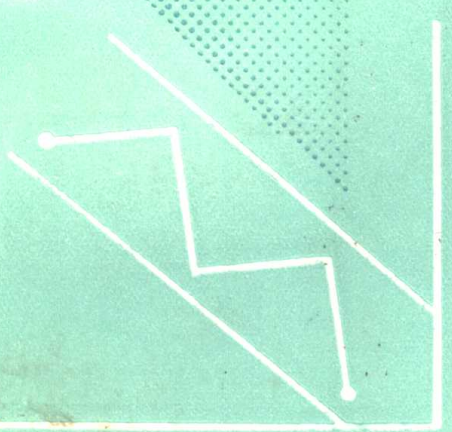
应用数理统计

李金平 编



YINGYONG
SHULITONGJI

河南大学出版社

应用数理统计

李金平　编

河南大学出版社

内 容 提 要

本书是在假定读者已学习了概率论以及数理统计学中的抽样分布、参数估计和假设检验的基础上，作为一门后续课程而编写的。

全书共分五章：相关与回归、方差分析、试验设计、抽样技术和非参数方法。本书侧重应用，同时注意了对问题背景的介绍、统计思想的分析、模型与定理的直观解释、方法间内在联系的剖析，以及应用中出现的一些情况的讨论。

本书配有适量习题并附有答案，可作为理工类各专业本、专科学生和非数学类研究生教材或教学参考书。本书还兼顾到了工程技术人员的需要。

应用数理统计

李金平 编

责任编辑 程庆

河南大学出版社出版

（开封市明伦街85号）

河南省新华书店发行

中国科学院开封印刷厂印刷

开本：850×1168毫米 1/32 印张：12.625 字数：318千字

1992年11月第1版 1992年11月第1次印刷

印数：1—2000 定价：4.10元

ISBN 7-81018-791-0／O·44

（豫）新登字第09号

序　言

数理统计学的建立与发展被认为是人类在本世纪内取得的重大科技成就之一，在许多领域中它已获得广泛的应用.近年来，在我国高等学校中数理统计教育受到普遍的重视．许多专业，不仅理、工、医、农，而且有的文科专业，都开设了数理统计课程.为适应教学的需要，一批各具特色的数理统计著作已陆续出版，丰富了我国数理统计的教材．现在李金平副教授向读者奉献的《应用数理统计》一书，也颇具自己的特色.

数理统计研究的是如何科学有效地收集、整理和分析数据，因此这是一门实用性很强的学科．但其思想丰富、哲理深刻、方法多样、体系完整，要想掌握好这门课程并能灵活运用并非易事.现在的数理统计教材，一般都是先用一定的篇幅讲解概率论准备知识，再花主要篇幅讲解参数估计、假设检验等数理统计中的基本理论，最后简要地介绍一些统计的实用分支．这样写法的好处是内容完整、便于教学；不足之处是由于囿于篇幅，往往对统计中比较实用的部分，未能展开论述，深入介绍．近年来，虽然国内也出版了一些应用统计有关分支的专著，但由于内容专深，不宜于在大学专科或本科教学中使用．因此，设法编些教材，在适当的水准上，在理论与实践的结合上，比较系统地介绍应用数理统计的有关分支，对于统计教学是很需要的．这本《应用数理统计》的出版，正符合了当前时宜的需求.

该书的撰写是假定读者已具有概率论、参数估计、假设检验等基本知识．它集中篇幅较系统地讲解了回归分析、相关分析、方差分析、试验设计、抽样技术、非参数方法等应用统计中一些重要内

