



理工类本科生

21世纪高等学校数学系列教材

应用数理统计

■ 主编 赵喜林 李德宜 龚谊承



WUHAN UNIVERSITY PRESS

武汉大学出版社



清华大学出版社

TSINGHUA UNIVERSITY PRESS

应用数理统计

王松桂 编著 王松桂 王松桂 王松桂



清华大学出版社
TSINGHUA UNIVERSITY PRESS



理工类本科生

Mathematics

21世纪高等学校数学系列教材

应用数理统计

- 主 编 赵喜林 李德宜 龚谊承
- 副主编 尹水仿 熊 丹 李春丽
- 编 委 丁咏梅 何晓霞 张 强



WUHAN UNIVERSITY PRESS

武汉大学出版社

图书在版编目(CIP)数据

应用数理统计/赵喜林,李德宜,龚谊承主编. —武汉:武汉大学出版社, 2009. 8

21 世纪高等学校数学系列教材

理工类本科生

ISBN 978-7-307-07245-9

I. 应… II. ①赵… ②李… ③龚… III. 数理统计—高等学校—教材 IV. O212

中国版本图书馆 CIP 数据核字(2009)第 134790 号

责任编辑:李汉保 责任校对:刘 欣 版式设计:支 笛

出版发行:武汉大学出版社 (430072 武昌 珞珈山)

(电子邮件:cbs22@whu.edu.cn 网址:www.wdp.whu.edu.cn)

印刷:通山金地印务有限公司

开本:787×1092 1/16 印张:22.75 字数:452 千字 插页:1

版次:2009 年 8 月第 1 版 2009 年 8 月第 1 次印刷

ISBN 978-7-307-07245-9/O·408 定价:30.00 元

版权所有,不得翻印;凡购买我社的图书,如有缺页、倒页、脱页等质量问题,请与当地图书销售部门联系调换。

21 世纪高等学校数学系列教材

编 委 会

- | | | |
|-------|-----------|--------------------------|
| 主 任 | 羿旭明 | 武汉大学数学与统计学院，副院长，教授 |
| 副 主 任 | 何 穗 | 华中师范大学数学与统计学院，副院长，教授 |
| | 蹇 明 | 华中科技大学数学学院，副院长，教授 |
| | 曾祥金 | 武汉理工大学理学院，数学系主任，教授、博导 |
| | 李玉华 | 云南师范大学数学学院，副院长，教授 |
| | 杨文茂 | 仰恩大学（福建泉州），教授 |
| 编 委 | （按姓氏笔画为序） | |
| | 王绍恒 | 重庆三峡学院数学与计算机学院，教研室主任，副教授 |
| | 叶牡才 | 中国地质大学（武汉）数理学院，教授 |
| | 叶子祥 | 武汉科技学院东湖校区，副教授 |
| | 刘 俊 | 曲靖师范学院数学系，系主任，教授 |
| | 全惠云 | 湖南师范大学数学与计算机学院，系主任，教授 |
| | 何 斌 | 红河师范学院数学系，副院长，教授 |
| | 李学峰 | 仰恩大学（福建泉州），副教授 |
| | 李逢高 | 湖北工业大学理学院，副教授 |
| | 杨柱元 | 云南民族大学数学与计算机学院，院长，教授 |
| | 杨汉春 | 云南大学数学与统计学院，数学系主任，教授 |
| | 杨泽恒 | 大理学院数学系，系主任，教授 |
| | 张金玲 | 襄樊学院，讲师 |
| | 张惠丽 | 昆明学院数学系，系副主任，副教授 |
| | 陈圣滔 | 长江大学数学系，教授 |
| | 邹庭荣 | 华中农业大学理学院，教授 |
| | 吴又胜 | 咸宁学院数学系，系副主任，副教授 |
| | 肖建海 | 孝感学院数学系，系主任 |
| | 沈远彤 | 中国地质大学（武汉）数理学院，教授 |
| | 欧贵兵 | 武汉科技学院理学院，副教授 |

赵喜林	武汉科技大学理学院, 副教授
徐荣聪	福州大学数学与计算机学院, 副院长
高遵海	武汉工业学院数理系, 副教授
梁 林	楚雄师范学院数学系, 系主任, 副教授
梅汇海	湖北第二师范学院数学系, 副主任
熊新斌	华中科技大学数学学院, 副教授
蔡光程	昆明理工大学理学院数学系, 系主任, 教授
蔡炯辉	玉溪师范学院数学系, 系副主任, 副教授
执行编委 李汉保	武汉大学出版社, 副编审
黄金文	武汉大学出版社, 副编审

内 容 简 介

本书系统地介绍了数理统计的基本内容和在 SPSS 中相应的操作方法. 主要内容分为六章: 数理统计的基本概念, 参数估计, 假设检验, 方差分析与正交试验设计, 回归分析, 多元统计分析初步.

