多元统计分析

（第 2 版）

陈钰芬　陈骥　主编

清华大学出版社

北　京

内 容 简 介

本书系统地介绍了多元统计分析技术的基本思想和方法原理，以社会、经济、商务等领域的实际问题为案例，结合SAS软件，介绍各种方法的SAS操作、实现过程与结果解释，帮助读者理解并掌握多元统计分析的基本方法，熟练应用软件进行数据分析，提高对实际数据的分析挖掘能力。

本书内容重点突出、习题设置合理、教学资源丰富，适合作为普通高等学校统计学专业或大数据相关专业本科生、经济管理类或社科类专业研究生的教材，也可作为从事社会、经济、管理等研究和实践工作的人士进行量化研究的参考书。

图书在版编目(CIP)数据

多元统计分析 / 陈钰芬，陈骥主编. -- 2版.

北京 ：清华大学出版社，2026. 3. -- ISBN 978-7-302

-70866-7

Ⅰ. O212.4

中国国家版本馆CIP数据核字第20265VS293号

责任编辑：高　岫

封面设计：马筱琨

版式设计：思创景点

责任校对：成凤进

责任印制：丛怀宇

出版发行：清华大学出版社

网　　址：https://www.tup.com.cn，https://www.wqxuetang.com

地　　址：北京清华大学学研大厦A座　　邮　　编：100084

社 总 机：010-83470000　　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：天津鑫丰华印务有限公司

经　　销：全国新华书店

开　　本：185mm×260mm　　**印　　张：**17　　**字　　数：**414千字

版　　次：2020年8月第1版　2026年4月第2版　　**印　　次：**2026年4月第1次印刷

定　　价：68.00元

产品编号：107298-01

前　言

数字化时代，数据成为一种战略性资产。无论是党政机关，还是众多企事业单位，都需要基于数据做出科学正确的决策。大数据正在加速渗透到各行各业，重塑未来战略格局。多元统计分析方法是处理多变量数据不可缺少的重要技术和方法，是大数据分析的重要工具。

多元统计分析是以概率统计为基础，应用线性代数的基本原理和方法，结合计算机对实际资料和信息进行分析挖掘的一种统计分析技术。它的应用性极强，在自然科学、社会科学、经济管理等各领域得到了越来越广泛的应用。

本书是浙江省登峰学科、浙江省优势特色学科、国家一流专业的建设成果之一，由作者结合二十多年的教学和科研工作经验编写而成，着重突出以下特点。

(1)注重统计基本思想。本书以贴近生活与社会实际的问题为引例，以深入浅出的方式简要阐述多元统计分析的基本思想，帮助学生理解，激发学习兴趣。

(2)阐明统计基本原理。多元统计方法的数学推导深奥繁杂，方法原理抽象难懂，为便于学生阅读并较好地理解，本书对各种方法的基本原理进行了详细的推导。在不失严谨的前提下，略过了一些复杂程度高但又不影响方法原理理解的数学推导，读者只需掌握初步的微积分、线性代数和概率统计知识，便能理解。

(3)突出实际案例应用。本着深入浅出的宗旨，在系统介绍多元分析基本理论和方法的同时，结合社会、经济、商务运行等领域的研究实例，把多元分析的方法与实际应用结合起来。立足国情省情，将习近平新时代中国特色社会主义思想和党的二十大、二十届三中全会精神融入教学案例。所有案例数据都是中国经济运行与改革实践中的真实数据，通过真实案例学习，学生可加深对现实问题的理解，提高解决实际问题的思维能力和分析能力，促使知识体系适应时代变革对统计人才的需求。

(4)结合 SAS 软件实现。多元统计分析的应用离不开计算机，本书案例主要运用 SAS 软件实现，在每种方法后结合实例介绍 SAS 软件的实现过程并解释结果。这有利于将 SAS 软件更好地融入各章内容，使读者深切体会多元统计分析的意义，便于读者进入应用领域。

(5)习题设置合理。为使读者掌握本书内容，又考虑到这门课程的应用性和实践性，每章给出一些思考与练习题，这些习题安排侧重于对基本概念的理解和知识点的实际应用，并不注重解题的数学技巧和难度。

(6)践行沁德启智使命。本书结合党的二十大精神、二十届三中全会部署和经济社会发展前沿，融入中国式现代化、新质生产力、经济高质量发展、提振消费等实际案例，从国家、区域、个人三个层面，将家国情怀、文化自信、生态文明、科学精神等德育元素有效融入教学案例，在知识传授中注重价值引领，助力学生实现知识、能力、素养、价值目标。

(7)数字教学资源丰富。本书提供丰富的数字教学资源，包括课程知识图谱、微课教学视频、教学案例与习题数据。课程团队构建了覆盖课程全领域、结构清晰的知识图谱，系统

梳理了课程83个重要知识点,绘制了144个知识点关系,匹配了322个教学资源,以结构化形式描述了课程所包含的知识点、教学资源、问题体系、能力体系之间的关系,具有知识管理、学习导航、学习评估等功能,帮助学生深入理解课程的知识点及其相互联系,为学生提供更加高效、精准的知识学习路径。通过扫描书中相应的二维码,可获取教学资源,观看相应的学习视频,链接知识图谱。

本书旨在让学生理解并掌握多元统计分析的基本方法,熟练应用软件进行数据分析,适合作为统计学专业本科生和非统计学专业研究生的教材,也可作为大数据或其他专业学生学习多元统计分析的教材或教学参考书,还可作为从事社会、经济、管理等研究和实践的人士进行量化研究的参考书。

本书共分10章。第1章、第2章主要介绍一元统计推广到多元统计的内容,阐述了多元正态分布的基本概念及其统计推断。第3章至第10章介绍了各种多元统计分析技术,这部分内容具有很强的实用性,特别是介绍了各种降维技术,将原始的多个指标化为少数几个综合指标,便于学生对数据进行分析挖掘。

本书由浙江工商大学陈钰芬教授和陈骥教授担任主编,具体编写分工为:张荣茂编写第1～2章,陈钰芬编写第3～8章,陈骥编写第9～10章。在本书的编写过程中,陈思超博士、硕士生苏楚文在数据处理和案例资料搜集方面以细致严谨的态度完成了大量繁琐的工作。我们也参考和吸收了一些同类教材的成果,在此一并感谢!

由于编者水平有限,书中谬误之处在所难免,恳请读者批评指正。


教学资源


案例与习题数据


SAS在线安装视频

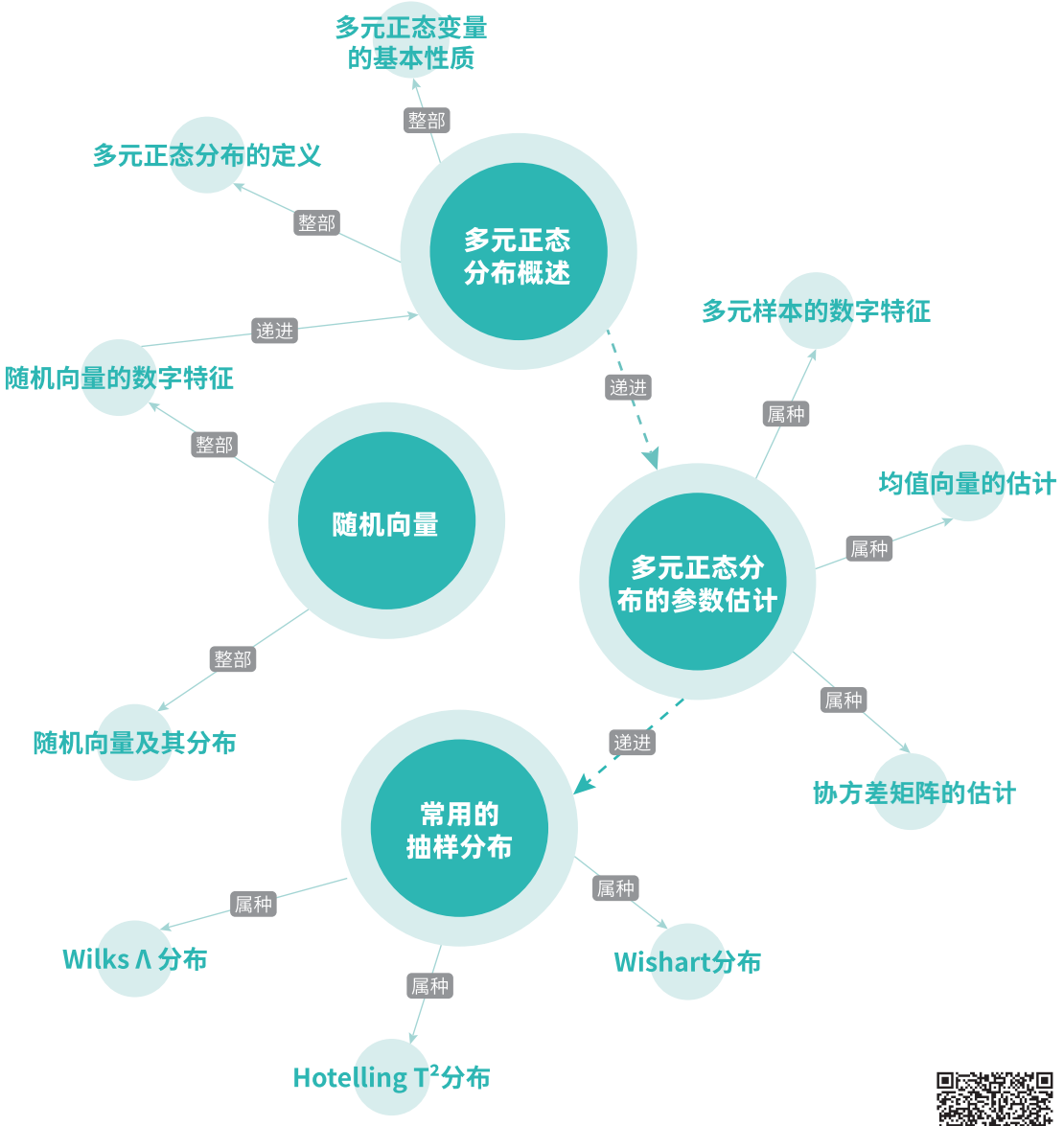
陈钰芬

2026年1月

目　录

第 1 章

多元正态分布



多元正态分布

1.1 随机向量

在多元统计分析中，多元正态分布占有相当重要的地位。就理论而言，多元正态分布有相当优良的性质，因此多元统计分析的许多重要理论和方法或直接或间接地建立在正态分布的基础上，而围绕多元正态分布，已经建立了一套行之有效的统计推断方法。就实践而言，在实际中遇到的许多随机向量都服从或近似服从正态分布。

在研究许多实际问题时，往往会遇到多指标问题，即在一个问题中涉及多个随机变量。由于这些指标之间往往有某种联系，因此需要把这些指标作为一个总体来研究。多元统计分析研究的就是多指标的总体。

1.1.1 随机向量的定义

假定我们每次同时观测一个个体的 p 个指标，将这 p 个指标(即变量)放在一起得到一个 p 维随机向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$，表示同一次观测的 p 个变量，而由这 p 个需要观测的指标的个体所构成的总体，我们称为 p 元总体。每次观测得到一个样品，全体 n 个样品形成一个样本。

定义1.1 p 个随机变量$X_1, X_2, \cdots, X_p$所组成的向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$称为随机向量。

注：如无特殊说明，本书中所称向量均指列向量。

假定我们一共进行了 n 次观测，得到的数据放在一起排成一个 $n\times p$ 矩阵，称为样本数据矩阵(或样本资料矩阵)，记为

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

横看矩阵的第 i 行，

$$\boldsymbol{X}'_{(i)}=(x_{i1}, x_{i2}, \cdots, x_{ip}),(i=1, \cdots, n)$$

表示第 i 个样品的观测值。在具体观测之前，它是一个 p 维的随机向量。

竖看矩阵的第 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。在具体观测之前，它是一个 n 维的随机向量。

利用这样的记号，我们可以将样本数据矩阵表示为

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{X}'_{(1)}\\ \boldsymbol{X}'_{(2)}\\ \vdots\\ \boldsymbol{X}'_{(n)}\end{bmatrix}=(\boldsymbol{X}_1,\ \boldsymbol{X}_2,\ \cdots,\ \boldsymbol{X}_p)$$

在观测之前，它是一个随机矩阵。而一旦观测值取定，$\boldsymbol{X}$ 就是一个数据矩阵。

多元统计分析中所涉及的很多方法都是充分运用各种手段从样本资料矩阵中提取信息，因此本书中需要运用随机向量或是多个随机向量构成的随机矩阵的一些性质。需要注意的是，本章中的多元样本是指简单随机样本，即不同样品的观测值之间是相互独立的，但是对多元样本中的每个样品而言，p 个指标的观测值之间往往是有相依关系的。不同样品的观测值之间有相依关系的一般属于多元时间序列分析研究的范畴。

