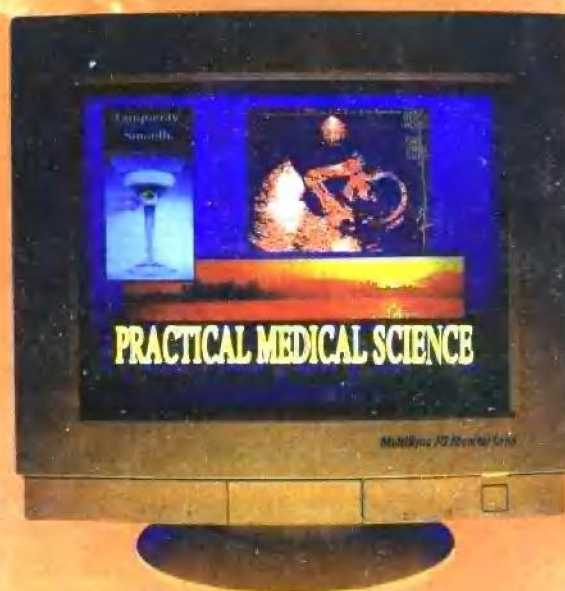




实用医学多因素统计方法

曹素华 主编

上海医科大学出版社

实用医学多因素统计方法

主 编	曹素华	
副主编	周家仪	
编写者	孙伟民	赵耐青
	李江晖	

上海医科大学出版社

责任编辑 沈 雁
封面设计 章向雄
责任校对 谢玲玉

实用医学多因素统计方法

曹素华 主编

上海医科大学出版社出版发行

上海市医学院路 138 号

邮政编码 200032

新华书店上海发行所经销

昆山亭林印刷总厂印刷

开本 787×1092 1/16 印张 15 字数 365 000

1998 年 7 月第 1 版 1998 年 7 月第 1 次印刷

印数 1—5 000

ISBN 7—5627-0422-8/R·400

定价：27.00 元

如遇印、装质量问题,请直接与印刷厂联系

(地址:昆山市中山路 293 号 邮编:215300)

前 言

当今时代,科学技术发展突飞猛进,计算机走进了千家万户,已成为信息时代的主要特征。随着计算机应用技术的发展,多因素统计分析方法在各领域中的应用日益广泛深入,掌握和应用多因素统计分析方法,处理科学研究数据,提高科研水平,已成为医药卫生研究人员的基本需求。

本书是在作者们多年来从事研究生教学和科研数据处理的经验基础上编写的,它系统地介绍医学卫生领域常用的多因素统计方法,内容有:多因子方差分析,协方差分析,聚类分析,判别分析,主成分分析,典型相关分析,多元线性回归,logistic 回归,Cox 回归,Poisson 回归等。

为便于实际工作者的仿照应用,本书在介绍各类统计方法时,均有相应的实例,并结合介绍计算机程序的应用及相应的输出结果解释,本书不仅可作为医学院校研究生教学用书,还可作为医药、卫生及生物科学等有关领域研究人员及实际工作者学习应用多元统计方法的参考之用。

本书所用的 STATA 软件小巧、灵活,不仅占用的计算机硬盘与内存空间均小,而且输出结果简单、易懂,故给实际使用带来很多方便。

为满足统计工作者及部分对统计原理感兴趣者的需要,本书还有部分标上“★”的章节(或段落)简单地讲述了各种方法的数学模型及计算方法等。对于这部分内容,初学者或不感兴趣者可以跳过。

为便于理解多元统计方法的有关概念,在附录中介绍矩阵代数的基本知识和一些多元统计中常用的计算方法及参数估计方法,供读者参考。

为方便使用 SAS 软件,本书的附录中还列出了有关的 SAS 程序。

本书的编写过程中得到上海医科大学卫生统计及社会医学教研室诸多老师的大力支持,顾杏元教授关心、督促整个编写过程,陈启明教授认真、仔细审阅了本书的全部内容,金丕焕教授也在百忙中抽空审阅本书的大部分章节,他们均对本书的内容、编排等提出了宝贵的意见。詹绍康教授、谢玲玉老师、翁仲华老师等为本书的编写、打印等工作予以积极地支持和帮助。上海医科大学出版社的编辑对书稿进行认真、仔细的审阅。特此一并感谢。

