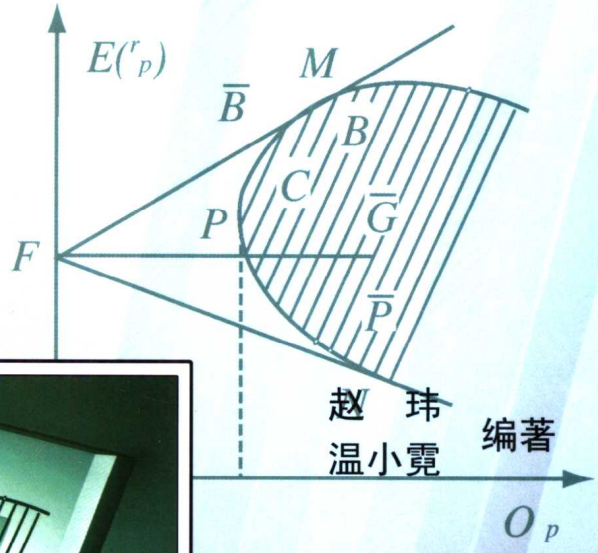
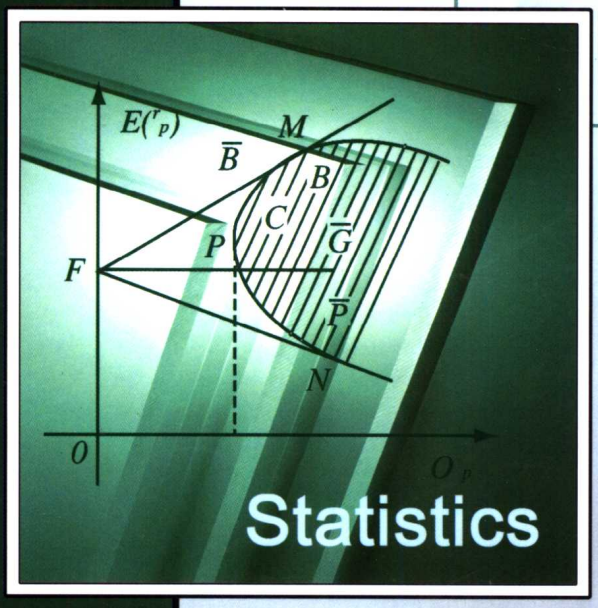
研究生系列教材

应用统计学教程

（下册）



赵 玮
温小霓 编著



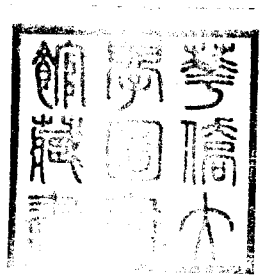
西安电子科技大学出版社
http://www.xduph.com

研究生系列教材

应用统计学教程

(下　册)

赵玮　温小霓　编著



西安电子科技大学出版社
2003

内 容 简 介

《应用统计学教程》是西安电子科技大学研究生系列教材之一，全书共3篇14章，系统地介绍了应用统计学的各主要分支，如数理统计、过程统计与统计决策分析，以及在工程、经济、金融和管理等领域中的应用统计分析。全书分上、下两册出版。

《教程》上册是本书的第一篇——数理统计基础，共5章，主要介绍数理统计的基础理论及其应用，内容包括概率论基础、抽样调查与抽样分布、参数估计、假设检验、方差分析与回归分析等。

《教程》下册共9章，包括本书的第二篇(过程统计与统计决策分析，第6章至第9章)和第三篇(应用统计分析，第10章至第14章)。第二篇重点介绍了过程统计与统计决策分析的有关部分，如各种常用随机过程(齐次与非齐次泊松过程、齐次马氏链、更新过程、平稳过程与正态过程)的统计推断和贝叶斯决策、概率排序型决策、混合策略分析等统计决策分析的有关理论与方法。第三篇较系统地介绍了数理统计与过程统计在软件可靠性、企业管理、宏观经济、社会保险和证券投资等领域进行统计分析的有关理论、专门知识、解决实际问题的基本思路与方法以及应用案例。

本书可供管理、经济、金融、工程和应用数学等各类研究生(硕士、博士)及本科生作教学用书，也可供有关教师和研究人员作参考书。各类研究生、本科生如何选用本教材内容，请参阅前言中附表的建议。

图书在版编目(CIP)数据

应用统计学教程·下册/赵玮等编著.
—西安：西安电子科技大学出版社，2003.1
(研究生系列教材)
ISBN 7-5606-1187-7

Ⅰ.应… Ⅱ.赵… Ⅲ.应用统计学—研究生—教材 Ⅳ.C8

中国版本图书馆CIP数据核字(2002)第087577号

策　　划　夏大平
责任编辑　夏大平　张晓燕　龙晖
出版发行　西安电子科技大学出版社(西安市太白南路2号)
电　　话　(029)8227828　　邮　编　710071
http://www.xduph.com　　E-mail：xdupfxb@pub.xaonline.com
经　　销　新华书店
印　　刷　陕西画报社印刷厂
版　　次　2003年3月第1版　2003年3月第1次印刷
开　　本　787毫米×1092毫米　1/16　印张 25.375
字　　数　602千字
印　　数　1～2 000册
定　　价　28.00元
ISBN 7-5606-1187-7/O·0057(课)