容．书中不仅收集了有关经典的统计方法，而且还介绍了一些具有实用价值的近些年来文献中的新成果，如非参数回归、稳健回归、U-统计量、密度估计、最近邻判别等．该书比较强调应用，在正文和习题中收集了丰富的应用实例与数据．这对于培养学生分析问题、解决问题的能力是很有帮助的．但该书也不是对一些公式方法、数据素材的简单罗列．它融合了作者多年来教学和实践的经验体会，努力讲解有关统计思想．在统计问题、模型引入之前，一般都先介绍有关历史渊源和问题背景，着重阐述基本想法．许多历史典籍和应用实例叙述得生动有趣、引人入胜，富有启发性．书中也不回避难点，对一些基本概念和方法常常反复解释，以便加深读者的印象．而对于一些基本理论，一般也作了必要的推导．至于不便证明的，也尽量给予直观阐述，尽力讲清原理．书中还注意各部分内容之间的联系，适时加以对比分析．如通过对回归系数显著性检验与方差分析的比较，帮助读者了解它们的基本原理的相通性；对各处出现的离差平方和分解的归纳总结，帮助读者掌握这一方法的精髓与技巧；对各处秩统计量方法的分析与比较，以帮助读者理解非参数秩方法的基本原理等．该书在讲述各部分具体内容时，还注意帮助读者了解统计学科的基本特征及整体全貌．

总之，作者在撰写中在许多方面作了有益的尝试，使得该书内容丰富、水平适中、深入浅出、通俗易懂．它可用作理工科本科、专科学生及非数学类研究生的教材或教学参考书，想必会收到较好的效果．也可用作应用统计研究和推广工作的参考书，对于广大工程技术人员，想必也会有所帮助．

陈桂景

目　录

绪　论

“统计学”一词，在英文中为 statistics，它源于state(国家、状态)，意思是一个国家所收集的国情资料.

统计学最初的萌芽是户口调查，其目的大抵是为了征兵和税收的需要. 在(基督教)旧约第二十四章中就记载了大约公元前1500年对以色列士兵进行的一次人口调查的很有趣的一段叙述. 而早在公元前3000年，古巴比伦、中国和埃及都已进行过人口调查. 收集、记录和整理各种数据这一人类社会活动，在历史上可以追溯到很远. 翻开中国史书，其中不乏关于钱粮人口、洪水、地震等记载.

16世纪早叶，伦敦开始出现死亡公报. 开始时，公报只是公布死于瘟疫的人数，后来又扩充到包括受洗礼的人数. 到该世纪末期，还包括死于其它疾病的人数.

19世纪，约翰·贝内特·拉维斯(John Bennet Lawes)根据罗瑟姆斯坦特五块田地上的小麦年产量，估计了1852～1879年整个英格兰和威尔士每英亩产量的变化.

随着人类社会活动和经济活动的进步，提出了对各种数据资料的需要，例如国民收入情况、教育状况等等.

大约在70～100年前，人们对统计学一词的意义通常理解为对数据资料的收集、整理以及将其编制成图或表格的形式. 直到今天，这项工作仍然是统计学的一个重要方面，称为“描述性统计”.

近几十年来，一方面，数学，特别是概率论的发展，为统计学发展提供了必要的工具；另一方面，生产和科学技术的飞速进

步，对统计学提出了更高、更迫切的需要，成为统计学发展的强大推动力，使得统计学理论不断完善，方法不断发展，逐渐成为许多科学技术领域的有力的以至不可缺少的工具．特别自第二次世界大战以来，统计学的进步尤为显著，逐步形成了今天这样一门分支众多、内容广泛、理论严谨的“数理统计学”．

数理统计学的最大特点是研究内容庞杂．它是面对“不确定性”(即随机性)的情况作出决定的一门学问．而在自然科学、工农业生产、社会、经济等方面，随机因素的影响普遍存在．因此可以说，几乎在人类活动的一切领域中都程度不同地用到数理统计学的理论和方法．下面，我们对数理统计学的基本研究内容予以扼要介绍，并结合实际例子加以说明和解释．

凡用观察或试验的方法研究问题时，都离不开收集数据和对数据进行科学的分析，在此基础上对所研究问题作出某种形式的结论．粗略地讲，收集数据和进行统计推断，这构成了数理统计学的两大基本内容．

