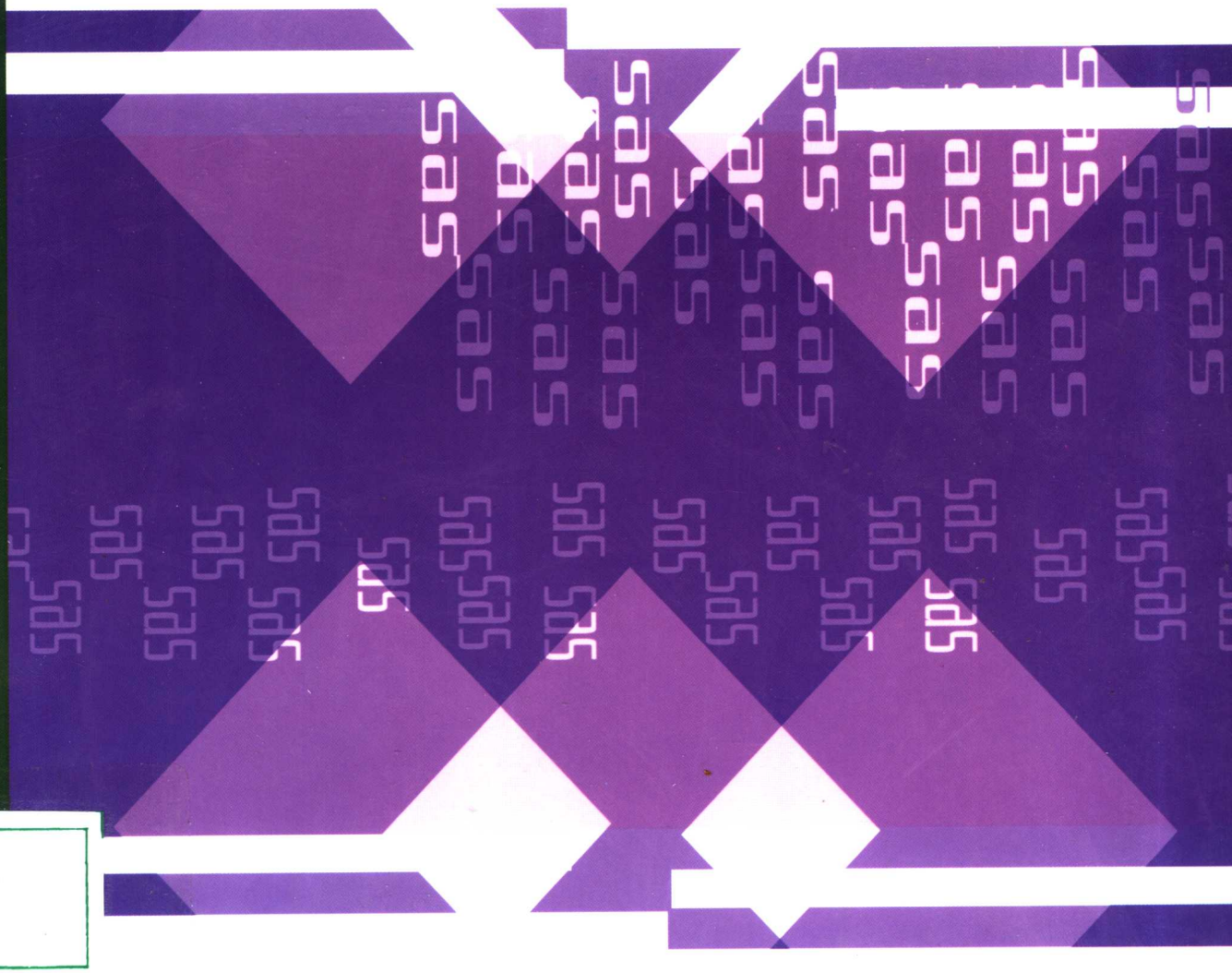
重复测量资料分析方法与SAS程序

余松林 向惠云 编著



科学出版社
www.sciencep.com

重复测量资料分析方法与SAS程序

华中科技大学同济医学院　余松林
俄亥俄州立大学医学与公共卫生学院　向惠云
编著

科学出版社
北京

Analysis of Repeated Measures Data with SAS Procedures

Songlin Yu
School of Public Health, Tongji Medical College
Huazhong University of Science and Technology
The People's Republic of China

Huiyun Xiang
School of Medicine and Public Health
Ohio State University
The United States of America

Science Press
Beijing

内 容 简 介

重复测量是指对同一观察对象的同一观察指标在不同时间或环境下进行的多次测量，用于分析观察指标的变化趋势及有关的影响因素。重复测量资料在不同科学领域内都很常见。本书对这类资料的分析方法做了系统介绍。内容包括单组和多组重复测量资料的方差分析方法，带有协变量的重复测量资料的分析方法，分类反应变量的重复测量资料的分析方法，以及复发事件的生存分析方法等先进统计学分析技术。

本书以读者为中心。在写作方法上由浅入深，学以致用。对每一种分析方法都有理论介绍、示例引导和结果评价。由于分析重复测量资料的计算量一般比较大，对每一种方法都系统地介绍了 SAS(Statistical Analysis System)程序。读者用自己的资料，在示例的引导下，运行 SAS 程序，即可得到所需要的结果。

本书的读者主要是自然科学与社会科学领域的研究人员、统计学工作者，也可作为大专院校统计学硕士和博士研究生的教材。

图书在版编目(CIP)数据

重复测量资料分析方法与 SAS 程序/余松林，向惠云编著. —北京：科学出版社，2004.3

ISBN 7-03-012985-7

Ⅰ.重… Ⅱ.①余…②向… Ⅲ.①重复性－统计资料－分析方法－研究②统计分析－应用软件，SAS－程序设计 Ⅳ.C81

中国版本图书馆 CIP 数据核字（2004）第 014292 号

责任编辑：李国红 / 责任校对：包志虹

责任印制：刘士平 / 封面设计：卢秋红

科学出版社 出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2004 年 3 月第 一 版 开本：787×1092 1/16