1.1.2　随机向量的分布

随机向量可以由它的分布函数来完全描述。

定义1.2　设 $\boldsymbol{X}=(X_1,\ X_2,\ \cdots,\ X_p)'$ 为 p 维随机向量，其联合分布函数为

$$F(x_1,\ \cdots,\ x_p)=P(X_1\leqslant x_1,\ \cdots,\ X_p\leqslant x_p)$$

记为 $\boldsymbol{X}\sim F$。

定义1.3　如果存在非负函数 $f(x_1,\ \cdots,\ x_p)$，使得对一切 $(x_1,\ \cdots,\ x_p)\in\mathbf{R}^p$，联合分布函数均可表示为

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

则称 $\boldsymbol{X}$ 为连续型随机向量，称 $f(x_1,\ \cdots,\ x_p)$ 为 $\boldsymbol{X}$ 的联合概率密度函数，简称为密度函数或者分布密度。

密度函数有以下两条重要性质：

(1) $\forall\ (x_1,\ \cdots,\ x_p)\in\mathbf{R}^p$，$f(x_1,\ \cdots,\ x_p)\geqslant 0$；

(2) $\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}f(t_1,\ \cdots,\ t_p)\,\mathrm{d}t_1\cdots\mathrm{d}t_p=1$。

事实上，一个 p 维变量的函数 $f(x_1,\ \cdots,\ x_p)$ 能作为 p 中某个随机向量的分布密度，当且仅当以上两条性质成立时。

对于随机向量，有时我们关注的是部分分量的分布信息，因此还需要定义边缘分布。

定义1.4　设 $\boldsymbol{X}=(X_1,\ X_2,\ \cdots,\ X_p)'$ 为 p 维随机向量，其联合分布函数为 $F(x_1,\ \cdots,\ x_p)$。$\boldsymbol{X}$ 的 q 个分量所组成的子向量 $(X_{i_1},\ \cdots,\ X_{i_q})'$ 的分布称为 $\boldsymbol{X}$ 的边缘(或边际)分布。

如果我们将 $\boldsymbol{X}$ 划分为 q 维子向量 $\boldsymbol{X}^{(1)}$ 与 $p-q$ 维子向量 $\boldsymbol{X}^{(2)}$，那么 $\boldsymbol{X}^{(1)}$ 的边缘分布为

$$\begin{aligned}F^{(1)}(x_1,\ \cdots,\ x_q)&=P(X_1\leqslant x_1,\ \cdots,\ X_q\leqslant x_q)\\&=P(X_1\leqslant x_1,\ \cdots,\ X_q\leqslant x_q,\ X_{q+1}\leqslant\infty,\ \cdots,\ X_p\leqslant\infty)\\&=F(x_1,\ \cdots,\ x_q,\ \infty,\ \cdots,\ \infty)\end{aligned}$$

当 $\boldsymbol{X}$ 有分布密度时，$\boldsymbol{X}^{(1)}$ 也有分布密度，其边缘密度为

$$f^{(1)}(x_1,\ \cdots,\ x_q)=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}f(x_1,\ \cdots,\ x_p)\,\mathrm{d}t_{q+1}\cdots\mathrm{d}t_p$$

在概率论中，我们学习过随机变量的条件分布与独立性等相关概念，随机向量中也有类似概念。

定义1.5 如果我们将 $\boldsymbol{X}$ 划分为 q 维子向量 $\boldsymbol{X}^{(1)}$ 与 $p-q$ 维子向量 $\boldsymbol{X}^{(2)}$，那么在给定 $\boldsymbol{X}^{(2)}$ 时，$\boldsymbol{X}^{(1)}$ 的分布称为条件分布。如果 $\boldsymbol{X}$ 有密度函数 $f(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$，那么给定 $\boldsymbol{X}^{(2)}$ 时，$\boldsymbol{X}^{(1)}$ 的密度函数为

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

其中，$f_2(\boldsymbol{x}^{(2)})$ 是 $\boldsymbol{X}^{(2)}$ 的边缘密度。

定义1.6 若 p 个随机向量 $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_p$ 的联合分布等于各自边缘分布的乘积，则称 $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_p$ 是相互独立的。需要注意的是，如果 $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_p$ 相互独立，那么其中任意两个随机向量两两独立，但是反之不真。

1.1.3 随机向量的数字特征

设 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$，$\boldsymbol{Y}=(Y_1, Y_2, \cdots, Y_q)'$ 为两个随机向量。

若 $E(X_i)=\mu_i$ 存在，则称

$$E(\boldsymbol{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$$

为随机向量 $\boldsymbol{X}$ 的均值向量。

根据定义容易验证均值向量具有以下性质：

$$E(\boldsymbol{AX}) = \boldsymbol{A}E(\boldsymbol{X})$$

$$E(\boldsymbol{AXB}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}$$

其中 $\boldsymbol{A}$，$\boldsymbol{B}$ 为大小适合矩阵运算的常数矩阵。

若 X_i 与 X_j 的协方差存在 $(i, j=1, \cdots, p)$，则称

$$D(\boldsymbol{X}) = E\left[\left(\boldsymbol{X}-E(\boldsymbol{X})\right)\left(\boldsymbol{X}-E(\boldsymbol{X})\right)'\right]$$

$$= \begin{bmatrix} D(X_1) & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_p) \\ \operatorname{cov}(X_2, X_1) & D(X_2) & \cdots & \operatorname{cov}(X_2, X_p) \\ \vdots & \vdots & \vdots & \vdots \\ \operatorname{cov}(X_p, X_1) & \operatorname{cov}(X_p, X_2) & \cdots & D(X_p) \end{bmatrix}$$

为随机向量 $\boldsymbol{X}$ 的协方差矩阵。

若 X_i 与 Y_j 的协方差存在 $(i=1, \cdots, p;\ j=1, \cdots, q)$，则称

$$\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y}) = E\left[\left(\boldsymbol{X}-E(\boldsymbol{X})\right)\left(\boldsymbol{Y}-E(\boldsymbol{Y})\right)'\right]$$

$$= \begin{bmatrix} \operatorname{cov}(X_1, Y_1) & \operatorname{cov}(X_1, Y_2) & \cdots & \operatorname{cov}(X_1, Y_q) \\ \operatorname{cov}(X_2, Y_1) & \operatorname{cov}(X_2, Y_2) & \cdots & \operatorname{cov}(X_2, Y_q) \\ \vdots & \vdots & \vdots & \vdots \\ \operatorname{cov}(X_p, Y_1) & \operatorname{cov}(X_p, Y_2) & \cdots & \operatorname{cov}(X_p, Y_q) \end{bmatrix}$$

为随机向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的协方差矩阵。当 $\boldsymbol{X}=\boldsymbol{Y}$ 时，$\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})$ 即为 $D(\boldsymbol{X})$。当 $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})=0$ 时，称 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 不相关。如果 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 独立，则 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 不相关。反之不真。

若X_i与X_j的协方差存在$(i, j=1, \cdots, p)$，则可以计算X_i与X_j的相关系数

$$r_{ij}=\frac{\operatorname{cov}(X_i, X_j)}{\sqrt{D(X_i)D(X_j)}}$$

将这 $p\times p$ 个相关系数排列成一个方阵$\boldsymbol{R}=(r_{ij})_{p\times p}$，称为 $\boldsymbol{X}$ 的相关矩阵。

若记X_i的方差 $D(X_i)$为σ_{ii}，则我们称$\boldsymbol{V}^{1/2}=\operatorname{diag}(\sqrt{\sigma_{11}}, \cdots, \sqrt{\sigma_{pp}})$为标准差矩阵。协方差矩阵与相关矩阵有如下关系：

$$\boldsymbol{\Sigma}=\boldsymbol{V}^{1/2}\boldsymbol{R}\,\boldsymbol{V}^{1/2}\text{或 }\boldsymbol{R}=(\boldsymbol{V}^{1/2})^{-1}\boldsymbol{\Sigma}\,(\boldsymbol{V}^{1/2})^{-1}。$$

根据协方差矩阵的定义，可以验证其具有以下性质：

(1) 随机向量 $\boldsymbol{X}$ 的协方差矩阵是对称非负定矩阵；

(2) $\operatorname{cov}(\boldsymbol{AX}, \boldsymbol{BY})=\boldsymbol{A}\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{B}$，其中 $\boldsymbol{A}$，$\boldsymbol{B}$ 为大小适合矩阵运算的常数矩阵。

1.2　多元正态分布概述

1.2.1　多元正态分布的定义

在一元统计中，我们知道若 $X\sim N(0, 1)$，则 X 的任意线性变换为$Y=\sigma X+\mu\sim N(\mu, \sigma^2)$。利用这一性质，我们可以由标准正态分布来定义一般正态分布。事实上，我们将这种方式推广到多元情况，可以得到多元正态分布的一种定义。

设$X_1, \cdots, X_m$为 m 个相互独立标准正态变量，$\boldsymbol{X}=(X_1, \cdots, X_m)$为这 m 个随机变量构成的随机向量；设 $\boldsymbol{\mu}$ 为 p 维常数向量，$\boldsymbol{A}$ 为 $p\times m$ 维常数矩阵，则称 $\boldsymbol{Y}=\boldsymbol{AX}+\boldsymbol{\mu}$ 的分布为 p 元正态分布，或称 $\boldsymbol{Y}$ 为 p 维正态随机向量，记为$\boldsymbol{Y}\sim N_p(\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{A}')$。

大家知道一元正态随机变量的密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{\frac{-(x-\mu)^2}{2\sigma^2}}\quad(\sigma>0,\ -\infty<x<\infty)\tag{1.1}$$

我们可以将式(1.1)改写为

$$f(x)=\frac{1}{(2\pi)^{1/2}\left|\sigma^2\right|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'(\sigma^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

类似一元情况，我们给出 p 维正态分布的密度函数。

设 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，且 $\boldsymbol{\Sigma}$ 正定(为了保证$\boldsymbol{\Sigma}^{-1}$存在)，那么 $\boldsymbol{X}$ 的联合密度函数为

$$f(\boldsymbol{x})=\frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'(\boldsymbol{\Sigma})^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

例 1.1　设 $\boldsymbol{X}=\begin{bmatrix}X_1\\X_2\end{bmatrix}$服从二元正态分布，利用参数$\mu_1=E(X_1)$，$\mu_2=E(X_2)$，$\sigma_1=\sqrt{D(X_1)}$，$\sigma_2=\sqrt{D(X_2)}$，$\rho=\dfrac{\operatorname{cov}(X_1, X_2)}{\sigma_1\sigma_2}$来表示 $\boldsymbol{X}$ 的联合密度。

解：我们可以将协方差矩阵写作

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$$

从而其行列式为

$$|\boldsymbol{\Sigma}|=\sigma_1^2\sigma_2^2(1-\rho^2)$$

其逆矩阵为

$$\boldsymbol{\Sigma}^{-1}=\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{bmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{bmatrix}$$

将其带入密度公式中，可以得到 $\boldsymbol{X}$ 的联合密度为

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

从密度函数的表达式可以看出，此密度函数的最高点坐标是(μ_1,μ_2)。如果用与 xy 平面平行的平面去截二元正态密度函数曲面，所得截面为一个椭圆，称为概率密度等高线。

我们可以利用 SAS 系统绘制二维正态分布曲面的图形及等高线图，具体如图 1-1～图 1-6所示。

图 1-1 与图 1-3 给出不同方差大小的正态图形，可以看出：方差较大时，密度函数曲面较为平缓；而方差较小时，(x_1,x_2)的取值更加集中在均值附近。这一点可以通过对比图 1-2 和图 1-4 的相应等高线图看出。图 1-5 给出当x_1与x_2有较强的相关性时密度函数曲面较为陡立，从图 1-6 的等高线图可以看出强相关性的等高线比弱相关性的等高线的离心率要大。

