. Tiedemann 在校对方面、B. Hutchings 在手稿的打字和事务管理方面的帮助一并表示感谢。

蒋庆琅, 伯克利
凌莉 方积乾 译
1993. 10. 7

目 录

前言	1
第一章 统计描述	1
第一节 概述	1
第二节 位置的度量	2
2.1 均数	3
2.2 中位数	4
2.3 众数	4
第三节 离散程度的度量	6
3.1 极差	6
3.2 方差	6
第四节 分组资料	8
4.1 分组资料的计算公式	9
第五节 图形	10
5.1 直方图	10
5.2 相对频数多边形	10
5.3 累计频率多边形图	12
5.4 百分位数、四分位数、四分位数间距	12
第六节 举例	13
6.1 频数分布	13
6.2 均数、众数和标准差	13
6.3 中位数	14
6.4 百分位数和四分位数和四分位数间距	14
6.5 直方图、频数多边形图、累计频数多边形图	14
第七节 本章结果证明	15
第八节 练习与问题	15
第二章 概率	16
第一节 基本概念	16
1.1 概率的三个组成部分	16
1.2 概率的定义	16
1.3 概率的取值范围	17
1.4 必然事件与不可能事件	18
1.5 补事件 (对立事件)	18
第二节 复合事件 “ A 和 B ”	18
2.1 条件概率	19

2.2	乘法定理	20
2.3	Bayes 定理	21
2.4	独立事件	22
2.5	相互独立	22
2.6	独立事件的乘法定理	23
第三节	复合事件“ A 和 B ”	23
3.1	加法定理	23
3.2	补事件的复合事件	24
3.3	De Morgan 定律	25
3.4	互不相容事件的加法定理	25
第四节	关于加法定理和乘法定理的说明	25
4.1	概述	26
4.2	结合律和分配律	26
第五节	阶乘、排列与组合	26
5.1	排列	27
5.2	组合	28
第六节	本章结果证明	29
第七节	练习与问题	30
第三章	随机变量	32
第一节	随机变量的定义	32
1.1	示例	32
第二节	联合概率分布	33
2.1	边缘概率	34
2.2	条件概率	34
2.3	随机变量的独立性	34
第三节	随机变量的期望	36
3.1	随机变量期望的一些性质	37
3.2	多个随机变量的线性函数的期望	38
3.3	样本均值的期望	38
第四节	随机变量的方差	39
4.1	方差的两个性质	40
4.2	相互独立随机变量的线性函数的方差	40
4.3	样本均值的方差	40
4.4	样本方差的期望	41
第五节	二维随机变量分布的协方差	41
5.1	协方差的性质	41
5.2	随机变量线性函数的方差	42
第六节	连续型随机变量	42
6.1	两随机变量的联合分布	43
第七节	本章结果证明	44

第八节 练习与问题	45
第四章 概率分布	47
第一节 概述	47
第二节 均匀分布 (离散型)	47
第三节 二项分布, $B(n; p)$	48
第四节 超几何分布	53
4.1 与二项分布的关系	54
4.2 超几何分布的应用	55
第五节 Poisson 分布	56
第六节 多项分布	58
6.1 X_i 与 X_j 之间的协方差	59
第七节 均匀分布 (连续型)	60
第八节 指数分布	61
第九节 正态分布	62
9.1 密度函数和分布函数	63
9.2 标准正态分布 $N(0, 1)$	65
第十节 本章结果证明	66
第十一节 练习与问题	71
第五章 统计推断——区间估计	74
第一节 概述	74
1.1 中心极限定理	74
1.2 二项分布的正态近似	77
第二节 总体均数的区间估计	78
2.1 标准差已知时总体均数的区间估计	78
2.2 标准差未知时总体均数的区间估计	81
2.3 总体方差已知时, 两总体均数差值的置信区间	82
2.4 总体方差未知时, 两总体均数差值的置信区间	82
第三节 总体率的置信区间	83
3.1 单个总体率 p 的置信区间	83
3.2 两总体率差别的置信区间	84
第四节 练习与问题	84
第六章 假设检验——基本概念	86
第一节 概述	86
第二节 假设检验的基本要素	87
2.1 统计学的假设	87
2.2 统计检验和两种假设	87
2.3 两类错误	88
2.4 检验有意义的水准和检验功效	89
2.5 统计量 (或检验统计量) 和临界值	89
2.6 拒绝域和接受域	90

2.7	功效 $1-\beta$ 的计算	90
第三节	假设检验的示例	91
第四节	检验功效	95
4.1	决定样本含量 n	96
第五节	关于单个概率检验的功效	97
第六节	练习与问题	98
第七章	关于总体均数和总体比率的假设检验	100
第一节	概述	100
第二节	t 分布	100
2.1	单样本情形	102
2.2	两样本情形	103
第三节	假设检验的步骤	103
第四节	关于总体均数的假设检验	103
4.1	关于单个总体均数的假设检验	104
4.2	关于两个总体均数的假设检验	104
4.3	两个总体——处理前后的比较	105
4.4	关于假设检验的几个注意点	107
第五节	关于概率的假设检验	108
5.1	有关单个概率的假设检验	108
5.2	关于两个概率的假设检验	109
第六节	练习与问题	111
第八章	χ^2 检验	115
第一节	概述	115
1.1	χ^2 分布	115
第二节	关于 χ^2 分布的一些定理	116
第三节	χ^2 检验	118
3.1	单向分类	118
3.2	双向分类	121
3.3	拟合优度	129
第四节	关于 χ^2 检验的另一些注记	131
第五节	本章结果证明	133
第六节	练习与问题	134
第九章	线性回归	139
第一节	概述	139
第二节	a 和 b 的估计	140
第三节	基础假设	144
第四节	有关的定理	145
第五节	线性回归中的统计推断	147
5.1	关于回归系数 b 的假设检验	147
5.2	期望值 $E(Y X_0) = a + bX_0$ 的置信区间	148

