返回总目录

目　　录

第 14 章　SAS 系统内两种处理类别数据程序概述 3

14.1　两种分析类别数据的程序 3

14.2　名 词 解 释 3

14.3　简单的随机抽样：一个母群 4

14.4　分等的简单随机抽样：多个母群 5

14.5　整个母群的观察与分析 5

14.6　采用随机分派的实验 6

14.7　等值变异数的检定：一个自变量 6

14.8　等值变异数的检定：两个或两个以上自变量 8

14.9　独立性检验 9

14.10　重复观察的实验 11

第 15 章　类别数据的归纳：统计程序 PROC FREQ 13

15.1　名 词 解 释 13

15.2　PROC FREQ 程序概述 13

15.3　如何撰写 PROC FREQ 程序 14

15.4　范　　例 20

第 16 章　类别数据的线性模型：统计程序 PROCCATMOD 26

16.1　PROC CATMOD 程序概述 26

16.2　CATMOD 程序在分析中所用到的统计模型 26

16.3　如何撰写 PROC CATMOD 程序 30

16.4　如何产生实验设计的矩阵 47

16.5　十四个范例 55

新世纪电子出版社版权所有

第三部分
类别数据的处理

希望电子出版社版权所有

第 14 章　SAS 系统内两种处理类别数据程序概述

14.1　两种分析类别数据的程序

本章介绍两个可用来分析类别数据的统计程序：一是 PROC FREQ，另一个是 PROC CATMOD，现分别简介其功能如下：

■ PROC FREQ 程序

- ●采用的文件是一个样本，一个或多个类别变量。
- ●画次数分配表或列联表 (Contingency Tables)。
- ●算出几种 (检定的) 统计值，如：χ^2 检定、费契尔的精确性测试 (Fisher's Exact Test)，相关系数。
- ●若读者指定执行分等简单的随机抽样 (Stratified Simple Random Sampling)，则 FREQ 程序可算出 CMH 统计值 (Cochran-Mantel-Haenszel)。

■ PROC CATMOD 程序

- ● 采用的文件是两个或两个以上的样本，两个或两个以上的类别变量。
- ● 将各种线性模型 (如：变异数分析、回归分析、对数线性分析、对数回归分析及重复观察分析) 应用在类别数据的函数上。
  [除了以上这两个统计程序外，第 19 章所介绍的 PROBIT 以及第 20 章介绍的 LOGISTIC 均可用来处理类别的数据。然而，因后者所用的模型属回归的线性模型，因此归入第四部分再介绍。有兴趣的读者可先行翻阅第 19，20 章，以便与 PROC FREQ，PROC CATMOD 等分析方法同时考虑。]

14.2　名 词 解 释

本节解释两个重要的名词：

■类别变量

类别变量产生的测量结果是一些不连贯的数字，这些数字之间有空隙。收集这一类数字的测量工具可以由最粗浅的名义变量 Nominal Variable (如：性别)，到最细致的绝对变量 Absolute Variable (如：家庭人口数目)；中间涵盖了次序变量 Ordinal Variable (如：对某一事件的看法分不赞成、无意见、反对) 及区间变量 Interval Variable (如：温度)。

■类别数据

类别数据来自多个观察体在一个 (或多个) 类别变量上的数值。

- ● 当类别数据来自观察体在一个类别变量上的数值时，我们可以画出次数分配表。

- 当类别数据来自观察体在两个类别变量上的数据时，我们可用矩阵的行表示第一个变量的等级 (如，性别、男、女)；以列表示另一变量的等级 (如：国籍里的中、美、日)，而形成一个二维的列联表 (Contingency Table)。
- 当类别数据来自两个以上的类别变量时，我们必须以多维的列联表表示。
- 当类别数据来自两个或两个以上的类别变量时，我们也可以将每一个观察体的值做成一个剖面图 (Profile) 如下：

	性　别	国　籍	对妇女解放运动的意见	家庭的人口数据
张三	1(男)	1 (中)	2 (不赞成)	5
李四	1 (男)	2 (美)	5 (很赞成)	3
王五	2 (女)	1 (中)	4 (赞成)	6
赵六	2 (女)	3 (日)	3 (无意见)	4

请读者注意：

类别数据多半以整数表示 (如：1 =男、2 =女)，很少有含小数的值。所以类别数据的值之间有空隙，并不连贯。

14.3　简单的随机抽样：一个母群

首先让我们以一个例子来解释：假如你希望知道一般中国人 [一个母群(One Population)] 对五种口味的喜好程度如何？你从北京市的电话簿上随手翻到两百个人 (样本) 的名字与电话号码。经过电话一一查询之后 (问："你最喜欢下列那一种口味 ?")，你发现这两百人的口味如下 (见次页)：

	北方菜	粤菜	湘菜	川菜	西餐
人数据	30	60	20	80	10
比例	15%	30%	10%	40%	5%

从上面的比例看来，川菜是最受欢迎的口味，其次是粤菜，再其次是北方菜，然后是湘菜与西餐。但上例的百分比例只是根据一组随机样本得来的资料。因此它们称为样本的统计值 (或参数的估计值)。

假设我们以 π_1，π_2，π_3，π_4 与 π_5 来代表整个中国对于这五种口味喜好的百分比，则样本统计数的大小是这五个参数的函数。统计上称此函数分配为多项式分配 (Multinomial Distribution)，它的函数表示如下：

$$\text{Prob}\,(n_1,\ n_2,\ n_3,\ n_4,\ n_5) = \frac{n!\{\pi_1^{n1}\pi_2^{n2}\pi_3^{n3}\pi_4^{n4}\pi_5^{n5}\}}{n_1!\,n_2!\,n_3!\,n_4!}$$

依上例的情况，　$n = 200$，

$n_1 = 30$，

$n_2 = 60$，

n_3 =20，

n_4 =80，

n_5 =10。

若我们限制口味成二选一的情况：北方菜或粤菜。则多项式分配会简化成二项式分配 (Binomial Distribution)。

14.4　分等的简单随机抽样：多个母群

现在假设你希望知道在中国的男 / 女 (两个母群) 对上述五种口味的喜好程度如何，则你必须修改抽样的计划；取一百位男士与一百位女士然后调查他 (或她) 们最喜好的口味。这种抽样的方法称为分等简单随机抽样 (Stratified Simple Random Sampling)。假如说男女的口味不尽相同，我们或许会得到下页的调查结果：