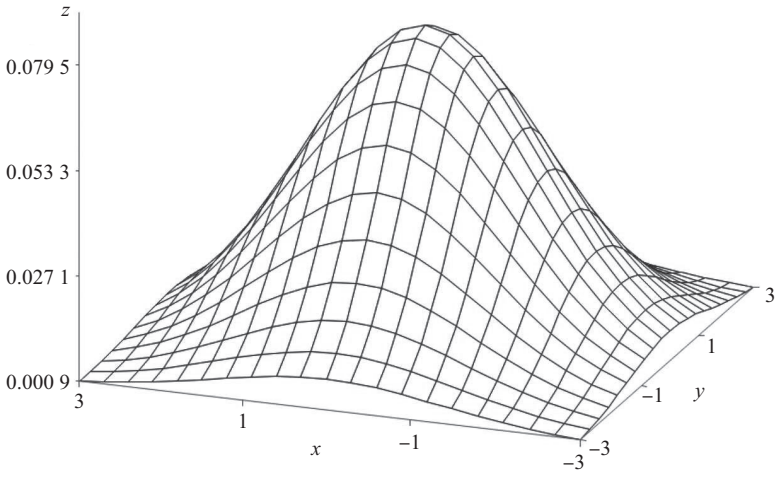


图 1-1　曲面图($\sigma_{11}^2=\sigma_{22}^2=2$，$\rho_{12}=0$)

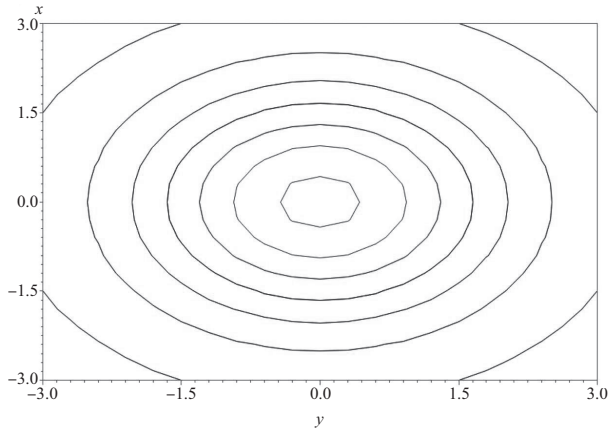


图 1-2　等高线图($\sigma_{11}^2=\sigma_{22}^2=2$，$\rho_{12}=0$)

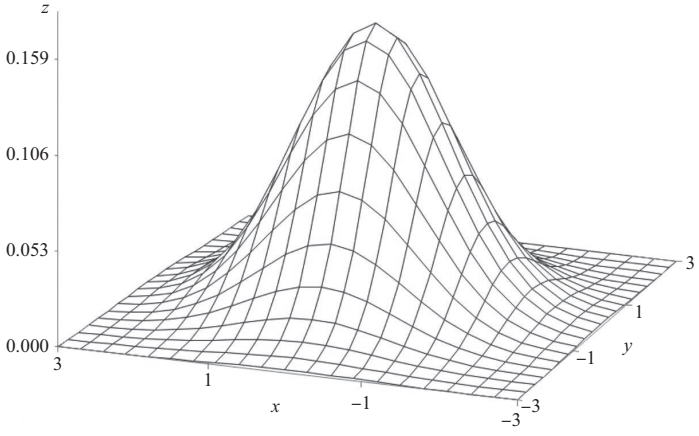


图 1-3　曲面图($\sigma_{11}^2=\sigma_{22}^2=1$，$\rho_{12}=0$)

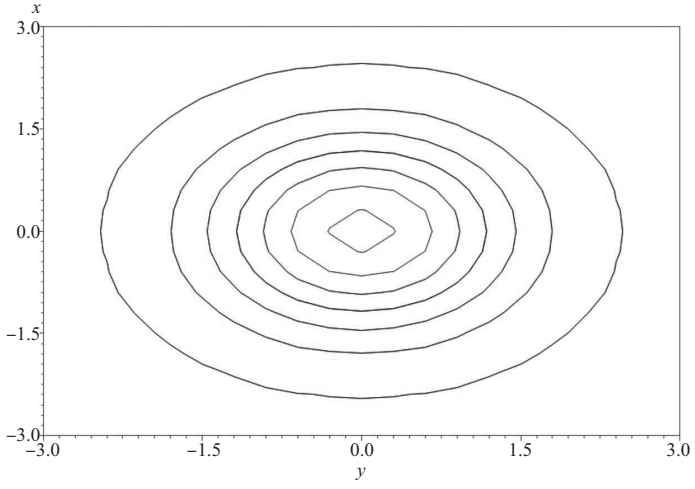


图 1-4　等高线图($\sigma_{11}^2=\sigma_{22}^2=1$，$\rho_{12}=0$)

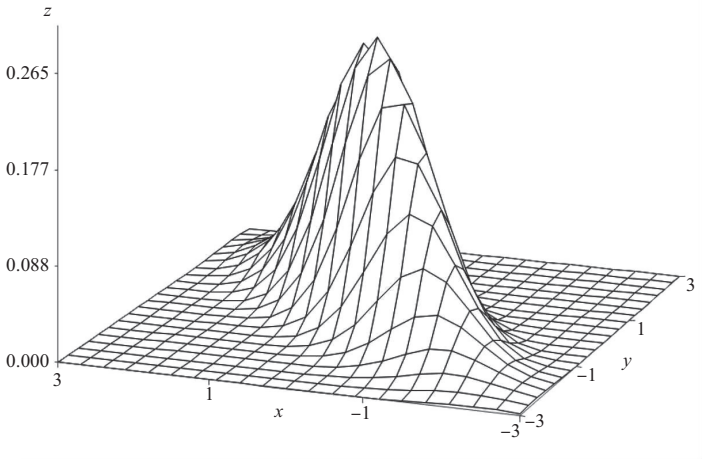


图 1-5　曲面图($\sigma_{11}^2=\sigma_{22}^2=1$，$\rho_{12}=0.8$)

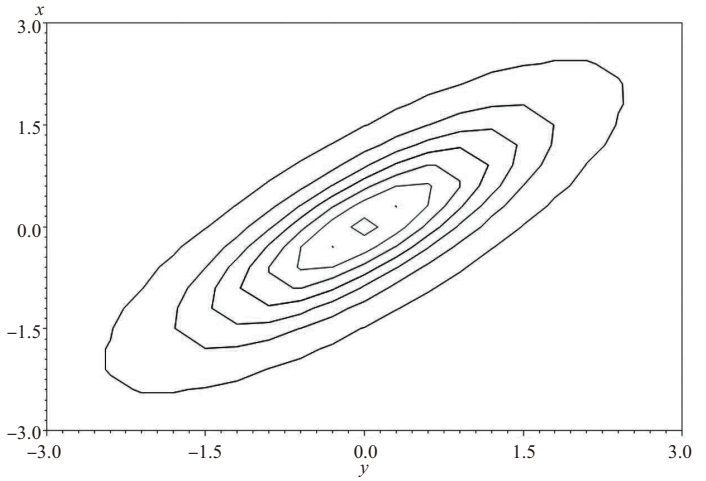


图 1-6　等高线图($\sigma_{11}^2=\sigma_{22}^2=1$，$\rho_{12}=0.8$)

1.2.2　多元正态变量的基本性质

多元正态分布在多元统计中占有十分重要的地位，许多重要理论与方法都建立在多元正态分布的性质之上。本节不加证明地给出多元正态变量的一些基本性质，以方便后文对正态分布及相关分布的处理。

设 $\boldsymbol{X}=(X_1,\cdots,X_p)'\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$。

(1) 若 $\boldsymbol{\Sigma}$ 为对角矩阵，则 $X_1,\cdots,X_p$ 相互独立。

(2) $\boldsymbol{X}$ 的任意边缘分布仍然为正态分布。特别地，如果将 $\boldsymbol{X}$，$\boldsymbol{\mu}$，$\boldsymbol{\Sigma}$ 做如下划分：

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{X}^{(1)}\\ \boldsymbol{X}^{(2)}\end{bmatrix}_{p-q}^{q}\qquad \boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}^{(1)}\\ \boldsymbol{\mu}^{(2)}\end{bmatrix}\qquad \boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}$$

其中，$\boldsymbol{X}^{(1)}$ 与 $\boldsymbol{\mu}^{(1)}$ 为 q 维向量，$\boldsymbol{X}^{(2)}$ 与 $\boldsymbol{\mu}^{(2)}$ 为 $p-q$ 维向量，$\boldsymbol{\Sigma}_{11}$ 为 $q\times q$ 维矩阵，$\boldsymbol{\Sigma}_{12}$ 为

$q\times(p-q)$ 维矩阵，$\boldsymbol{\Sigma}_{21}$ 为 $(p-q)\times q$ 维矩阵，$\boldsymbol{\Sigma}_{22}$ 为 $(p-q)\times(p-q)$ 维矩阵，则 $\boldsymbol{X}^{(1)}\sim N_q(\boldsymbol{\mu}^{(1)},\boldsymbol{\Sigma}_{11})$，$\boldsymbol{X}^{(2)}\sim N_{p-q}(\boldsymbol{\mu}^{(2)},\boldsymbol{\Sigma}_{22})$。顺便指出，$\boldsymbol{X}^{(1)}$ 与 $\boldsymbol{X}^{(2)}$ 相互独立，当且仅当 $\boldsymbol{\Sigma}_{12}$ 为零矩阵。

注意 如果一个随机向量的任意边缘分布都是正态分布，并不能推出它本身是多元正态分布。例如，考虑密度函数

$$f(x_1,x_2)=\frac{1}{2\pi}\mathrm{e}^{-\frac{1}{2}(x_1^2+x_2^2)}\left[1+x_1x_2\mathrm{e}^{-\frac{1}{2}(x_1^2+x_2^2)}\right]$$

所对应的随机向量 $(\boldsymbol{X}_1,\boldsymbol{X}_2)$。经过计算可以得出，$\boldsymbol{X}_1\sim N(0,1)$，$\boldsymbol{X}_2\sim N(0,1)$，但是它们的联合密度显然不是正态的。

(3) 设 $\boldsymbol{A}$ 是 $s\times p$ 阶常数矩阵，$\boldsymbol{d}$ 为 s 维常数向量，则 $\boldsymbol{AX}+\boldsymbol{d}$ 也服从正态分布，且 $\boldsymbol{AX}+\boldsymbol{d}\sim N_s(\boldsymbol{A\mu}+\boldsymbol{d},\boldsymbol{A\Sigma A}')$。

(4) 若 $\boldsymbol{\Sigma}$ 为正定矩阵，则 $(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\sim\chi^2(p)$。

1.3　多元正态分布的参数估计

在第 1.1.3 小节中我们给出了随机向量的数字特征。在实际应用中，均值向量和协方差矩阵等数字特征通常是未知的，需要利用样本来估计。本节考察多元正态总体的均值向量和协方差矩阵的估计，采用最常见的，也是具有较好性质的极大似然估计法给出其估计量，并给出极大似然估计法的性质。

1.3.1　多元样本的数字特征

考虑 p 元正态总体 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，设 $\boldsymbol{X}_{(1)},\cdots,\boldsymbol{X}_{(n)}$ 为来自这个 p 元正态总体的简单随机样本，其中 $\boldsymbol{X}_{(i)}=(x_{i1},\cdots,x_{ip})'(i=1,\cdots,n)$。

样本均值向量 $\bar{\boldsymbol{X}}$ 的定义为

$$\bar{\boldsymbol{X}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{(i)}=(\bar{x}_1,\cdots,\bar{x}_p)'=\frac{1}{n}\boldsymbol{X}'\boldsymbol{1}_n \tag{1.2}$$

在式(1.2)中，$\bar{x}_i=\frac{1}{n}\sum_{b=1}^{n}x_{bi}(i=1,\cdots,p)$，$\boldsymbol{1}_n$ 是一个 n 维的分量全为 1 的向量。

样本离差矩阵的定义为

$$\begin{aligned}\boldsymbol{A}&=\sum_{b=1}^{n}(\boldsymbol{X}_{(b)}-\bar{\boldsymbol{X}})(\boldsymbol{X}_{(b)}-\bar{\boldsymbol{X}})'\\&=\boldsymbol{X}'\boldsymbol{X}-n\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}'\\&=\boldsymbol{X}'\left[\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n'\right]\boldsymbol{X}\\&=(a_{ij})_{p\times p}\end{aligned} \tag{1.3}$$

在式(1.3)中，$a_{ij}=\sum_{b=1}^{n}(x_{bi}-\bar{x}_i)(x_{bj}-\bar{x}_j)(i,j=1,\cdots,p)$。

样本协方差阵的定义为

$$\boldsymbol{S}=\frac{1}{n-1}\boldsymbol{A}=(s_{ij})_{p\times p}\left(\text{或者}\boldsymbol{S}^{*}=\frac{1}{n}\boldsymbol{A}\right)$$

此时，$s_{ij}=\dfrac{1}{n-1}\sum_{b=1}^{n}(x_{bi}-\bar{x}_i)(x_{bj}-\bar{x}_j)$ $(i,j=1,\cdots,p)$。