本书读者对象为理工类非统计专业硕士研究生, 也可以作为理工类专业本科生、高校教师、科技工作者和工程技术人员的参考书.

序

数学是研究现实世界中数量关系和空间形式的科学. 长期以来, 人们在认识世界和改造世界的过程中, 数学作为一种精确的语言和一个有力的工具, 在人类文明的进步和发展中, 甚至在文化的层面上, 一直发挥着重要的作用. 作为各门科学的重要基础, 作为人类文明的重要支柱, 数学科学在很多重要的领域中已起到关键性、甚至决定性的作用. 数学在当代科技、文化、社会、经济和国防等诸多领域中的特殊地位是不可忽视的. 发展数学科学, 是推进我国科学研究和技术发展, 保障我国在各个重要领域中可持续发展的战略需要. 高等学校作为人才培养的摇篮和基地, 对大学生的数学教育, 是所有的专业教育和文化教育中非常基础、非常重要的一个方面, 而教材建设是课程建设的重要内容, 是教学思想与教学内容的重要载体, 因此显得尤为重要.

为了提高高等学校数学课程教材建设水平, 由武汉大学数学与统计学院与武汉大学出版社联合倡议、策划, 组建 21 世纪高等学校数学课程系列教材编委会, 在一定范围内, 联合多所高校合作编写数学课程系列教材, 为高等学校从事数学教学和科研的教师, 特别是长期从事教学且具有丰富教学经验的广大教师搭建一个交流和编写数学教材的平台. 通过该平台, 联合编写教材, 交流教学经验, 确保教材的编写质量, 同时提高教材的编写与出版速度, 有利于教材的不断更新, 极力打造精品教材.

本着上述指导思想, 我们组织编撰出版了这套 21 世纪高等学校数学课程系列教材. 旨在提高高等学校数学课程的教育质量和教材建设水平.

参加 21 世纪高等学校数学课程系列教材编委会的高校有: 武汉大学、华中科技大学、云南大学、云南民族大学、云南师范大学、昆明理工大学、武汉理工大学、湖南师范大学、重庆三峡学院、襄樊学院、华中农业大学、福州大学、长江大学、咸宁学院、中国地质大学、孝感学院、湖北第二师范学院、武汉工业学院、武汉科技学院、武汉科技大学、仰恩大学(福建泉州)、华中师范大学、湖北工业大学等 20 余所院校.

高等学校数学课程系列教材涵盖面很广, 为了便于区分, 我们约定在封首上以汉语拼音首写字母缩写注明教材类别, 如: 数学类本科生教材, 注明: SB; 理工类本科生教材, 注明: LGB; 文科与经济类教材, 注明: WJ; 理工类硕士生教材, 注明: LGS, 如此等等, 以便于读者区分.

武汉大学出版社是中共中央宣传部与国家新闻出版署联合授予的全国优秀出

版社之一. 在国内有较高的知名度和社会影响力、武汉大学出版社愿尽其所能为国内高校的教学与科研服务. 我们愿与各位朋友真诚合作, 力争使该系列教材打造成为国内同类教材中的精品教材, 为高等教育的发展贡献力量!