XDUP 1458A01-1

前　言

“统计”这个名词起源于拉丁文的“status”，其原意是“情况”，它是随着社会生产的发展和适应国家管理的需要而产生与发展起来的。在公元前1000多年的夏朝，我国就开始了人口、土地等方面的统计。在古埃及、罗马帝国也有人口调查的记载。当18世纪初统计学初步形成一种科学体系时，统计这个名词仍然是和表示国家情况的事实记录制度相联系的。随着世界社会经济发展的迫切需要，统计学开始飞速发展。进入18世纪与19世纪后，以法国数学家P. S. Laplace、波兰统计学家J. Neyman等为代表的一些学者将具有严密的逻辑推理基础的概率论引入了统计学，从而形成了数理统计的基本内容，如参数估计、假设检验、方差分析、相关与回归分析、抽样理论等。进入20世纪后，愈来愈多的应用数学方法，特别是以解决随机动态系统的分析为对象的统计数学方法，如多元统计、时间序列分析、统计决策分析、各种随机过程及其统计推断以及相应的数值计算方法的引入，使统计学的数学理论日趋系统、完善与实用化，进而使其应用领域逐步扩大到天文、地理、社会、经济、工程、军事作战等众多领域，并已经成为这些领域的科学试验、生产管理与工程设计不可缺少的助手，且由于其应用的广泛性而被冠之以“应用统计学”的名称。目前，尽管应用统计学的内涵仍有争议，但通常人们认为它应包括如下四个方面：① 统计情报的搜集，即搜集表示任何大量总体中各单位特征的情报(信息)；② 所得资料(信息)的统计研究，即阐明那些可以在大量观察资料基础上建立起来的规律的统计决策分析；③ 统计观察方法的拟定及统计资料的分析，它构成了数理统计与过程统计的基本内容；④ 统计学与各种领域中专门学科相结合的边缘分支，如天文统计学、地质统计学、工程统计学、经济统计学和社会统计学(人口统计学、消费统计学、卫生统计学等)等。

随着我国改革开放及社会主义市场经济的日益深入，应用统计学在工程、经济、社会及企业管理中的作用日益显著，这不仅由于未来面对日益激烈的市场竞争，人们对各国(地区)经济、社会、科技进步及企业产品市场需求的分析与预测的需要增加，而且由于在工程设计、科学实验及现代化管理的各种问题中不确定因素的增加而导致提高定量化水平的需要增加。然而迄今为止，国内尚缺乏一本内容(理论、方法与应用)较为广泛，数学基础较为适中的，以培养研究生独立科研工作能力为主要目标的论著与教材。为此，作者根据自己长期从事各种工程、社会经济与军事作战系统的随机分析与设计的科研实践及从事统计数学、随机运筹学的理论教学与研究生培养实践与经验编写了这本书，以期为国内的研究生教学贡献自己的一份微薄力量。

与国内的应用统计学与数理统计学及其有关论著相比，本书具有如下特点：

(1) 本书的主要目标是培养研究生在进行各种随机动态系统分析与设计过程中的独立科研工作能力，为此本书在第二篇(共四章)较为系统、全面地介绍了各种随机过程(泊松过程、齐次马氏过程、更新过程、平稳过程、正态过程)的统计推断的基本理论与方法及统计决策理论的主要基础，这些内容恰恰是国内有关论著与教材所缺乏而科研实践又必须具

备的理论基础。

(2) 为了解决工程、经济、社会、企业管理中的各种随机动态系统的分析与决策问题，往往需要上述各领域的专门知识及目前前沿研究领域的有关内容，为此本书第三篇(共五章)较为系统及全面地介绍了应用统计学的理论与方法在企业管理、软件工程分析与设计、宏观经济系统分析、证券投资分析及保险统计分析中的应用，以及有关的研究专题及主要的参考文献，其中有不少内容是作者多年来的科研与论文成果。阅读与讲解这部分内容既可使读者通过对这些实际案例的学习加深对数理统计、过程统计与统计决策分析理论与方法的理解，还可以使读者了解解决这些实际问题的思路与方法以及应用背景与专门知识，从而在较短的时间内进入有关领域的研究专题。

(3) 考虑到读者对象的数学基础，故本书在写作中力求在不涉及测度论的基础上以尽可能少的篇幅、较为浅显的数学描述来介绍应用统计学中内容较为深入的有关部分，以供研究生从事科研工作之用。

(4) 本书采用积木式结构，全书3篇共14章，经过有关章节的各种组合，可供管理类、经济类、工程(含系统工程)类及应用数学类(含运筹与控制)研究生(硕士、博士)及本科生的教学使用或供教师参考使用，也可供相关领域从事统计工作及有关科研与管理人员参考使用。同时由于本书的第篇内容较为广泛且安排相对独立，因此只要稍作补充，还可作为“软件可靠性分析与测试”、“企业管理统计分析”、“宏观经济统计分析”、“证券投资分析”、“保险精算学”等课程的教材使用，从而做到了一书多用的目的。有关各类专业及各层次的教学时数及所选篇章建议见下附表。教授全书约需60～80学时。

附表　各类专业选用内容建议

	内容篇章	管理类研究生	经济类研究生	工程类研究生	应用数学及系统工程研究生	本科生
上册	第一篇　数理统计基础					
	第1章　随机变量、随机向量及其统计特性	选	选	选	—	√
	第2章　抽样调查与抽样分布	讲	讲	讲	—	√
	第3章　参数估计				—	√
	第4章　假设检验				—	√
	第5章　方差分析与回归分析				—	√
下册	第二篇　过程统计与统计决策分析					
	第6章　多元统计分析	√	√	√	√	√
	第7章　随机过程统计推断	√	√	√	√	—
	第8章　平稳时间序列分析	√	√	√	√	—
	第9章　统计决策分析	√	√	√	√	√
	第三篇　应用统计分析					
	第10章　软件可靠性及其统计分析	—	—	√	√	案例选讲
	第11章　企业管理统计分析	√	√	√	√	
	第12章　宏观经济统计分析	√	√	√	√	
	第13章　保险统计分析	√	√	—	√	
	第14章　证券投资统计分析	√	√	—	√	