	北方菜	粤菜	湘菜	川菜	西餐	总数	
男	10	50	20	15	5	100	=n1
女	5	10	5	60	20	100	=n2
总数	15	60	25	75	25	200	=n

根据既定的实验设计，男女各有一百人参与了这个调查。所以上列表格的列总数 (100) 是固定的。然而行总数的值却不固定，其大小视实际调查结果而决定。若上述的数量代表实际调查的结果，则我们可说男士们一般偏爱中国传统的口味，如粤菜与湘菜。而女士们多偏向川菜与西餐等。到底男女的口味有多大的差异呢？这个问题仍可用同样的多项式分配来处理：

$$\text{Prob}\,(n_{11},n_{12},\ldots,n_{15},n_{21},n_{22},\ldots,n_{25})$$

$$= \frac{n_1!n_2!\,\pi_{11}^{n_{11}}\,\pi_{12}^{n_{12}}\ldots\,\pi_{25}^{n_{25}}}{n_{11}!n_{12}!\ldots n_{25}!}$$

每一个 π 都代表母群的参数，n 代表样本大小，下标则表示性别 (1，2) 与菜名 (1，2，3，4，5) 的组合。

分等取样所得的结果最适用于 CATMOD 程序的处理。特别是当样本的数据够多时，上述的多项式分配可以用常态分配来取代 (这个理论的基础是中央极限理论 Central Limit Theorem)。由此理论基础可导出统计的检定值称为 Wald 值。这个统计检定的样本分配近似 χ^2 分配。

14.5　整个母群的观察与分析

在这一节内，我们将讨论整个母群的观察与分析。举例来说，上海市的某商业电台准备对听众作民意调查，所用的方法是邀请听众在某个预订的周六下午主动打电话到电台，表明他们最喜欢的电台节目与他们的年龄。 没有来电话的听众则自动放弃参与的机会。所以这个调查的对象就是那些热心且主动参与这次电话调查的所有听众。从统计的角

度来看，这些人就是整个母群了。

调查的结果可用上述的行列表表示如下：

节目年龄	新闻与气象	综艺节目	小说广播	心理辅导	总数
年长的	75	300	250	12	637
年轻的	100	25	50	200	375
总数	175	325	300	212	1012

这类的资料处理通常是用 PROC FREQ，其统计的模型是超几何分布 (Hypergeometric Distribution)。在样本够大的情况下，这个分配可用常态分配来取代，因此最后导出的统计检定是用卡平方 (χ^2) 当作随机的样本分配。但我们所尝试要回答的问题是：听众的年龄到底跟他们喜欢收听的节目有何关系?

同样的民意调查可推广到哈尔滨、北京市、天津、南京、广州等地区。这么多地区所共同收集的资料同样地可利用 PROC FREQ 来分析。其理论基础是 CMH 测试 (Cochran-Mantel-Haenszel Test)，并且随机样本的分析也是利用卡平方分配。

14.6　采用随机分派的实验

假设某市立高中把自然科学组的学生分成实验班与普通班，任课的老师们完全一样；唯一不同的是，实验班的教学著重实验室的操作与自我发现，普通班的教法则采用传统式的授课、抄笔记、做家庭作业等。所有的学生以抽签的方式编入这两班之一。经过高考之后，升学的人数百分比如下：

教 学 法	上 榜 人 数	落 榜 人 数	总数
实 验 班	36	14	50
普 通 班	22	22	44
总 数	58	36	94

在这类的实验中，通常实验班与普通班人数是预订的 (所以列总数各为 50 及 44)，然而上榜与落榜的人数则不可预知。由于数据本身是类别数据，上述的表格可用 PROC FREQ 来分析，以便探讨实验班的教学方法是否比普通班更有效。

14.7　等值变异数的检定：一个自变量

此种检定可以用皮尔森的 χ^2 检定或用可能比 χ^2 检定 (Likelihood Ratio χ^2 Test) 来执行。上述的两种 χ^2 检定都包含在 PROC FREQ 统计程序里，读者只须选用 CHISQ 选项即可。现在我们试用上一节 (14.6) 的例子来说明等值变异数的检定：这个例子中有一个自变量，即教学法。此自变量下分两级：实验班及普通班。我们研究的问题是：这两班学生的上榜人数是否有显著的不同? 我们的虚无假设是：这两班学生的上榜人数是相同的。因此我们撰写如下的程序来检验这一个虚无假设：

程　序

```
DATA ONEVAR;
    INPUT METHOD $ RESULT $ COUNT @@;
    CARDS;
EXP  PASS 36     EXP FAIL 14
CTRL PASS 22    CTRL FAIL 22
;
PROC FREQ;
    WEIGHT COUNT;
    TABLES METHOD * RESULT / CHISQ;
RUN;
```

结　果

根据检定的结果，χ^2 值足以推翻虚无假设 (P≈0.029)。所以我们的结论是实验班与普通班上榜的人数在统计上有显著的不同。具体地说，实验班上榜的人数是落榜人数的两倍半，然而在普通班里，学生上榜或落榜的比例则完全一样。

报表 14.1　等值变异数的检定：一个自变量

```
    TABLE OF METHOD BY RESULT
METHOD     RESULT
Frequency |
Percent   |
Row Pct   |
Col Pct   |FAIL    |PASS    |  Total
 ---------+--------+--------+
CTRL      |     22 |     22 |     44
          |  23.40 |  23.40 |  46.81
          |  50.00 |  50.00 |
          |  61.11 |  37.93 |
----------+--------+--------+
EXP       |     14 |     36 |     50
          |  14.89 |  38.30 |  53.19
          |  28.00 |  72.00 |
          |  38.89 |  62.07 |
----------+--------+--------+
Total         36       58       94
            38.30    61.70   100.00

        STATISTICS FOR TABLE OF METHOD BY RESULT

 Statistic                      DF     Value        Prob
 Chi-Square                      1     4.794       0.029
 Likelihood Ratio Chi-Square     1     4.822       0.028
 Continuity Adj. Chi-Square      1     3.908       0.048
 Mantel-Haenszel Chi-Square      1     4.743       0.029
 Fisher's Exact Test (Left)                        0.992
 (Right)                                        2.39E-02
 (2-Tail)                                       3.48E-02
 Phi Coefficient                        0.226
 Contingency Coefficient                0.220
 Cramer's V                             0.226

 Sample Size = 94
```