21 世纪高等学校数学系列教材编委会

2007 年 7 月

前 言

数理统计是工科院校硕士研究生的一门公共基础课,作者多年来一直从事这门课程的教学.作为一门应用性很强,并借助计算机技术广泛渗透到各个应用领域的课程,越来越受到重视.这本教材是根据工科学生的特点和注重课程的应用性而编写的.全书将统计软件 SPSS 和数理统计的内容有机地结合在一起,书中的例题除了给出手工计算结果外,同时也给出了在 SPSS 中的操作步骤,书中显示的 SPSS 操作界面是 17.0 的版本.编写这部分内容的目的是,一方面使学生会用统计软件解决实际问题,不至于理论与实际脱节;另一方面,又明白统计软件背后的理论基础是什么.针对工科硕士生的数学基础,一些繁琐而复杂的理论推导给予省略,取而代之以简单直观的描述.统计模型的引入、统计方法的介绍,尽量做到从实例出发,循序渐进,简明易懂.考虑到多元统计分析方法的重要性,最后一章介绍了多元正态分布,聚类分析,判别分析,主成分分析和因子分析的基本内容.为了提高读者学习的兴趣,每章前面都给出了一个有趣的应用实例.

全书共分六章,分别介绍了数理统计的基本概念、参数估计、假设检验、方差分析与正交试验设计、回归分析、多元统计分析初步.各章配有适量习题,书末附有答案,关键词汇给出了英文拼写.全教材的理论学时大致需要 52 个学时,可以适当配备一些实践学时.

本书可以作为理工类非统计专业硕士研究生,高年级本科生的数理统计课程教材、也可以作为高校教师、科技工作者和工程技术人员的参考书.

本书由赵喜林、李德宜、龚谊承主编,尹水仿、熊丹、李春丽任副主编.由赵喜林、李德宜和龚谊承提出编写思路.各章的编写人员为:第一章由龚谊承编写,第二章由赵喜林编写,第三章由李春丽编写,第四章由丁咏梅编写,第五章由何晓霞编写,第六章由熊丹和张强编写,全书的 SPSS 操作部分由熊丹编写,英文关键词由龚谊承整理,赵喜林对全书做修改和统稿,熊丹和李春丽做了全书的校对工作.

由于作者水平有限,书中的不妥之处在所难免,恳请读者提出批评和建议.

作 者

2008 年 12 月于武汉

目 录

第 1 章 数理统计的基本概念	1
§ 1.1 导论	2
§ 1.2 数理统计的基本概念	5
§ 1.3 抽样分布	18
本章小结	26
习题一	28
第 2 章 参数估计	30
§ 2.1 点估计	31
§ 2.2 点估计的评价标准	37
§ 2.3 区间估计(interval estimation)	45
本章小结	60
习题二	62
第 3 章 假设检验	64
§ 3.1 假设检验的思想方法	65
§ 3.2 正态总体均值和方差的假设检验	70
§ 3.3 分布拟合检验	86
本章小结	116
习题三	117
第 4 章 方差分析与正交试验设计	121
§ 4.1 单因素方差分析	125
§ 4.2 两因素方差分析	138
§ 4.3 正交试验设计	155
本章小结	166
习题四	168
第 5 章 回归分析	172

§ 5.1 一元线性回归	174
§ 5.2 一元非线性回归	193
§ 5.3 多元回归分析	201
本章小结	211
习题五	213
第 6 章 多元统计分析初步	216
§ 6.1 多元正态分布	218
§ 6.2 聚类分析	224
§ 6.3 判别分析	246
§ 6.4 主成分分析	263
§ 6.5 因子分析	273
本章小结	290
习题六	291
附件 1 SPSS 简介	300
附件 2 关键词	313
附件 3 常用分布表	318
习题答案	348
参考文献	352

第1章 数理统计的基本概念

看不见的价值

1941年,第二次世界大战期间.有一天,美国哥伦比亚大学著名的统计学家沃尔德(Abraham Wald,1902—1950)教授,来了一个意外的访客,那是英国皇家空军的作战指挥官.他说:沃尔德教授,每次飞行员出发去执行轰炸任务,我们最怕听到的回报是:“呼叫总部,我中弹了”.请协助我们改善这个关系着飞行员生死的难题吧!沃尔德接下这个紧急研究案,分析德国地面炮火击中联军轰炸机的资料,建议机体装甲应该如何加强,才能降低被炮火击落的机会.但依照当时的航空技术,机体装甲只能局部加强,否则机体过重,会导致起飞困难及操控迟钝.

沃尔德的研究发现,机翼是最容易被击中的部位,而飞行员的座舱与机尾,则是最少被击中的部位.