样本相关矩阵的定义为

$$\boldsymbol{R}=(r_{ij})_{p\times p} \tag{1.4}$$

在式(1.4)中，$r_{ij}=\dfrac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}=\dfrac{a_{ij}}{\sqrt{a_{ii}}\sqrt{a_{jj}}}$ $(i,j=1,\cdots,p)$。

1.3.2 均值向量和协方差矩阵的极大似然估计

设$\boldsymbol{X}_{(1)},\cdots,\boldsymbol{X}_{(n)}$为来自$p$元正态总体$N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$的简单随机样本，利用极大似然法可以求出$\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$的参数估计分别为$\hat{\boldsymbol{\mu}}=\bar{\boldsymbol{X}}$，$\hat{\boldsymbol{\Sigma}}=\boldsymbol{S}^{*}$。

$\hat{\boldsymbol{\mu}}$和$\hat{\boldsymbol{\Sigma}}$具有如下基本性质：

$E\bar{\boldsymbol{X}}=\boldsymbol{\mu}$，即$\bar{\boldsymbol{X}}$是$\boldsymbol{\mu}$的无偏估计。但是$E\boldsymbol{S}^{*}=\dfrac{n-1}{n}\boldsymbol{\Sigma}$，因此$\boldsymbol{\Sigma}$的极大似然估计不是无偏估计。在上文的定义中，我们将$\boldsymbol{S}=\dfrac{1}{n-1}\boldsymbol{A}$定义为样本协方差矩阵，就是因为$\boldsymbol{S}$是$\boldsymbol{\Sigma}$的无偏估计。

可以证明$\bar{\boldsymbol{X}}$，$\boldsymbol{S}$是$\boldsymbol{\mu}$，$\boldsymbol{\Sigma}$的具有最小方差的无偏估计，也即$\bar{\boldsymbol{X}}$，$\boldsymbol{S}$是$\boldsymbol{\mu}$，$\boldsymbol{\Sigma}$的有效估计。此外，$\bar{\boldsymbol{X}}$，$\boldsymbol{S}$还是$\boldsymbol{\mu}$，$\boldsymbol{\Sigma}$的相合估计及充分统计量。关于如何利用极大似然法求得$\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$的参数估计，以及$\bar{\boldsymbol{X}}$，$\boldsymbol{S}$的统计性质的证明，有兴趣的读者可以阅读相关文献了解[①]。

样本均值向量和样本离差矩阵在正态总体下还有一些重要性质。

定理1.1 设$\bar{\boldsymbol{X}}$和$\boldsymbol{A}$分别为p元正态总体$N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$的样本均值向量和样本离差矩阵，则

(1) $\bar{\boldsymbol{X}}\sim N_p\left(\boldsymbol{\mu},\dfrac{1}{n}\boldsymbol{\Sigma}\right)$；

(2) 若设$\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_{n-1}$独立同$N_p(\boldsymbol{0},\boldsymbol{\Sigma})$分布，则$\boldsymbol{A}$与$\sum_{t=1}^{n-1}\boldsymbol{Z}_t\boldsymbol{Z}_t'$同分布；

(3) $\bar{\boldsymbol{X}}$与$\boldsymbol{A}$相互独立；

(4) $\boldsymbol{A}$为正定矩阵的充要条件是$n>p$。

注意 这时$\boldsymbol{A}$是随机矩阵，因此“$\boldsymbol{A}$为正定矩阵”这句话的含义实际上是“$\boldsymbol{A}$为正定矩阵”这个事件的概率为1。

1.4 常用分布与抽样分布

在数理统计中我们学习过，为了了解总体，我们对总体抽样得到样本，然后对样本进

① 高惠璇. 应用多元统计分析[M]. 北京：北京大学出版社，2005.

行加工，得到一个不包含未知量的样本的函数，这个样本函数我们一般称为统计量。在多元统计中也有类似的概念，比如我们前面介绍的样本均值向量 $\bar{\boldsymbol{X}}$ 和样本离差矩阵 $\boldsymbol{A}$ 等都是不含未知量的样本的函数，因此它们都是统计量。统计量的分布称为抽样分布。

在一元正态总体中，用于检验参数 μ，σ^2 的抽样分布有 χ^2 分布、t 分布及 F 分布。这些抽样分布推广到多元正态总体中，与之对应的分布为 Wishart 分布、Hotelling T^2 分布及 Wilks 分布。

1.4.1　Wishart 分布

如果从一元正态总体 $N(\mu, \sigma^2)$ 中抽取 n 个简单随机样本 $X_1, \cdots, X_n$，我们用样本方差

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})$$

来估计 σ^2，此时 $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}) \sim \chi^2(n-1)$。因此，可以得到 $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$。那么对 p 元正态总体，样本协方差矩阵 $\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{A}$ 又有怎样的分布呢？

这里先简要介绍一下如何定义随机矩阵的分布。设随机矩阵

$$\boldsymbol{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

将该矩阵的列向量（或者行向量）一个接一个地连接起来，组成一个长向量，这种操作一般称为将矩阵拉直为向量。这个拉直向量的分布就定义为随机矩阵 $\boldsymbol{X}$ 的分布。随机矩阵的分布还有其他不同的定义，本书中所指的随机矩阵的分布都是以拉直向量的分布来定义的。当 $\boldsymbol{X}$ 为对称矩阵时，只需要考虑下三角部分组成的长向量的分布，即 $(X_{11}, X_{21}, \cdots, X_{n1}, X_{22}, \cdots, X_{n2}, \cdots, X_{nn})$ 的分布。

定义1.7　设 $\boldsymbol{X}_{(b)} \sim N_p(\boldsymbol{\mu}_b, \boldsymbol{\Sigma})(b=1, \cdots, n)$ 是相互独立的 n 个 p 维正态变量，记 $\boldsymbol{X} = (\boldsymbol{X}_{(1)}, \cdots, \boldsymbol{X}_{(n)})'$ 为一个 $n \times p$ 矩阵，则称随机矩阵 $\boldsymbol{W} = \sum_{b=1}^{n}\boldsymbol{X}_{(b)}\boldsymbol{X}_{(b)}' = \boldsymbol{X}'\boldsymbol{X}$ 的分布为自由度为 n 的 p 维非中心 Wishart 分布，记为 $\boldsymbol{W} \sim W_p(n, \boldsymbol{\Sigma}, \Delta)$。其中，$\Delta$ 一般称为非中心参数，$\Delta = \sum_{b=1}^{n}\boldsymbol{\mu}_b\boldsymbol{\mu}_b'$。当 $\boldsymbol{\mu}_b = 0$ 时，我们一般称为中心 Wishart 分布，记为 $\boldsymbol{W} \sim W_p(n, \boldsymbol{\Sigma})$。

当 $p=1$，$\boldsymbol{\mu}_b = 0$ 时，$\boldsymbol{X}_{(b)} \sim N(0, \sigma^2)$，此时 $\boldsymbol{W} = W_1(n, \sigma^2) = \sum_{b=1}^{n}X_{(b)}^2 \sim \sigma^2\chi^2(n)$。也就是说，$W_1(n, 1)$ 就是 $\chi^2(n)$。因此，Wishart 分布是 χ^2 分布在多元正态情形下的推广。

下面我们不加证明地给出 Wishart 分布的几条性质。

（1）设 $\boldsymbol{X}_{(b)} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})(b=1, \cdots, n)$ 相互独立，则样本离差矩阵 $\boldsymbol{A}$ 服从 Wishart 分布，即

$$\boldsymbol{A} = \sum_{b=1}^{n}(\boldsymbol{X}_{(b)} - \bar{\boldsymbol{X}})(\boldsymbol{X}_{(b)} - \bar{\boldsymbol{X}})' \sim W_p(n-1, \boldsymbol{\Sigma})$$

(2) 设$\boldsymbol{W}_i \sim W_p(n_i, \boldsymbol{\Sigma})$ $(i=1, \cdots, k)$相互独立，若令$n=n_1+\cdots+n_k$，则有

$$\sum_{i=1}^{k} \boldsymbol{W}_i \sim W_p(n, \boldsymbol{\Sigma})$$

这个性质一般称为Wishart分布关于自由度n具有可加性，这点与χ^2分布类似。

(3) 设p阶随机矩阵$\boldsymbol{W} \sim W_p(n, \boldsymbol{\Sigma})$，$\boldsymbol{C}_{m\times p}$为常数矩阵，则

$$\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}' \sim W_m(n, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}')$$

特别地，如果取$\boldsymbol{C}$为向量$\boldsymbol{l}=(l_1, \cdots, l_p)'$，则有$\boldsymbol{l}'\boldsymbol{W}\boldsymbol{l} \sim W_1(n, \boldsymbol{l}'\boldsymbol{\Sigma}l)$，也即$\dfrac{\boldsymbol{l}'\boldsymbol{W}\boldsymbol{l}}{\boldsymbol{l}'\boldsymbol{\Sigma}l} \sim \chi^2(n)$。

1.4.2 Hotelling T^2分布

在一元统计中我们学过，若$X \sim N(0, 1)$，$Y \sim \chi^2(n)$，且X与Y独立，则随机变量$t=\dfrac{X}{\sqrt{Y/n}}$服从自由度为n的t分布，也称为学生分布。我们还学过，如果将t平方，就得到

$$t^2=\frac{nX^2}{Y} \sim F(1, n)$$

即$t^2(n)$服从第一自由度为1、第二自由度为n的中心F分布。下面仿照一元情形将t^2的分布推广到p元总体的情形。

定义1.8 设$\boldsymbol{W} \sim W_p(n, \boldsymbol{\Sigma})$，$\boldsymbol{X} \sim N_p(0, \boldsymbol{\Sigma})$，$n \geqslant p$，$\boldsymbol{\Sigma}>0$，且$\boldsymbol{W}$与$\boldsymbol{X}$相互独立，则称随机变量$T^2=n\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X}$所服从的分布为第一自由度为$p$，第二自由度为$n$的Hotelling T^2分布，记为

$$T^2 \sim T^2(p, n)$$

注意 我们可以证明T^2分布只与n，p有关，与$\boldsymbol{\Sigma}$无关，因此在表示T^2分布的记号中没有$\boldsymbol{\Sigma}$。

T^2分布与F分布也有一定的关系。在一元统计中，如果$t=\dfrac{X}{\sqrt{Y/n}} \sim t(n)$，则有$t^2=\dfrac{X^2}{Y/n} \sim F(1, n)$。推广到$p$元情形，这个关系是$\dfrac{n-p+1}{pn}T^2(p, n)=F(p, n-p+1)$。这一点的证明及更多相关性质的介绍可以参看相关文献[①]。下面我们不加证明，给出T^2分布的两条重要性质。这两条性质在多元正态总体的假设检验中将会用到。

(1) 设$\boldsymbol{X}_{(b)}$$(b=1, \cdots, n)$是从$p$维正态总体$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$中抽取的$n$个随机样本，$\bar{\boldsymbol{X}}$为样本均值向量，$\boldsymbol{A}$为样本离差矩阵，则统计量

$$\begin{aligned} T^2 &= (n-1)\left[\sqrt{n}\,(\bar{\boldsymbol{X}}-\boldsymbol{\mu})\right]'\boldsymbol{A}^{-1}\left[\sqrt{n}\,(\bar{\boldsymbol{X}}-\boldsymbol{\mu})\right] \\ &= n(n-1)(\bar{\boldsymbol{X}}-\boldsymbol{\mu})'\boldsymbol{A}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}) \sim T^2(p, n-1) \end{aligned}$$

(2) 设有两个p维正态总体$N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$，$N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$，从这两个总体中抽出容量分别为$n_1$和$n_2$的两个样本。记$\bar{\boldsymbol{X}}_1$，$\bar{\boldsymbol{X}}_2$为两样本的均值向量，$\boldsymbol{S}_1$，$\boldsymbol{S}_2$为两样本协方差矩阵，并记

① 高惠璇. 应用多元统计分析[M]. 北京：北京大学出版社，2005.

$$\boldsymbol{S}_p=\frac{n_1\boldsymbol{S}_1+n_2\boldsymbol{S}_2}{n_1+n_2-2}$$

若$\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$，则

$$\frac{n_1n_2}{n_1+n_2}(\bar{\boldsymbol{X}}_1-\bar{\boldsymbol{X}}_2)'\boldsymbol{S}_p^{-1}(\bar{\boldsymbol{X}}_1-\bar{\boldsymbol{X}}_2)\sim T^2(p,\ n_1+n_2-2)$$