一、有效地收集数据

所谓有效，一是要求所得到的数据便于数学上的处理，就是能用尽量简单方便的数学模型来描述这批数据；二是用尽可能少的人力、物力、财力上的代价，使所获得的数据尽量多地携带所研究对象的信息．比如要进行某项社会调查，如果被调查对象是经过有意识“挑选”的，这样的数据就可能没有“代表性”．基于这样的数据所作的结论就会不可靠，甚至会把我们引入歧途，这种数据就不是有效的．收集数据大致有如下两种途径：

1．抽样观察

比如要考察一批产品的质量情况，一是产品批量可能很大，不可能逐件进行检查；二是有些检查是破坏性的，比方考查灯泡寿命，全面观察就使得生产和检查都失去了意义．总而言之，我们经常需要抽取一部分个体，称为样本．这里有两个问题需要考

虑：①抽多少？抽太多会造成不必要的浪费，抽太少又会“代表性”不够．②如何抽？统计学中研究这类问题的分支称为抽样技术．

2．试验设计

假定生产某种药品，有3个影响质量指标的因素：①反应温度——3个水平：100℃，110℃，120℃；②反应时间——3个水平：6小时，8小时，10小时；③某二成分的摩尔比——3个水平：1:1.2，1:1.6，1:2.0．目的：选择最佳工艺条件．

三个因素有$3^3=27$种不同水平的搭配，要进行全面试验，需要做27个试验．如果说这么多次的试验人们还可以接受的话，那么我们看一个复杂的例子：6个因素，每因素各有5个水平，全面试验的次数多达$5^6=15\,625$！

这就向人们提出一个问题：如何安排试验，以便能用尽可能少的试验次数尽可能多地反映出所研究问题的概况，达到预期目的．对这类数学问题的研究构成了统计学的另一重要分支——试验设计．

二、有效地使用数据

如上收集到的数据是一批杂乱无章的数字，比方是一批灯泡的寿命，我们不能一目了然地从这批数字中直接看出所需结果．这就需要对数据进行一些加工整理，把样本中包含的有关总体的信息“浓缩”在为数不多的统计量里．假如要对总体的某一个量进行估计，虽然使用的是同一个样本，但由于考虑问题的角度不同，可以提出不同的估计量．哪一个估计量好呢？在什么意义下好呢？这些问题的探讨属于统计推断这一分支，它是统计学的主要部分．

三、得出某种形式的结论

1．推断

推断就是从样本推知总体的某些方面．由于数据资料受随机

因素影响，这就使得我们的结论不可能百分之百正确.还是从一个例子说起：我们想要比较A，B两种药物对某种疾病的疗效，假定实际情况是B比A好，但这一点我们并不知道．甲选了200个病人做试验，结果发现，服A药的100人中60人痊愈，服B药的100人中80人痊愈，甲于是得出结论：B比A好．当然，甲的结论是正确的．若某乙也选200个病人做试验，完全有可能得出A比B好的结论．当然，乙的结论是错误的．这个例子也许有点极端，但我们不能排除这种可能性．事实上，影响一个人服药效果的很多随机因素是我们无法定量地加以控制的．比如：一个从来不吃药的人，第一次吃药可能特别见效；对多数人有效的药物可能对少数人效果不大；甚至某种关怀和安慰产生的精神作用也会导致症状明显减轻等等．这样，当我们得出一个结论时，同时也需要知道这个结论的可靠程度有多大．这是统计推断的特点．

2．预测

如果我们要研究的是某地区十岁儿童的身高情况，如上所说，这是一个推断问题．但若我们的目的是研究这批儿童成年后的身高情况，这就是个预测问题．此类问题甚多，比如：明年长江的最高水位将达到多少？明年市场对某种商品的需求量多大？更典型的，每天的天气预报等等．在不少情况下，预测是统计学研究的最终目的，而推断只是为这一目的服务的中间环节。

3．统计决策

推断的目的在于客观地弄清事实，而决策就牵涉到采取行动的问题．比如某公司委托一位统计学家做一项市场调查，统计分析的结果表明，某种商品未来的市场销售量可达10 000件．但公司决定生产多少，这就是另外一个问题了，它还需考虑少生产将会带来的利润损失，多生产可能造成的积压损失以及资金周转情况等等．通过综合比较，最后决定生产8 000件还是12 000件，这就是决策问题．

以上我们从应用的角度对统计学的性质和任务作了一个粗略的介绍，当然，这些还不是统计学的全部.

