河北省省级精品课教材

应用多元统计分析

李春林　陈旭红　编著

清华大学出版社

应用多元统计分析

李春林　陈旭红　编著

清华大学出版社
北京

内 容 简 介

本书是在河北省精品课“多元统计分析”课程建设的基础上，贴近省属院校实际，以学生的应用分析技能为主要培养目标，以方法、案例引导，对学生开展方法学习、案例分析、数据处理、结果讨论、文献阅读和论文撰写全方位的应用分析技能训练，是一本主要面向省属院校统计学各专业和其他相关专业的高年级本科生或研究生的应用型教材。

图书在版编目(CIP)数据

应用多元统计分析 / 李春林，陈旭红编著. —北京：清华大学出版社，2013
ISBN 978-7-302-33296-1

Ⅰ. ①应… Ⅱ. ①李… ②陈… Ⅲ. ①多元分析—统计分析—高等学校—教材
Ⅳ. ①O212.4

中国版本图书馆 CIP 数据核字(2013)第 168807 号

责任编辑：陈　明　洪　英
封面设计：傅瑞学
责任校对：刘玉霞
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：170mm×230mm　　印　　张：14.5　　字　　数：268 千字
版　　次：2013 年 8 月第 1 版　　印　　次：2013 年 8 月第1 次印刷
印　　数：1～2000
定　　价：29.00 元

产品编号：051546-01

前言

多元统计分析是统计学科中的一个重要分支，在自然科学、社会科学等领域具有广泛的应用，是探索多元世界强有力的工具。河北经贸大学的“多元统计分析”课程是统计学各专业的主干课程，是河北省的省级精品课程。在精品课程建设的过程中，我们结合丰富的教学、科研实践和大量鲜活的案例，贴近省属院校实际，以学生的应用分析技能为主要培养目标，以方法、案例引导进行多元统计分析方法的学习，对学生开展方法学习、案例分析、数据处理、结果讨论、文献阅读和论文撰写全方位的应用分析技能训练。

作为省属院校，我们切身体会到应用分析能力的培养对学生未来发展的重要性，也切实感受到国内纯应用性专业教材匮乏的无奈。因此，我们在建设省级精品课程的同时，结合科研和教学经验，紧贴应用分析技能培养这条省属院校学生培养与就业的生命线，编写了这本以应用为主线、以方法与软件相结合更好地解决实际问题为核心的《应用多元统计分析》教材。

本书用浅显的语言阐明各种多元统计方法的功能和原理，针对具体的案例，通过在国内广泛使用的统计分析软件 SPSS，讲授方法的上机实现和应用，尽可能详尽地介绍统计软件的各种操作选项和提供数据处理结果的解释，结合文献阅读和论文撰写对学生进行应用分析技能的培养。

本书涵盖了常用的多元统计分析方法，是一本主要面向省属院校统计学和经济学、管理学、生物医学统计等有关专业的高年级本科生或研究生的应用型教材和教学参考书，也可作为社会统计工作者和数据分析人员的实用参考书。

本书在编写过程中，研究生孟杰、刘扬、冯丽红、李圣瑜、俱翠、胡一帆、王洪彪做了大量的基础性工作，清华大学出版社对教材的编写和出版给予了大力支持，陈明编辑为本书做了大量的组织工作，在此一并表示感谢！由于作者水平有限，书中难免出现疏漏和错误，希望广大读者提出宝贵意见，以便进一步修改。

李春林

2013 年 7 月于石家庄

目录

CHAPTER 1

第1章 多元统计分析的理论基础

1.1 多元分布

1.1.1 随机向量

多元统计分析讨论的是多变量总体。把 p 个随机变量放在一起得 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$ 是一个 p 维随机向量，如果同时对 p 个变量作一次观测，则得到一个样本 $(x_{11},x_{12},\cdots,x_{1p})$，$n$ 次观测得到的样本排成一个 $n\times p$ 矩阵，称为**样本数据矩阵**(或**样本资料矩阵**)，记为