本书在编写过程中,全体编写者均尽心尽力,勤奋工作,但由于我们知识和经验有限,书中难免存在一些问题,敬请读者批评指正,以便日臻完善。

曹素华

1997 年 11 月

使用说明

1. 本书为适应医学统计学课程改革而编写,适用于高等医学院校本科生、研究生及具有一定工作经验的临床医师和医学科研工作者。

2. 书中打★部分讲述的是各种统计方法的数学模型和计算方法,初学者或者对此不感兴趣者,可以跳过。

3. 本书共有十二章,除第十二章外,其他各章均介绍一种多元统计方法或一类数据的多元统计方法,有关这些方法的 STATA(4.0)程序已写在各章的正文中,相应的 SAS(6.04)程序写在附录中,由于 STATA(4.0)中没有聚类分析和判别分析程序,故第八章正文中没有 STATA 程序,而用相应的 SAS 程序替代。又由于 SAS(6.04)中没有 Poisson 回归的程序,故在附录中缺少第五章 Poisson 回归的程序。

4. 请注意,STATA 与 SAS 软件中有关对数线性模型的输出结果的参数估计是不同的。这是因为这两个软件中所提及的参数有不同的含义。

5. 本书正文中的 STATA 程序为与其他内容相区别,语句下均加下划线“_____”。STATA 软件要求语句用小写英文字母。

6. 全书的例子统一编为[例 i-j]或 exam i-j,表示第 i 章第 j 个例子,如附录五中的程序 exam i-j. sas 是第 i 章中[例 i-j]的 SAS 程序,附录六中的 exam i-j. txt 是[例 i-j]的数据文件。

目 录

第一章 多因子方差分析	(1)
第一节 概述.....	(1)
第二节 二因子方差分析.....	(1)
第三节 三因子方差分析.....	(7)
第四节 方差分析的随机效应模型	(14)
第五节 方差分析的混合效应模型	(17)
第六节 几点说明	(20)
第二章 协方差分析	(22)
第一节 概述	(22)
第二节 完全随机设计、一个协变量的协方差分析.....	(23)
第三节 一个协变量、二因子析因实验设计协方差分析.....	(28)
第四节 几点说明	(35)
第三章 多元线性回归分析	(36)
第一节 概述	(36)
第二节 实例分析	(37)
第三节 多重共线性	(42)
第四节 自变量的选择	(45)
第五节 几点说明	(56)
第四章 多元 logistic 回归	(57)
第一节 概述	(57)
第二节 非条件 logistic 回归	(58)
第三节 条件 logistic 回归	(63)
第四节 几点说明	(66)
第五章 Poisson 回归	(68)
第一节 概述	(68)
第二节 单个协变量的 Poisson 回归模型	(68)

第三节	多个协变量的 Poisson 回归模型	(75)
第四节*	模型的参数估计与假设检验	(81)
第六章	对数线性模型	(83)
第一节	概述	(83)
第二节	二维列联表对数线性模型	(84)
第三节	三维列联表对数线性模型	(91)
第四节*	模型的参数估计与假设检验	(97)
第五节	几点说明	(100)
第七章	生存分析	(101)
第一节	概述	(101)
第二节	Cox 模型	(104)
第三节	指数分布模型	(108)
第四节	Weibull 分布模型	(113)
第五节	几点说明	(116)
第八章	判别分析	(122)
第一节	概述	(122)
第二节	距离判别分析	(122)
第三节	Bayes 判别分析	(127)
第四节	判别能力的假设检验	(131)
第五节	判别效果的评价	(133)
第六节	逐步判别分析	(134)
第七节	SAS 程序及输出结果	(135)
第九章	聚类分析	(149)
第一节	概述	(149)
第二节	系统聚类法	(151)
第三节	应用实例	(152)
第十章	主成分分析	(161)
第一节	概述	(161)
第二节	主成分的求法	(162)

第三节	主成分个数的确定及实际意义的解释·····	(164)
第四节	主成分分析的一般步骤·····	(166)
第五节	主成分分析的应用·····	(168)
第十一章	典型相关分析·····	(174)
第一节	概述·····	(174)
第二节	典型相关变量与典型相关系数的求法·····	(175)
第三节	典型变量的性质及其实际意义的解释·····	(178)
第四节★	冗余度分析 ·····	(179)
第五节★	典型相关系数的显著性检验 ·····	(183)
第六节	利用 STATA 软件实施典型相关分析的一般步骤 ·····	(183)
第七节	应用举例·····	(186)
第十二章	多元回归分析策略·····	(188)
第一节	全回归分析·····	(188)
第二节	主要影响因素分析·····	(190)
第三节	某因素是否与某疾病发病(或死亡)有关的多因素分析·····	(190)
第四节	统计分析结果解释的要点和应注意的问题·····	(191)
附录	·····	(193)
一、	矩阵及其基本运算 ·····	(193)
二、	参数估计法 ·····	(205)
三、	迭代法 ·····	(208)
四、	STATA 入门 ·····	(212)
五、	SAS 程序 ·····	(214)
六、	正文中某些实例的数据 ·····	(220)