沃尔德详尽的资料分析,令英国皇家空军十分满意.但在研究成果报告的会议上,却发生一场激辩.负责该项目的作战指挥官说:“沃尔德教授的研究清楚地显示,联军轰炸机的机翼,弹孔密密麻麻,最容易中弹.因此,我们应该加强机翼的装甲.”

但沃尔德却坚定而客气的说:“将军,我尊敬你在飞行上的专业,但我有完全不同的看法,我建议加强飞行员座舱与机尾发动机部位的装甲,因为那儿最少发现弹孔.”

在全场错愕怀疑的眼光中,沃尔德解释说:“我所分析的样本中,只包含顺利返回基地的轰炸机.从统计的观点来看,我认为被多次击中机翼的轰炸机,似乎还是能够安全返航.而飞机很少发现弹着点的部位,并不是真的不会中弹,而是一旦中弹,根本就无法返航.”指挥官反驳说:“我很佩服沃尔德教授没有任何飞行经验,就敢做这么大胆的推论.以我个人而言,过去在执行任务时,也曾多次机翼中弹严重受创.要不是我飞行技术老到,运气也不错,早就机毁人亡了.所以,我依然强烈主张应该加强机翼的装甲.”

这两种意见坚持不下,皇家空军部部长陷入苦思.他到底要相信这个作战经验丰富的飞行将军,还是要相信一个独排众议的统计学家?

由于战况紧急,无法做更进一步的研究,部长决定接受沃尔德的建议,立刻加强驾驶舱与机尾发动机的防御装甲.不久之后,联军轰炸机被击落的比例,果然显著降低.为了确认这个决策的正确性,一段时间后,英国军方动用了敌后工作人员,

收集了部分坠毁在德国境内的联军飞机残骸. 其中弹的部位, 果真如沃尔德所预料, 主要集中在驾驶舱与机尾的位置.

乍看之下, 作战指挥官加强机翼装甲的决定十分合理, 但他忽略了这个事实: 弹着点的分布, 是一种严重偏误的资料. 因为最关键的数据, 是在被击落的飞机身上, 但这些飞机却无法被观察到. 因此, 布满了弹痕的机翼, 反而是飞机最强韧的部位. 空军作战指挥官差点因为太重视看得见的弹痕, 反而做出错误的决策. 这个案例有两个特别值得警惕的地方.

第一, 收集更多数据, 并不会改善决策质量. 由于弹痕资料的来源本身就有严重的偏误, 努力收集更多的资料, 恐怕只会更加深原有的误解.

第二, 召集更多作战经验丰富的飞行员来提供专业意见, 也不能改善决策质量. 因为这些飞行员, 正是产生偏误数据过程中的一环. 他们都是安全回航的飞行员, 虽然可能有机翼中弹的经验, 但都不是驾驶舱或发动机中弹的“烈士”. 简单地说, 当他们愈认真凝视那些“看得到”的弹痕, 他们离真相就愈远.

资料来源: <http://www.readingtimes.com.tw/timeshtml/ad/DH0185/P001.html>

样本是统计分析的根本, 样本要满足什么条件, 读完本章就会有答案.

§ 1.1 导 论

假想你是一位企业工作人员, 你需要为你所在的企业搜集一些相关资料, 此时你会发现统计信息是一个企业在市场上立足的根本; 而历届政府工作报告中占据相当篇幅也最有说服力的就是统计数据; 历届人大代表会后的记者招待会上, 记者们提问最多的也是相关的统计数据. 那么, 你接下来面临的问题自然是这些数据是怎么得来的? 经过了哪些处理和分析? 该如何解读这些数据呢?

你的思路是想“透过数据看世界”, 此时, 你实际上已经站在了“数理统计”这座宫殿的大门口了.

1.1.1 数理统计学的研究对象

在我们的身边有形和无形的数据形成了一个茫茫数海, 这些数据蕴含着什么信息? 可以给我们什么启示呢? 我们如何从中归纳出精辟的结论, 为决策提供依据? 这就构成了数理统计学的研究对象.