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \overset{\text{def}}{=\!=} \begin{bmatrix} \boldsymbol{X}_{(1)} \\ \boldsymbol{X}_{(2)} \\ \vdots \\ \boldsymbol{X}_{(n)} \end{bmatrix} \overset{\text{def}}{=\!=} (\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_p) \tag{1-1}$$

矩阵 $\boldsymbol{X}$ 的第 i 行：$\boldsymbol{X}_{(i)}=(x_{i1},x_{i2},\cdots,x_{ip})(i=1,2,\cdots,n)$ 表示对第 i 个样本的观测值，是一个 p 维的随机向量。矩阵 $\boldsymbol{X}$ 的第 j 列：$\boldsymbol{X}_j=\begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}(j=1,2,\cdots,p)$ 表示对第 j 个变量的 n 次观测，是一个 n 维随机向量。

1.1.2 多元分布函数与密度函数

1. 联合分布

设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$ 是 p 维随机向量，称 p 元函数

$$F(x_1,x_2,\cdots,x_p)=P\{X_1\leqslant x_1,x_2,\cdots,X_p\leqslant x_p\} \tag{1-2}$$

为 $\boldsymbol{X}$ 的**联合分布函数**。

若存在非负函数 $f(x_1,x_2\cdots,x_p)$，使得随机向量 $\boldsymbol{X}$ 的联合分布函数对一切 $(x_1,x_2,\cdots,x_p)\in\mathbb{R}^p$ 均可表示为

$$F(x_1,x_2,\cdots,x_p)=\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_p}f(t_1,t_2,\cdots,t_p)\mathrm{d}t_1\mathrm{d}t_2\cdots\mathrm{d}t_p \tag{1-3}$$

则称 $f(x_1,x_2,\cdots,x_p)$ 为**连续型随机向量 $\boldsymbol{X}$ 的联合概率密度函数**，简称为**多元密度函数**或**密度函数**。

多元密度函数 $f(x_1,x_2,\cdots,x_p)$ 满足以下两条性质：

(1) 对一切实数 $x_1,x_2,\cdots,x_p$，$f(x_1,x_2\cdots,x_p)\geqslant 0$；

(2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,x_2,\cdots,x_p)\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_p=1$。

2. 边缘分布

称随机向量 $\boldsymbol{X}$ 的部分分量 $(X_{i_1},\cdots,X_{i_m})(1\leqslant m<p)$ 的分布为**边缘分布**。

设 $\boldsymbol{X}^{(1)}=(X_{i_1},\cdots,X_{i_r})$ 为 r 维随机向量，$\boldsymbol{X}^{(2)}=(X_{i_{r+1}},\cdots,X_{i_p})$ 为 $p-r$ 维随机向量。若 $\boldsymbol{X}=(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)})$，则 $\boldsymbol{X}^{(1)}$ 的边缘分布密度函数为

$$\begin{aligned}f_1(x^{(1)})&=f_1(x_{i_1},\cdots,x_{i_r})\\&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,x_2,\cdots,x_p)\mathrm{d}x_{i_{r+1}}\cdots\mathrm{d}x_{i_p}\end{aligned} \tag{1-4}$$

$\boldsymbol{X}^{(2)}$ 的边缘分布密度函数为

$$\begin{aligned}f_2(x^{(2)})&=f_2(x_{i_{r+1}},\cdots,x_{i_p})\\&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,x_2,\cdots,x_p)\mathrm{d}x_{i_1}\cdots\mathrm{d}x_{i_r}\end{aligned} \tag{1-5}$$