2004 年 3 月第一次印刷 印张：16 1/2

印数：1－2 000 字数：389 000

定价：48.00 元

（如有印装质量问题，我社负责调换〈环伟〉）

前　言

重复测量是指对同一事物或受试对象的同一观察指标进行的多次测量。重复测量资料在各个科学领域内都很常见，例如，在临床医学研究中，对患高血压的病人，需定期监测其血压水平，以分析该病人的病情发展情况；在教育学研究中，观察在不同教育环境下学生的学习进步情况；在社会学研究中研究人群的就业和失业情况等等。由于这类资料的统计学特性与通常的独立测量资料不同，需要有特殊的统计学方法进行分析。对重复测量资料的分析方法的研究是当前统计学中的一个热点领域。自20世纪90年代以来，在统计学上取得了很多成果，国际上发表了许多重要论著。但由于受到统计工具的限制，使得对重复测量资料的统计学分析方法的应用尚不够普及。在我国，有关这一领域的研究还处于早期阶段。本书的目的是向我国广大科技工作者介绍这一领域的主要研究成果，希望在现有统计软件支持下推广这一统计学分析方法的应用，并促进这一学科领域在我国的发展。

本书内容包括四个部分：第一部分介绍重复测量资料的统计学性质（第一章）；第二部分介绍反应变量为连续型测量值的重复测量资料分析方法（第二章至第九章）；第三部分介绍反应变量为离散型观察值的重复测量资料分析方法（第十章和第十一章）；最后的第四部分（第十二章）介绍重复事件或多终点事件的生存分析方法。本书采用国际通用的统计软件SAS（Statistical Analysis System）作为统计计算工具，以实现统计运算的全过程。在写作过程中用例子进行说明，采用理论解释与SAS程序相结合、同时并进的方法，以期取得更为直观的阅读和理解效果。尚无大众软件支持的方法，未纳入本书的内容。

本书是一本有关重复测量资料分析方法的统计学专著，是科学研究、特别是医学和生物学研究必备的统计学参考书，本书也适于作研究生的教材。

在本书的写作过程中，得到了华中科技大学同济医学院和公共卫生学院各级领导的关心和支持，得到了董时富教授、王增珍教授、张家放教授、熊光炼教授和流行病学与卫生统计学系的其他老师和同事们的鼓励和帮助，特致以谢意。特别要感谢方亚教授和王家春教授，他们仔细地阅读了部分书稿，提出了宝贵的修改意见，并给予了技术方面的大力支持和帮助。还有很多没有在这里提到名字的朋友和同事，都对本书的编写给予了热情的关怀和鼓励，特别是我的亲人，为本书的问世，一直默默奉献，谨向他们致以深深的谢意。

还要特别感谢Dr. Hervey W. Gimbel和他所领导的USA-China Health Project组织，特别感谢陈国庆博士。在他们的帮助下，使我有机会再次到美国并参加Loma Linda大学和Wake Forest大学医学院的教学和科学研究工作。在那里有机会接触并学习有关统计学方面的前沿科学技术理论。没有这次访美的机会，就不会有本书的问世。

由于作者水平有限，缺点和错误在所难免。不妥之处，敬请批评指正。

余松林

2003年9月30日于武汉

目　录

第一章　重复测量资料的特点

重复测量(repeated measure)是指对同一观察对象(受试者、病人、动物、植物、机器等)的同一观察指标在不同时间点上进行的多次测量。这类资料在几乎所有科学研究领域内都可见到。例如在生物学中,为了解某种生物生长发育规律对其生长过程进行追踪观察所得到的资料。在临床医学研究中,为了解病人的病情变化对有关病情监测指标进行连续的监测所得到的资料。在环境科学中,对同一个地区在不同时间点上的污染物所测量的资料。在毒物致畸试验中,将某种化合物给予受孕雌鼠以观察每一胎中胎鼠的畸胎数或死亡数,也属于重复测量资料。因为同一胎的胎鼠对毒物的反应素质,较不同胎的胎鼠对毒物的反应素质更具有一致性。对某些疾病的家庭聚集性研究中,同一家系的成员间对该种疾病较不同家系的成员间具有比较相近的素质。按重复测量设计所收集的资料用于分析该观察指标在不同时间上的变化特点。

一些传统的统计方法,如 t 检验、方差分析、线性回归模型等,都要求各次观察是相互独立的。而重复测量资料由于是对同一受试者的某项观察指标进行的多次测量,在同一受试者的多次测量之间可能存在某种相关性,用通常的统计方法就不能充分揭示出其内在的特点,有时甚至会得出错误的结论。因此,有关重复测量资料的分析方法是近代统计学研究的热点之一。

第一节　重复测量数据结构与独立数据结构的区别及其优缺点

为了解重复测量数据与独立观察数据之间的区别,下面用一个完全随机设计的独立数据结构与具有一个受试者内因素(时间)的重复测量数据结构进行比较。

完全随机设计的独立数据结构:从正常人、可疑硅沉着病者及一期硅沉着病病人中各随机抽取 5 人,测量他们的血清黏蛋白含量(mg/L),结果列于表 1.1 中第一部分。