数理统计学研究怎样有效地收集、处理、分析、解释带有随机性的数据, 以对所考察的问题作出推断或预测, 直至为采取一定的决策和行动提供依据和建议. 若在这句文字中去掉“随机性”这几个字, 那就是统计学的研究范围. 统计学也就是数据科学. 由于实践中人们收集到的数据大都受到随机性的影响, 因而数理统计学是统计学的主要组成部分. 在许多西方国家(如美国), 这两个名词的含义基本

相同. 许多大学里统计系就是研究数理统计学的, 使用“数理统计”一词时只是强调统计学中用到很多现代数学知识. 在我国由于历史的原因, 这两个名词在实际使用时有较大差别. 笔者相信, 在若干年后, 这两个名词所指的范围将相差不大.

1.1.2 统计工作诸环节

用数理统计方法解决一个实际问题时, 一般有如下几个步骤:

建立数学模型; 收集整理数据; 进行统计推断、做预测和决策.

但需要注意, 这些环节不能截然分开, 也不一定按上述次序, 有时是互相交错的. 具体的各个步骤的内容如下:

(1) 模型的选择和建立. 模型是指关于所研究总体的某种假定, 一般是给总体分布规定某个类型. 另外, 建立模型要依据概率的知识、所研究问题的专业知识、以往的经验以及从总体中抽取的样本(数据).

(2) 数据的收集. 一般有三种收集数据的方式, 包括全面观测、抽样观测和安排特定的实验. 全面观测又称为普查, 即对总体中每个个体都加以观测, 测定所需要的指标. 抽样观测又称为抽查, 是指从总体中抽取一部分, 测定其相关的指标值. 这方面的研究内容构成数理统计的一个分支学科, 称为抽样调查. 安排特定实验以收集数据, 这些特定的实验要有代表性, 并使所得数据便于进行分析. 这里面所包含的数学问题, 构成数理统计学的又一支学科, 即实验设计的内容.

(3) 数据整理. 其目的是把包含在数据中的有用信息提取出来. 整理数据通常有两种形式, 一种形式是制定适当的图表(如散点图), 以反映隐含在数据中的粗略的规律性或一般趋势; 另一种形式是计算若干数字特征, 以刻画样本某些方面的性质, 如样本均值、样本方差等简单描述性统计量.

(4) 统计推断. 是指根据总体模型以及由总体中抽出的样本, 作出有关总体分布的某种论断. 数据的收集和整理是进行统计推断的必要准备, 统计推断是数理统计学的主要任务.

(5) 统计预测. 统计预测的对象是随机变量在未来某个时刻所取的值, 或设想在某种条件下对该变量进行观测时将取的值. 例如, 预测一种产品在未来3年内的市场销售量, 某个10岁男孩在3年后的身高、体重, 等等.

(6) 统计决策. 依据所做的统计推断或预测, 并考虑到行动的后果(以经济损失的形式表示)而制定的一种行动方案. 其目的是使损失尽可能小, 或反过来说, 使收益尽可能大. 例如, 一个商店要决定今年内某种产品的进货数量, 商店的统计学者根据抽样调查, 预测该产品本店今年销售量为1 000件. 假定每积压一件产品损失20元, 而少销售一件产品则损失10元, 要据此作出关于进货数量的决策.

1.1.3 数理统计的一些渊源

数理统计学是伴随着概率论的发展而发展起来的. 19世纪中叶以前已出现了

若干重要的工作,如 C. F. 高斯和 A. M. 勒让德关于观测数据误差分析和最小二乘法的研究. 到 19 世纪末期,经过包括 K. 皮尔森在内的一些学者的努力,这门学科已开始形成. 但数理统计学发展成一门成熟的学科,则是 20 世纪上半叶的事,这在很大程度上要归功于 K. 皮尔森、R. A. 费歇尔等学者的工作. 特别是费歇尔的贡献,对这门学科的建立起了决定性的作用. 1946 年 H. 克拉默发表的《统计学数学方法》是第一部严谨且比较系统的数理统计著作,可以把这部著作作为数理统计学进入成熟阶段的标志.