3. 条件分布

称给定 $\boldsymbol{X}^{(2)}$ 时 $\boldsymbol{X}^{(1)}$ 的分布为**条件分布**。当 $\boldsymbol{X}$ 的密度函数为 $f(x^{(1)},x^{(2)})$、$\boldsymbol{X}^{(2)}$ 的密度函数为 $f_2(x^{(2)})$ 时，则给定 $\boldsymbol{X}^{(2)}$ 时 $\boldsymbol{X}^{(1)}$ 的条件密度函数为

$$f_1(x^{(1)}\mid x^{(2)})=f(x^{(1)},x^{(2)})/f_2(x^{(2)}) \tag{1-6}$$

4. 独立性

设 $X_1,\cdots,X_p$ 是 p 个随机变量，X_i 的分布函数记为 $F_i(x_i)(i=1,\cdots,p)$，

$F(x_1,\cdots,x_p)$是 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$的联合分布函数。若对一切实数 x_1，$x_2,\cdots,x_p$，

$$F(x_1,\cdots,x_p)=F_1(x_1)\cdots F_p(x_p) \tag{1-7}$$

均成立，则称 $X_1,\cdots,X_p$ **相互独立**。

在连续型随机变量的情况下，$X_1,\cdots,X_p$ 相互独立，当且仅当 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$的联合密度函数 $f(x_1,x_2,\cdots,x_p)$满足：对一切实数 $x_1,x_2,\cdots,x_p$，

$$f(x_1,x_2,\cdots,x_p)=f_1(x_1)f_2(x_2)\cdots f_p(x_p) \tag{1-8}$$

均成立，其中 $f_i(x_i)$是 $\boldsymbol{X}_i$ 的密度函数($i=1,\cdots,p$)。

1.1.3　随机向量的数字特征

设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$，$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_q)$是两个随机向量。

1. 均值向量

若 $E(X_i)=\mu_i$ 存在，则称 $E(\boldsymbol{X})=(E(X_1),E(X_2),\cdots,E(X_p))=(\mu_1,\mu_2,\cdots,\mu_p)$为随机向量 $\boldsymbol{X}$ 的**均值向量**。

2. 协方差阵

若 X_i 和 Y_j 的协方差 $\mathrm{cov}(X_i,Y_j)$存在($i=1,2,\cdots,p;\ j=1,2,\cdots,q$)，则称

$$\begin{aligned}\mathrm{cov}(\boldsymbol{X},\boldsymbol{Y})&=E[(\boldsymbol{X}-E\boldsymbol{X})(\boldsymbol{Y}-E\boldsymbol{Y})]\\&=\begin{bmatrix}\mathrm{cov}(X_1,Y_1) & \mathrm{cov}(X_1,Y_2) & \cdots & \mathrm{cov}(X_1,Y_q)\\ \mathrm{cov}(X_2,Y_1) & \mathrm{cov}(X_2,Y_2) & \cdots & \mathrm{cov}(X_2,Y_q)\\ \vdots & \vdots & & \vdots\\ \mathrm{cov}(X_p,Y_1) & \mathrm{cov}(X_p,Y_2) & \cdots & \mathrm{cov}(X_p,Y_q)\end{bmatrix}\end{aligned} \tag{1-9}$$

为随机向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的协方差阵。若 $\mathrm{cov}(\boldsymbol{X},\boldsymbol{Y})=\boldsymbol{O}$(其中 $\boldsymbol{O}$ 表示零矩阵)，则称 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ **不相关**。若 $\boldsymbol{Y}=\boldsymbol{X}$，则 $\mathrm{cov}(\boldsymbol{X},\boldsymbol{X})=D(\boldsymbol{X})$为 $\boldsymbol{X}$ 的(协)方差阵。

3. 相关阵

记 $r_{ij}=\dfrac{\mathrm{cov}(X_i,X_j)}{\sqrt{\mathrm{var}(X_i)}\sqrt{\mathrm{var}(X_j)}}=\dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$ $(i,j=1,2,\cdots,p)$，称 $\boldsymbol{R}=(r_{ij})_{p\times p}$ 为 $\boldsymbol{X}$ 的**相关阵**，r_{ij} 也称为分量 X_i 与 X_j 之间的(线性)相关系数。