有人称统计学为一门艺术，因为它的方法在实用中可以有很大的灵活性和创造性，科学家们可以使用任何他所能想到的方法和技术，远非简单地套用公式问题. 这就给学习统计学的人提出了一个要求：必须正确、深入地理解它的概念和方法，掌握统计思想，以求达到融汇贯通.

第一章　相关与回归

在自然科学、工程技术以至社会科学等领域，常碰到同处于一个统一体中的若干变量之间存在着某种互相联系、互相制约的依存关系.这种关系不是那种严格的所谓函数关系，而是一类非确定性的关系，统计学上称为相关关系．这种例子俯拾皆是，比如：气温、降雨量与农作物产量，某矿石中A成分含量与B成分含量，一个化学反应过程中某物质的回收率与反应时间、容器内的温度和压力，纤维强度与拉伸倍数.

回归分析是研究一个变量和其它若干个变量之间的相关关系的一种数学工具，它是在一组试验或观测数据的基础上，寻找被随机性掩盖了的变量之间的依赖关系．粗略一点讲，可以理解为用一种确定的函数关系去近似代替比较复杂的相关关系．这个函数称为回归函数，实用中称为经验公式．搞清楚诸变量之间的依赖关系，我们可以根据几个变量的取值，有效地估计另一个变量的取值或进行其它统计分析．这在实际中是很有用的．例如：要估计一个地区某种农作物的单位面积产量，可以根据该地区的降雨量、单位面积播种量、施肥量以及单位面积产量与它们之间的回归关系式对产量作出满意的估计.

“回归”一词是高尔顿(F. Galton)于1886年首先提出的．他在研究家族成员之间的遗传规律时发现：虽然高个子的父亲确有生高个子儿子的趋向，但一群高个子父亲的儿子们的平均身高却低于父亲们的平均身高；反之，一群低个子父亲的儿子们的平均身高高于父亲们的平均身高．高尔顿称这一现象为“向平均高度的回归”，也即回归到“平均祖先型”．今天人们对“回归”这一概

念的理解与高尔顿的原意已有很大不同，但这一名词一直沿用下来，成为统计学中最常用的概念之一.

相关分析和回归分析作为研究多个变量之间的非确定性关系的一种有力工具，在理论和应用上都有了广泛而深入的发展，成为统计学中最活跃的分支之一. 在这一章中，我们主要讨论两个变量间的线性相关关系和线性回归问题，因为它不仅概括了一大类实际统计问题，而且由于其结构简单、处理方便，从而成为近似地处理其它类问题的较合适的模型和处理较复杂问题的基础. 对近年来发展起来的非参数回归和稳健回归，也作了基本介绍.

§1.1 相关系数及其检验

一、散点图

设(x,y)是一个二维随机向量，我们作了n次观测，得到n对数据(x_1,y_1)，(x_2,y_2)，…，(x_n,y_n). 把这n个点画在一个平面坐标系内，得到的图形称为散点图.

两个随机变量之间相关关系的强弱和类型不同，散点图的形状也各不相同(见图1-1).