5.3	回归线 $E(Y X) = a + bX$ 的置信带	148
第六节	回归直线的假设检验	151
6.1	F 分布	153
6.2	线性回归的 F 检验	153
6.3	$b=0$ 的假设检验	154
第七节	注记	156
7.1	线性模型	156
7.2	X 值的选择	156
7.3	个体 Y 值的预测区间	157
7.4	内插与外推	157
第八节	一些其它形式的回归	157
8.1	多项式回归	157
8.2	Logistic 回归	157
第九节	本章结果证明	158
第十节	练习与问题	160
第十章	相关	161
第一节	相关系数 ρ_{XY}	161
1.1	样本相关系数 r_{XY}	162
第二节	相关和回归的关系	163
第三节	二变量正态分布	165
第四节	ρ_{XY} 的统计推断	168
4.1	零相关的检验	168
4.2	ρ_{XY} 的置信区间	169
4.3	假性相关?	170
第五节	其它类型的相关	173
5.1	等级相关系数 (Spearman 相关系数) r_s	173
5.2	点双序列相关系数 r_{pb}	174
5.3	Φ 系数	176
第六节	本章结果证明	177
第七节	练习与问题	181
第十一章	多重回归与相关	185
第一节	回归方程	185
第二节	a, b_1, b_2 的估计	185
第三节	定理和推论	187
第四节	复相关系数 R	191
4.1	关于总体复相关系数 ρ_{y-12} 的假设检验	191
第五节	偏相关	193
第六节	一般多重回归	194
6.1	回归方程和常数的估计	194
6.2	统计推断的某些定理	196

6.3 多重相关系数, R	198
第七节 多重 Logistic 回归	198
第八节 本章结果证明	199
第九节 练习与问题	201
第十二章 单因素方差分析	203
第一节 概述	203
第二节 固定效应模型	203
2.1 线性模型	203
2.2 最小二乘估计、平方和 (SSq)、平均平方和 (MSq)	204
2.3 零假设和检验统计量	206
第三节 示例	207
第四节 多重比较	210
4.1 最小有意义差 (LSD) 法	210
4.2 t 化极差法	210
4.3 对比	211
4.4 Bonferroni 不等式	212
4.5 t 检验	213
第五节 随机效应模型	216
5.1 平方和的分解	217
5.2 零假设	217
第六节 两种模型的比较	219
第七节 本章结果证明	220
第八节 练习与问题	223
第十三章 两因素方差分析——固定效应模型	225
第一节 概述	225
第二节 线性模型	226
第三节 每个格子一个观察值	227
3.1 最小二乘估计与平方和 (SSq)	228
3.2 零假设和检验的统计量	229
3.3 多重比较	232
第四节 每个格子 n 个观察值	233
4.1 线性模型	234
4.2 最小二乘估计和平方和 (SSq)	235
4.3 零假设和检验的统计量	236
第五节 本章结果证明	242
第六节 练习与问题	244
第十四章 两因素方差分析——随机效应模型和混合模型	246
第一节 概述	246
第二节 每个格子一个观察值——随机效应模型	246
2.1 平方和与平均平方和	247

2.2 零假设和检验统计量	248
第三节 每个格子 n 个观察值——随机效应模型	248
3.1 平方和与平均平方和	250
3.2 零假设和检验的统计量	251
第四节 两个模型的比较——固定效应模型与随机效应模型	252
4.1 期望和方差	252
4.2 统计推断和统计量	253
第五节 混合模型	256
5.1 主效应和交互效应	256
5.2 效应的估计量和平方和	257
5.3 平均平方和的期望	258
5.4 假设和统计量	258
第六节 本章结果证明	262
第七节 练习与问题	265
第十五章 实验设计	267
第一节 概述	267
1.1 随机化	267
1.2 重复	268
第二节 完全随机化设计	268
2.1 随机数字表	268
2.2 分析方法	269
第三节 随机化区组——单区组	269
3.1 注意事项	270
3.2 实例	270
3.3 线性模型和 F 比	271
3.4 区组效率	271
第四节 拉丁方——双区组	273
4.1 实例	274
4.2 三因素的匀称性	275
4.3 拉丁方的重复	276
4.4 拉丁方中的观测值	277
4.5 线性模型和假设	278
4.6 离差平方和与检验统计量	278
4.7 资料分析：一个实例	280
4.8 拉丁方的效率	282
第五节 希腊拉丁方——三因素区组设计	286
第六节 本章结果证明	287
第七节 练习与问题	289
第十六章 协方差分析	292
第一节 概述	292