第一章 多因子方差分析

第一节 概 述

单因子方差分析是检验多个样本均数间差别有无统计意义的统计方法。在单因子方差分析问题中涉及的影响观察指标的因素只有一个,例如研究化合物十硼氢对棉布、府绸、的确凉、尼龙四种衣料的渗透性能,分别将四种衣料做成小口袋,袋内装棉花,然后在相同的条件下进行实验,记录各口袋内棉花中吸附的十硼氢的含量。在此研究中衣料是唯一被涉及的影响因素,四种衣料是此因素的四个水平。又如,石棉肺患者、可疑患者、非患者三组对象的用力肺活量比较问题中,涉及的影响因素是患病程度,比较三组对象用力肺活量的差别有无统计意义,是为了回答患病(患石棉肺)是否会影响用力肺活量。

在医学卫生领域中,还常会碰到研究多个因素对某个观察指标的作用问题,例如,药物疗效研究中,欲了解 A、B 两药的使用能否引起缺铁性贫血病人的红细胞增加?还想了解两药一起使用是否会引起协同或拮抗作用?据此目的,某研究者将 12 例缺铁性贫血病人随机分成四组,分别给予四种治疗方法(一般疗法、一般疗法加 A 药、一般疗法加 B 药、一般疗法加 A 药加 B 药),一个月后观察红细胞增加数($10^{12}/L$)。在此研究中反映疗效的观察指标是红细胞增加数,涉及的影响因素有两个,一个是 A 药使用情况,另一个是 B 药使用情况。

多因子方差分析(multifactor analysis of variance)是分析两个或两个以上因素对观察指标影响的统计方法。

方差分析中,称影响观察指标的因素为因子(factor),称因子所处的一个状态为因子的一个水平(level of factor),称因子水平的组合为处理(treatment)。如前面 A、B 药疗效一例中,有两个因子,每个因子有两个水平(用与不用),构成四个组合(A、B 均不用,A 用 B 不用,A 不用 B 用,A、B 均用),每个组合为一个处理。

多因子方差分析与单因子方差分析一样,根据造成观察指标变异的原因,分解离均差平方和及自由度,计算均方,借助 F 分布作出统计推断。应用多因子方差分析的条件也与单因子方差分析相同,要求各观察独立,满足正态分布,且各处理有相同的方差。

第二节 二因子方差分析

一、实例

[例 1-1]^[1] 前述 A、B 药治疗缺铁性贫血一例,试验结果如表 1-1:

表 1-1 四种疗法治疗缺铁性贫血后红细胞增加数($10^{12}/L$)

疗 法	红细胞增加数 (y)	总体均数的记号
疗法 1(一般疗法)	0.8 0.9 0.7	μ_{11}
疗法 2(一般疗法加 A 药)	1.3 1.2 1.1	μ_{21}
疗法 3(一般疗法加 B 药)	0.9 1.1 1.0	μ_{12}
疗法 4(一般疗法加 A 药加 B 药)	2.1 2.2 2.0	μ_{22}

本例 2 个因子,每个因子有 2 个水平,上表中的 4 个疗法就是由它们组成的 4 个处理。分别记相应的 4 个总体均数为 $\mu_{11}, \mu_{21}, \mu_{12}, \mu_{22}$ 。2 个因子可用 2 个变量 a, b 数量化如下:

$a=1$ 表示不用 A 药 $a=2$ 表示用 A 药

$b=1$ 表示不用 B 药 $b=2$ 表示用 B 药

则上表数据可表达如表 1-2:

表 1-2 a, b 数量化表

a	b	y
1	1	0.8
1	1	0.9
1	1	0.7
2	1	1.3
2	1	1.2
2	1	1.1
1	2	0.9
1	2	1.1
1	2	1.0
2	2	2.1
2	2	2.2
2	2	2.0

本例研究目的之一为 A 药的使用是否会引起病人的红细胞变化,此问题归结到统计上是检验不用 A 药(疗法 1 与疗法 3)和用 A 药(疗法 2 与疗法 4)的总体均数是否相等,即检验:

$$H_{01}: (\mu_{11} + \mu_{12}) = (\mu_{21} + \mu_{22})$$

研究目的之二是,用 B 药是否有作用,统计上是检验不用 B 药(疗法 1 与疗法 2)和用 B 药(疗法 3 与疗法 4)的总体均数是否相等,即检验:

$$H_{02}: (\mu_{11} + \mu_{21}) = (\mu_{12} + \mu_{22})$$

研究目的之三是, A 药与 B 药是否有交互作用。所谓 A, B 有协同作用,是指同时用 A、B 药(疗法 4)起的作用大于单独用 A 药(疗法 2)和单独用 B 药(疗法 3)的作用之和,即:

$$\mu_{22} - \mu_{11} > (\mu_{21} - \mu_{11}) + (\mu_{12} - \mu_{11})$$

所谓 A, B 有拮抗作用, 是指同时用 A, B 药的作用小于单独用 A 药与单独用 B 药的作用之和, 即

$$\mu_{22} - \mu_{11} < (\mu_{21} - \mu_{11}) + (\mu_{12} - \mu_{11})$$

不论协同或拮抗作用均意味着 A, B 药同时使用的作用不等于单独作用之和

$$\mu_{22} - \mu_{11} \neq (\mu_{21} - \mu_{11}) + (\mu_{12} - \mu_{11})$$

由 $\mu_{22} - \mu_{11} = \mu_{22} - \mu_{21} + \mu_{21} - \mu_{11} = \mu_{22} - \mu_{12} + \mu_{12} - \mu_{11}$, 可以证明 $\mu_{22} - \mu_{11} \neq (\mu_{21} - \mu_{11}) + (\mu_{12} - \mu_{11})$ 与 $\mu_{22} - \mu_{21} \neq \mu_{12} - \mu_{11}$ 与 $\mu_{22} - \mu_{12} \neq \mu_{21} - \mu_{11}$ 等价。

$\mu_{22} - \mu_{21}$ 表示在用 A 药条件下, 用 B 药与不用 B 药的差别, $\mu_{12} - \mu_{11}$ 表示不用 A 药的条件下用 B 药与不用 B 药的差别; $\mu_{22} - \mu_{21} \neq \mu_{12} - \mu_{11}$ 意味着在 A 药使用的条件下 B 药所起的作用不同于不用 A 药的条件下 B 药的作用, 即 B 药的作用大小受 A 药使用与否的影响; $\mu_{22} - \mu_{12}$ 表示在用 B 药的条件下, 用 A 药与不用 A 药的差别, $\mu_{21} - \mu_{11}$ 表示在不用 B 药的条件下, 用 A 药与不用 A 药的差别; $\mu_{22} - \mu_{12} \neq \mu_{21} - \mu_{11}$ 意味着 A 药的作用大小受 B 药的使用与否的影响, 统计上称这种 A 与 B 的相互影响为交互作用。

据以上分析, 研究两药有无协同或拮抗作用, 只要检验假设:

$$H_{03} : \mu_{22} - \mu_{21} = \mu_{12} - \mu_{11}$$

或

$$H_{03} : \mu_{22} - \mu_{12} = \mu_{21} - \mu_{11}$$

二、STATA 程序及输出结果的解释

STATA 程序

. infile a b y using d:\mbook\exam1-1.txt (由文件 'd:\m book\exam1-1.txt' 读入数据) (12 observations read)

. anova y a b a * b

方差分析, 语句格式为: anova 观察变量名 因子名表, (其中“a * b”表示 A, B 交互作用项) 输出结果:

Number of obs	=	12	R-squared	=	0.9737
Root MSE	=	0.10	Adj R-squared	=	0.9638
①	②	③	④	⑤	⑥
Source	Partial SS	df	MS	F	Prob > F
Model	2.96249994	3	0.98749998	98.75	0.0000
a	1.6875	1	1.6875	168.75	0.0000
b	0.907499974	1	0.907499974	90.75	0.0000
a * b	0.367499967	1	0.367499967	36.75	0.0003
Residual	0.080000002	8	0.01		
Total	3.04249994	11	0.276590904		

上面输出结果是一张方差分析表。表中第①列是引起观察指标变异的来源, 分为 Model 和 Residual 两部分。Model 部分又分为 3 小部分, a、b 分别表示因子 A、因子 B 的作用; a * b 表示因子 A 与 B 的交互作用。Residual 即误差部分。②描述各变异来源的离均差平方和。③

描述相应的自由度。④为均方。⑤为相应的 F 值。⑥为 P 值。据此结果,在 $\alpha=0.05$ 水平上, $a, b, a * b$ 三项均有统计意义,即因子 A 、因子 B 、因子 A 与 B 的交互作用均有统计意义,据题意, A 药有作用, B 药有作用, A 药与 B 药还有交互作用。至于此交互作用是协同作用还是拮抗作用?可用以下 STATA 语句,求得四种处理均数后作出估计。

. tab a b ,sum(y) means

按变量 a, b 分类后求变量 y 的均数

输出结果:

Means of y			
a	1	2	Total
1	0.8	1.0	0.9
2	1.2	2.1	1.65
Total	1	1.55	1.275

由上表,估计 $\mu_{11}=0.8, \mu_{12}=1.0, \mu_{21}=1.2, \mu_{22}=2.1$,所以估计同时用 A, B 两药,病人红细胞平均增加为 $\mu_{22}-\mu_{11}=2.1-0.8=1.3$,单用 A 药,病人红细胞平均增加为 $\mu_{21}-\mu_{11}=1.2-0.8=0.4$,而单用 B 药,病人红细胞平均增加 $\mu_{12}-\mu_{11}=1.0-0.8=0.2, 1.3>0.4+0.2$,即 A, B 两药同时使用,引起的红细胞增加数比单独用 A 药和单独用 B 药的增加数之和还大,故认为 A 药与 B 药的交互作用是协同作用。

注意:当因子 A 与 B 间的交互作用有统计意义时,对 A (或 B)的单独作用的解释须小心。本例,用 B 药时,同时用 A 药比不同时用 A 药的病人的红细胞增加均数大,不用 B 药时,用 A 药比不用 A 药病人的红细胞增加均数也大,故可以说 A 药有利于病人的红细胞的增加。但有时可能出现这种情况,用 B 药时,用 A 药比不用 A 药病人的红细胞平均增加数大,而不用 B 药时,用 A 药比不用 A 药的病人的红细胞平均增加数小,此时就不能简单地说 A 药有利于病人的红细胞增加,需分别就用 B 药和不用 B 药两种情况述说 A 药的作用。对 B 作用的解释也是如此。

三★、模型

二因子方差分析模型为

$$I. y_{abk} = \mu_{ab} + \epsilon_{abk}, a=1, \dots, n_a; b=1, \dots, n_b; k=1, \dots, n$$

其中: y_{abk} : 为 A 因子的第 a 水平与 B 因子的第 b 水平构成的处理的第 k 个观察

μ_{ab} : 为 A 因子的第 a 水平与 B 因子的第 b 水平构成的处理的总体均数

ϵ_{abk} : 为误差项,相互独立,服从正态分布 $N(0, \sigma^2)$

或

$$II. y_{abk} = \mu_{..} + \alpha_a + \beta_b + (\alpha\beta)_{ab} + \epsilon_{abk}$$

其中: $\mu_{..}$: 表示所有观察的总体均数

α_a : 参数,描述因子 A 的第 a 个水平的作用,满足 $\sum_a \alpha_a = 0$

β_b : 参数,描述因子 B 的第 b 个水平的作用,满足 $\sum_b \beta_b = 0$

$(\alpha\beta)_{ab}$: 参数, 描述因子 A 的第 a 水平与因子 B 的第 b 水平的交互作用, 满足 $\sum_a (\alpha\beta)_{ab} = \sum_b (\alpha\beta)_{ab} = 0$

ϵ_{abk} : 相互独立, 服从正态分布 $N(0, \sigma^2)$

I、II 两模型有关系: $\mu_{ab} = \mu_{..} + \alpha_a + \beta_b + (\alpha\beta)_{ab}$, 即后者进一步分解处理均数为总均数, 因子 A 的作用, 因子 B 的作用和因子 A 与 B 的交互作用之和。

模型中的 n_a, n_b 分别表示因子 A 与 B 的水平数, n 表示每个处理的观察例数。例 1-1 中, $n_a = 2, n_b = 2, n = 3$ 。

据前面分析, 研究 A 药是否有作用, 对于模型 I 只要检验

$$H_{01}: \mu_{11} + \mu_{12} = \mu_{21} + \mu_{22}$$

对于模型 II 即检验

$$H_{01}: \alpha_1 = \alpha_2 = 0$$

同样检验 B 药是否起作用, 对于模型 I 只要检验

$$H_{02}: \mu_{11} + \mu_{21} = \mu_{12} + \mu_{22}$$

对于模型 II 即检验

$$H_{02}: \beta_1 = \beta_2 = 0$$

检验 A 与 B 有无交互作用的无效假设, 对于模型 I 为

$$H_{03}: \mu_{12} - \mu_{11} = \mu_{22} - \mu_{21} \text{ 或 } H_{03}: \mu_{21} - \mu_{11} = \mu_{22} - \mu_{12}$$

对于模型 II 即检验

$$H_{03}: (\alpha\beta)_{22} - (\alpha\beta)_{21} = (\alpha\beta)_{12} - (\alpha\beta)_{11}$$

即 $H_{03}: (\alpha\beta)_{22} = (\alpha\beta)_{21} = (\alpha\beta)_{12} = (\alpha\beta)_{11} = 0$