1.2 多元正态分布

1.2.1 多元正态分布的定义和性质

1. 多元正态分布的定义

定义 1(密度定义) 若 p 元随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$ 的概率密度函数为

$$f(x_1,x_2,\cdots,x_p)=\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right\} \tag{1-10}$$

其中,$\boldsymbol{\mu}$ 为 $p\times 1$ 的常数向量,$\boldsymbol{\Sigma}$ 为正定阵($\boldsymbol{\Sigma}>\boldsymbol{O}$),$|\boldsymbol{\Sigma}|$ 为协差阵$\boldsymbol{\Sigma}$ 的行列式,则称 $\boldsymbol{X}$ 服从 p 元正态分布,也称 $\boldsymbol{X}$ 为 p 元正态变量。记为 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$。

当 $p=2$ 时,可以得到二元正态分布的密度公式。

2. 多元正态分布的性质

性质 1 若 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$,则 $E(\boldsymbol{X})=\boldsymbol{\mu}$,$D(\boldsymbol{X})=\boldsymbol{\Sigma}$。

本性质将正态分布的参数$\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 赋予了明确的统计意义。

性质 2 设 $Y_1,Y_2,\cdots,Y_k$ 相互独立,$Y_i\sim N_p(\mu_i,\Sigma_i)(\forall i)$,则

$$\sum_{i=1}^{k}Y_i\sim N_p\left(\sum_{i=1}^{k}\mu_i,\sum_{i=1}^{k}\Sigma_i\right) \tag{1-11}$$

性质 3 如果正态随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$ 的协方差阵$\boldsymbol{\Sigma}$ 是对角阵,则 $\boldsymbol{X}$ 的各分量是相互独立的随机变量。

性质 4 多元正态向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$的任意线性变换仍然服从多元正态分布。

1.2.2 多元正态分布均值向量和协方差阵的估计

1. 均值向量的估计

设样本 $\boldsymbol{X}_{(1)},\boldsymbol{X}_{(2)},\cdots,\boldsymbol{X}_{(n)}$ 间相互独立,且服从于 p 元正态分布 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$,$n>p$,$\boldsymbol{\Sigma}>\boldsymbol{O}$,则总体参数均值$\boldsymbol{\mu}$ 的估计量是

$$\hat{\boldsymbol{\mu}}=\overline{\boldsymbol{X}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{(i)}=\frac{1}{n}\left(\sum_{i=1}^{n}X_{i1},\sum_{i=1}^{n}X_{i2},\cdots,\sum_{i=1}^{n}X_{ip}\right)\boldsymbol{X}$$
$$=(\overline{X}_1,\overline{X}_2,\cdots,\overline{X}_p) \tag{1-12}$$

均值向量 $\boldsymbol{\mu}$ 的估计量就是样本均值向量，且 $\hat{\boldsymbol{\mu}}$ 是 $\boldsymbol{\mu}$ 的无偏估计。

2. 协方差阵的估计

总体参数协方差阵 $\boldsymbol{\Sigma}$ 的极大似然估计是

$$\hat{\boldsymbol{\Sigma}}_p=\frac{1}{n}\boldsymbol{L}=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_{(i)}-\overline{\boldsymbol{X}})(\boldsymbol{X}_{(i)}-\overline{\boldsymbol{X}})^{\mathrm{T}}$$

$$=\frac{1}{n}\begin{bmatrix}\sum_{i=1}^{n}(X_{i1}-\overline{X}_1)^2 & \sum_{i=1}^{n}(X_{i1}-\overline{X}_1)(X_{i2}-\overline{X}_2) & \cdots & \sum_{i=1}^{n}(X_{i1}-\overline{X}_1)(X_{ip}-\overline{X}_p)\\ & \sum_{i=1}^{n}(X_{i2}-\overline{X}_2)^2 & \cdots & \sum_{i=1}^{n}(X_{i2}-\overline{X}_2)(X_{ip}-\overline{X}_p)\\ & & \ddots & \vdots\\ & & & \sum_{i=1}^{n}(X_{ip}-\overline{X}_p)^2\end{bmatrix} \tag{1-13}$$