希望电子出版社版权所有

14.8　等值变异数的检定：两个或两个以上自变量

假若我们将上述高考的例子推广至重点与非重点的学校。无论公非重点学校均设置实验班与普通班，则等值变异数的检定将牵涉两个自变量：一是教学法 (分实验班及普通班)，另一个是学校类别 (分重点及非重点)。

四个高中班参与实验，两班是从重点学校中选出，另两班则由非重点学校中选出。其实验的结果可能如下所示：

样 本	学校	教 学 法	上榜人数	落榜人数	总数
1	非重点	实验班	30	20	50
2	非重点	普通班	23	27	50
3	重点	实验班	29	21	50
4	重点	普通班	24	26	50
总数			106	94	200

在这个教学法研究中，我们所提出的问题比以前更详细，即：这四种班级学生的上榜人数是否有显著的不同? 我们的虚无假设是：这四种班级学生的上榜人数是相同的。同时可以预测，如果四班上榜人数有显著的不同，可能有下列四种原因：

A. 重点、非重点学校之别

B. 教学法之别

C. 学校与教学法的交互效果

D. 学校，教学法以及这二者之间的交互因素。

(最后的这种解释法又称为饱和或满秩的模型，因它把所有可能的因素都考虑进去了!)

利用 PROC CATMOD 的程序，上述 A-D 项的解释可转换成下面的 SAS 指令：

a. MODEL RESULT = SCHOOL;

b. MODEL RESULT = METHOD;

c. MODEL RESULT = SCHOOL*METHOD;

d. MODEL RESULT = SCHOOL METHOD SCHOOL*METHOD;

程　序

```
DATA A;
    INPUT SCHOOL $ METHOD $ PASS FAIL;
    CARDS;
PRIVATE EXP     30   20
PRIVATE CTRL    23   27
PUBLIC  EXP     29   21
PUBLIC  CTRL    24   26
;
DATA B;
    SET A;
    RESULT='PASS'; COUNT=PASS;OUTPUT;
    RESULT='FAIL'; COUNT=FAIL;OUTPUT;
```

希望图书出版社版权所有

```
PROC CATMOD DATA=B;
    WEIGHT COUNT;
    RESPONSE JOINT;
    MODEL RESULT=SCHOOL METHOD SCHOOL*METHOD/NOPARM;
RUN;
```

结　果

对截距与教学法 (主效果) 的 χ^2 检定均达到显著的程度 (P<0.10)；然而，就读之学校或教学法与学校的交互作用则未达显著水准。 由此可下结论说：数据显示，上 (落) 榜人数仍多半取决于老师的教学法，与学生就读的学校似乎无关。

报表 14.2　等值变异数的检定：两个或两个以上自变量

CATMOD PROCEDURE

Response: RESULT	Response Levels	(R)=	2
Weight Variable: COUNT	Populations	(S)=	4
Data Set: B	Total Frequency	(N)=	200
	Observations	(Obs)=	8

POPULATION PROFILES

Sample	SCHOOL	METHOD	Sample Size
1	PRIVATE	CTRL	50
2	PRIVATE	EXP	50
3	PUBLIC	CTRL	50
4	PUBLIC	EXP	50

RESPONSE PROFILES

Response	RESULT
1	FAIL
2	PASS

Sample	Response Function	DESIGN MATRIX 1	2	3	4
1	0.54000	1	1	1	1
2	0.40000	1	1	-1	-1
3	0.52000	1	-1	1	-1
4	0.42000	1	-1	-1	1

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	180.03	0.0000
SCHOOL	1	0.00	1.0000
METHOD	1	2.93	0.0867
SCHOOL*METHOD	1	0.08	0.7753
RESIDUAL	0	.	.

14.9　独立性检验

在这一节里，我们将介绍如何利用皮尔森的 χ^2 检定来检验一组样本里两个类别变量的关系；即：这两个变量到底是独立无关或者是互相有关连的？举例说，我们想知道加

科学出版社版权所有

工出口区的工人对妇女解放及独身子女政策这两件事的看法是否互相有关连？我们可抽样调查两百名工人，问他们两个问题。问题一：您是否赞成妇女解放运动？(是、否)。问题二：您是否认为取消独身子女政策？(应该、不应该)。请读者注意，这一类型的统计检定一般均无法预知到底有多少人会赞成 (或不赞成) 妇解运动或独身子女政策所以边际频率将由抽样调查的结果获得。总人数 (N=200) 是唯一一个可事先控制的数据。假设获得的数据如下表所列：

	独身子女政策应该保留	独身子女政策应该取消	总数
赞成妇解	98	22	120
不赞成妇解	0	80	80
总数	98	102	200

此例的虚无假设是：工人对这两件事情的看法是无关连的。现用皮尔森 χ^2 检定来决定是否应接受或拒绝此虚无假设。

程　序

```
DATA WOMEN;
    INPUT LIBERTY $ ONECHILD $ COUNT;
    CARDS;
AGREE YES 98
AGREE NO  22
DISAGREE YES 0
DISAGREE NO  80
;
PROC FREQ;
    WEIGHT COUNT;
    TABLES LIBERTY*onechild/CHISQ;
RUN;
```

结　果

χ^2 检定达到非常显著的程度 (P<0.001)，而且呈极高的关系 (绝对值=.80 或 .625)。所以，根据数据的次数分布，我们可下结论说：赞成妇解的受试者也倾向赞成无条件的堕胎 (98 人)。反之，不赞作妇解的受试者也同时不支持无条件的堕胎 (80 人)。

报表 14.3　独立性检验

```
   TABLE OF LIBERTY BY ABORTION

LIBERTY    onechild
Frequency |
Percent   |
Row Pct   |
Col Pct   |NO      |YES     |  Total
----------+--------+--------+
AGREE     |     22 |     98 |    120
          |  11.00 |  49.00 |  60.00
          |  18.33 |  81.67 |
          |  21.57 | 100.00 |
----------+--------+--------+
DISAGREE  |     80 |      0 |     80
```