注：“√”表示讲授内容，“—”表示参考。

本书在写作过程中力求做到思路清晰，内容层次分明，概念准确无误，数学分析由浅入深。推理论证既考虑到各类专业的理论基础，又力求做到严密正确。对于一些由于篇幅与时间限制而未能给出证明的结论均尽可能地指明参考文献与出处。

本书的第2章至第6章及11.3节的初稿由温小霓完成，并由赵玮作了修改、补充与加工。第7章至第14章、前言、第1章和附表及全书的定稿由赵玮完成。全书承叶正麟教授审阅，提出了宝贵意见，研究生彭少波、陈久梅、冯光兰、孙晓琳、郭彦超、杨绪红等为全书的打印付出了辛勤的劳动，在此表示深切的感谢。由于时间及作者水平的限制，书中肯定有不少疏漏与错误之处，敬请广大专家与读者提出批评与宝贵意见。

此书的出版得到了西安电子科技大学研究生教材建设基金的资助。

赵　玮

2002年6月于西安

目　录

第二篇　过程统计与统计决策分析

第三篇　应用统计分析

第二篇

过程统计与统计决策分析

第 6 章　多元统计分析

本章介绍统计分析与统计决策理论中的三个重要部分：判别分析、聚类分析与主成分分析。由于它们的研究对象均可用随机向量来描述，故它们是多元统计分析的重要内容。此外，由于随机向量是一类特殊的随机过程，即参数集 T 为离散且有限的随机过程，故它们又构成了随机过程统计推断理论的一个组成部分。判别分析、聚类分析与主成分分析自诞生之日起就有着各自深刻的物理背景，且随着计算机科学与随机过程理论研究的日益深入，使得它们在生物学、心理学、航天学、经济学及工程技术中获得了广泛的应用。目前，它们已经成为这些学科领域中应用最为广泛与活跃的分支之一，并已成为计算机科学中的“人工智能(AI)”与“专家系统(ES)”的重要理论基础之一。

6.1　判 别 分 析

人们在对自然科学与社会经济问题的研究中常遇到这样一类问题：根据对某个研究对象(或系统)的特征观测来判定该对象的基本类别。例如，在考古学与刑事犯罪学中，人们根据所挖掘出来的人头盖骨的高、宽等特征来判定其性别类别(男性或女性)；在医疗诊断中，医生根据病人的消瘦程度及血液化验中的血糖等指标来判定该病人是否患有糖尿病；在经济学中，人们根据各个国家或地区的人均国民收入、人均工农业产值、人均消费水平等多项指标来判定该国家(地区)的经济发展水平所属类型(如发达国家、中等发达国家、发展中国家等)；在航空测量中，人们根据高空拍摄到的遥测遥感照片(往往是模糊的)的多项基本特征来进行地形识别(如分辨公路、桥梁、房屋、树林、河流、山坡等)；其他方面，如人们熟知的犯罪嫌疑人的指纹识别，计算机输入的汉字识别等均属于此类问题。上述问题人们通称为分类问题或模式识别(Pattern Recognition)问题，并根据所研究对象 $\boldsymbol{X}(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 所属的总体类别 $\pi_1,\pi_2,\cdots,\pi_s$ 是已知还是未知的前提来研究模式识别，分别称为判别分析与聚类分析。本节介绍在总体类别已知的前提下的模式识别(即判别分析)的基本概念、理论与方法。

设所研究的对象(或系统)可用 p 个特征来刻画，故该研究对象可用随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 来描述，该研究对象所属的总体有 s 类，可分别记为 $\pi_1,\pi_2,\cdots,\pi_s$。若该研究对象属于第 j 类总体，则可记作 $\boldsymbol{X}\in\pi_j$。而判别分析的主要内容即为根据不同的判别准则来寻找一个 $\boldsymbol{X}$ 的判决函数 $g(\boldsymbol{X})$，并根据 $g(\boldsymbol{X})$ 的属性来判定 $\boldsymbol{X}$ 的所属总体类别。构造判别函数的主要准则有最短距离准则、最小期望损失准则、费歇(Fisher)准则、最小平方准则(LMS)、库尔贝克(Kullback)准则、不确定型(信息熵)准则、最大后验概率准则、最大似然函数准则等。限于篇幅，本节主要介绍由前三种准则来构成判决函数所形成的判别方法，相应的方法称为最小距离法(几何分类法)、贝叶斯(Bayes)判别法与费歇判别法。

6.1.1　最小距离法

这一方法的基本思想是认为各类研究对象都比较均匀分布在“代表”各类研究对象的一个向量的周围，对于任意一个未知类别的研究对象 $\boldsymbol{X}$，最直观的方法是通过比较它与各类研究对象 π_j 的“代表”之间的距离，并根据它属于最小距离所对应的那一类来作为其所属母体类别 π_j 的依据。即若设 $\mu^{(j)}$ 作为母体 $\pi^{(j)}$ 的“代表”，$d(\boldsymbol{X},\mu^{(j)})$ 表示研究对象 $\boldsymbol{X}$ 与总体 π_j“代表”μ_j 的距离，并若有 $d(\boldsymbol{X},\mu^{(i)})<d(\boldsymbol{X},\mu^{(j)})$，$j\neq i$，$j=1,2,\cdots,s$，则判定 $\boldsymbol{X}\in\pi_i$，或定义判决函数 $g_{ij}(\boldsymbol{X})=d(\boldsymbol{X},\mu^{(i)})-d(\boldsymbol{X},\mu^{(j)})$，若有 $g_{ij}(\boldsymbol{X})<0$，$j\neq i$，$j=1,2,\cdots,s$，则 $\boldsymbol{X}\in\pi_i$。