重复测量的数据结构:对 5 名粉尘作业工人的血清黏蛋白含量(mg/L)连续 3 年的测量结果列于表 1.1 中第二部分。

表 1.1　独立数据结构与重复测量数据结构比较

第一部分:完全随机设计的独立数据结构						第二部分:重复测量的数据结构			
血清黏蛋白含量(mg/L)						受试者号	血清黏蛋白含量(mg/L)		
正常人		可疑硅沉着病者		一期硅沉着病病人			上岗前测量值	第一年测量值	第二年测量值
受试者号	测量值	受试者号	测量值	受试者号	测量值				
1	64.2	1	94.8	1	69.6	1	69.6	78.9	85.9
2	42.8	2	70.4	2	69.7	2	48.2	67.3	80.4
3	52.5	3	85.7	3	65.4	3	58.9	65.2	74.9
4	48.2	4	85.7	4	96.4	4	47.6	71.8	71.2
5	80.2	5	91.1	5	95.2	5	61.0	82.0	93.7

从表 1.1 的例子可以看出，独立数据结构的各个观察值是彼此独立的，它适宜于用通常的方差分析方法做统计分析。重复测量数据结构是对每一受试者的同一观察指标（血清黏蛋白含量）进行的多次测量。由于这种多次测量之间可能存在相关性，就需要用特殊的统计方法进行分析。

重复测量设计的主要优点是可以减少样本含量，其次是能够控制个体变异，即个体差异。例如在单因素实验中，可以用随机区组（或称配伍组）设计方法来缩小随机误差。而重复测量设计是以同一个受试者作为一个区组，故可以把它看成为是随机区组设计的一种极端形式。但在随机区组设计下的每一测量都是在不同受试者身上进行的，它们对某种处理因素的反应是独立的，符合独立性的假定。而在重复测量设计下的测量是在同一受试者身上进行的，它们对同一处理因素在不同时间上的反应可能是不独立的，后一次的测量结果可能受到前一次测量结果的影响。因此，对同一个体在不同时间上的测量值之间就可能存在相关关系。这给分析工作带来了一定的复杂性。

在实际工作中，重复测量资料比独立观察资料往往更为多见。如在临床研究中，需要观察病人在不同时间的某些生理、生化或病理指标的变化趋势，不同时间或疗程的治疗效果。在流行病学研究中观察队列人群在不同时间上的发病情况。在卫生学研究中，纵向观察儿童的生长发育规律等。重复测量资料在自然科学和社会科学的很多领域内都可见到。

第二节　重复测量值的基本模型

用 y_{ij} 表示第 i 受试者在第 j 时间点上的测量值。可以将 y_{ij} 分解为三个分量：

$$y_{ij} = \mu_j + \alpha_{ij} + e_{ij} \tag{1.1}$$

式中：μ_j 代表总体在时间点 j 的平均水平，为一不变常数，称为固定效应；α_{ij} 代表第 i 受试者在时间点 j 的系统效应，它描述了个体 i 在时间点的抗蜕变特性，称为随机效应。因此，个体 i 在时间点的平均值为 $\mu_j + \alpha_{ij}$。e_{ij} 为随机测量误差，代表了测量值 y_{ij} 对平均值 $\mu_j + \alpha_{ij}$ 的偏离程度大小。

用 E、Var 和 Cov 分别表示期望值、方差和协方差。为了导出重复测量值的协方差结构，首先给出如下假定：

1. 在给定 j 下，随机效应 α_{ij} 在同一总体内随受试者而不同，具有均值为 0，方差为 δ_{jj} 的正态分布，即 $E(\alpha_{ij})=0$，$Var(\alpha_{ij})=\delta_{jj}$。由于随机效应的平均值都被吸收到总平均值 $\mu_j + \alpha_{ij}$ 中，故只剩下随机部分。

对于不同的受试者 i 和 $l(i \neq l)$，在时间点 j 的系统效应 α_{ij} 与在时间点 k 的随机效应 α_{lk} 之间无相关关系。对于同一受试者（$i = l$），在时间点 j 的随机效应 α_{ij} 与时间点 k 的随机效应 α_{ik} 之间有相关关系。可将受试者 i 在时间点 j 的随机效应与受试者 l 在时间点 k 的随机效应之间的协方差表示为：

$$Cov(\alpha_{ij}, \alpha_{lk}) = \begin{cases} 0 & \text{当 } i \neq l \\ \delta_{jk} & \text{当 } i = l \end{cases}$$

2. 在给定 j 下，随机误差 e_{ij} 也是在同一总体内随受试者而不同，具有均值为 0，方差为 ϕ_{jj} 的正态分布。即 $E(e_{ij})=0$，$Var(e_{ij})=\phi_{jj}$。

所有随机误差都相互独立。即当 $i \neq l$ 或 $j \neq k$ 时，相关系数 $\rho(e_{ij}, e_{lk}) = 0$；当 $i = l$ 及 $j = k$ 时，$\rho(e_{ij}, e_{lk}) = 1$，或将它们之间的协方差表示为：

$$Cov(e_{ij}, e_{lk}) = \begin{cases} 0 & \text{当 } i \neq l \text{ 或 } j \neq k \\ \phi_{jk} & \text{当 } i = l \text{ 和 } j = k \end{cases}$$

3．α_{ij} 与 e_{ij} 彼此独立　即对所有 i、l、j、k 来说，相关系数 $\rho(\alpha_{ij}, e_{lk}) = 0$，或者它们之间的协方差 $Cov(\alpha_{ij}, e_{lk}) = 0$。

在上述假定下，对重复测量值 y_{ij} 与 y_{lk} 的协方差结构推导如下：

$$\begin{aligned} Cov(y_{ij}, y_{lk}) &= E[(y_{ij} - \mu_j)(y_{lk} - \mu_k)] \\ &= E[(\alpha_{ij} + e_{ij})(\alpha_{lk} + e_{lk})] \\ &= E[(\alpha_{ij}\alpha_{lk} + \alpha_{ij}e_{lk} + \alpha_{lk}e_{ij} + e_{ij}e_{lk})] \end{aligned}$$