希望电子出版社版权所有

```
          |  40.00 |   0.00 | 40.00
          | 100.00 |   0.00 |
          |  78.43 |   0.00 |
----------+--------+--------+
Total        102      98     200
           51.00    49.00   100.00
      STATISTICS FOR TABLE OF LIBERTY BY onechild
 Statistic                      DF       Value        Prob
 Chi-Square                      1     128.105       0.000
 Likelihood Ratio Chi-Square     1     162.840       0.000
 Continuity Adj. Chi-Square      1     124.857       0.000
 Mantel-Haenszel Chi-Square      1     127.464       0.000
 Fisher's Exact Test (Left)                        7.26E-36
 (Right)                                              1.000
 (2-Tail)                                          1.26E-35
 Phi Coefficient                          -0.800
 Contingency Coefficient                   0.625
 Cramer's V                               -0.800
 Sample Size = 200
```

14.10　重复观察的实验

假如我们设计一个实验，探讨到底是退休妇女还是职业妇女的生活比较忙碌？利用随机取样的方式，我们各取七十五名退休妇女与职业妇女作样本。选定某一周的第一天 (周一) 与最末一天 (周六) 的中午十二点来取资料。每一位受试必须诚实地回答：她在那一个时刻累不累？答案是二分的：累或不累。经过两次不同时日的观察之后，结果可能如下所示：

时　间		身　份	
周一	周六	退休妇女	职业妇女
累	累	45	35
累	不累	30	40
不累	累	27	55
不累	不累	48	20
总　人　数		75	75

在这个调查中，唯一的自变量是妇女的身份。其因变量是周一的调查结果与周六的调查结果，因有两个关系的因变量，所以这类的实验称为重复观察的实验。以上的数据经过 PROC CATMOD 处理后，分析的结果如下：

程　序

```
DATA REPEAT;
    INPUT STATUS $ MONDAY $ SATURDAY $ COUNT;
    CARDS;
RETIRE TIRED TIRED 45
WORK  TIRED TIRED 35
RETIRE TIRED NOTIRED 30
WORK  TIRED NOTIRED 40
RETIRE NOTIRED TIRED 27
WORK  NOTIRED TIRED 55
RETIRE NOTIRED NOTIRED 48
WORK  NOTIRED NOTIRED 20
```

```
;
PROC CATMOD;  WEIGHT COUNT;
    RESPONSE MARGINALS;
    MODEL MONDAY*SATURDAY=STATUS _RESPONSE_ STATUS*_RESPONSE_;
    REPEATED TIME 2/_RESPONSE_=TIME;
RUN;
```

结　果

分析的结果显示，周一与周六的数据没有显著的差异。退休妇女与职业妇女间也无明显的差异。最后，退休妇女与职业妇女在周一 (或周六) 感觉累或不累的百分比近乎相同。

报表 14.4　重复观察的实验

CATMOD

Response: MONDAY*SATURDAY	Response Levels	(R)=	4
Weight Variable: COUNT	Populations	(S)=	2
Data Set: REPEAT	Total Frequency	(N)=	300
	Observations	(Obs)=	8

POPULATION PROFILES

Sample	STATUS	Sample Size
1	RETIRE	150
2	WORK	150

RESPONSE PROFILES

Response	MONDAY	SATURDAY
1	NOTIRED	NOTIRED
2	NOTIRED	TIRED
3	TIRED	NOTIRED
4	TIRED	TIRED

Sample	Function Number	Response Function	DESIGN MATRIX 1	2	3	4
1	1	0.50000	1	1	1	1
	2	0.52000	1	1	-1	-1
2	1	0.50000	1	-1	1	-1
	2	0.40000	1	-1	-1	1

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	566.40	0.0000
STATUS	1	2.21	0.1369
TIME	1	0.96	0.3279
STATUS*_RESPONSE_	1	2.15	0.1422
RESIDUAL	0	.	.

NOTE: _RESPONSE_ = TIME

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.4800	0.0202	566.40	0.0000
STATUS	2	0.0300	0.0202	2.21	0.1369
TIME	3	0.0200	0.0204	0.96	0.3279
STATUS*_RESPONSE_	4	-0.0300	0.0204	2.15	0.1422

NOTE: _RESPONSE_ = TIME

教育电子出版社版权所有

第 15 章　类别数据的归纳：统计程序 PROC FREQ

15.1　名 词 解 释

首先，我们要用一个例子，解释本章常用到的三个名词：类别变量，类别及类别数据。假设某一个大学有两百名教员，其中男性一百二十人，女性八十人。那么「性别」就是一个类别变量，这个变量下含两个类别：男、女。若我们以 1 代表男性，2 代表女性，则这一百二十个 “1” 及八十个 “2” 就是类别数据。

15.2　PROC FREQ 程序概述

PROC FREQ 最适合用来计算类别数据出现的次数，以及类别变量之间的关系。一般而言，最简单的统计分析是建立一个次数分配表 (又称频率表)。这个次数分配表可由一个或多个类别变量所产生。下面我们用例子示范如何利用 PROC FREQ 来建立不同的次数分配表。 假设一间大学有教员两百人，他们的四个类别变量分别是：性别 SEX (下分男、女)、教育程度 EDUC (下分两个类别，即：硕士、博士)、职位 STATUS (下分讲师、副教授、教授三类别) 及年论文篇数 PUB (下分两个类别：三篇及以下，三篇以上)。下面将分别说明一元、二元、三元 (或三元以上) 的次数分配表是如何产生的。

■ 单元变量的次数分配表

现就性别这一个变量做次数分配：

程　序

```
PROC FREQ;
     TABLES SEX;
```

则你将得到一个次数分配表如下：

性别	男	女	总数
人数	120	80	200

■ 二元变量的次数分配表

现就性别及教育程度两个变量来做次数分配请读者注意：我们用星号连接两个变量：

程　序

```
PROC FREQ;
     TABLES SEX*EDUC;
```

则你将得到一个次数分配表如下：

数据电子版权所有

性别	教育程度		列次数
	硕士	博士以上	
男	40	80	120
女	40	40	80
行次数	80	120	总数=200

请读者注意：第一个变量的类别是列，第二个变量的类别是行。读者也可以要求执行 'x^2' 测试，看这两个变量之间是否有关系？虚无假设是两变量间无关系。若检验结果发现有关系，则读者可进一步要求计算它们的相关程度。