在采用上述思路作统计判决(决策)时需解决如下两个问题：

(1) 总体 $\pi^{(j)}$ 的代表 $\mu^{(j)}$ 如何得出？

(2) 距离 $d(\boldsymbol{X},\boldsymbol{Y})$ 如何定义？

关于 π_j 的“代表”是容易求得的，这只需从 π_j 的总体中相互独立地抽取 n_j 个个体 $x_1^{(j)}$，$x_2^{(j)}$，$\cdots$，$x_{n_j}^{(j)}$，$j=1,2,\cdots,s$，其中每个个体样本需要刻画其 p 个特征，故有 $\boldsymbol{x}_i^{(j)}=(x_{i1}^{(j)},x_{i2}^{(j)},\cdots,x_{ip}^{(j)})^{\mathrm{T}}$，$i=1,2,\cdots,n_j$，$j=1,2,\cdots,s$。$x_{ik}^{(j)}$ 表示从总体 π_j 中抽取的第 i 个样本的第 k 个特征量。显然可用总体 π_j 的 n_j 个样本个体的均值向量 $\mu^{(j)}=(\mu_1^{(j)},\mu_2^{(j)},\cdots,\mu_p^{(j)})^{\mathrm{T}}$ 来作为总体 π_j 类的“代表”，其中

$$\mu_k^{(j)}=\frac{1}{n_j}\sum_{i=1}^{n_j}x_{ik}^{(j)}\qquad k=1,2,\cdots,p;\ j=1,2,\cdots,s$$

关于在 $\mathbf{R}^p$ 中的距离概念容易从泛函分析中得到。

定义 6.1.1　设 $d(\cdot,\cdot)$ 是从 $\mathbf{R}^p\times\mathbf{R}^p$ 到 $[0,\infty)$ 的函数，对 $\mathbf{R}^p$ 中的任意点 $\boldsymbol{x}_i$，$\boldsymbol{x}_j$ 和 $\boldsymbol{x}_k$，若 $d(\boldsymbol{x}_i,\boldsymbol{x}_j)$(简记 d_{ij})满足如下三条件，则称 d_{ij} 为点 $\boldsymbol{x}_i$ 与点 $\boldsymbol{x}_j$ 的距离，其中三条件为：

(1) $d_{ij}\geqslant 0$，当且仅当 $\boldsymbol{x}_i=\boldsymbol{x}_j$ 时，$d_{ij}=0$；

(2) $d_{ij}=d_{ji}$；

(3) $d_{ij}\leqslant d_{ik}+d_{kj}$(三角不等式)。

设 $\boldsymbol{x}_i\in\mathbf{R}^p$ 或 $\boldsymbol{x}_j\in\mathbf{R}^p$ 或 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$，$\boldsymbol{x}_j=(x_{j1},x_{j2},\cdots,x_{jp})^{\mathrm{T}}$，则常用的距离有如下几种：

(1) 明考夫斯基距离为

$$d_{ij}(m)=\Big[\sum_{\alpha=1}^{p}|x_{\alpha i}-x_{\alpha j}|^m\Big]^{1/m}\qquad m>0\tag{6.1.1}$$

当 $m=1,2,\infty$ 时，该距离分别为：

$$d_{ij}(1)=\sum_{\alpha=1}^{p}|x_{\alpha i}-x_{\alpha j}|\text{——绝对距离；}$$

$$d_{ij}(2)=\Big[\sum_{\alpha=1}^{p}|x_{\alpha i}-x_{\alpha j}|^2\Big]^{1/2}\text{——欧氏距离；}$$

$$d_{ij}(\infty)=\max_{1\leqslant\alpha\leqslant p}|x_{\alpha i}-x_{\alpha j}|\text{——契比雪夫距离。}$$

(2) $\boldsymbol{B}$ 模距离。对任给的正定矩阵 $\boldsymbol{B}$，称 d_{ij} 为 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 的 $\boldsymbol{B}$ 模距离，其中

$$d_{ij}=[(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{x}_i-\boldsymbol{x}_j)]^{1/2}\tag{6.1.2}$$

若对任意点 $\boldsymbol{x}_i$ 有数学期望 $\mu=E(\boldsymbol{x}_i)$和协方差矩阵 $\boldsymbol{\Sigma}_i=\mathrm{cov}(\boldsymbol{x}_i)$存在时，若 $\boldsymbol{B}=\boldsymbol{\Sigma}_j^{-1}$，得马哈拉诺比斯(Maholanobis)距离(简称马氏距离)$d_{ij}=[(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)]^{1/2}$，其优点是可以克服变量之间的相关性干扰，并消除各变量间的影响。当 $\boldsymbol{B}=\mathrm{diag}\left(\frac{1}{\sigma_1^2},\cdots,\frac{1}{\sigma_p^2}\right)$时，便得到了方差加权距离：

$$d_{ij}=\left[\sum_{\alpha=1}^{p}\frac{(x_{\alpha i}-x_{\alpha j})^2}{\sigma_\alpha^2}\right]^{1/2}$$