因为有：$E(\alpha_{ij}e_{lk}) = 0$，$E(\alpha_{lk}e_{ij}) = 0$，$E(\alpha_{ij}\alpha_{lk}) = \Delta_{il}\delta_{jk}$，$E(e_{ij}e_{lk}) = \Delta_{il}\Delta_{jk}\phi_{jk}$。

其中：Δ_{il} 及 Δ_{jk} 为克罗内克(Kronecker)符号，其赋值规则为：

$$\Delta_{il} = \begin{cases} 0 & \text{当 } i \neq l \\ 1 & \text{当 } i = l \end{cases}, \Delta_{jk} = \begin{cases} 0 & \text{当 } j \neq k \\ 1 & \text{当 } j = k \end{cases}$$

所以：

$$\begin{aligned} Cov(y_{ij}, y_{lk}) &= E[(\alpha_{ij}\alpha_{lk} + \alpha_{ij}e_{lk} + \alpha_{lk}e_{ij} + e_{ij}e_{lk})] \\ &= E(\alpha_{ij}\alpha_{lk}) + E(e_{ij}e_{lk}) \\ &= \Delta_{il}\delta_{jk} + \Delta_{il}\Delta_{jk}\phi_{jk} \\ &= \Delta_{il}(\delta_{jk} + \Delta_{jk}\phi_{jk}) \end{aligned}$$

从而得到每一对测量值 (y_{ij}, y_{lk}) 之间的协方差为：

$$Cov(y_{ij}, y_{lk}) = \begin{cases} 0 & i \neq l, \quad \text{有 } \Delta_{il} = 0 \\ \delta_{jk} & i = l, j \neq k, \text{有 } \Delta_{il} = 1, \Delta_{jk} = 0 \\ \delta_{jk} + \phi_{jk} & i = l, j = k, \text{有 } \Delta_{il} = 1, \Delta_{jk} = 1 \end{cases} \tag{1.2}$$

根据式(1.2)得到不同受试者($i \neq l$)的测量值之间无相关关系，同一受试者($i = l$)的测量值之间有相关关系。相关系数为：

$$\rho_{jk} = \frac{\delta_{jk}}{\sqrt{(\delta_{jj} + \phi_{jj})(\delta_{kk} + \phi_{kk})}} = \frac{\delta_{jk}}{\sqrt{\sigma_{jj} \times \sigma_{kk}}} \tag{1.3}$$

式中：$\sigma_{jj} = \delta_{jj} + \phi_{jj}$，$\sigma_{kk} = \delta_{kk} + \phi_{kk}$。这一相关系数又称为类内相关(intraclass correlation)，它测量同一受试者的各测量值之间的相关强度。ρ_{jk} 的取值在(0,1)之间。

在特定情况下，如果令式中的 $\delta_{jj} = \delta_{kk} = \delta_{jk} = \delta^2$ 以及 $\phi_{jj} = \phi_{kk} = \phi^2$，则可将式(1.3)写为：

$$\rho = \frac{\delta^2}{\delta^2 + \phi^2} \tag{1.4}$$

根据上式，可将协方差写为相关系数与方差的函数：

$$\delta^2 = \rho \times (\delta^2 + \phi^2) \tag{1.5}$$

对 δ_{jk} 的不同规定就产生出不同的相关结构。

第三节　重复测量资料的相关类型

为表达方便起见，令 y_{ij} 的方差为 $Var(y_{ij}) = \sigma_{jj}$，$y_{ik}$ 的方差为 $Var(y_{ik}) = \sigma_{kk}$。$y_{ij}$ 和 y_{ik} 之

间的协方差为 $Cov(y_{ij}, y_{ik}) = \sigma_{jk}$。从式(1.3)得到协方差的表达式为：

$$Cov(y_{ij}, y_{ik}) = \rho_{jk}[Var(y_{ij})Var(y_{ik})]^{1/2}, \text{或 } \sigma_{jk} = \rho_{jk}(\sigma_{jj}\sigma_{kk})^{1/2}$$

协方差与相关系数只相差一个比例常数 $[Var(y_{ij})Var(y_{ik})]^{1/2}$ [或表示为 $(\sigma_{jj}\sigma_{kk})^{1/2}$]，两者都是测量两变量之间相关密切程度的统计指标。

在独立数据结构下的 y_{ij} 与 y_{lk}，由于恒有 $i \neq l$ 及 $j \neq k$，它们之间的相关系数 $\rho = 0$，故协方差 $Cov(y_{ij}, y_{lk})$ 也为 0。并假设该两变量分别服从具有公共方差 σ^2 的正态分布。

在重复测量情况下，对第 i 受试者在时间 j 的观察值与在时间 k 的观察值之间的相关系数可能不为 0，因此它们之间的协方差 $Cov(y_{ij}, y_{ik})$ 也就不为 0。

假设有 n 例受试者，有 p 个观察时间点，用 y_{ij} 与 y_{ik} 表示受试者 i 分别在时间点 $j(j = 1, \cdots, p)$ 与 $k(k = 1, \cdots, p)$ 的观察值，对受试者 i 的 p 个时间点的测量值之间的相关系数用相关矩阵 $\boldsymbol{R}_i$ 及相应的协方差矩阵 $\sum_i$ 分别表示为：

$$\boldsymbol{R}_i = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}, \sum_i = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

矩阵中共需计算 $p(p-1)/2$ 个元素。设有 n 例受试者，对第 $i(i = 1, 2, \ldots, n)$ 个受试者有一个 $(p \times p)$ 维的相关矩阵 $\boldsymbol{R}_i$，全部 n 例受试者资料的相关矩阵 R 为一个 $(n \times p)(n \times p)$ 维的对角矩阵，其第 i 主对角线上的元素(矩阵块)为 $\boldsymbol{R}_i$。