1.4.3　Wilks Λ 分布

我们在数理统计中学过，若 $X\sim\chi^2(m)$，$Y\sim\chi^2(n)$，且 X 与 Y 相互独立，则

$$F=\frac{X/m}{Y/n}\sim F(m,\ n)$$

在一元统计中 F 分布主要用来做方差齐性检验，两个总体的样本方差的比在原假设下是服从 F 分布的。在多元总体中，样本的协方差矩阵是一个矩阵，不能再简单相除得到统计量了。因此，我们考虑用与协方差矩阵有关的一个量来描述总体的离散程度(或者称为变异度)。这样的参数我们一般称为广义方差。用哪些数量指标来描述广义方差呢？一般而言，多用矩阵的行列式、迹或者特征值来描述。目前最常用的是利用行列式定义的。有了广义方差的定义，再仿照 F 分布的定义，称两个广义方差之比的统计量为 Wilks Λ 统计量。

定义1.9　设 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$，则称协方差矩阵的行列式$|\boldsymbol{\Sigma}|$为 $\boldsymbol{X}$ 的广义方差。再设 $\boldsymbol{A}_1\sim W_p(n_1,\ \boldsymbol{\Sigma})$，$\boldsymbol{A}_2\sim W_p(n_2,\ \boldsymbol{\Sigma})$ $(\boldsymbol{\Sigma}>0,\ n_1\geqslant p)$，且$\boldsymbol{A}_1$与$\boldsymbol{A}_2$独立，则称

$$\Lambda=\frac{|\boldsymbol{A}_1|}{|\boldsymbol{A}_1+\boldsymbol{A}_2|}$$

为 Wilks 统计量或 Λ 统计量，其所遵从的分布称为 Wilks 分布，记为 $\Lambda\sim\Lambda(p,\ n_1,\ n_2)$。

Wilks 分布比较复杂，在不同的情形下许多学者对其精确分布及近似分布都进行了深入的研究。当 p 或者n_2比较小而$n_1>p$ 时，可以通过 F 分布得到 Λ 统计量的精确分布，具体情况如表 1-1 所示。

表 1-1　$\Lambda\sim\Lambda(p,\ n_1,\ n_2)$与 F 分布的关系$(n_1>p)$

p	n_2	统计量 F	F 的自由度
任意	1	$\frac{(n_1-p+1)}{p}\cdot\frac{(1-\Lambda)}{\Lambda}$	$p,\ n_1-p+1$
任意	2	$\frac{(n_1-p)}{p}\cdot\frac{(1-\sqrt{\Lambda})}{\sqrt{\Lambda}}$	$2p,\ 2(n_1-p)$
1	任意	$\frac{(1-\Lambda)}{\Lambda}\cdot\frac{n_1}{n_2}$	$n_2,\ n_1$
2	任意	$\frac{(1-\sqrt{\Lambda})}{\sqrt{\Lambda}}\cdot\frac{(n_1-1)}{n_2}$	$2n_2,\ 2(n_1-1)$

当$n_2>2$，$p>2$ 时，我们有这样的近似分布：

$$\text{当}\ n_1\to\infty,\ -\left[n_1-\frac{1}{2}(p-n_2+1)\right]\ln\Lambda\sim\chi^2(p\ n_2)。$$

此外，类似于 F 分布中 $F(n, m)$ 与 $\frac{1}{F(m, n)}$ 同分布，Λ 分布也有一个类似的性质：若 $n_2 < p$，则 $\Lambda(p, n_1, n_2) = \Lambda(n_2, p, n_1 + n_2 - p)$。

【课后练习】

1. 设 (X_1, X_2, X_3) 的联合密度为

$$f(x_1, x_2, x_3) = \begin{cases} \dfrac{1 - \sin x \sin y \sin z}{8\pi^2} & 0 \leqslant x \leqslant 2\pi,\ 0 \leqslant y \leqslant 2\pi,\ 0 \leqslant z \leqslant 2\pi \\ 0 & \text{其他} \end{cases}$$

(1) 求 X_1 的边缘密度。

(2) 求 (X_1, X_2) 的边缘密度。

(3) 试证明 X_1, X_2, X_3 两两独立但不互相独立。

2. 设

$$\boldsymbol{A} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} \end{bmatrix}$$

(1) 试证明 $\boldsymbol{A}$ 是一个正交矩阵(即 $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}_3$)。

(2) 已知 $X \sim N_3(\mu \boldsymbol{I}_3, \sigma^2 \boldsymbol{I}_3)$，设 $\boldsymbol{Y} = (Y_1, Y_2, Y_3)' = \boldsymbol{A}\boldsymbol{X}$，试证明

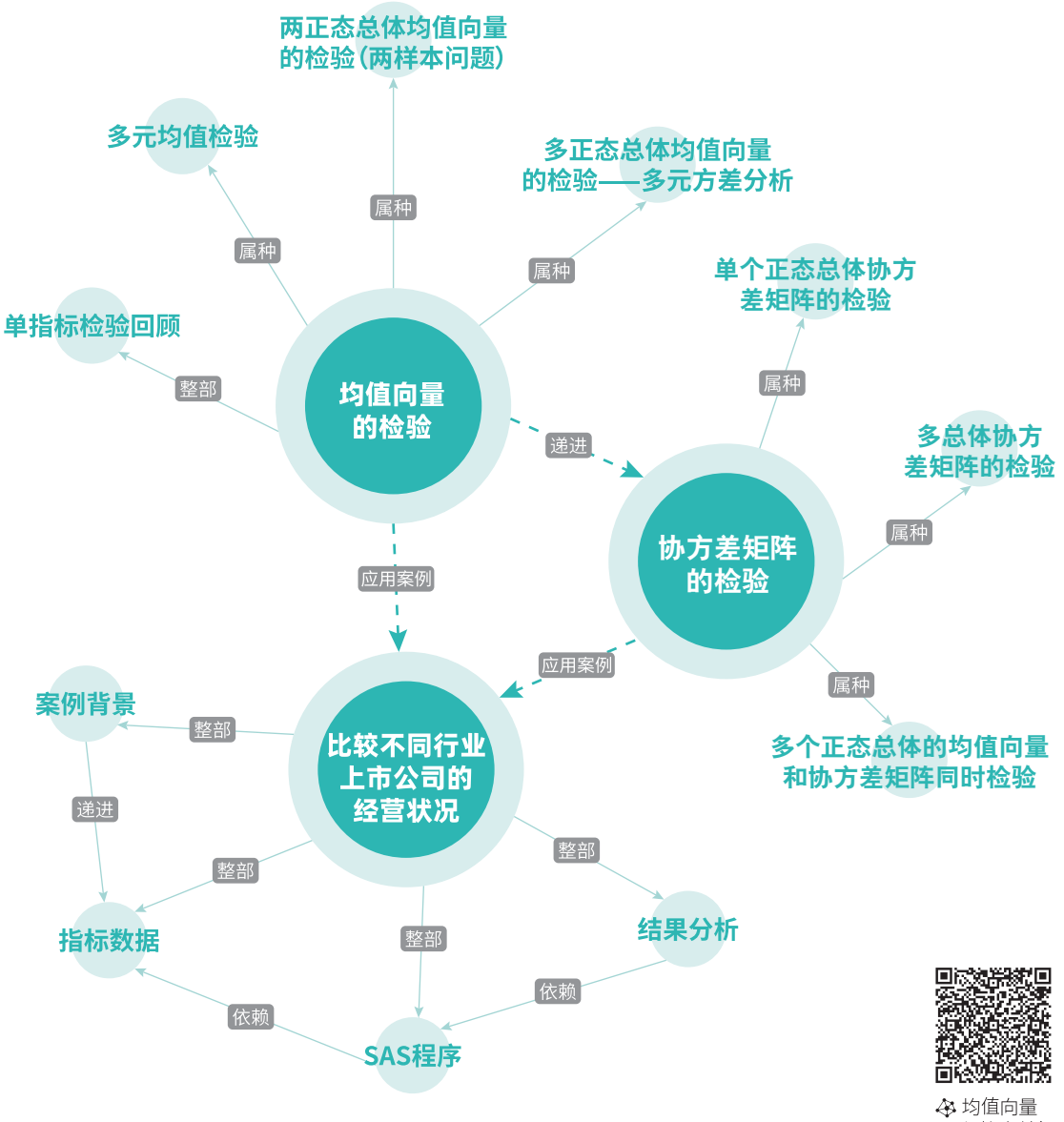
① $Y_1^2 + Y_2^2 + Y_3^2 = \sum_{i=1}^{3} (X_i - \overline{X})^2$，其中 $\overline{X} = \frac{1}{3}(X_1 + X_2 + X_3)$；

② $Y_1 \sim N(\sqrt{3}\mu, \sigma^2)$，$Y_2, Y_3 \sim N(0, \sigma^2)$；

③ Y_1, Y_2, Y_3 相互独立。

第2章

均值向量与协方差矩阵的检验



均值向量与协方差矩阵的检验

在一元统计中，我们已经学习过关于正态总体 $N(\mu,\sigma^2)$ 的均值与方差的检验，了解到常用的检验方法有 μ 检验、t 检验、F 检验和 χ^2 检验等。在实际应用中，对某一客观事物的考察往往需要多个指标。例如，为了考察某企业的生产经营状况，需要综合考察其资本结构、盈利能力、成长能力等多个维度。对于多指标的正态总体 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，有些实际问题需要对 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 进行统计推断。本章会介绍 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 在不同情形下的假设检验。虽然本章中的方法多是单指标情形的直接推广，但是由于多指标问题的复杂性，本章只重点介绍检验统计量的形式，以及如何利用这些统计量做检验，对于这些检验问题的理论推证全部省略，有兴趣的读者可以参看相关的参考文献。本章最后还将介绍有关检验的 SAS 上机实现方法。

2.1 均值向量的检验

2.1.1 单指标检验回顾

我们先回顾一下假设检验的基本步骤：

(1) 根据问题提出待检验的统计假设 H_0 和 H_1。

(2) 选取一个合适的统计量并得出它的抽样分布。

(3) 给定显著性水平 α，通过统计量的抽样分布确定临界值，进而得到拒绝域，建立判别准则。

(4) 根据样本观测值计算统计量的值，看是否落在拒绝域中，从而对假设检验做出统计判断，并给出具体的解释。

下面我们具体回顾一下单指标均值检验是怎么做的。假定从总体 $N(\mu,\sigma^2)$ 中抽出样本 $x_1, x_2, \cdots, x_n$，要做如下假设检验：

$$H_0:\mu=\mu_0 \quad H_1:\mu\neq\mu_0$$

当 σ^2 已知时，检验统计量为

$$\mu=\sqrt{n}\,\frac{(\bar{x}-\mu_0)}{\sigma} \tag{2.1}$$

在式(2.1)中，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 为样本均值。如果原假设成立，则 μ 服从标准正态分布，进一步得到拒绝域为 $|\mu|>z_{\alpha/2}$，$z_{\alpha/2}$ 为标准正态分布的上 $\alpha/2$ 分位数。我们也可以选用

$$\mu^2=n(\bar{x}-\mu_0)'(\sigma^2)^{-1}(\bar{x}-\mu_0)$$

作为检验统计量，由于在原假设下 μ 服从标准正态分布，因此 μ^2 服从自由度为1的 χ^2 分布，从而拒绝域为 $\mu^2>\chi_1^2(\alpha/2)$。

当 σ^2 未知时，先用

$$S^2=\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n-1}$$

作为 σ^2 的估计，然后用统计量

$$t=\sqrt{n}\,\frac{(\bar{x}-\mu_0)}{S}$$

作为检验统计量。如果原假设成立，则 t 服从自由度为 $n-1$ 的 t 分布，拒绝域为 $|t|>t_{n-1}(\alpha/2)$，$t_{n-1}(\alpha/2)$ 为 t_{n-1} 的上 $\alpha/2$ 分位数。类似于 σ^2 已知的情形，我们也可以选用

$$t^2=n\,(\bar{x}-\mu_0)'\,(S^2)^{-1}(\bar{x}-\mu_0)$$

作为检验统计量。当原假设为真时，t^2 服从第一自由度为 1、第二自由度为 $n-1$ 的 F 分布，简记为 $t^2\sim F_{1,n-1}$，拒绝域为 $t^2>F_{1,n-1}(\alpha)$，$F_{1,n-1}(\alpha)$ 为 $F_{1,n-1}$ 的上 $\alpha/2$ 分位数。

2.1.2　多元均值检验

设总体 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，随机样本为 $\boldsymbol{X}_{(\alpha)}(\alpha=1,\cdots,n)$，我们要检验这组随机样本的均值是否为某一个指定的向量 $\boldsymbol{\mu}_0$，即

$$H_0:\boldsymbol{\mu}=\boldsymbol{\mu}_0$$

$$H_1:\boldsymbol{\mu}\neq\boldsymbol{\mu}_0$$

检验方法与单指标均值检验类似，也根据协方差矩阵是否已知分两种情况讨论。

1. 协方差矩阵 $\boldsymbol{\Sigma}$ 已知

类似于一元的形式，我们采用的检验统计量为

$$\chi_0^2=n\,(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)'\,\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)$$