当 $\boldsymbol{B}=\boldsymbol{I}$ 单位阵时，$\boldsymbol{B}$ 模距离成为欧氏距离(6.1.1)式。

1. 两总体判别

(1) 设两个总体 π_1、π_2 有相同的协方差阵 $\boldsymbol{\Sigma}(\boldsymbol{\Sigma}>0)$，而均值向量分别为 $\mu^{(1)}$、$\mu^{(2)}$，对于一个给定的样本 $\boldsymbol{x}$ 要判断它是来自于哪一个总体，根据前述思路可先计算 $\boldsymbol{x}$ 与两总体“代表”的距离 $d(\boldsymbol{x},\mu^{(i)})$，$i=1,2$。判别规则可设计为

$$\begin{cases}\boldsymbol{x}\in\pi_1 & \text{当 } d(\boldsymbol{x},\mu^{(1)})<d(\boldsymbol{x},\mu^{(2)})\\ \boldsymbol{x}\in\pi_2 & \text{当 } d(\boldsymbol{x},\mu^{(1)})>d(\boldsymbol{x},\mu^{(2)})\end{cases}\tag{6.1.3}$$

当 $d(\boldsymbol{x},\mu^{(1)})=d(\boldsymbol{x},\mu^{(2)})$时，$\boldsymbol{x}$ 可归属于 π_1、π_2中任何一个。在采用马氏距离计算距离平方差时，利用协方差阵 $\boldsymbol{\Sigma}$ 的对称性及$(\boldsymbol{\Sigma}^{-1})^{\mathrm{T}}=(\boldsymbol{\Sigma}^{\mathrm{T}})^{-1}=\boldsymbol{\Sigma}^{-1}$可得

$$\begin{aligned}d^2(\boldsymbol{x},\mu^{(1)})-d^2(\boldsymbol{x},\mu^{(2)})&=(\boldsymbol{x}-\mu^{(1)})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu^{(1)})-(\boldsymbol{x}-\mu^{(2)})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu^{(2)})\\&=(\boldsymbol{x}-\mu^{(1)})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu^{(1)})-(\boldsymbol{x}-\mu^{(2)})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu^{(2)})\\&\quad-(\boldsymbol{x}-\mu^{(1)})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu^{(2)})+(\boldsymbol{x}-\mu^{(2)})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu^{(1)})\\&=-2\left[\boldsymbol{x}-\frac{\mu^{(1)}+\mu^{(2)}}{2}\right]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mu^{(1)}-\mu^{(2)})\\&=-2W(\boldsymbol{x})\end{aligned}\tag{6.1.4}$$

令

$$\bar{\mu}=\frac{\mu^{(1)}+\mu^{(2)}}{2}$$

则 $W(\boldsymbol{x})=(\boldsymbol{x}-\bar{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mu^{(1)}-\mu^{(2)})$，称 $W(\boldsymbol{x})$为判别函数。利用(6.1.4)式容易得知判别规则(6.1.3)式变为

$$\begin{cases}\boldsymbol{x}\in\pi_1 & \text{当 } W(\boldsymbol{x})>0\\ \boldsymbol{x}\in\pi_2 & \text{当 } W(\boldsymbol{x})<0\end{cases}\tag{6.1.5}$$

显然，$W(\boldsymbol{x})$是 $\boldsymbol{x}$ 的一个线性函数，判别规则取决于 $W(\boldsymbol{x})$大于 0 还是小于 0，这里 0 值被称为阈值点。

(2) 当协方差阵不同时，设总体 π_1，π_2分别有协方差阵 $\boldsymbol{\Sigma}_1$ 与 $\boldsymbol{\Sigma}_2(\boldsymbol{\Sigma}_1\neq\boldsymbol{\Sigma}_2)$，均值向量仍设为 $\mu^{(1)}$，$\mu^{(2)}$，则由前思路知可定义判别函数为

$$\begin{aligned}W_{ij}(\boldsymbol{x})&=d^2(\boldsymbol{X},\mu^{(j)})-d^2(\boldsymbol{X},\mu^{(i)})\\&=(\boldsymbol{x}-\mu^{(j)})^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\mu^{(j)})-(\boldsymbol{x}-\mu^{(i)})^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\mu^{(i)})\end{aligned}$$

判别规则为

$$\begin{cases} \boldsymbol{x} \in \pi_i & \text{若 } W_{ij}(\boldsymbol{x}) > 0 \\ \text{待判} & \text{若 } W_{ij}(\boldsymbol{x}) = 0 \\ \boldsymbol{x} \in \pi_j & \text{若 } W_{ij}(\boldsymbol{x}) < 0 \end{cases} \quad ; i \neq j;\ i,j = 1,2 \tag{6.1.6}$$

(3) 若两总体均值向量及协方差阵均未知，则可用总体的样本来估计均值与协方差阵；若两总体样本协方差基本接近，则设 $\boldsymbol{x}_1^{(1)},\boldsymbol{x}_2^{(1)},\cdots,\boldsymbol{x}_{n_1}^{(1)}$ 和 $\boldsymbol{x}_1^{(2)},\boldsymbol{x}_2^{(2)},\cdots,\boldsymbol{x}_{n_2}^{(2)}$ 分别为总体 π_1 和 π_2 的样本，其中

$$\boldsymbol{x}_i^{(1)} = (x_{i1}^{(1)},x_{i2}^{(1)},\cdots,x_{ip}^{(1)})^{\mathrm{T}} \qquad i = 1,2,\cdots,n_1$$

$$\boldsymbol{x}_i^{(2)} = (x_{i1}^{(2)},x_{i2}^{(2)},\cdots,x_{ip}^{(2)})^{\mathrm{T}} \qquad i = 1,2,\cdots,n_2$$