下面为表达方便，将省去下标 i。把第 $i(i = 1, 2, \ldots, n)$ 例受试者的 $(p \times p)$ 维相关矩阵 $\boldsymbol{R}_i$ 表示为 $\boldsymbol{R}$，把协方差矩阵 $\sum_i$ 记为 $\sum$。

相关类型主要有下列几种：

1. 独立结构(independence structure)　即无相关关系。相关矩阵主对角线上的元素为 1，非主对角线上的元素为 0。它表示不同时间点上的测量值之间彼此独立，无相关关系。与独立结构相关系数相对应的协方差矩阵结构称球性(sphericity)结构。即各时间点测量值的方差相等，即 $\sigma_{jj} = \sigma^2, (j = 1, 2, \cdots, p)$，不同时间点测量值之间的协方差为 0，即 $\sigma_{jk} = 0 (j \neq k)$。独立结构的相关矩阵 R 及球性协方差矩阵 $\sum$ 分别表示为：

$$R = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}, \sum = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I$$

矩阵 $\sum$ 中的 σ_{jk} 当 $j \neq k$ 时为 0。$\sum$ 的等号最右边的 I 为 $p \times p$ 维单位矩阵。

2. 可互换相关结构(exchangeable correlation structure)　在这种结构下，主对角线上的元素为 1，非主对角线上的元素为 ρ。它表示不同时间点上的测量值之间彼此不独立，存在一定的相关关系。但这种相关关系对任何两个时间点的测量值来说都是相等的，不随两个时间点之间的间隔大小而改变。与可互换相关结构相对应的协方差矩阵结构称复合对称结构(compound symmetry structure)。这时各时间点测量值的方差相等，即 $\sigma_{jj} = \sigma^2 (j = 1, 2, \cdots, p)$，不同时间点测量值之间的协方差为 δ^2，即 $\sigma_{jk} = \delta^2 (j \neq k)$。可互换结构的相关矩

阵 R 及复合对称协方差矩阵 $\sum$ 分别表示为：

$$R=\begin{bmatrix}1 & \rho & \cdots & \rho\\ \rho & 1 & \cdots & \rho\\ \vdots & \vdots & \vdots & \vdots\\ \rho & \rho & \cdots & 1\end{bmatrix},\sum=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p}\\ \vdots & \vdots & \vdots & \vdots\\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}\end{bmatrix}=\begin{bmatrix}\sigma^2 & \delta^2 & \cdots & \delta^2\\ \delta^2 & \sigma^2 & \cdots & \delta^2\\ \vdots & \vdots & \vdots & \vdots\\ \delta^2 & \delta^2 & \cdots & \sigma^2\end{bmatrix}$$

这里 $\sigma^2=\phi^2+\delta^2$，ϕ^2 是随机误差的方差，而 δ^2 是协方差。ρ、ϕ^2 和 δ^2 三者的关系是

$$\rho=\frac{\delta^2}{\phi^2+\delta^2}\quad 或\ \delta^2=\rho(\phi^2+\delta^2)\qquad [见式(1.4)及式(1.5)]$$

上面的协方差矩阵 $\sum$ 可简写为：

$$\sum=\phi^2 I+\delta^2 ee'$$

式中 I 为 $p\times p$ 维单位矩阵，e 是所有元素都为 1 的 $p\times p$ 维矩阵。

3．一阶相关结构(one-dependent structure)　又称一阶自回归结构(first-order regressive structure)。在时间点 j 的测量值只受其前一时间点 $j-1$ 测量值的影响，而与再前面的测量值无关。一阶相关结构的相关矩阵 R 及对应的协方差矩阵 $\sum$ 分别表示为：

$$R=\begin{bmatrix}1 & \rho & \cdots & \rho^{p-2} & \rho^{p-1}\\ \rho & 1 & \cdots & \rho^{p-3} & \rho^{p-2}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \rho^{p-2} & \rho^{p-3} & \cdots & 1 & \rho^{p-1}\\ \rho^{p-1} & \rho^{p-2} & \cdots & \rho & 1\end{bmatrix},\sum=\sigma^2\begin{bmatrix}1 & \rho & \cdots & \rho^{p-2} & \rho^{p-1}\\ \rho & 1 & \cdots & \rho^{p-3} & \rho^{p-1}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \rho^{p-2} & \rho^{p-3} & \cdots & 1 & \rho^{p-1}\\ \rho^{p-1} & \rho^{p-2} & \cdots & \rho & 1\end{bmatrix}$$

4．循环相关结构(circular correlation structure)　又称 Toeplitz 相关结构。在一定空间范围内或地域范围内所获取的重复测量资料，毗邻之间的相关性与非毗邻之间的相关性是不同的。如测量植物叶片的叶绿素含量时，两相邻方向(如东与南，西与北)上的叶绿素测量值之间的相关强度与两相对方向(如东与西，南与北)上的叶绿素测量值之间的相关强度不同。如果把两相邻方向上的相关系数记为 ρ_1，两相对方向上的相关系数记为 ρ_2，则可以把四个方向的相关矩阵 R 及协方差矩阵 $\sum$ 分别表示为：

$$R=\begin{bmatrix}1 & \rho_1 & \rho_2 & \rho_1\\ \rho_1 & 1 & \rho_1 & \rho_2\\ \rho_2 & \rho_1 & 1 & \rho_1\\ \rho_1 & \rho_2 & \rho_1 & 1\end{bmatrix},\sum=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14}\\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24}\\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34}\\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44}\end{bmatrix}=\begin{bmatrix}\sigma^2 & \delta_1^2 & \delta_2^2 & \delta_1^2\\ \delta_1^2 & \sigma^2 & \delta_1^2 & \delta_2^2\\ \delta_2^2 & \delta_1^2 & \sigma^2 & \delta_1^2\\ \delta_1^2 & \delta_2^2 & \delta_1^2 & \sigma^2\end{bmatrix}$$