数理统计学的发展大致可以分为三个时期. 其一,20 世纪以前. 这个时期又可以分成两段,大致上可以把高斯和勒让德关于最小二乘法用于观测数据的误差分析的工作作为分界线,前一阶段属萌芽时期,基本上没有超出描述性统计的范围. 后一阶段可以算做是数理统计学的幼年阶段. 首先,强调了推断的地位,而摆脱了单纯描述的性质. 由于高斯等的工作揭示了正态分布的重要性,学者们普遍认为,在实际问题中遇见的几乎所有的连续变量,都可以满意地用正态分布来刻画. 这种观点使关于正态分布的统计得到了深入的发展,但延缓了非参数统计的发展. 19 世纪末, K. 皮尔森给出了以他的名字命名的分布,并给出了估计参数的一种方法——矩法估计. 德国的 F. 赫尔梅特发现了统计上十分重要的 χ^2 分布. 其二,20 世纪初到第二次世界大战结束. 这是数理统计学蓬勃发展达到成熟的时期. 许多重要的基本观点和方法,以及数理统计学的主要分支学科,都是在这个时期建立和发展起来的. 这个时期的成就,包含了至今仍在广泛使用的大多数统计方法. 在其发展中,以英国统计学家、生物学家费歇为代表的英国学派起了主导作用. 其三,第二次世界大战战后时期. 这一时期中,数理统计学在应用和理论两方面继续获得很大的进展.

1.1.4 数理统计的应用领域

数理统计的应用涉及工农业生产、自然科学和技术科学以及社会经济等众多广泛的领域,只要有数据的地方就会用到数理统计方法. 具体地说,大致可以分为如下几个方面:

(1) 在农业中,对田间试验进行适当的设计和统计分析.

(2) 在工业生产中,实验设计法、回归设计和回归分析、方差分析、多元分析等统计方法,对于试制新产品和改进老产品、改革工艺流程、寻求适当的配方等问题起着广泛的作用,对于统计质量管理在控制工业产品的质量中起着十分重要的作用.

(3) 医学是较早使用数理统计方法的领域之一. 在防治一种疾病时,需要找出导致这种疾病的种种因素. 统计方法在发现和验证这些因素上,是一个重要工具. 另一方面的应用是,用统计方法确定一种药物对治疗某种疾病是否有效,用处多大,以及比较几种药物或治疗方法的效力.

(4)在自然科学和技术科学中,统计方法可以用于地震、气象和水文方面的预报、地质资源的评测等.

(5)在社会、经济领域方面,如人口调查和预测,心理学中能力方面的分析等.

◎思考题

1. 数理统计是研究什么的?
2. 数理统计与概率论、统计学有什么样的关系?

§ 1.2 数理统计的基本概念

1.2.1 总体 样本 参数

1. 总体(population)

在数理统计学中,把研究对象的全体称为总体或母体,构成总体的每个元素称为个体. 如要研究某大学的学生身高情况,则该大学的全体学生就是该问题的总体,每个学生是一个个体;要研究某厂生产的显像管的质量时,总体是该厂生产的所有显像管,每一个显像管是一个个体.

每个个体有很多特征,比如学生有:性别、年龄、身高、体重、籍贯等,我们并不是关心个体的所有特征,而只是关心某些数量指标值,比如身高,显像管的寿命等. 抛开实际背景,总体就是一堆数,这堆数有其自身的分布规律. 若考察的数量指标用 X 表示,则 X 是一个随机变量, X 的可能取值就是总体中的数,研究总体的分布规律实际上就是研究随机变量 X 的分布规律,所以,总体就是一个随机变量.

定义 1.1 一个随机变量 X 或其相应的分布函数 $F(X)$ 称为一个总体.

如果我们对每一研究对象要观测两个或多个数量指标,此时用多维随机向量来表示总体,这是多元分析研究的对象.

根据总体中所含元素的个数是有限的还是无限的,又可以分为有限总体和无限总体. 我们讨论的主要是无限总体.

2. 样本(sample)

为了研究总体的分布规律或某些特征,我们从总体中随机地抽取 n 个个体,观测它们的数量指标,分别记为 $X_1, X_2, \dots, X_n$, 这些通过观察或试验得到的数据称为总体的一个样本或子样本, n 称为样本容量(sample size). 这些观察或试验过程称为抽样(sampling).

例如用同一架天平称某重物 5 次,得到一组 5 个数据 x_1, x_2, x_3, x_4, x_5 就称这一组数据是一个样本,样本容量为 5.

每个容量为 n 的样本都可以认为是 n 维空间的一个点,样本所有可能的取值构成了 n 维空间的一个子集,称为样本空间(sample space). 注意“数据”一词在这