令

$$\bar{\boldsymbol{x}}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{x}_i^{(1)}, \quad \bar{\boldsymbol{x}}^{(2)} = \frac{1}{n_2}\sum_{i=1}^{n_2}\boldsymbol{x}_i^{(2)}$$

$$\boldsymbol{A}_1 = \sum_{i=1}^{n_1}(\boldsymbol{x}_i^{(1)} - \bar{\boldsymbol{x}}^{(1)})(\boldsymbol{x}_i^{(1)} - \bar{\boldsymbol{x}}^{(1)})^{\mathrm{T}}$$

$$\boldsymbol{A}_2 = \sum_{i=1}^{n_2}(\boldsymbol{x}_i^{(2)} - \bar{\boldsymbol{x}}^{(2)})(\boldsymbol{x}_i^{(2)} - \bar{\boldsymbol{x}}^{(2)})^{\mathrm{T}}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n_1 + n_2 - 2}(\boldsymbol{A}_1 + \boldsymbol{A}_2)$$

利用(6.1.4)式结论，可设计判别函数为

$$W(\boldsymbol{x}) = \left[\boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})\right]^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}) \tag{6.1.7}$$

判别规则同(6.1.5)式。注意到上述判别函数设计的前提是 $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$，，然而对于一个实际问题此前提是否能满足应根据样本序列来作如下的假设检验：H_0：$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$；$H_1$：$\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$。有关检验方法详见文献[5]或[8]。

其中需要说明的是，无论什么判别方法都不可避免地会产生误判，同样按最小距离判别规则判别也会产生误判。例如当 $p = 1$ 时，若设 $\pi_1 \sim N(\mu_1,\sigma^2)$，$\pi_2 \sim N(\mu_2,\sigma^2)$，$\bar{\mu} = \frac{\mu_1 + \mu_2}{2}$，且 $x \in \pi_1$ 时，从图 6.1.1(a)可看出，当 x 的观察值落在 $\bar{\mu}$ 的右边时，按最小距离判别规则应判 x 属于 π_2，从而产生误判。误判概率为阴影部分面积。若把判别限定为 γ (见图 6.1.1(b))，虽然把属于 π_1 而误判给 π_2 的概率减小，但同时却把属于 π_2 而误判给 π_1 的概率增加。因此当 π_1 很接近 π_2 时，即 μ_1 与 μ_2 相近时，误判的概率增大，在这种情况下判别将失去效果。故在进行判别分析前，应首先检验总体是否有显著差异。若有显著差异，然后再进行判别分析。

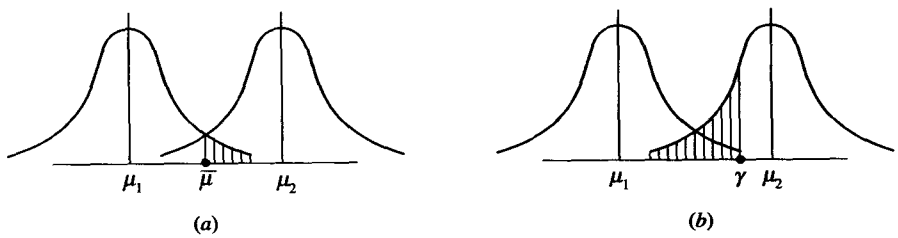


图 6.1.1

一般地还可给出一个待定区域，如规定 $(c,d)=\left(\bar{\mu}-\frac{1}{3}|\mu_1-\mu_2|,\bar{\mu}+\frac{1}{3}|\mu_1-\mu_2|\right)$ 为待判区域，则有

判别规则：

$$\begin{cases} x\in\pi_1 & \text{当 } x\leqslant c \\ x\in\pi_2 & \text{当 } x\geqslant d \\ \text{待判} & x\in(c,d) \end{cases}$$

2. 多总体判别

当事物本身分类大于两类时，就成为多总体判别问题。例如当病人肺部有阴影时，需要判别他是肺结核、肺水肿、肺部良性肿瘤或肺癌，这时就成为四个总体的判别问题。

设有 s 个总体 $\pi_1,\pi_2,\cdots,\pi_s$，具有相同的正定协方差矩阵 $\boldsymbol{\Sigma}$ 和不同的均值向量 $\boldsymbol{\mu}^{(i)}=(\mu_1^{(i)},\mu_2^{(i)},\cdots,\mu_p^{(i)})^{\mathrm{T}}$，$i=1,2,\cdots,s$；类似地，判别函数可取为

$$W_{ij}(\boldsymbol{x})=\left[\boldsymbol{x}-\frac{1}{2}(\boldsymbol{\mu}^{(i)}+\boldsymbol{\mu}^{(j)})\right]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(i)}-\boldsymbol{\mu}^{(j)})\qquad i,j=1,2,\cdots,s \tag{6.1.8}$$

判别规则为

$$\boldsymbol{x}\in\pi_i,\ W_{ij}(\boldsymbol{x})>0\qquad i\neq j;\ j=1,2,\cdots,s$$