协方差矩阵 $\sum$ 中的双下标 1、2、3、4 分别表示东南西北四个方向。如(11)表示正东，(12)表示东南，(13)表示东西等。

一坐城市不同测量点的大气污染重复测量资料也具有循环相关结构的特点。

5．带状主对角结构(banded main diagonal structure)　各次测量值之间的相关系数为 0，但方差不等。协方差矩阵的主对角元素不相等，非主对角元素为 0。其相关矩阵 R 及协方差矩阵 $\sum$ 分别表示为：

$$R=\begin{bmatrix}1&0&\cdots&0\\0&1&\cdots&0\\\vdots&\vdots&\vdots&\vdots\\0&0&\cdots&1\end{bmatrix},\sum=\begin{bmatrix}\sigma_{11}&0&0&0\\0&\sigma_{22}&0&0\\0&0&\sigma_{33}&0\\0&0&0&\sigma_{44}\end{bmatrix}=\begin{bmatrix}\sigma_1^2&0&0&0\\0&\sigma_2^2&0&0\\0&0&\sigma_3^2&0\\0&0&0&\sigma_4^2\end{bmatrix}$$

式中：$\sigma_i^2\neq\sigma_j^2, i\neq j$。

6. 空间幂相关(spatial power correlation)　测量值之间的相关关系随距离的加大而减弱。相关矩阵 R 及协方差矩阵 $\sum$ 分别为：

$$R=\begin{bmatrix}1&\rho^{d_{12}}&\cdots&\rho^{d_{1p-1}}&\rho^{d_{1p}}\\\rho^{d_{21}}&1&\cdots&\rho^{d_{2p-1}}&\rho^{d_{2p}}\\\vdots&\vdots&\vdots&\vdots&\vdots\\\rho^{d_{p-1,1}}&\rho^{d_{p-1,2}}&\cdots&1&\rho^{d_{p-1,p}}\\\rho^{d_{p1}}&\rho^{d_{p2}}&\cdots&\rho^{d_{p,p-1}}&1\end{bmatrix},\sum=\sigma^2\begin{bmatrix}1&\rho^{d_{12}}&\cdots&\rho^{d_1p-1}&\rho^{d_{1p}}\\\rho^{d_{21}}&1&\cdots&\rho^{d_{2p-1}}&\rho^{d_{2p}}\\\vdots&\vdots&\vdots&\vdots&\vdots\\\rho^{d_{p-1,1}}&\rho^{d_{p-1,2}}&\cdots&1&\rho^{d_{p-1,p}}\\\rho^{d_{p1}}&\rho^{d_{p2}}&\cdots&\rho^{d_{p,p-1}}&1\end{bmatrix}$$

式中的 $d_{ij}=d_{ji}$ 表示两点之间的距离。这种空间协方差结构又称 Markov 协方差结构。

实际资料的相关结构是比较复杂的。上面介绍的几种只是在资料分析中常用的几种类型。最后是无结构的相关关系。

7. 无结构相关(unstructured correlation)　其相关结构无规律可循。可以把无结构相关矩阵 R 及协方差矩阵 $\sum$ 分别表示为：

$$R=\begin{bmatrix}1&\rho_{12}&\rho_{13}&\rho_{14}\\\rho_{21}&1&\rho_{23}&\rho_{24}\\\rho_{31}&\rho_{32}&1&\rho_{34}\\\rho_{41}&\rho_{42}&\rho_{43}&1\end{bmatrix},\sum=\begin{bmatrix}\sigma_{11}&\sigma_{12}&\sigma_{13}&\sigma_{14}\\\sigma_{21}&\sigma_{22}&\sigma_{23}&\sigma_{24}\\\sigma_{31}&\sigma_{32}&\sigma_{33}&\sigma_{34}\\\sigma_{41}&\sigma_{42}&\sigma_{43}&\sigma_{44}\end{bmatrix}$$

在协方差矩阵 $\sum$ 中的主对角元素不相等，对角线的上(或下)三角中的元素不相等。

常用的统计分析方法如方差分析方法要求协方差为球性或复合对称性，否则，会造成过多地拒绝本来是真的无效假设。在多因素分析中，如果不考虑这种相关性，也会使得参数估计值的方差变小，其结果也会导致过多地拒绝本来是真的无效假设。

第四节　关于协方差阵的假设检验

本节介绍关于协方差矩阵的球性检验和复合对称性检验。

一、资料表达方式

对 n 个受试者在 p 个时间点上的反应变量 y 进行测量，得到一组样本观察值。令受试者 $i(i=1,2,\cdots,n)$ 在 p 个时间点上的测量值为 $\mathbf{y}_i=(y_{i1},y_{i2},\cdots,y_{ip})'$，$y_i$ 为 $p\times1$ 维观察值向量。记平均值向量为：

$$\bar{\mathbf{y}}=(\bar{y}_1,\bar{y}_2,\cdots,\bar{y}_p)' \tag{1.6}$$

其中 $\bar{y}_j=\sum_{i=1}^{n}y_{ij}$ 为第 $j(j=1,2,\cdots,p)$ 个时点的平均值。

记 p 个时点测量值之间的样本协方差矩阵 $\boldsymbol{S}$ 为：

$$\boldsymbol{S}=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}, s_{jk}=\begin{cases} \sum_{i=1}^{n}(y_{ij}-\bar{y}_j)(y_{ik}-\bar{y}_k), j\neq k \text{ 为协方差} \\ \sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2, \qquad j=k \text{ 为方差} \end{cases} \tag{1.7}$$