我们可以证明当原假设为真时，χ_0^2 服从自由度为 p 的 χ^2 分布。直观上来看，χ_0^2 越大，意味着样本均值 $\bar{\boldsymbol{X}}$ 与 $\boldsymbol{\mu}_0$ 的差距越大，因此拒绝域应取 χ_0^2 较大的部分。

当给定显著性水平 α 后，根据样本计算出 χ_0^2 的值，然后查 χ^2 分布表得到临界值 $\chi_p^2(\alpha)$ 满足

$$P\left(\chi_0^2>\chi_p^2(\alpha)\right)=\alpha$$

拒绝域即为 $\{\chi_0^2>\chi_p^2(\alpha)\}$。

利用统计计算软件(如 SAS)还可以计算出显著性概率值(也称为 p 值)。在许多统计计算软件中，主要是利用 p 值来给出假设检验的结果。因此下面要介绍 p 值的概念，并具体指出如何利用 p 值给出检验结果。

由于在原假设为真时，$\chi_0^2\sim\chi^2(p)$，而由样本值可计算出 χ_0^2 的值，记为 d。我们称概率值

$$p=P(\chi_0^2\geqslant d)$$

为显著性概率值，或简称为 p 值。

在这个例子中，p 值的直观意义可以这样理解：检验统计量 χ_0^2 反映了 $\bar{\boldsymbol{X}}$ 与 $\boldsymbol{\mu}_0$ 的偏差大小，而 χ_0^2 较大时，我们倾向于拒绝原假设。现在根据观测数据我们计算得到 χ_0^2 的值是 d，而如果原假设成立，可以根据 χ_0^2 的分布计算出 $p=P(\chi_0^2\geqslant d)$ 的值。如果 p 值很小，比如 $p=0.02<\alpha=0.05$，则说明相对于 $\chi^2(p)$ 分布，根据观测数据计算得到 χ_0^2 的值 d 是偏大的，这也就意味着在 $\alpha=0.05$ 的显著性水平下有足够的证据拒绝原假设，即认为 $\boldsymbol{\mu}$ 与 $\boldsymbol{\mu}_0$ 有显著差异；如果 p 值较大，比如 $p=0.24\geqslant\alpha=0.05$，则说明相对于 $\chi^2(p)$ 分布，根据观测数据计算得到 χ_0^2 的值 d 并不大，这也就意味着在 $\alpha=0.05$ 的显著性水平下没有足够的证据拒

绝原假设，即认为 $\boldsymbol{\mu}$ 与$\boldsymbol{\mu}_0$ 没有显著差异。

2. 协方差矩阵 $\boldsymbol{\Sigma}$ 未知

由于协方差矩阵未知，我们需要先估计协方差矩阵。因为$\hat{\boldsymbol{\Sigma}}^{-1}=\dfrac{1}{n-1}\boldsymbol{A}$ 是$\boldsymbol{\Sigma}$ 的无偏估计，因此类似于单指标检验时的思路，我们先考察

$$\begin{aligned}T^2&=n(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)'\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)\\&=n(n-1)(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)'\boldsymbol{A}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)\end{aligned}$$

的分布。根据第1.4.2 小节 Hotelling T^2 分布的性质(1)

$$T^2\sim T^2(p,n-1)$$

再利用T^2 分布与 F 分布的性质(参见第1.4.2 小节)，我们将检验统计量取为

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

在原假设下，$F\sim F(p,n-p)$，拒绝域为$\{F>F_\alpha(p,n-p)\}$，$F_\alpha(p,n-p)$ 为分布 $F(p,n-p)$ 的上 α 分位数。

也许有读者会有疑问，既然多元正态随机向量的每一个分量都是一元正态随机变量，那么对这 p 个分量分别做均值的假设检验问题与本节介绍的对均值向量直接做假设检验问题有什么区别呢？由于 p 个分量之间往往有相互依赖的关系，分开做假设检验会忽略这种依赖信息。因此，分开做假设检验会由于信息缺失而不准确。在实际工作中，我们往往会将一元检验与多元检验联合使用。当实际工作要求做全面检查时，我们多考虑使用多元检验；当实际工作要求着重检查某个指标或者要求分析各指标之间的关系与差异时，我们多考虑使用一元检验。

2.1.3 两正态总体均值向量的检验

在许多实际问题中，除了第2.1.2 小节中介绍的纵向比较，有时也需要横向比较。比如，两所大学的新生录取成绩是否有明显差异；不同行业之间的工资水平是否有明显差异；不同地区之间的物价水平是否有明显差异。这些问题都可以归结为检验两个总体的均值向量是否相等的问题。我们一般称这种问题为两样本问题。两样本问题又可以根据协方差矩阵是否相等分为两种情形。

1. 两样本协方差矩阵相等(但未知)时均值向量的检验

设$\boldsymbol{X}_{(\alpha)}(\alpha=1,\cdots,n_1)$ 为来自 p 元正态总体$N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$ 的容量为n_1 的样本，$\boldsymbol{Y}_{(\alpha)}(\alpha=1,\cdots,n_2)$ 为来自 p 元正态总体$N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$ 的容量为n_2 的样本，两样本相互独立，$n_1>p$，$n_2>p$，$\boldsymbol{\Sigma}$ 未知。需要进行的假设检验是

$$H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$$

$$H_1:\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$$

与单总体均值检验类似，我们采用的检验统计量形式为

$$T^2=\frac{1}{1/n_1+1/n_2}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})'\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}) \tag{2.2}$$

在式(2.2)中，$\bar{\boldsymbol{X}}=\frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{X}_i$，$\bar{\boldsymbol{Y}}=\frac{1}{n_1}\sum_{i=1}^{n_2}\boldsymbol{Y}_i$。协方差矩阵$\boldsymbol{\Sigma}$的估计采用$\hat{\boldsymbol{\Sigma}}=\frac{\boldsymbol{A}_x+\boldsymbol{A}_y}{n_1+n_2-2}$，其中，$\boldsymbol{A}_x$与$\boldsymbol{A}_y$分别是两个总体的样本离差矩阵。

当原假设成立时，利用第1.4.1小节 Wishart 分布的性质(1)与T^2统计量的定义可以得到

$$T^2 \sim T^2_{p,\,n_1+n_2-2}$$

再利用T^2与F分布的关系(参看第1.4.2小节)我们可以得到

$$F^* = \frac{n_1+n_2-p-1}{(n_1+n_2-2)\,p}T^2 \sim F_{p,\,n_1+n_2-p-1}$$

我们取F^*作为检验统计量。如果F^*的值较大，意味着T^2的值较大，也就说明两个总体的距离较远，倾向于拒绝原假设，因此拒绝域取为F^*的值较大的区域，即当给定显著性水平α后，若$F^* > F_{p,\,n_1+n_2-p-1}(\alpha)$，则拒绝原假设，否则没有充分的理由拒绝原假设。

2. 协方差矩阵不相等的情形

假设从两个正态总体$N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$和$N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$中分别抽取容量为$n_1$和$n_2$的两个样本。当这两个样本协方差不相等时，并没有统一的处理办法。下面介绍两种较为简单的情形。

如果$n_1=n_2$，我们可以采取成对数据处理的技巧。令

$$\boldsymbol{Z}_{(i)}=\boldsymbol{X}_{(i)}-\boldsymbol{Y}_{(i)}$$

这样可以将两样本均值检验问题化为单样本均值检验问题，即我们做假设检验

$$H_0:\boldsymbol{\mu}_Z=\boldsymbol{0} \qquad H_1:\boldsymbol{\mu}_Z\neq\boldsymbol{0}$$

然后对$\boldsymbol{Z}$采用上一节的方法进行假设检验。

如果$\boldsymbol{\Sigma}_1$和$\boldsymbol{\Sigma}_2$相差很大时，我们考虑利用T^2统计量的近似分布来构造检验统计量。T^2统计量的形式是

$$T^2=(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})'\left[\frac{\boldsymbol{A}_x}{n_1(n_1-1)}+\frac{\boldsymbol{A}_y}{n_2(n_2-1)}\right]^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}) \tag{2.3}$$

在式(2.3)中，$\bar{\boldsymbol{X}}$，$\bar{\boldsymbol{Y}}$，$\boldsymbol{A}_x$，$\boldsymbol{A}_y$的含义与前文相同。记

$$\boldsymbol{S}_*=\frac{\boldsymbol{A}_x}{n_1(n_1-1)}+\frac{\boldsymbol{A}_y}{n_2(n_2-1)}$$

再令

$$f^{-1}=(n_1^3-n_1^2)^{-1}\left[(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})'\boldsymbol{S}_*^{-1}\left(\frac{\boldsymbol{A}_x}{n_1-1}\right)\boldsymbol{S}_*^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})\right]^2 T^{-4}$$
$$+(n_2^3-n_2^2)^{-1}\left[(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})'\boldsymbol{S}_*^{-1}\left(\frac{\boldsymbol{A}_y}{n_2-1}\right)\boldsymbol{S}_*^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})\right]^2 T^{-4}$$

可以证明：当原假设为真时，$\left(\frac{f-p+1}{fp}\right)T^2$近似服从$F_{p,\,f-p+1}$分布，进而可以做假设检验。

2.1.4　多正态总体均值向量的检验 —— 多元方差分析

在许多实际问题中，我们要研究的总体往往不止两个。例如要研究不同地区物价水平

时，如果将一个地区看作一个总体，此时要研究的总体可以多达几十甚至上百个，此时就需要运用多元方差分析的知识。为了更好地理解多元方差分析，我们先回顾一元方差分析的相关知识。

假定有 r 个正态总体 $N(\boldsymbol{\mu}_1, \sigma^2)$，…，$N(\boldsymbol{\mu}_r, \sigma^2)$，从各个正态总体中抽取样本如下：

$$\begin{aligned}
&\boldsymbol{X}_1^{(1)}, \cdots, \boldsymbol{X}_{n_1}^{(1)} \sim N(\boldsymbol{\mu}_1, \sigma^2)\\
&\boldsymbol{X}_1^{(2)}, \cdots, \boldsymbol{X}_{n_2}^{(2)} \sim N(\boldsymbol{\mu}_2, \sigma^2)\\
&\vdots\\
&\boldsymbol{X}_1^{(r)}, \cdots, \boldsymbol{X}_{n_r}^{(r)} \sim N(\boldsymbol{\mu}_r, \sigma^2)
\end{aligned}$$

在方差分析的问题中，我们假定 r 个正态总体的方差相等。需要检验的假设是

$$\begin{aligned}
&H_0: \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_r\\
&H_1: \text{存在 } i \neq j, \text{使得} \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j
\end{aligned}$$

为了构造检验统计量，我们先定义以下平方和

$$\text{总偏差平方和 SST} = \sum_{i=1}^{r}\sum_{j=1}^{n_i} (\boldsymbol{X}_j^{(i)} - \bar{\boldsymbol{X}})^2$$

$$\text{组内偏差平方和 SSE} = \sum_{i=1}^{r}\sum_{j=1}^{n_i} (\boldsymbol{X}_j^{(i)} - \bar{\boldsymbol{X}}_i)^2$$

$$\text{组间偏差平方和 SSA} = \sum_{i=1}^{r} n_i (\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})^2$$

其中，$\bar{\boldsymbol{X}}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i} \boldsymbol{X}_j^{(i)}$ 是第 i 组的样本均值，$\bar{\boldsymbol{X}} = \dfrac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} \boldsymbol{X}_j^{(i)}$ 是总均值，总样本量 $n = n_1 + \cdots + n_r$。此时，通过代数运算我们得知如下平方和分解公式成立：

$$\text{SST} = \text{SSE} + \text{SSA}$$

从直观上考察，如果原假设成立，在总偏差平方和 SST 不变的条件下，组间偏差平方和相对于组内偏差平方和应该偏小，因此检验统计量取为

$$F = \frac{\text{SSA}/(r-1)}{\text{SSE}/(n-r)}$$

拒绝域为 $\{F > F_\alpha\}$，其中的 F_α 通过 $P(F > F_\alpha) = \alpha$ 确定。

我们将上述方法推广到 r 个 p 元正态总体 $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$，…，$N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma})$，从这 r 个总体中抽取的独立样本为

$$\begin{aligned}
&\boldsymbol{X}_1^{(1)}, \cdots, \boldsymbol{X}_{n_1}^{(1)} \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\\
&\boldsymbol{X}_1^{(2)}, \cdots, \boldsymbol{X}_{n_2}^{(2)} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})\\
&\vdots\\
&\boldsymbol{X}_1^{(r)}, \cdots, \boldsymbol{X}_{n_r}^{(r)} \sim N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma})
\end{aligned}$$