如果有 $\boldsymbol{x}$ 使得 $W_{ij_1}(\boldsymbol{x})=W_{ij_2}(\boldsymbol{x})=\cdots=W_{ij_r}(\boldsymbol{x})=0(1\leqslant r\leqslant s)$，则 $\boldsymbol{x}$ 可判定属于 $\pi_i,\pi_{j_1},\pi_{j_2},\cdots,\pi_{j_r}$ 中的任一个，即在边界上的点可判断为相邻区域的任一个。

当 $\boldsymbol{\mu}^{(i)}(i=1,2,\cdots,s)$ 和相同的协方差阵 $\boldsymbol{\Sigma}$ 都未知时，可用估计量代替。设 $\boldsymbol{x}_k^{(i)}$ 是从总体 π_i 中抽得的样本容量为 n_i 的第 k 个样本$(i=1,2,\cdots,s;\ k=1,2,\cdots,n_i)$，记

$$\bar{\boldsymbol{x}}^{(i)}=\frac{1}{n_i}\sum_{k=1}^{n_i}\boldsymbol{x}_k^{(i)},\ \boldsymbol{A}_i=\sum_{k=1}^{n_i}(\boldsymbol{x}_k^{(i)}-\bar{\boldsymbol{x}}^{(i)})(\boldsymbol{x}_k^{(i)}-\bar{\boldsymbol{x}}^{(i)})^{\mathrm{T}}\qquad i=1,2,\cdots,s$$

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{n-s}\sum_{i=1}^{s}\boldsymbol{A}_i$$

其中 $n=\sum_{i=1}^{s}n_i$，则类似地有判别函数

$$W_{ij}(\boldsymbol{x})=\left[\boldsymbol{x}-\frac{1}{2}(\bar{\boldsymbol{x}}^{(i)}+\bar{\boldsymbol{x}}^{(j)})\right]^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{x}}^{(i)}-\bar{\boldsymbol{x}}^{(j)}) \tag{6.1.9}$$

判别规则同(6.1.9)式。

当协方差阵不同且 $\boldsymbol{\mu}^{(1)},\boldsymbol{\mu}^{(2)},\cdots,\boldsymbol{\mu}^{(s)}$ 与 $\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\cdots,\boldsymbol{\Sigma}_s$ 均未知时，有估计量如下：

$$\hat{\boldsymbol{\mu}}=\frac{1}{n_i}\sum_{k=1}^{n_i}\boldsymbol{x}_k^{(i)}\qquad i=1,2,\cdots,s$$

$$\hat{\boldsymbol{\Sigma}}^{(i)}=\sum_{k=1}^{n_i}(\boldsymbol{x}_k^{(i)}-\bar{\boldsymbol{x}}^{(i)})(\boldsymbol{x}_k^{(i)}-\bar{\boldsymbol{x}}^{(i)})^{\mathrm{T}}$$

其判别函数与判别规则同(6.1.8)式与(6.1.9)式。

例 6.1.1　试讨论两总体判别分析中当 $p=1$ 时的判别规则及其几何意义。此中设两总体 π_1 与 π_2 有总体均值与总体方差分别为 $\mu^{(j)}$ 与 $\sigma_j^2(j=1,2)$，子样均值与子样方差分别为 $\bar{x}^{(j)}$ 与 $s_j^2(j=1,2)$。

解　注意到当 $p=1$ 时，即 x 为一维随机变量，此时显然有 $\Sigma_j=\sigma_j^2\approx s_j^2(j=1,2)$。不失一般性，可设 $\mu^{(2)}<\mu^{(1)}$，此时有马氏距离为

$$d(x,\mu^{(j)})=\frac{|x-\mu^{(j)}|}{s_j}\qquad j=1,2$$

当 $\bar{x}^{(2)}<x<\bar{x}^{(1)}$ 时，用此样本均值 $\bar{x}^{(1)}$ 与 $\bar{x}^{(2)}$ 代替 $\mu^{(1)}$ 与 $\mu^{(2)}$，则可得

$$d(x,\mu^{(2)})-d(x,\mu^{(1)})=\frac{x-\bar{x}^{(2)}}{s_2}-\frac{-(x-\bar{x}^{(1)})}{s_1}=\frac{s_1+s_2}{s_1s_2}(x-\mu^*)$$

其中 μ^* 称为阈值，且有

$$\mu^*=\frac{s_1\bar{x}^{(2)}+s_2\bar{x}^{(1)}}{s_1+s_2}$$

其对应的判别规则为

$$\begin{cases}x\in\pi_1 & \text{若 } x>\mu^*\\ x\in\pi_2 & \text{若 } x<\mu^*\\ \text{待判} & \text{若 } x=\mu^*\end{cases}$$

其中阈值 μ^* 将 x 的样本空间 Ω 划为两个互不相交的子区间 $R_1=(\mu^*,\infty)$，$R_2=(-\infty,\mu^*)$，并有 $R_1\cup R_2\cup\{\mu^*\}=\Omega$。此外，还可设计线性判别函数 $W(x)=x-\mu^*$，显然，此时对应的判别规则为

$$\begin{cases}x\in\pi_1 & \text{若 } W(x)>0\\ x\in\pi_2 & \text{若 } W(x)<0\\ \text{待判} & \text{若 } W(x)=0\end{cases}$$

例 6.1.2　现有两种检测精神、心理正常与否的心理测试方法，从正常与不正常人群中各抽 25 人经测试得数据如表 6.1.1 所示。试建立判别函数以判断一新学生的心理正常与否。