二、球性检验

球性检验的假设写为：无效假设：H_0：$\sum=\sigma^2 I$，备择假设：H_1：$\sum\neq\sigma^2 I$。其中的 I 为 $p\times p$ 单位矩阵。所用检验统计方法为似然比检验法，此法为 Mauchly 于 1940 年提出，故称为 Mauchly 球性检验。所用统计量为 λ。λ 的计算公式为：

$$\lambda=\frac{|\boldsymbol{S}|}{[tr(\boldsymbol{S})/p]^p} \tag{1.8}$$

式中 $\boldsymbol{S}$ 为式(1.7)所定义的样本协方差矩阵。$|\boldsymbol{S}|$ 为 $\boldsymbol{S}$ 的行列式的值，$tr(\boldsymbol{S})=\sum_{j=1}^{p}s_{jj}$ 是协方差矩阵的主对角线元素之和，称为 S 的迹，p 为测量的时间点个数。检验用 χ^2 统计量的计算公式为：

$$\chi^2=-\left[(n-1)-\frac{n-(2p^2+2)}{6p}\right]\ln\lambda \tag{1.9}$$

在无效假设条件下，χ^2 服从自由度 $\nu=[p(p-1)/2]-1$ 的 χ^2 分布。当 $\chi^2>\chi^2_{\alpha(\nu)}$ 时，在 α 水平上拒绝无效假设。

【例 1.1】 对表 1.1 中第二部分：5 名受试者的血清黏蛋白含量(mg/L)重复测量值的协方差矩阵结构作球性检验。本例的 $n=5, p=3$。球性检验的 SAS 程序如下：

```
OPTIONS LS=64 PS=45 NODATE NONUMBER;
PROC IML;                    /*  例 1.1 资料的第 2 部分  */
y={69.6 78.9 85.9,
   48.2 67.0 80.4,
   58.9 65.2 74.9,
   47.6 71.8 71.2,
   61.0 82.0 93.7};
p=NCOL(y);
n=NROW(y);
PRINT n;
xx=y'*(I(n)-(1/n)*j(n,n))*y;      /*  xx=未校正平方和与交叉乘积和矩阵  */
PRINT xx;
RUN;
s=xx/(n-1);                  /*      s=协方差矩阵      */
PRINT s;
TITLE 'Test for sphericity';
const1=-((n-1)-(2*p*p+p+2)/(6*p));
lamda=(det(s)/((trace(s)/p)**p));
```

```
df=(p*(p+1)/2)-1;
chi_sq=const1*LOG(lamda);
p_value=1-PROBCHI(chi_sq,df);   /*   计算 p 值   */
PRINT const1 lamda chi_sq df p_value;
RUN;
```

输出结果有:样本协方差矩阵 $\boldsymbol{S}$、$\boldsymbol{S}$ 的行列式的值$|\boldsymbol{S}|$及 $\boldsymbol{S}$ 的迹 $tr(\boldsymbol{S})$分别为:

$$\boldsymbol{S}=\begin{bmatrix}86.04 & 39.90 & 49.57\\ 39.90 & 53.52 & 51.54\\ 49.57 & 51.54 & 79.67\end{bmatrix},|\boldsymbol{S}|=83825.87,tr(\boldsymbol{S})=219.23$$

将以上结果代入式(1.8)及式(1.9)得到

$$\lambda=\frac{83825.87}{(219.23/3)^3}=0.2148$$

$$\chi^2=-\left[(5-1)-\frac{(2\times3^2+3+2)}{6(3)}\right]\ln(0.2148)=4.19,\nu=[3*(3+1)]/2-1=5$$

查附表 3:χ^2 界值表得 $\chi^2_{0.05(5)}=11.07>4.19$。用 SAS 程序计算的实际概率 $P=0.5228$。因此,不拒绝球性假设。

在 SAS 软件的重复测量资料方差分析模块中有 Mauchly 协方差矩阵球形检验的语句,将在本章最后部分介绍。

三、复合对称性检验

复合对称性检验的假设写为:

无效假设:H_0: $\sum=\phi^2I+\delta^2ee'$,备择假设:H_1: $\sum\neq\phi^2I+\delta^2ee'$

所用检验统计方法为似然比检验法,所用统计量为 λ。λ 的计算公式为:

$$\lambda=\frac{|\boldsymbol{S}|}{[tr(\boldsymbol{S})/p]^p(1-r)^{p-1}[1+(p-1)r]} \tag{1.10}$$

式中 S、$|S|$及 $tr(S)$的定义见式(1.8)的说明。r 的计算公式为:

$$r=\frac{\sum_{j=1}^{p}\sum_{k=j+1}^{p}s_{jk}}{2\{p(p-1)[tr(\boldsymbol{S})/p]\}} \tag{1.11}$$

检验用 χ^2 统计量的计算公式为:

$$\chi^2=-\left\{(n-1)-\frac{[p(p+1)^2(2p-3)]}{6(p-1)(p^2+p-4)}\right\}\ln\lambda \tag{1.12}$$

在无效假设条件下,χ^2 服从自由度 $\nu=[p(p+1)/2]-2$ 的 χ^2 分布。当 $\chi^2>\chi^2_{\alpha(\nu)}$时,则在 α 水平上拒绝无效假设。