■ **三元或三元以上变量的次数分配表**

现就所有变量 (四个) 做次数分配：

程　序

```
PROC FREQ;
    TABLES SEX*EDUC*STATUS*PUB;
```

请读者注意：我们仍用星号连接各变量。这个程序将会产生四个次数分配表：每一个分配表中性别与教育程度的值均固定在一个值上，即：男硕士、男博士、女硕士、女博士。而职位 (倒数第二个变量) 的类别成列，年论文篇数 (倒数第一个变量) 的类别成行。如：

男硕士

职位	年论文篇数		列次数
	三篇及以下	三篇以上	
讲师	40	10	50
副教授	15	20	35
教授	15	0	15
行次数	70	30	总数=100

和上述二元变量的次数分配表类似的是：读者也可要求在四个二元次数分配表内 (即：男硕士、男博士、女硕士、女博士的次数表)，执行 x^2 测试，以判断职位与年论文篇数两者之间的关系。

最后，请读者注意：多元变量的次数分析极容易产生大量的报表。例如，假设有五个变量 (每一变量各有十个类别)，则这个多元变量次数分配即可能产生四千多页的报表。到底是不是每一页输出资料都是必要的？读者最好事先想清楚再分析，以免浪费时间、精力。

15.3　如何撰写 PROC FREQ 程序

PROC FREQ 含五道指令，它们的格式如下：

```
PROC FREQ    选项串;
    TABLES 次数分配表的设计 / 选项串;
    WEIGHT 变量名称;
```

```
        BY        变量名称串;
    OUTPUT OUT=输出资料文件名称    关键字符串;
```

指令 #1　PROC FREQ 选项串;

此指令有四个选项：

(1) DATA=输入文件

为输入文件命名。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 文件，对它执行分析。

(2) ORDER=FREQ

ORDER=DATA

ORDER=INTERNAL (内设值)

ORDER=FORMATTED

界定某一变量下各类别的输出次序。当 ORDER=FREQ 时，次序先后依各类别次数多少而定。次数最多的那一个类别最先，次数第二多的那一类别第二，以此类推。此方说，一组内男有一百二十人，女八十人，则此选项规定男生组是第一组，女生组是第二组。当 ORDER=DATA 时，类别次序就是它们在输入文件内出现的顺序。当 ORDER=INTERNAL 时，类别次序由英文字母先后顺序决定。如性别的两个类别，男以 MALE 代表，女以 FEMALE 代表。此选项会定女生组为第一组，男生组为第二组。当 ORDER=FORMATTED 时，类别次序由外在格式决定。当省略此选项时，内设值是 ORDER=INTERNAL。另外，遗漏数据总是排在最前面。

(3) FORMCHAR (1，2，7)=三个英文字母

指定三个字母用来画次数分配表上的各种线段。例如：读者选用 ABC，则分配表的直线将由 A (第一个字母) 构成；表上的横线将由 B (第二个字母) 构成；直线与横线交会的地方由 C (第三个字母) 构成。若选用三个空格 (即 FORMCHAR (1，2，7)= '　']，则分配表上没有线段分隔行与列。若省略此选项，则 SAS 自动订 FORMCHAR (1，2，7)='¦ - +'。

(4) PAGE

要求 FREQ 程序每页只打印一个次数分配表。当不界定此选项时，FREQ 会考虑每页空间而打印多个次数分配表。

指令 #2　TABLES 次数分配表的设计 / 选项串;

删除号 (/) 以前的次数分配表设计指一元、二元或多元变量的安排。最普通的方法是用星号连接各变量，如：A*B*C*D*E。如果须界定一个以上分配表时，读者可用括号及两横线 (--) 来简化这一部分程序的撰写，如：

TABLES A*(B C);　　等于 TABLES A*B A*C;

TABLES (A B)*(C D);　等于 TABLES A*C A*D B*C B*D;

TABLES (A B C)*D;　　等于 TABLES A*D B*D C*D;

TABLES (A--C)*D;　　也等于 TABLES A*D B*D C*D;

TABLES A--C*D;　　语意含糊，应避免使用。

改善的方法是加括号成 (A--C)*D;

TABLES (A--C);　　等于 TABLES A B C; (产生三个单元变量的次数分配表)。

若省略此选项，则 PROC FREQ 对输入文件内的变量逐一做次数分配 (即一元变量的次数分配)。

若读者要求一元变量的次数分配,然而未在删除号 (/) 后选用任何选项 (如：TABLES A;)，则 PROC FREQ 会自动计算该变量 (A) 下各类别的次数，累积次数，次数百分比，及累积次数百分比；并将这些数值包括在次数分配表内。若读者要求多元 (指二元及以上) 变量的次数分配，而且没有在删除号 (/) 后选用任何选项 (如：TABLES A*B;)，则 PROC FREQ 会自动计算两变量交集的次数，两变量交集的次数百分比 (分别以全体总次数为分母、以列次数为分母、以行次数为分母)，并将这些数值包括在次数分配表内。

现在我们来谈有关遗漏资料的处理。假设有一输入文件，内含两名受试，类别变量是性别(男=M，女=F)：

OBS	SEX
1	M
2	.

在此例中，遗漏数据的个数等于 1。这个值会印在分配表的下方。由于遗漏类别等于 F，分配表将不会包括这个遗漏类别，所以这个文件的次数分配表应是：

M	Total
1	
	2

missing subject=1

删除号 (/) 以后的选项可分下列五大类讨论。第一类选项界定统计分析的种类；第二类选项要求制造更多次数分配的数值，第三类选项控制报表的印出，第四类选项界定有关统计分析的细节，第五类选项涵盖其它选项。

第一类选项　下列选项可用来指示统计分析的种类：

(1) CHISQ

要求三个卡平方 (χ^2) 测试。这些测试包括皮尔森的 χ^2 测试，可能比 χ^2 测试及孟德-韩金 (Mantel-Haenszel) χ^2 测试。输出资料包括 ϕ 系数、列联系数 (Coefficient of Contingency)、克尔姆的 V 系数 (Cramer's V) 及费契尔的精确性测试 (Fisher's Exact Test)-- 只适用于二行乘二列的次数分配表。

(2) MEASURES

要求一系列的线性关系指标以及它们的标准误。这些指标包括皮尔森系数、斯皮尔门系数、甘玛系数、堪氏系数 (Kendall's tau-b)、史都尔系数 (Stuart's tau-c)、素摩尔系数 (Somer's D)、蓝姆达系数 (Lambda)－对称与不对称的、不定系数 (UncertaintyCoefficient)－对称与不对称的。若次数分配表是二行乘二列的，则此指令还包含奇数比 (Odds Ratio)、风险率 (Risk Ratio) 以及它们的信赖区间等。