其中 $\boldsymbol{L}$ 是离差阵，它是每一个样本（向量）与样本均值（向量）的离差积形成的 n 个 p 阶对称阵的和。同一元相似，$\hat{\boldsymbol{\Sigma}}_p$ 不是 $\boldsymbol{\Sigma}$ 的无偏估计，为了得到无偏估计我们常用样本协差阵 $\hat{\boldsymbol{\Sigma}}=\frac{1}{n-1}\boldsymbol{L}$ 作为总体协差阵的估计。

1.3　多元正态分布均值向量和协方差阵的检验

1.3.1　单总体均值向量的检验

设总体 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，随机样本 $\boldsymbol{X}_{(\alpha)}(\alpha=1,\cdots,n)$。检验

$$H_0:\boldsymbol{\mu}=\boldsymbol{\mu}_0(\boldsymbol{\mu}_0\text{为已知向量}),\quad H_1:\boldsymbol{\mu}\neq\boldsymbol{\mu}_0$$

1. 当 $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_0$ 已知时均值向量的检验

取检验统计量为

$$T_0^2=n(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_0)\overset{H_0}{\sim}\chi^2(p) \tag{1-14}$$

按传统的检验方法，对给定的显著性水平 α，查 χ^2 分布临界值表得 λ_α，使 $P\{T_0^2>\lambda_\alpha\}=\alpha$，则否定域为 $\{T_0^2>\lambda_\alpha\}$。

由样本值 $x_{(\alpha)}(\alpha=1,\cdots,n)$，计算 $\bar{\boldsymbol{X}}$ 及 T_0^2 值，若 $T_0^2>\lambda_\alpha$，则否定 H_0，否则 H_0 相容。

利用统计软件还可以通过计算显著性概率值（p 值）给出检验结果，且由此得出的结果更丰富。

假设在 H_0 成立情况下，随机变量 $T_0^2\sim\chi^2(p)$，由样本值计算得到 T_0^2 的值为 d，同时可以计算以下概率值

$$p=P\{T_0^2\geqslant d\} \tag{1-15}$$

常称此概率值为**显著性概率值**，或简称为 p 值。

对给定的显著性水平 α，当 $p<\alpha$ 时，则在显著性水平 α 下否定假设 H_0。在这种情况下，可能犯“以真当假”的第一类错误，且 α 就是犯第一类错误的概率。

当 $p\geqslant\alpha$ 时，在显著性水平 α 下 H_0 相容。在这种情况下，可能犯“以假当真”的第二类错误，且犯下第二类错误的概率 β 为

$$\beta=P\{T_0^2\leqslant\lambda_\alpha\mid\boldsymbol{\mu}=\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_0\} \tag{1-16}$$

其中检验统计量 $T_0^2\sim\chi^2(p,\delta)$，非中心参数

$$\delta=n(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) \tag{1-17}$$

p 值的直观含义可以这样看，检验统计量 T_0^2 的大小反映 $\bar{\boldsymbol{X}}$ 与 $\boldsymbol{\mu}_0$ 的偏差大小，当 H_0 成立时 T_0^2 的值应该较小。现由观测数据计算 T_0^2 值为 d；当 H_0 成立时统计量 $T_0^2\sim\chi^2(p)$，由 χ^2 分布可以计算该统计量$\geqslant d$ 的概率值（即 p 值）。比如 $p=0.02<\alpha=0.05$，这时出现一个比较小的概率标准（$\alpha=0.05$）还要小的事件 $\{T_0^2\geqslant d\}$。也就是说，在 $\boldsymbol{\mu}=\boldsymbol{\mu}_0$ 假设下，观测数据中极少情况会出现 T_0^2 的值大于等于 d 值，故在 0.05 显著性水平下有足够的证据否定原假设，即认为 $\boldsymbol{\mu}$ 与 $\boldsymbol{\mu}_0$ 有显著的差异。