总样本数 $n = n_1 + n_2 + \cdots + n_r$。需要检验的假设是

$$H_0: \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_r \qquad H_1: \text{存在 } i \neq j, \text{使得} \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$$

前文所叙述的 3 个平方和现在成为矩阵的形式：

$$\text{总离差矩阵 } \boldsymbol{T} = \sum_{i=1}^{r}\sum_{j=1}^{n_i} (\boldsymbol{X}_j^{(i)} - \bar{\boldsymbol{X}})^2$$

$$\text{组内离差矩阵 } \boldsymbol{A}=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\boldsymbol{X}_j^{(i)}-\bar{\boldsymbol{X}}_i)^2$$

$$\text{组间离差矩阵 } \boldsymbol{B}=\sum_{i=1}^{r}n_i(\bar{\boldsymbol{X}}_i-\bar{\boldsymbol{X}})^2$$

这三者之间仍然有 $\boldsymbol{T}=\boldsymbol{A}+\boldsymbol{B}$ 成立。

由于 $\boldsymbol{T}$，$\boldsymbol{A}$，$\boldsymbol{B}$ 三者都是矩阵，我们采用第1.4.3小节所用的广义方差来度量矩阵大小。类似一元情形，我们取检验统计量为

$$\Lambda=\frac{|\boldsymbol{A}|}{|\boldsymbol{A}+\boldsymbol{B}|}$$

可以证明在原假设成立的情况下，Λ 的分布就是Wilks Λ 分布，即 $\Lambda\sim\Lambda(p,n-r,r-1)$。

注意› 此处，分母是组内离差矩阵的行列式，因此拒绝域为 $\{\Lambda<\lambda_\alpha\}$，其中，$\lambda_\alpha$ 通过 $P(\Lambda<\lambda_\alpha)=\alpha$ 确定。

由于Wilks Λ 分布本身很复杂，我们也可以采用第1.4.3小节的方法用 χ^2 分布或 F 分布来近似。具体在哪些情形下如何近似已经在第1.4.3小节中总结，这里不再赘述。

2.2　协方差矩阵的检验

前一节讨论了多元正态分布均值的检验问题，这一类问题主要是考察不同总体的平均水平是否有不同。本节主要考查不同总体平均水平波动大小的问题。协方差矩阵可以反映波动程度大小，因此这类问题可以转化为协方差矩阵的检验问题。

2.2.1　单个正态总体协方差矩阵的检验

假设 $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n$ 是来自 p 元正态总体 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的一个样本，$\boldsymbol{\Sigma}_0$ 是已知的给定矩阵，且 $\boldsymbol{\Sigma}_0>0$。考虑假设检验问题：

$$H_0:\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_0$$
$$H_1:\boldsymbol{\Sigma}\neq\boldsymbol{\Sigma}_0$$

我们用的检验统计量是

$$\lambda=\exp\left[\operatorname{tr}\left(-\frac{1}{2}\boldsymbol{A}\boldsymbol{\Sigma}_0^{-1}\right)\right]\left|\boldsymbol{A}\boldsymbol{\Sigma}_0^{-1}\right|\left(\frac{\mathrm{e}}{n}\right)^{np/2}$$

这个检验统计量的抽样分布很难得到，通常我们采用 λ 的相关近似分布来得到拒绝域。当样本容量 n 很大时，如果原假设成立，那么 $-2\ln\lambda$ 的极限分布是 $\chi^2\left[\frac{p(p+1)}{2}\right]$。同时，如果给定检验水平 α，当样本容量 n 很大时，我们可以由样本值计算出 λ 的值。当 $-2\ln\lambda>\chi_\alpha^2\left[\frac{p(p+1)}{2}\right]$，即 $\lambda<\exp\left(-\frac{\chi_\alpha^2}{2}\right)$ 时，拒绝 H_0。

2.2.2　多总体协方差矩阵的检验

与均值检验类似，在实际应用中也有横向比较多个总体的协方差矩阵是否相同的需

求。因此,本节我们考虑多总体协方差矩阵检验的问题。

与均值向量的检验类似,假定 r 个 p 元正态总体 $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \cdots, N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$,从这 r 个总体中抽取的独立样本为

$$\begin{aligned}
\boldsymbol{X}_1^{(1)}, \cdots, \boldsymbol{X}_{n_1}^{(1)} &\sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\
\boldsymbol{X}_1^{(2)}, \cdots, \boldsymbol{X}_{n_2}^{(2)} &\sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \\
&\vdots \\
\boldsymbol{X}_1^{(r)}, \cdots, \boldsymbol{X}_{n_r}^{(r)} &\sim N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)
\end{aligned}$$

总样本数 $n = n_1 + n_2 + \cdots + n_r$。需要检验的假设是

$$H_0: \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_r$$

$$H_1: \text{存在 } i \neq j, \text{ 使得} \boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j$$

我们所采用的检验统计量是

$$M = (n-r)\ln\left|\frac{\boldsymbol{A}}{n-r}\right| - \sum_{t=1}^{r}(n_t-1)\ln\left|\frac{\boldsymbol{A}_t}{n_t-1}\right|$$

其中,$\boldsymbol{A}_t = \sum_{i=1}^{n_t}(\boldsymbol{X}_i^{(t)} - \bar{\boldsymbol{X}}_t)(\boldsymbol{X}_i^{(t)} - \bar{\boldsymbol{X}}_t)'$ 为第 t 个总体的组内样本离差矩阵,$\bar{\boldsymbol{X}}_t = \frac{1}{n_t}\sum_{i=1}^{n_t}\boldsymbol{X}_i^{(t)}$,$t = 1, 2, \cdots, r$ 为第 t 个总体的样本均值,$\boldsymbol{A} = \sum_{t=1}^{r}\boldsymbol{A}_t$ 为 r 个总体的样本离差矩阵的和。

当样本容量 n 很大时,如果原假设 H_0 为真,M 的近似分布为

$$(1-d)M \sim \chi^2(f) \tag{2.4}$$

在式(2.4)中,$f = \frac{1}{2}p(p+1)(r-1)$

$$d = \begin{cases} \dfrac{2p^2+3p-1}{6(p+1)(r-1)}\left[\sum\limits_{i=1}^{r}\dfrac{1}{(n_i-1)} - \dfrac{1}{(n-r)}\right] & \text{当 } n_i \text{ 不全相等} \\ \dfrac{(2p^2+3p-1)(r-1)}{6(p+1)(n-r)} & \text{当 } n_i \text{ 全部相等} \end{cases}$$

2.2.3 多个正态总体的均值向量和协方差矩阵同时检验

本小节我们考虑介绍一种稍微复杂但是比较常见的情形,即同时检验 r 个总体的均值和协方差是否相同。具体说来,问题背景与前两小节类似。假定有 r 个 p 元正态总体 $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \cdots, N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$,$\boldsymbol{X}_i^{(t)}$ 为来自第 t 个总体的随机样本($t = 1, \cdots, r$; $i = 1, \cdots, n_t$)。检验问题为

$$H_0: \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)} = \cdots = \boldsymbol{\mu}^{(r)}, \text{ 且} \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_r$$

$$H_1: \boldsymbol{\mu}^{(i)} (i = 1, \cdots, r) \text{ 与 } \boldsymbol{\Sigma}_i (i = 1, \cdots, r) \text{ 至少有一组不全相等}$$

我们所采用的统计量与上一节略有不同,形式为

$$M^* = (n-r)\ln\left|\frac{\boldsymbol{T}}{(n-r)^p}\right| - \sum_{t=1}^{r}(n_t-1)\ln\left|\frac{\boldsymbol{A}_t}{(n_t-1)^p}\right|$$

$\boldsymbol{A}_t$ 与上一节相同,仍为第 t 个总体的样本离差矩阵,$\boldsymbol{T} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\boldsymbol{X}_j^{(i)} - \bar{\boldsymbol{X}})^2$ 为总离差矩阵

（参见2.1.4 节）。当样本容量 n 很大时，在原假设成立时有这样的近似分布：

$$(1-b)M^{*}\sim\chi^{2}(f)$$

其中，

$$f=\frac{1}{2}p(p+3)(k-1)$$

$$b=\left(\sum_{i=1}^{r}\frac{1}{n_i-1}-\frac{1}{n-r}\right)\left[\frac{2p^{2}+3p-1}{6(p+3)(r-1)}\right]-\frac{p-r+2}{(n-r)(p+3)}$$

2.3　SAS 实现与应用案例

1. 案例背景

视频

SAS 实现与应用案例

全面评价上市公司的经营状况有很多方法，本例采用《国有资本金效绩评价规则》中竞争性工商企业的评价指标体系，即考察上市公司 8 大基本指标：净资产收益率、总资产报酬率、总资产周转率、流动资产周转率、资产负债率、已获利息倍数、销售增长率和资本积累率。表 2-1 的数据来自 3 个行业 35 家上市公司2018年年报，均以合并会计报表为依据计算得到。净资产收益率与资产负债率直接取自会计年报，其余各指标计算公式如下：

$$\text{总资产报酬率}=\frac{\text{利润总额}+\text{利息支出}}{(\text{年初总资产}+\text{年末总资产})/2}\times100\%$$

$$\text{总资产周转率}=\frac{\text{主营业务收入}}{(\text{年初总资产}+\text{年末总资产})/2}$$

$$\text{流动资产周转率}=\frac{\text{主营业务收入}}{(\text{年初流动资产}+\text{年末流动资产})/2}$$

$$\text{已获利息倍数}=\frac{\text{利润总额}+\text{利息支出}}{\text{利息支出}}$$

$$\text{销售增长率}=\frac{\text{本年主营业务收入}-\text{上年主营业务收入}}{\text{上年主营业务收入}}\times100\%$$

$$\text{资本积累率}=\frac{\text{年末所有者权益}-\text{年初所有者权益}}{\text{年初所有者权益}}\times100\%$$

如果将不同的行业看作不同的总体，那么上述 35 家上市公司的数据就可以认为来自 3 个总体，下面我们尝试对这 3 个不同行业上市公司的经营状况进行比较。

2. 指标数据

35 家上市公司 2018 年的年报数据如表 2-1 所示。

表 2-1　35 家上市公司2018年年报数据

行业	公司简称	股票代码	净资产收益率	总资产报酬率	资产负债率	总资产周转率	流动资产周转率	已获利息倍数	销售增长率	资本积累率
交通运输业	盐田港	000 088	7.31	5.79	24.82	0.04	0.35	124.59	17.38	9.82
	五洲交通	600 368	12.43	7.44	66.68	0.16	0.51	2.52	0.76	−0.85
	山东高速	600 350	10.38	8.07	57.26	0.11	0.74	5.98	−15.09	0.16
	东莞控股	000 828	17.27	12.55	40.72	0.15	0.77	−42.05	11.87	14.83
	中原高速	600 020	4.06	5.38	77.64	0.11	0.98	1.65	−7.79	−24.20
	宁沪高速	600 377	17.71	13.31	39.05	0.20	1.70	15.72	0.20	13.36
	深高速	600 548	22.85	13.16	52.46	0.03	0.21	5.31	1.51	23.75
	申通地铁	600 834	2.08	3.02	45.94	0.28	1.75	2.02	−0.03	0.82
	铁龙物流	600 125	9.29	8.02	40.60	1.70	3.08	16.20	33.85	7.71
	广深铁路	601 333	2.72	3.16	18.60	0.54	3.03	38.29	8.84	0.56
	中远海能	600 026	0.37	2.68	53.84	0.20	1.71	1.37	27.13	3.56
	宁波海运	600 798	7.05	8.28	39.86	0.34	2.63	5.16	20.29	18.86
	天津港	600 717	3.81	4.86	38.63	0.34	1.45	5.45	−7.89	0.86
信息传输、软件和信息技术服务业	中国联通	600 050	2.86	2.19	41.50	0.06	0.43	88.64	−17.03	3.23
	天威视讯	002 238	7.17	4.75	26.40	0.38	0.86	−10.96	−4.94	1.64
	科大讯飞	002 230	6.94	4.48	46.34	0.55	1.05	−38.53	45.54	3.26
	富瀚微	300 613	5.52	2.30	13.04	0.36	0.44	−0.99	−8.28	9.25
	同花顺	300 033	20.23	14.66	19.15	0.33	0.38	−8.56	−1.62	5.39
	汇纳科技	300 609	13.36	12.30	13.59	0.43	0.53	−29.46	22.67	11.71
	远光软件	002 063	9.10	7.15	13.05	0.49	0.67	−20.52	8.47	11.79
	长亮科技	300 348	4.50	2.28	31.96	0.56	1.03	5.36	23.63	4.97
	久远银海	002 777	12.51	9.05	10.59	0.53	0.63	−23.11	24.78	92.85
	东华软件	002 065	8.86	6.13	42.08	0.56	0.66	9.35	16.66	2.77
	四维图新	002 405	6.96	5.01	19.67	0.22	0.53	−15.59	−1.30	7.69
	华东电脑	600 850	13.63	6.09	59.97	1.22	1.25	−52.37	10.70	12.24
电气机械和器材制造业	格力电器	000 651	33.36	13.01	63.10	0.73	0.92	−31.98	29.05	38.68
	美的集团	000 333	25.66	10.78	64.94	0.94	1.37	15.14	7.83	11.49
	飞科电器	603 868	34.46	31.89	29.58	1.14	1.60	−72.16	3.10	8.02
	亿纬锂能	300 014	17.07	7.99	63.10	0.50	1.16	7.49	31.13	15.79
	欧普照明	603 515	22.67	15.08	40.85	1.17	1.53	−76.71	15.60	19.35
	九阳股份	002 242	20.70	14.32	42.50	1.35	1.92	−79.70	12.22	6.91
	杭电股份	603 618	4.75	4.05	59.64	0.83	1.19	2.50	4.96	10.49
	海信家电	000 921	19.79	7.37	63.86	1.51	2.25	46.21	7.76	11.32
	白云电器	603 861	7.45	4.84	50.92	0.52	0.83	9.29	17.77	−7.56
	特变电工	600 089	6.38	3.81	57.90	0.44	0.81	5.93	4.42	14.59