(3) CMH

要求执行克伦-孟德-韩金 (Cochran-Mantel-Haenszel) 的统计检验。这是用来测量行变量与列变量的关系程度 (并同时控制 TABLES 指令内所有其它变量)。在二行乘二列分配表内，PROC FREQ 同时预测共同的相对风险率 (Common Relative Risk)，以及比斯罗测试 (Breslow's Test)。

(4) ALL

要求上面三种 (CHISQ，MEASURES 及 CMH) 统计测试及关系指标的计算。

(5) EXACT

要求执行费氏的标准测试 (Fisher's Exact Test)。计算的公式是根据 Mehta 及 Patel (1983) 的文献。分析的结果含双尾以及左右单尾的显著度，以利读者的解释。若观察体个数太多，则此选项不太适用。决定观察体个数的公式如下：

$\frac{n}{(r-1)(c-1)}$，此处 n = 观察体个数 r = 分配表的列数 c = 分配表的行数

若计算的结果超过 5，则读者最好不选用此选项。若结果不大于 5，则读者仍可使用此选项。当读者另外界定 ALL 的选项时，此选项不包括在内，须另外界定。

(6) AGREE

此选项计算且检定次数分类表内行分类与列分类吻合的程度，此适用于正方的分类表，若行、列数均等于 2，则此选项自动执行麦克内码检定（Mc Nemar’s test）。若行、列数大于 2，则此选项执行鲍克尔的对称检定（Bowker’s test of symmetry）。除此之外，选项 AGREE 还计算 Kappa 值及其标准误、加权后的 Kappa 值以及其标准误等。有关这些统计值的定义以及其均等的检定，读者可参考 Agresti(1990) 与 Fleiss (1981) 的文献。

第二类选项　下列六选项与分配表所含的数值有关：

(1) EXPECTED

根据独立性测试的虚无假设要求印出各细格 (即行与列的交集) 的期待次数。

(2) DEVIATION

要求印出各细格内期待次数与实际次数的差异。

(3) CELLCHI2

要求印出每一细格对卡平方统计值的贡献，这个贡献的定义如下：

$$\frac{(\text{期待次数} - \text{实际次数})^2}{(\text{期待次数})}$$

(4) CUMCOL

要求印出各细格内以列次数为分母的累积百分比。

(5)PLCORR

计算多元关系系数（ Polychoric Correlation Coefficient）。这个关系系数的统计假设是行变相值与列变相值的母群分配是一个二元的常态分配，当行、列数等于 2 时，多元关系系数也四分关系系数 (Tetrachoric Correlation Coefficient)。若二元常态分配的假设成立，则多元关系系数等于皮尔森系数的最大可能率估计值。此估

计值由牛顿-罗福森计算系统来推算，推算过程采用循环式。若循环次数达到 20 次（内设值）或读者界定的 MAXITER=次数，或收敛值小于 0.0001（内设值）或读者自定的 CON VER GE=值，则循环推算的过程完成。

(6)TOTTCT

要求列印次数的百分比，此选项只适用于三元以上的次数分配表。

(7)MISSPRINT

在多元 (指二元变量及以上) 次数分配表中，印出遗漏数据的总数。

(8) SPARSE

促使 PROC FREQ 印出 TABLES 指令中全部变量之类别间所有可能的排列组合。此选项与 LIST 选项 (见下述) 及输出文件有关。

第三类选项　下列六个选项控制报表的印出：

(1) NOFREQ

抑止印出各细格的次数。

(2) NOPERCENT

抑止印出各细格的次数百分比。

(3) NOROW

抑止印出各细格内以列次数为分母的次数百分比。

(4) NOCOL

抑止印出各细格内以行次数为分母的次数百分比。

(5) NOCUM

抑止印出各细格内的累积次数及累积百分比。

(6) NOPRINT

抑止印出次数分配表，但允许印出各种统计分析的结果，即：CHISQ，MEASURES，CMH 及 ALL 等选项所产生的分析结果。

第四类选项　下列选项界定有关统计分析的细节：

(1) SCORES=RANK

SCORES=TABLE

SCORES=RIDIT

SCORES=MODRIDIT

指明到底对何种数据执行 CMH 统计检验或计算皮尔逊关系。若 SCORES=TABLE，则指的是分配表上行与列的次数。另外三种 (RANK，RIDIT，MODRIDIT) 都是用来做无参数分析的。内设值是 SCORES=TABLE。 有关 RIDIT 的定义请见 Bross (1958)，Mack 和 Skillings (1980) 的参考文献。有关 MODRIDIT 请参考 Elteren (1960) 和 Lehmann (1975) 的著作。

(2) ALPHA=概率值

此值介于 0 与 1 之间，内设值是 .05，代表犯第一类型错误的概率。

(3) CONVERGE =极小的正值

此选项控制循环推算多元相关系数的收敛指挥，与 PLCORR 选项联用，内设值是

0.0001。

(4) MAXITER=正整数

此选项控制循环推算多元关系系数的循环次数，与 PLCORR 选项联用，内设值是 20 次。

第五类选项　下列选项属其他类：

(1) MISSING

要求 PROC FREQ 将遗漏数据包括在样本总数里，然后根据这个总数来计算次数及累积次数的百分比。

(2) LIST

要求将多元变量的次数分配表以条列式表示。当 PROC FREQ 执行独立性检验时，不能用此选项。另外，若选用此选项，则各细格的期待次数不会被印出 (即使读者选用了 EXPECTED 选项)。

(3) OUTPCT

要求下列变量的值纳入上述界定的 OUT=SAS 输出数据文件内：

PCT_COL 直行的次数百分比

PCT_ROW 横列的次数百分比

PCT_TABL 每二元次数分配表的总百分比

(4) OUTEXPECT

此选项要求 OUT=SAS 输出数据文件内含各细格的期待次数，此期待数属变量 EXPECTED 的值。

指令 #3　WEIGHT 变量名称；

这个变量的值代表每个观察体的加权值 (一般而言，每个观察体应只代表一个数据点)。这个变量的值必须是正实数，带小数点亦可。假如有一个文件，内含三个变量：种族 (RACE)、性别 (SEX) 及工作时数 (HRSWORK)；而且假设种族下有两个类别：白人、黑人。若程序是

```
PROC FREQ;
     TABLES RACE*SEX;
```