又比如当 $p=0.22\geqslant\alpha=0.05$ 时，表示在 $\boldsymbol{\mu}=\boldsymbol{\mu}_0$ 假设下，观测数据中经常会出现 T_0^2 的值大于等于 d 值的情况，故在 0.05 显著性水平下没有足够的证据否定原假设，即认为 $\boldsymbol{\mu}$ 与 $\boldsymbol{\mu}_0$ 没有显著的差异。

2. 当 $\boldsymbol{\Sigma}$ 未知时均值向量的检验

考虑统计量

$$\begin{aligned}T^2&=(n-1)[\sqrt{n}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)]^{\mathrm{T}}\boldsymbol{A}^{-1}[\sqrt{n}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)]\\&=(n-1)n(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{A}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)\sim T^2(p,n-1)\end{aligned} \tag{1-18}$$

再利用 T^2 与 F 分布的关系，统计量取为

$$F=\frac{(n-1)-p+1}{(n-1)p}T^2\sim F(p,(n-1)-p+1)$$
$$=F(p,n-p) \tag{1-19}$$

进行检验。

3. 似然比统计量

在数理统计中关于总体参数的假设检验，通常是利用最大似然原理导出似然比统计量来进行检验。在多元统计分析中几乎所有重要的检验都是利用最大似然比原理给出的。下面我们回顾一下最大似然比原理。

设 p 元总体的密度函数为 $f(x,\theta)$，其中 θ 是未知参数，且 $\theta\in\boldsymbol{\Theta}$（参数空间），又设 $\boldsymbol{\Theta}_0$ 是 $\boldsymbol{\Theta}$ 的子集，我们希望对下列假设：

$$H_0:\theta\in\boldsymbol{\Theta}_0,\quad H_1:\theta\overline{\in}\boldsymbol{\Theta}_0$$

作出判断，这就是假设检验问题。称 H_0 为**原假设**（或**零假设**），H_1 为**对立假设**（或**备择假设**）。

从总体 $\boldsymbol{X}$ 抽取容量为 n 的样本 $\boldsymbol{X}_{(t)}(t=1,\cdots,n)$。把样本的联合密度函数

$$L(x_{(1)},\cdots,x_{(n)};\theta)=\prod_{t=1}^{n}f(x_{(t)};\theta) \tag{1-20}$$

记为 $L(\boldsymbol{X};\theta)$，并称它为**样本的似然函数**。

引入统计量

$$\lambda=\max_{\theta\in\Theta_0}L(\boldsymbol{X};\theta)\mid\max_{\theta\in\Theta}L(\boldsymbol{X};\theta) \tag{1-21}$$

它是样本 $\boldsymbol{X}_{(t)}(t=1,\cdots,n)$ 的函数，常称 λ 为**似然比统计量**。由于 $\boldsymbol{\Theta}_0\subset\boldsymbol{\Theta}$，从而 $0\leqslant\lambda\leqslant1$。

由最大似然比原理知，如果 λ 取值太小，说明 H_0 为真时观测到此样本 $\boldsymbol{X}_{(t)}(t=1,\cdots,n)$ 的概率比 H_0 为不真时观测到此样本 $\boldsymbol{X}_{(t)}(t=1,\cdots,n)$ 的概率要小得多，故有理由认为假设 H_0 不成立，所以从似然比出发，以上检验问题的否定域为

$$\{\lambda(\boldsymbol{X}_{(1)},\cdots,\boldsymbol{X}_{(n)})<\lambda_\alpha\} \tag{1-22}$$