【例 1.2】 仍以表 1.1 中第二部分:5 名受试者的血清黏蛋白含量(mg/L)重复测量值资料为例,对协方差矩阵的复合对称性检验的步骤加以说明。本例的 $n=5$,$p=3$。复合对称性检验的 SAS 程序如下(本程序中的前 15 行与球性检验中的程序相同):

```
OPTIONS LS=64 PS=45 NODATE NONUMBER;
PROC IML; /* 例 1.1 资料中的第二部分   */
```

```
y= |69.6 78.9 85.9,
   48.2 67.0 80.4,
   58.9 65.2 74.9,
   47.6 71.8 71.2,
   61.0 82.0 93.7|;
p=NCOL(y);
n=NROW(y);
PRINT n;
xx=y'*(I(n)-(1/n)*j(n,n))*y; /*s=Matrix of uncorrected sum of squares */
PRINT xx;
RUN;
s=xx/(n-1);              /* covar=Covariance matrix */
PRINT s;
TITLE 'Test of compound symmetry';
detment=DET(s);
square=SUM(diag(s));
sumall=SUM(s);
ssquare=square/p;
r=(sumall-square)/(p*(p-1)*ssquare);
lamda=detment/((ssquare**p)*((1-r)**(p-1))*(1+(p-1)*r));
correct=(n-1)-(p*(p+1)**2*(2*p-3))/(6*(p-1)*(p*p+p-4));
chi_sq=-correct*LOG(lamda);
df=p*(p+1)/2-2;
p_value=1-PROBCHI(chi_sq,df);
PRINT ssquare r detment sumall;
PRINT lamda chi_sq df p_value;
RUN;
```

计算结果为(协方差矩阵及其行列式的值见例 1.1 的球性检验结果)：

$$\sum_{j=1}^{p}\sum_{k=j+1}^{p} s_{jk} = 564.062$$

$$r = \frac{564.062}{2[3(3-1)(219.23/3)]} = 0.6432$$

$$\lambda = \frac{83825.87}{(219.23/3)^3(1-0.6432)^{3-1}[1+(3-1)(0.6432)]} = 0.7381$$

代入式(1.12)有：

$$\chi^2 = -\left\{(5-1)-\frac{[3(3+1)^2(2\times 3-3)]}{6(3-1)(3^2+3-4)}\right\}\ln(0.7381) = 0.7592$$

$$\nu = 3(3+1)/2-2 = 4$$

查附表 3：χ^2 界值表得 $\chi^2_{0.05(4)} = 9.48 > 0.7592$。用 SAS 程序计算出的实际概率 $P = 0.9439$。故在 $\alpha = 0.05$ 水平上不拒绝复合对称结构的无效假设。

四、协方差矩阵的 H 型结构与 Mauchly 球性检验

Huynh 和 Feldt(1970)等导出了由两个均方的比值得到的 F 检验统计量满足精确 F 分

布的一组充分必要条件。在重复测量情况下的充分必要条件是,不管同一受试者的重复测量值之间的相关结构如何,只要原始协方差矩阵 $\sum = (\sigma_{jk})$ 满足下列条件:

$$\sigma_{jj} + \sigma_{kk} - 2\sigma_{jk} = c \qquad \text{对所有 } j,k \text{ 成立}$$

式中 c 为常数。把 $\sigma_{jj} + \sigma_{kk} - 2\sigma_{jk} = c$ 称为 H 型条件,把具有这一条件的协方差矩阵称为 H 型结构。H 型结构条件比球性结构和复合对称结构的条件要弱,故其包含的面要宽一些。但一阶相关结构和循环相关结构不满足这一条件。

为检验协方差矩阵是否具备 H 型结构,Huynh 和 Feldt 建议用$(p-1)$个正交对比集合的协方差矩阵作球性检验。设 $\boldsymbol{S}$ 为样本协方差矩阵,$\boldsymbol{C}$ 为$(p-1)$个正交对比所构成的矩阵,则似然比检验统计量 λ 的计算公式为:

$$\lambda = \frac{|\boldsymbol{CSC}'|}{\{[tr(\boldsymbol{CSC}')]/(p-1)\}^{p-1}} \tag{1.13}$$

λ 的值与对正交对比矩阵 $\boldsymbol{C}$ 的选取无关。正交对比矩阵 $\boldsymbol{C}$ 可以有多种选择。例如,在有 3 个时间点测量值的情况下可以构造出两个正交矩阵 $\boldsymbol{C}$:

$$\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}, \boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

可选用其中任何一个正交对比矩阵 $\boldsymbol{C}$。还可以选择正交多项式系数。

检验用 χ^2 统计量的计算公式为:

$$\chi^2 = -\left[(n-1) - \frac{2(p-1)^2 + (p-1) + 2}{6(p-1)}\right]\ln\lambda \tag{1.14}$$

在无效假设条件下,χ^2 服从自由度 $\nu = [p(p-1)/2] - 1$ 的 χ^2 分布。当 $\chi^2 > \chi^2_{\alpha(\nu)}$时,则在 α 水平上拒绝无效假设。在 SAS 的 PROC GLM 过程的 REPEATED 语句中的选项 PRINTE 就可产生应用正交对比的 Mauchly 球性检验统计量、χ^2 值及概率水准。如例 1.1 的表 1.1 中第二部分:5 名受试者的重复测量数据的协方差矩阵用下列 SAS 程序输出 λ,χ^2 及概率 P。

```
OPTIONS LS=64 PS=45 NODATE NONUMBER NOCENTER;
DATA exmp1 _ 1;
INPUT subj y1 y2 y3;
CARDS;
1  69.6  78.9  85.9
2  48.2  67.3  80.4
3  58.9  65.2  74.9
4  47.6  71.8  71.2
5  61.0  82.0  93.7
;
PROC GLM DATA=exmp1 _ 1;
MODEL y1 y2 y3= /NOUNI;
REPEATED year 3/PRINTE PRINTH;
RUN;
```