则我们会得到一个二元变量的次数分配表，它告诉我们文件内含多少位男性白人、男性黑人、女性白人、女性黑人。若再加上一道 WEIGHT 指令，其变量为各人的工作小时数 (HRSWORK)：

```
PROC FREQ;
     TABLES RACE*SEX;
     WEIGHT HRSWORK;
```

则所得的次数分配表将显示上面四种人的工作总时数。

指令 #4　BY 变量名称串；

SAS 依据此指令所列举的变量将文件分成几个小的文件，然后对每一个小的文件分别执行分析。当读者选用此指令时，文件内的数据必须先按照 BY 变量的值做由小到大的重新排列，这个步骤可借 PROC SORT 达成。

指令#5 OUTPUT OUT=输出数据文件名称 关键字符串

这个指令会产生一个输出数据文件，内含 FREQ 程序的分析结果。这些结果源自 TABLES 指令下的选项如 AGREE、ALL、CHISQ、CMH、EXACT、MEASURES 以及 PLCORR 等。此外，有效数据与遗漏数据的个数也可纳入此数据文件内。OUTPUT 指令在 FREQ 程序内只可执行一次，而且以最后一个 TABLES 指令的分析结果为主。

以下分别解释这个指令的两个选项：

（1）OUT=输出数据文件名称

这个选项界定 SAS 输出数据文件的档名。

（2）关键字符串

下列关键字也代表变量名称（前后加底线，如_AJCHI_），其意义如下：

关键字	意义
AJCHI	经过连续性纠正过后的卡平方值
ALL	由选项 CHISQ、CMH、MEASURES 导出的统计值，以及有效数据的个数
BDCHI	比斯罗（Breslow-Day）检定值
CHISQ	三个卡平方检定值：皮尔森的 x^2 检定，可能比 x^2 检定、孟德-韩金 x^2 检定，φ 系数、列联系数、克尔森的 V 系数，以及费切尔的正准检定值
CMH	克伦-孟德-韩金关系系数、相对风险率、对数奇数比以及纠正过后的奇数比
KAPPA	Kappa 系数值
EQKAP	Kappa 系数的等值检定
EQWKP	加权后 Kappa 系数的等值检定
MEASURES	一系列的关系系数指标如：皮尔森系数、斯皮尔门系数、甘玛系数、堪氏系数、史杜尔系数、索摩尔系数、蓝姆达系数（对称与不对称的）、不定系数（对称与不对称的）等。若次数分配表是二行乘二列的，则此关键字尚包括奇数比、风险率以及它们的信赖区间等。
PLCORR	多元相关系数

值得读者注意的是：在界定这些关键字之前，读者必须先在 TABLES 指令下界定其关系的选项（如 PLCORR），否则，这些关键字的值不存在。若列联表的向度是三元以上，如 TABLES A*B*C/KAPPA；则 Kappa 值是针对每一个 A 值下 B*C 二元列联表导出的。因此，变量 A 及其值亦被纳入 OUT=的输出数据文件内。同理，任何 By 变量及其值亦被纳入输出数据文件。

15.4　范　　例

例一：学历与公务人员鉴定考试的关系研究

九名公务人员参加鉴定考试，他们学历 (EDUC) 可大致分成三类：1=低，2=中，3=

高。考试的结果分成六等 (1—6)，等级愈高，成绩愈好。让我们看看这两个变量间有无关系。由于人数太少，所以指令 TABLES 后面加 Fisher's Exact Test (以选项 EXACT 表示)。

程　序

```
DATA CHI;
    DO EDUC=1 TO 3;
       DO LEVEL=1 TO 6;
          INPUT WT @@; OUTPUT;
       END;
    END;
   CARDS;
2 0 0 0 0 0
0 1 2 0 0 0
0 0 0 1 2 1
;
PROC FREQ; WEIGHT WT; TABLES EDUC*LEVEL/EXACT;
    TITLE 'FISHER"S EXACT TEST FOR 3 BY 6 TABLE';
RUN;
```

结　果

分析的结果显示，这九名公务员的学历与他们考试的结果是息息关系的 (Fisher's 测试值的统计显著度是 0.0238)。由于人数不足，所以卡平方的测试结果不可靠。

报表 15.1　学历与公务人员鉴定考试的关系研究

```
                    FISHER"S EXACT TEST FOR 3 BY 6 TABLE

                          TABLE OF EDUC BY LEVEL
EDUC     LEVEL
Frequency|
Percent  |
Row Pct  |
Col Pct  |       1 |      2 |      3 |      4 |      5 |      6 |  Total
---------+--------+--------+--------+--------+--------+--------+
      1  |      2 |      0 |      0 |      0 |      0 |      0 |      2
         |  22.22 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  22.22
         | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
         | 100.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
---------+--------+--------+--------+--------+--------+--------+
      2  |      0 |      1 |      2 |      0 |      0 |      0 |      3
         |   0.00 |  11.11 |  22.22 |   0.00 |   0.00 |   0.00 |  33.33
         |   0.00 |  33.33 |  66.67 |   0.00 |   0.00 |   0.00 |
         |   0.00 | 100.00 | 100.00 |   0.00 |   0.00 |   0.00 |
---------+--------+--------+--------+--------+--------+--------+
      3  |      0 |      0 |      0 |      1 |      2 |      1 |      4
         |   0.00 |   0.00 |   0.00 |  11.11 |  22.22 |  11.11 |  44.44
         |   0.00 |   0.00 |   0.00 |  25.00 |  50.00 |  25.00 |
         |   0.00 |   0.00 |   0.00 | 100.00 | 100.00 | 100.00 |
---------+--------+--------+--------+--------+--------+--------+
Total          2        1        2        1        2        1        9
           22.22    11.11    22.22    11.11    22.22    11.11   100.00
```

```
FISHER"S EXACT TEST FOR 3 BY 6 TABLE
          STATISTICS FOR TABLE OF EDUC BY LEVEL
Statistic                          DF      Value       Prob

Chi-Square                         10     18.000      0.055
Likelihood Ratio Chi-Square        10     19.095      0.039
Mantel-Haenszel Chi-Square          1      7.114      0.008
Fisher's Exact Test (2-Tail)                        2.38E-02
Phi Coefficient                             1.414
Contingency Coefficient                     0.816
Cramer's V                                  1.000
  Sample Size = 9
  WARNING: 100% of the cells have expected counts less
        than 5. Chi-Square may not be a valid test.
```