3. SAS 程序

```
proc univariate data=lizi normal;
 var jzcsyl zzcbcl zcfzl zzczzl ldzczzl yhlxbs xszzl zbjll;
 by a;
```

```
run;
proc GLM;
class a;
model jzcsyl zcfzl ldzczzl xszzl=a;
means a/bon;
run;
proc GLM;
class a;
model jzcsyl zcfzl ldzczzl xszzl=a;
means a/hovtest=bartlett;
run;
proc GLM;
class a;
model jzcsyl zcfzl ldzczzl xszzl=a;
means a/hovtest=levene;
run;
```

4. SAS 程序说明与输出说明

本部分内容省略了数据步的输入。第一部分是对各个变量进行正态性检验。“proc univariate”表示采用 univariate 过程步。var 语句是指定分析变量，by 语句指定了以 a 变量即行业变量进行分组。

前两节所介绍的假设检验方法都是基于正态性假设，因此我们需要先对各数据是否服从多元正态分布进行检验。遗憾的是，在常见的软件中往往只能进行单个变量的正态性检验，多元正态检验的实现较为困难。在实际工作中，往往通过对每个变量的正态性检验来对整体向量的分布做出判断。一般情况下，如果数据量较大且没有明显的证据表明所得数据不服从多元正态分布时，我们认为数据就来自多元正态总体。在本例中，表 2-2 汇总了对每一个变量进行正态性检验的结果。SAS 系统一共给出了 4 种正态性检验的统计量，它们分别是 Shapiro-Wilk 统计量、Kolmogorov-Smirnov 统计量、Cramer-von Mises 统计量和 Anderson-Darling 统计量。由于在本例中样本量比较小，3 个行业的样本量分别为$n_1=13$，$n_2=12$ 和$n_3=10$，因此我们选择 Shapiro-Wilk 统计量。

表 2-2　每个变量的正态性检验结果

指标		Shapiro-Wilk	
		统计量	*P* 值
净资产收益率	第一行业	0.931 809	0.359 8
	第二行业	0.927 556	0.354 9
	第三行业	0.928 107	0.429 5
总资产报酬率	第一行业	0.900 933	0.137 7
	第二行业	0.890 505	0.119 6
	第三行业	0.812 352	0.020 5
资产负债率	第一行业	0.963 277	0.803 3
	第二行业	0.899 233	0.155 0
	第三行业	0.854 927	0.066 5

（续表）

指标		Shapiro-Wilk	
		统计量	*P* 值
总资产周转率	第一行业	0.600 361	＜0.000 1
	第二行业	0.835 126	0.024 2
	第三行业	0.942 637	0.582 7
流动资产周转率	第一行业	0.914 545	0.211 6
	第二行业	0.907 388	0.197 5
	第三行业	0.938 619	0.537 7
已获利息倍数	第一行业	0.689 874	0.000 4
	第二行业	0.821 970	0.016 8
	第三行业	0.867 204	0.092 7
销售增长率	第一行业	0.961 605	0.778 2
	第二行业	0.966 308	0.868 6
	第三行业	0.870 140	0.100 3
资本积累率	第一行业	0.906 605	0.165 2
	第二行业	0.465 201	＜0.000 1
	第三行业	0.891 881	0.178 0

从 *P* 值可以看出，3 个行业均可以认为符合正态分布的指标是净资产收益率、资产负债率、流动资产周转率和销售增长率。因此，我们后面将只对这 4 个指标进行分析比较。这 4 个指标分别反映了企业的获利能力、资本结构、流动资产利用效率及企业的发展态势，可以认为是对企业经营状况的一个比较全面的反映。

接下来进入均值检验的步骤。“proc GLM”表示采用 GLM 过程步。class 语句是分类语句，model 语句是用来规定因素对结果的效应，这里是考察不同的分类下 4 个指标是否相同。means 语句表示希望得到不同分类的均值情况，选项“bon”表示进行两两比较的 Bonferroni 检验。运行程序后得到的结果很丰富。我们将部分结果汇总在表 2-3 中。

表 2-3　方差分析表

源	分析变量	平方和	均方	自由度	*F* 统计量	*P* 值
模型	净资产收益率	724.885 7	362.442 8	2	6.32	0.004 9
	资产负债率	3 862.369 7	1 931.184 9	2	8.52	0.001 1
	流动资产周转率	4.001 2	2.000 6	2	4.42	0.020 2
	销售增长率	230.400 0	115.200 0	2	0.54	0.589 5
误差	净资产收益率	1 834.616 2	57.331 8	32		
	资产负债率	7 257.204 0	226.787 6	32		
	流动资产周转率	14.489 6	0.452 8	32		
	销售增长率	6 861.860 3	214.433 1	32		

（续表）

源	分析变量	平方和	均方	自由度	*F* 统计量	*P* 值
校正合计	净资产收益率	2 559.501 8		34		
	资产负债率	11 119.573 7		34		
	流动资产周转率	18.490 8		34		
	销售增长率	7 092.260 3		34		

表 2-3 给出了 4 个指标的方差来源，包括模型(行业)、误差及校正的总的方差来源，还给出了自由度、均方、*F* 统计量的值及 *P* 值。从 *P* 值来看，3 个行业的净资产收益率、资产负债率及流动资产周转率都有显著差别，而销售增长率没有显著差别。那么这 3 个行业的 3 个指标究竟有怎样的差别呢？在“结果”窗口的“均值”选项卡下，可以查看到每个指标中 3 个行业两两比较是否有显著差异。我们将一些重要结果汇总在表 2-4 中。

表 2-4　3 个行业两两比较的结果

比较		分析变量			
		净资产收益率	资产负债率	流动资产周转率	销售增长率
3-2	均值间差值	9.926	7.785	0.653 0	3.444
	下限	1.735	−8.218	−0.074 9	−12.397
	上限	18.117	23.788	1.380 9	19.285
	是否显著	是	否	否	否
3-1	均值间差值	10.204	25.527	−0.096 6	6.382
	下限	2.158	9.237	−0.811 7	−9.180
	上限	18.250	41.818	0.618 5	21.943
	是否显著	是	是	否	否
2-1	均值间差值	0.278	−17.742	−0.749 6	2.938
	下限	−7.380	−32.973	−1.430 2	−11.872
	上限	7.936	−2.511	−0.069 1	17.748
	是否显著	否	是	是	否

表 2-4 中上、下限指的是 95%置信区间的上、下限，而是否显著指的是在 95%置信水平下两个行业相比是否有显著差异。因此，当这个置信区间包含 0 时，应认为两个行业的均值相比没有显著差异；反之则应认为有显著差异。以第三行业(电气机械和器材制造业)与第二行业(信息传输、软件和信息技术服务业)比较为例，两者的资产负债率、流动资产周转率与销售增长率均没有显著差异，而制造业的净资产收益率要高于信息服务业，似乎说明在2018年制造业的盈利能力要强于信息服务业，这从数据方面可能也说明了2018年信息服务业遭遇了寒冬。

将上文程序中 means 语句修改为“means a/hovtest=bartlett;”可以进行方差齐性的检验，得到的结果如表 2-5 所示。

表 2-5 Bartlett 方差齐性检验

净资产收益率	源	自由度	卡方	P>卡方
	a	2	5.780 2	0.055 6
资产负债率	源	自由度	卡方	P>卡方
	a	2	0.854 1	0.652 4
流动资产周转率	源	自由度	卡方	P>卡方
	a	2	15.713 5	0.000 4
销售增长率	源	自由度	卡方	P>卡方
	a	2	2.788 2	0.248 1

从表 2-5 可以看出，在0.05置信水平下，可以认为 3 个行业的净资产收益率、资产负债率及销售增长率的方差是相等的，但是流动资产周转率的方差在 3 个行业间不相等。再将 means 语句改为“means a/hovtest=levene;”可以得到如表 2-6 所示的结果。

表 2-6 Levene 方差齐性检验组均值的平方离差 ANOVA

净资产收益率	源	自由度	平方和	均方	F 值	P>F
	a	2	35 487.3	17 743.7	4.38	0.020 9
	误差	32	129 720	4 053.8		
资产负债率	源	自由度	平方和	均方	F 值	P> F
	a	2	75 556.4	37 778.2	0.53	0.595 3
	误差	32	2 292 983	71 655.7		
流动资产周转率	源	自由度	平方和	均方	F 值	P> F
	a	2	4.802 0	2.401 0	7.28	0.002 5
	误差	32	10.560 5	0.330 0		
销售增长率	源	自由度	平方和	均方	F 值	P> F
	a	2	207 308	103 654	1.55	0.228 0
	误差	32	2 141 412	66 919.1		

表 2-6 说明在0.05置信水平下，资产负债率与销售增长率的误差平方在 3 个行业间没有显著差异，而净资产收益率和流动资产周转率的误差平方在 3 个行业间有显著差异。结合表 2-5，这似乎说明，除了行业因素，还有别的因素对净资产收益率有影响。

5. 结果分析

从表 2-7 中可以看出，3 个行业中，电气机械和器材制造业表现稍好于交通运输业与信息传输、软件和信息技术服务业。原因可能在于信息技术服务业前几年发展迅猛，进入企业过多，导致企业所能获得的平均资本不足，造成了不良局面，以致整个行业不景气，获利能力不足。而表 2-7 中所列举的上市制造业企业都是具有成熟运营能力的企业，其获利能力与成长能力都比较稳定，整体运营能力更强。

表 2-7　3 个行业各个指标描述统计量的估计

行业	样本量	净资产收益率		资产负债率		流动资产周转率		销售增长率	
		均值	标准差	均值	标准差	均值	标准差	均值	标准差
1	13	9.026	6.890	45.854	15.956	1.455	0.983	7.002	14.520
2	12	9.303	4.858	28.112	16.132	0.705	0.281	9.940	17.647
3	10	19.229	10.569	53.639	12.199	1.358	0.476	13.384	10.030

【课后练习】

一、简答题

1. P 值是什么？试以两样本协方差矩阵相等(但未知)时均值向量的检验为例说明如何利用 P 值做假设检验。

2. 试谈 Wilks 统计量在多元方差分析中的重要意义。

二、上机分析题

1. 人均 GDP、三产比重、工业生产者出厂价格指数、人均可支配收入和人口增长这 5 个指标从不同侧面可以较好地说明一个地区的经济社会发展状况。数据 EXE2_1 选取内蒙古、广西、贵州、云南、西藏、甘肃、青海、宁夏和新疆 9 个边远省份的相关指标，试比较这些省份的社会经济发展状况与全国平均水平有无显著差异(假定这 5 个指标服从五元正态分布)。

2. 数据 EXE2_2 中选取某两个地区各 6 个单位的 5 项财务评价指标(分别记为 X_1，X_2，X_3，X_4 和 X_5)。在评分结果表中，序号 1～6 为第一个地区单位代号，7～12 为第二个地区单位代号。试比较两个地区基层单位的财务状况是否有差异(假定这两个地区的 5 个指标都服从正态分布且协方差矩阵相等)。