表 6.1.1　正常人和精神病患者的测试 1 和测试 2 数据

	正常人			精神病患者				正常人			精神病患者		
序号	$x_{k1}^{(1)}$	$x_{k2}^{(1)}$	$W^*(x)$	$x_{k1}^{(2)}$	$x_{k2}^{(2)}$	$W^*(x)$	序号	$x_{k1}^{(1)}$	$x_{k2}^{(1)}$	$W^*(x)$	$x_{k1}^{(2)}$	$x_{k2}^{(2)}$	$W^*(x)$
1	22	6	62	24	38	8	14	13	13	3	3	12	−45
2	20	14	36	19	36	−13	15	20	14	36	10	51	−88
3	23	9	61	11	43	−67	16	19	15	29	22	22	30
4	23	1	77	6	60	−126	17	20	11	42	11	30	−41
5	17	8	33	9	32	−55	18	18	17	20	6	30	−66
6	24	9	66	10	17	−20	19	20	7	50	20	61	−58
7	23	13	53	3	17	−55	20	23	6	67	20	43	−22
8	18	18	18	15	56	−73	21	23	23	33	15	48	−57
9	22	16	42	14	43	−52	22	25	9	71	5	53	−117
10	19	18	23	15	8	48	23	23	5	69	10	43	−72
11	20	17	30	14	46	−88	24	21	12	45	13	19	−9
12	20	31	2	20	62	−60	25	23	7	65	12	4	16
13	21	9	51	14	36	−38	新病人 x	28	36	22			

解　设 π_1 为正常人类，π_2 为精神病患者类。今从 π_1 中相互独立地抽取容量为 $n_1=25$ 的样本：$\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \cdots, \boldsymbol{x}_{25}^{(1)}$，其中，$\boldsymbol{x}_k^{(1)}=(x_{k1}^{(1)}, x_{k2}^{(1)})$，$k=1,2,\cdots,25$，并从 π_2 中相互独立地抽取容量为 $n_2=25$ 的样本 $\boldsymbol{x}_1^{(2)},\boldsymbol{x}_2^{(2)},\cdots,\boldsymbol{x}_{25}^{(2)}$，其中，$\boldsymbol{x}_k^{(2)}=(x_{k1}^{(2)},x_{k2}^{(2)})$，$k=1,2,\cdots,25$。经计算可得

$$\bar{\boldsymbol{x}}^{(1)}=\frac{1}{25}\sum_{k=1}^{25}\boldsymbol{x}_k^{(1)}=(20.80,\ 12.32)^{\mathrm{T}}$$

$$\bar{\boldsymbol{x}}^{(2)}=\frac{1}{25}\sum_{k=1}^{25}\boldsymbol{x}_k^{(2)}=(12.80,\ 36.40)^{\mathrm{T}}$$

$$\boldsymbol{A}_1=\begin{pmatrix}165.6 & -126.48\\ -126.48 & 981.36\end{pmatrix},\quad \boldsymbol{A}_2=\begin{pmatrix}882.24 & 334.32\\ 334.32 & 6909.84\end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{25+25-2}(\boldsymbol{A}_1+\boldsymbol{A}_2)=\begin{pmatrix}21.83 & 4.33\\ 4.33 & 164.40\end{pmatrix}$$

$$\hat{\boldsymbol{\mu}}^{(1)}-\hat{\boldsymbol{\mu}}^{(2)}=\hat{\boldsymbol{x}}^{(1)}-\hat{\boldsymbol{x}}^{(2)}=(8.00\quad 24.08)^{\mathrm{T}}$$

$$\hat{\boldsymbol{\Sigma}}^{-1}=\frac{1}{3570.1}\begin{pmatrix}164.40 & -4.33\\ -4.33 & 21.83\end{pmatrix}$$

由(6.1.7)式可得线性判别函数

$$W(\boldsymbol{x})=\frac{1}{3570.1}(1400x_1-560x_2-100\ 80)$$

$$\approx\frac{280}{3570.1}(5x_1-2x_2-36)=\frac{280}{3570.1}W^{(*)}(\boldsymbol{x})$$

式中，当 $W(\boldsymbol{x})\geqslant 0$，即 $W^{(*)}(\boldsymbol{x})\geqslant 0$ 时，判定新病人为正常。若 $W(\boldsymbol{x})<0$ 时，可判定新病人为不正常。用原始数据进行判断时，新参加测试的人的指标 $\boldsymbol{x}=(28\quad 36)$，代入 $W^{(*)}(\boldsymbol{x})$，得 $W^{(*)}(\boldsymbol{x})=22>0$，故此人属于正常人。此外，共将表 6.1.1 中 25 个正常人数据代入 $W^{(*)}(\boldsymbol{x})$ 都大于 0，而 25 个不正常的人数据代入后有 4 个大于 0，故误判概率为 8%。

6.1.2　贝叶斯(Bayes)判别

贝叶斯判别法的基本思想为在各总体的概率分布及先验概率已知的前提下分别计算待判对象属于各总体的后验概率，并以最大后验概率对应的总体来作为待判对象的所属总体。

1. 两总体贝叶斯判别

设有两总体 π_1 和 π_2 具有概率密度函数 $f_1(x)$ 和 $f_2(x)$，根据以往的统计分析，已知出现总体 π_i 的先验概率分别为 $p_i(i=1,2)$，即 $P(\pi_i)=p_i$。当 $\boldsymbol{x}$ 样本已知时，有后验概率

$$P(\pi_i|x)=\frac{p_i f_i(x)}{\sum_{i=1}^{2}p_i f_i(x)} \tag{6.1.10}$$

判别 $\boldsymbol{x}$ 属于哪个总体 π_j，可采用最大后验概率判别规则，即有