四*、方差分析计算过程

1. 分解离均差平方和

表 1-3 一些统计记号

记号	意 义
$y_{ab.} = \sum_k y_{abk}$	因子 A 的第 a 水平与因子 B 的第 b 水平构成的处理的观察之和
$\bar{y}_{ab.} = \frac{y_{ab.}}{n}$	因子 A 的第 a 水平与因子 B 的第 b 水平构成的处理的观察的均数
$y_{a..} = \sum_b \sum_k y_{abk}$	因子 A 的第 a 水平观察之和
$\bar{y}_{a..} = \frac{y_{a..}}{nn_b}$	因子 A 的第 a 水平的观察的均数
$y_{.b.} = \sum_a \sum_k y_{abk}$	因子 B 的第 b 个水平的观察之和
$\bar{y}_{.b.} = \frac{y_{.b.}}{nn_a}$	因子 B 的第 b 个水平的观察的均数
$y_{...} = \sum_a \sum_b \sum_k y_{abk}$	所有观察的总和
$\bar{y}_{...} = \frac{y_{...}}{N}$	所有观察的总均数, 其中 $N = n_a n_b n$ 为总例数

表 1-4 例 1-1 的统计量

A 药	B 药						合计
	不用 ($b=1$)			用 ($b=2$)			
不用 ($a=1$)	$y_{111}=0.8$	$y_{112}=0.9$	$y_{113}=0.7$	$y_{121}=0.9$	$y_{122}=1.1$	$y_{123}=1.0$	$y_{1..}=5.4$
		$y_{11.}=2.4$			$y_{12.}=3.0$		
		$\bar{y}_{11.}=0.8$			$\bar{y}_{12.}=1.0$		$\bar{y}_{1..}=0.9$
用 ($a=2$)	$y_{211}=1.3$	$y_{212}=1.2$	$y_{213}=1.1$	$y_{221}=2.1$	$y_{222}=2.2$	$y_{223}=2.0$	$y_{2..}=9.9$
		$y_{21.}=3.6$			$y_{22.}=6.3$		
		$\bar{y}_{21.}=1.2$			$\bar{y}_{22.}=2.1$		$\bar{y}_{2..}=1.65$
合计		$y_{.1.}=6.0$			$y_{.2.}=9.3$		$y_{...}=15.3$
		$\bar{y}_{.1.}=1.0$			$\bar{y}_{.2.}=1.55$		$\bar{y}_{...}=1.275$

根据引起观察值 y 变异的来源,各观察 y 与总均数的差可被分解如下:

$$y_{abk} - \bar{y}_{...} = (y_{abk} - \bar{y}_{ab.}) + (\bar{y}_{a..} - \bar{y}_{...}) + (\bar{y}_{.b.} - \bar{y}_{...}) + (\bar{y}_{ab.} - \bar{y}_{a..} - \bar{y}_{.b.} + \bar{y}_{...})$$

可以证明

$$\sum_{a,b,k} (y_{abk} - \bar{y}_{...})^2 = \sum_{a,b,k} (y_{abk} - \bar{y}_{ab.})^2 + \sum_{a,b,k} (\bar{y}_{a..} - \bar{y}_{...})^2 + \sum_{a,b,k} (\bar{y}_{.b.} - \bar{y}_{...})^2 + \sum_{a,b,k} (\bar{y}_{ab.} - \bar{y}_{a..} - \bar{y}_{.b.} + \bar{y}_{...})^2$$

上式等号左边为总的离均差平方和,记为 SS_T ;等号右边的第一项描述由随机因素引起的离均差平方和,记为 SS_e ;第二项描述由因子 A 引起的离均差平方和,记为 SS_A ;第三项则为由因子 B 引起的离均差平方和,记为 SS_B ;最后一项为由因子 A 与因子 B 的交互作用引起的离均差平方和,记为 SS_{AB} 。后三项之和(即 $SS_A + SS_B + SS_{AB}$)又称为处理平方和或模型平方和,记为 SS_{tr} 。

上面给出了各离均差平方和的定义,实际计算时可从定义出发,但为简化运算过程,一般按以下公式计算(本例举例):

$$\begin{aligned} C &= \frac{y_{...}^2}{N} & C &= (15.3)^2/12 = 19.5075 \\ SS_T &= \sum_{a,b,k} y_{abk}^2 - C & SS_T &= (0.8^2 + 0.9^2 + \dots + 2.0^2) - 19.5075 = 3.0425 \\ SS_{tr} &= \sum_{a,b} \frac{y_{ab.}^2}{n} - C & SS_{tr} &= (2.4^2 + 3.0^2 + 3.6^2 + 6.3^2)/3 - 19.5075 = 2.9625 \\ SS_A &= \sum_a \frac{y_{a..}^2}{nn_b} - C & SS_A &= (5.4^2 + 9.9^2)/(3 \times 2) - 19.5075 = 1.6875 \\ SS_B &= \sum_b \frac{y_{.b.}^2}{nn_a} - C & SS_B &= (6.0^2 + 9.3^2)/(3 \times 2) - 19.5075 = 0.9075 \\ SS_{AB} &= SS_{tr} - SS_A - SS_B & SS_{AB} &= 2.9625 - 1.6875 - 0.9075 = 0.3675 \\ SS_e &= SS_T - SS_{tr} & SS_e &= 3.0425 - 2.9625 = 0.0800 \end{aligned}$$