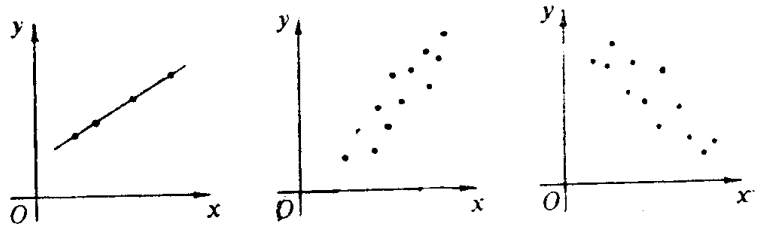
二、相关系数的概念

相关系数是刻划x，y间线性相关关系强弱的一个量. 设有容量为n的样本(x_1,y_1)，(x_2,y_2)，…，(x_n,y_n). x与y间的样本相关系数定义为：

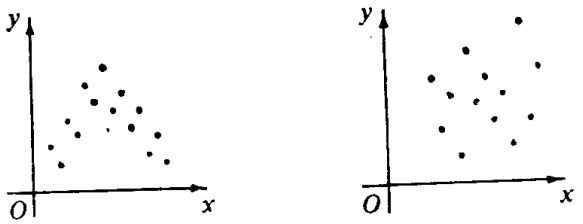
$$r=l_{xy}/\sqrt{l_{xx}\cdot l_{yy}},$$

其中
$$l_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2,\qquad l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2,$$

$$l_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}),$$



(a)严格线性关系　　(b)正相关　　(c)负相关



(d)曲线相关　　(e)无相关关系

图 1-1　各种相关关系的散点图

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\qquad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i.$$

在 l_{xy} 的表达式中，每一项既含有诸 x_i 与样本均值 $\bar{x}$ 的离差，又含有诸 y_i 与其样本均值 $\bar{y}$ 的离差．若两类离差均同号，说明 x 与 y 变化趋势是一致的，这时 l_{xy} 是一些正数的和，因而其值就比较大．反之，若两类离差均异号，x 和 y 的变化趋势正相反，l_{xy} 就是一个绝对值较大的负数．这两种情况都反映出 x 与 y 间有较强的相关关系．若 l_{xy} 的各项中，正、负分布均匀且它们的出现次数大体相当，说明 x 和 y 间的依存关系较松散，无一定变化趋势，l_{xy} 中的正、负项互相抵消一部分，使得 l_{xy} 为一个绝对值较小的数．另外我们看到，r 的值不依赖 x 和 y 的单位的选取，它的大小反映的是 x 和 y 之间关系的相对密切程度，从而不同问题的 r 值可以互相进行比较．

易见，$-1 \leqslant r \leqslant 1$．$r=0$，称 x 与 y 不相关；$r>0$，称 x 与 y 正相关；$r<0$，称 x 与 y 负相关；$r=\pm 1$，称 x 与 y 之间存在线性关系．需强调指出的是，当 $r \approx 0$ 时，我们即认为 x 与 y 之间几乎无线性关系，但完全有可能是某种曲线相关关系．

三、相关系数的计算

为了更方便地计算样本相关系数，可通过简单的代数运算，得如下公式：

$$l_{xx}=\sum_{i=1}^{n} x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2,$$

$$l_{yy}=\sum_{i=1}^{n} y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2,$$

$$l_{xy}=\sum_{i=1}^{n} x_i y_i-\frac{1}{n}\left(\sum_{i=1}^{n} x_i \cdot \sum_{j=1}^{n} y_j\right).$$

这样的表达式使我们计算 r 时的各步骤可以在表格上进行．另外，由 r 的表达式可以看出，当诸 x_i（或诸 y_j）同加、减、乘、除一个数时，r 值保持不变．利用这一特点，当一批数据普遍过大或过小时，可通过数据变换大大减少不必要的计算量．在后面的例子中可以看到这一点．

例 1 某矿石中含有 A，B 两种有用成分．我们发现，在 A 的含量较高的矿石中，B 的含量也较高．我们猜测 A 与 B 的含量之间存在一定的相关关系．今从不同区域和不同时间开采的该矿石中，取出一个大小为 10 的样本，数据如下表．求 A 与 B 的含量之间的相关系数．

样品编号	1	2	3	4	5	6	7	8	9	10
A含量 x_i(%)	67	54	72	64	39	22	58	43	46	34
B含量 y_i(%)	24	15	23	19	16	11	20	16	17	13