按传统的检验方法，λ_α 是由显著性水平 α 确定的临界值，它满足当 H_0 成立时使得

$$P\{\lambda(\boldsymbol{X}_{(1)},\cdots,\boldsymbol{X}_{(n)})<\lambda_\alpha\}=\alpha \tag{1-23}$$

为了得到 λ_α，必须研究似然比统计量 λ 的抽样分布。在一些特殊的情况下，可以得到 λ 的精确分布；但在很多情况下是得不到 λ 的精确分布的。当样本量很大且满足一定正则条件时，$-2\ln\lambda$ 的抽样分布与 χ^2 分布十分接近。下面不加证明地给出一条很有用的结论。

定理 1 当样本容量 n 很大时，

$$-2\ln\lambda = -2\ln\left[\max_{\theta\in\Theta_0} L(\boldsymbol{X};\theta) \Big/ \max_{\theta\in\Theta} L(\boldsymbol{X};\theta)\right] \tag{1-24}$$

近似服从自由度为 f 的 χ^2 分布，其中 $f=\Theta$ 的维数 $-\Theta_0$ 的维数。

本章将讨论的一些检验问题，就是利用似然比统计量的近似分布进行检验的方法。

当 $\boldsymbol{\Sigma}$ 未知时检验均值向量 $\boldsymbol{\mu}=\boldsymbol{\mu}_0$ 的似然比统计量及其分布。

设样本的似然函数为 $L(\boldsymbol{\mu},\boldsymbol{\Sigma})$。检验均值向量 $\boldsymbol{\mu}=\boldsymbol{\mu}_0$ 的似然比统计量为

$$\lambda = \max_{\mu=\mu_0,\Sigma>O} L(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}) \Big/ \max_{\mu,\Sigma>O} L(\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{1-25}$$

否定域为

$$\{\lambda<\lambda_\alpha\}\Leftrightarrow\{T^2>T_\alpha^2\}\Leftrightarrow\{F>F_\alpha\}$$

其中

$$F=\frac{n-p}{p}\frac{T^2}{n-1}\sim F(p,n-p) \tag{1-26}$$

1.3.2 多总体均值向量的检验

1. 两正态总体均值向量的检验

1) 两总体协方差阵相等(但未知)时均值向量的检验

设 $\boldsymbol{X}_{(\alpha)}(\alpha=1,\cdots,n)$ 为来自总体 $\boldsymbol{X}\sim N_{(p)}(\boldsymbol{\mu}^{(1)},\boldsymbol{\Sigma})$ 的随机样本；$\boldsymbol{Y}_{(\alpha)}(\alpha=1,\cdots,m)$ 为来自样本总体 $\boldsymbol{Y}\sim N_{(p)}(\boldsymbol{\mu}^{(2)},\boldsymbol{\Sigma})$ 的随机样本，且相互独立，$\boldsymbol{\Sigma}$ 未知。检验

$$H_0:\boldsymbol{\mu}^{(1)}=\boldsymbol{\mu}^{(2)},\quad H_1:\boldsymbol{\mu}^{(1)}\neq\boldsymbol{\mu}^{(2)}$$

当 $p=1$ 时，取检验统计量为

$$t=\frac{(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})\Big/\sqrt{\frac{1}{n}+\frac{1}{m}}}{\sqrt{\left[\sum_{i=1}^{n}(\boldsymbol{X}_{(i)}-\bar{\boldsymbol{X}})^2+\sum_{j=1}^{m}(\boldsymbol{Y}_{(j)}-\bar{\boldsymbol{Y}})^2\right]\Big/n+m-2}}\sim t(n+m-2) \tag{1-27}$$

即

$$t^2=\frac{nm}{m+n}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})^{\mathrm{T}}\left[\frac{\sum_{i=1}^{n}(\boldsymbol{X}_{(i)}-\bar{\boldsymbol{X}})^2+\sum_{j=1}^{m}(\boldsymbol{Y}_{(j)}-\bar{\boldsymbol{Y}})^2}{n+m-2}\right]^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})$$