2. 分解自由度

总的自由度	$df_T = N - 1$	本例	$df_T = 12 - 1 = 11$
处理的自由度	$df_{tr} = n_a n_b - 1$		$df_{tr} = 2 \times 2 - 1 = 3$
因子 A 的自由度	$df_A = n_a - 1$		$df_A = 2 - 1 = 1$
因子 B 的自由度	$df_B = n_b - 1$		$df_B = 2 - 1 = 1$

$$A \text{ 与 } B \text{ 的交互作用的自由度 } df_{AB} = (n_a - 1)(n_b - 1) \quad df_{AB} = (2 - 1)(2 - 1) = 1$$

$$\text{误差项的自由度 } df_e = df_T - df_{tr} \quad df_e = 11 - 3 = 8$$

3. 均方的计算及其数学期望

均方 = 离均差平方和 / 自由度

可以证明它们的数学期望分别为:

$E(MS_A) = n n_b \sum_{a=1}^{n_a} \alpha_a^2 \div (n_a - 1) + \sigma^2$ 说明当因子 A 不起作用(即 H_{01} 为真)时, MS_A 是 σ^2 的无偏估计。

$E(MS_B) = n n_a \sum_{b=1}^{n_b} \beta_b^2 \div (n_b - 1) + \sigma^2$ 说明, 当因子 B 不起作用(即 H_{02} 为真)时, MS_B 是 σ^2 的无偏估计。

$E(MS_{ab}) = n \sum_{a=1}^{n_a} \sum_{b=1}^{n_b} (\alpha\beta)_{ab}^2 \div [(n_a - 1)(n_b - 1)] + \sigma^2$ 说明, 当 A 与 B 无交互作用(即 H_{03} 为真)时, MS_{ab} 也是 σ^2 的无偏估计。

$E(MS_e) = \sigma^2$ 说明, 无论各因子是否有作用, 也不管是否有交互作用, 误差均方 MS_e 的总体均数总是 σ^2 , 即误差均方 MS_e 总是 σ^2 的无偏估计。

4. 统计量

因为当 H_{01} 为真时, $F_A = MS_A / MS_e$ 服从 $F(df_A, df_e)$ 分布。当 H_{02} 为真时, $F_B = MS_B / MS_e$ 服从 $F(df_B, df_e)$ 分布。当 H_{03} 为真时, $F_{AB} = MS_{AB} / MS_e$ 服从 $F(df_{AB}, df_e)$ 分布。所以, 可借助 F 分布分别用 F_A, F_B, F_{AB} 三个统计量检验因子 A , 因子 B 及因子 A 与因子 B 的交互作用有无统计意义。

像单因子方差分析一样, 在多因子方差分析中也可用一张方差分析表概括地描述方差分析的结果。下面是例 1-1 的方差分析表。

表 1-5 方差分析表

变异来源	SS	df	MS	F	P 值
处理间模型	2.9625	3	0.9875	98.75	0.0000
因子 A	1.6875	1	1.6875	168.75	0.0000
因子 B	0.9075	1	0.9075	90.75	0.0000
A 与 B 的交互作用	0.3675	1	0.3675	36.75	0.0003
误差	0.0800	8	0.0100		
总的	3.0425	11	0.2766		

第三节 三因子方差分析

一、实例

[例 1-2]^[1] 某研究者以大白鼠作试验, 观察指标是肝重与体重之比(%), 主要想了解正氟醚对观察指标的作用, 同时要考察用生理盐水和用戊巴比妥作为诱导药对正氟醚毒性

作用有无影响,对不同性别大白鼠诱导物的作用有何不同,以及对不同性别大白鼠正氟醚的作用是否相同。在此研究中,涉及的三个因子及其水平如下:

A 因子(正氟醚),二个水平:不用,用

B 因子(诱导药物),二个水平:生理盐水,戊巴比妥

C 因子(大白鼠性别),二个水平:雌,雄

三个二水平因子构成了 $2 \times 2 \times 2$ 个处理,每个处理用了 3 只大白鼠作试验,结果如下表:

表 1-6 八个处理组大白鼠的肝重与体重之比

A 因子	B 因子	C 因子	肝重与体重之比(%)			总体均数
不用正氟醚	生理盐水	雌	5.26	5.68	5.83	μ_{111}
不用正氟醚	生理盐水	雄	5.00	5.52	5.38	μ_{112}
不用正氟醚	戊巴比妥	雌	5.87	5.50	6.20	μ_{121}
不用正氟醚	戊巴比妥	雄	6.13	6.46	5.21	μ_{122}
用正氟醚	生理盐水	雌	5.42	5.60	5.70	μ_{211}
用正氟醚	生理盐水	雄	6.30	7.02	5.90	μ_{212}
用正氟醚	戊巴比妥	雌	4.64	4.60	5.44	μ_{221}
用正氟醚	戊巴比妥	雄	6.02	5.70	5.48	μ_{222}

用 a, b, c 三个变量将三个因子数量化如下表:

表 1-7 三个因子数量化

因子	变量名	数量化
A	a	不用正氟醚 $a=1$, 用正氟醚 $a=2$
B	b	用生理盐水 $b=1$, 用戊巴比妥 $b=2$
C	c	雌性 $c=1$, 雄性 $c=2$

再以变量 y 表示肝重与体重之比(%),则表 1-6 的资料可转换成下表形式:

表 1-8 数量化后的数据

a	b	c	y
1	1	1	5.26
1	1	1	5.68
...	...	...	...
2	2	2	5.48

本例有三个因子,每个因子有两个水平,它们构成八个处理,相应的总体均数分别记为 μ_{abc} (其中 $a=1,2$; $b=1,2$; $c=1,2$)。据题意,想了解正氟醚用与不用对观察指标是否有

影响,相应的检验假设为:

$$H_0: \mu_{111} + \mu_{112} + \mu_{121} + \mu_{122} = \mu_{211} + \mu_{212} + \mu_{221} + \mu_{222}$$

研究用生理盐水和用戊巴比妥作为诱导物对正氟醚毒性作用有何影响,即研究因子 B 与 A 是否有交互作用,相应的检验假设为:

$$H_0: (\mu_{221} + \mu_{222}) - (\mu_{211} + \mu_{212}) = (\mu_{121} + \mu_{122}) - (\mu_{111} + \mu_{112})$$

研究对不同性别大白鼠诱导物的作用有何不同,即研究因子 C 与因子 B 是否有交互作用,相应的检验假设为:

$$H_0: (\mu_{122} + \mu_{222}) - (\mu_{121} + \mu_{221}) = (\mu_{112} + \mu_{212}) - (\mu_{111} + \mu_{211})$$

研究对不同性别大白鼠正氟醚的作用是否相同,即研究因子 C 与因子 A 有无交互作用,相应的检验假设为:

$$H_0: (\mu_{212} + \mu_{222}) - (\mu_{211} + \mu_{221}) = (\mu_{112} + \mu_{122}) - (\mu_{111} + \mu_{121})$$

三因子方差模型还可分析 A, B, C 三个因子的二级交互作用,即可研究在 A 的不同水平下 B 与 C 的交互作用是否不同;或者说在 B 的不同水平下 A 与 C 的交互作用是否不同;也可以说在 C 的不同水平下 A 与 B 的交互作用是否不同。如上例可分析对不同性别的白鼠,诱导物对正氟醚毒性作用的影响是否不同,即分析 C 的不同水平下, A, B 的交互作用是否相同,相应的检验假设为:

$$H_0: (\mu_{221} - \mu_{211}) - (\mu_{121} - \mu_{111}) = (\mu_{222} - \mu_{212}) - (\mu_{122} - \mu_{112})$$

二、STATA 程序及相应输出结果

. infile a b c y using d:exam1-2.txt (读入数据)

(24 observations read)

. anova y a b c a * b a * c b * c a * b * c (执行方差分析)

语句格式为:anova 观察变量名 因子和交互作用变量名表

输出结果:

Number of obs =		24	R-squared =		0.6110
Root MSE =		0.409649	Adj R-squared =		0.4409
①	②	③	④	⑤	⑥
Source	Partial SS	df	MS	F	Prob > F
Model	4.21818392	7	0.602597703	3.59	0.0162
a	0.002016672	1	0.002016672	0.01	0.9141
b	0.077066736	1	0.077066736	0.46	0.5077
c	0.79935039	1	0.79935039	4.76	0.0443
a * b	1.9040668	1	1.9040668	11.35	0.0039
a * c	1.33481659	1	1.33481659	7.95	0.0123
b * c	0.052266656	1	0.052266656	0.31	0.5845
a * b * c	0.048600079	1	0.048600079	0.29	0.5979
Residual	2.68499963	16	0.167812477		
Total	6.90318355	23	0.300138415		