例二：心脏动脉病的研究

本文件由 Kleinbaum，Kupper 及 Morgenstern 于 1982 年提出，内含五个变量。年龄(AGE)，大致分两组：55 岁以下，55 岁及以上；心电图的不正常状况 (ECG)：YES=不正常，NO=正常；是否得心脏病 (CHD)：YES=得病，NO=没得病，血脂肪 (CAT)：YES=高，NO=低。

此例主要在探讨 CAT 与 CHD 的关系，同时控制受试者的年龄及 ECG。分析结果显示两者确有关系 (显著度=0.04)。血脂肪高的人比血脂肪低的人更容易得心脏病 (1.7 倍的概率)。

程　序

```
DATA CHD;
    INPUT AGE $ ECG $ CHD $ CAT $ WT;
    CARDS;
<55  0  YES  YES   1
<55  0  YES   NO  17
<55  0   NO  YES   7
<55  0   NO   NO 257
<55  1  YES  YES   3
<55  1  YES   NO   7
<55  1   NO  YES  14
<55  1   NO   NO  52
55+  0  YES  YES   9
55+  0  YES   NO  15
55+  0   NO  YES  30
55+  0   NO   NO 107
55+  1  YES  YES  14
55+  1  YES   NO   5
55+  1   NO  YES  44
55+  1   NO   NO  27
;
PROC FREQ ORDER=DATA;
```

东南大学出版社版权所有

```
        WEIGHT WT;
        TABLES AGE*ECG*CAT*CHD/NOPRINT CMH;
        TITLE 'EXAMPLE 17.9 FROM KLEINBAUM, KUPPER,AND MORGENSTERN,
        P.353';
    RUN;
```

结　果

报表 15.2　心脏动脉病的研究

EXAMPLE 17.9 FROM KLEINBAUM, KUPPER, AND MORGENSTERN,P.353

SUMMARY STATISTICS FOR CAT BY CHD

CONTROLLING FOR AGE AND ECG

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	4.153	0.042
2	Row Mean Scores Differ	1	4.153	0.042
3	General Association	1	4.153	0.042

EXAMPLE 17.9 FROM KLEINBAUM, KUPPER, AND MORGENSTERN,P.353

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Bounds	
Case-Control	Mantel-Haenszel	1.891	1.025	3.490
(Odds Ratio)	Logit	1.906	1.030	3.526
Cohort	Mantel-Haenszel	1.696	1.020	2.818
(Col1 Risk)	Logit	1.712	1.032	2.840
Cohort	Mantel-Haenszel	0.900	0.814	0.996
(Col2 Risk)	Logit	0.904	0.811	1.008

The confidence bounds for the M-H estimates are test-based.

EXAMPLE 17.9 FROM KLEINBAUM, KUPPER, AND MORGENSTERN,P.353

Breslow-Day Test for Homogeneity of the Odds Ratios

Chi-Square =　0.164　　DF =　3　　Prob = 0.983

Total Sample Size = 609

例三：2×2 次数分配表的分析

本例旨在示范 2×2 次数分配表的分类一致性。数据来自 17 位大学生的学习与背景资料，存在磁盘 DATA 子目录下的 MYDATA.DAT 文件。每一位学生的资料以下列的变项值代表：

ID=学生的识别号

SEX=性别

AGE=年龄

GPA=学业平均成绩，以 1-4 点表之

CRITICAL=分析能力

POLPREF=赞成自由党或民主党的政治立场

SATV=SAT 的语文测试分数

GRADE=前述 GPA 变项的二分变项，GPA≧3 时，GRADE=’HIGH’（高分），GRADE=’LOW’（低分）。

VERBAL=前述 SATV 变项的二分变项，SATV≧500 时，VERBAL=’HIGH’（高分），VERBAL=’LOW’（低分）。

根据 GRADE 与 VERBAL 二分变项的分数，我们将十七位大学生一一扫入 2×2 表中的一细格内，然后界定选项 AGREE 以便计算这两个二分变项分类的一致性，并执行麦克内玛检定。

程　式

```
OPTIONS NODATE LS=80;
TITLE ‘EXAMPLE 15.3 HOW MUCH AGREEMENT ES THERE IN A 2 X 2 TABLE?’;
DATA AGREE;
    INFILE ‘A:\DATA\MYDATA.DAT’;
INPUT ID SEX $ AGE GPA CRITICAL POLPREF $ SATV;
    IF GPA=. OR SATV=. THEN DELETE;
    IF GPA GE 3 THEN GRADE=’HIGH’;
    IF GPA LT 3 THEN GRADE=’LOW’;
    IF SATV GE 500 THEN VERBAL=’HIGH’;
    IF SATV LT 500 THEN VERBAL=’LOW’;

PROC FREQ DATA=AGREE;
    TABLES GRADE*VERBAL/AGREE;
RUN;
```

结　果

报表显示，麦克内玛检定值=5，其显著度=0.025。分类一致性以 Kappa 值表之（=0.459），其标差误（0.171）以及 95%的信赖区间（0.123~0.794）均显示大学生的平均成绩与其语文能力关系甚高。

一般而言，若两个二分变项对一组数据的分类是完全一致的，亦即所有的数据都分入 2×2 表的对角线上，则 Kappa=1.00。反之，若对角线上两细格{（High，High），（Low，Low）组}内的频率低于边际频率相乘后所导出的期待次数，则 Kappa 值下跌至 0 与-1 之间。有关 Kappa 的定义，读者可参阅 Cohen 1960 年的文章。

报　表 15.3　2×2 次数分配表的分析

```
Example 15.3 How much agreement is there in a 2 X 2 table?        2

                    TABLE OF GRADE BY VERBAL

          GRADE     VERBAL

          Frequency
          Percent
          Row Pct
          Col Pct   IGH      LOW       Total
          HIGH            6        0        6
                      35.29     0.00    35.29
                     100.00     0.00
                      54.55     0.00
          LOW             5        6       11
                      29.41    35.29    64.71
                      45.45    54.55
                      45.45   100.00
          Total          11        6       17
                      64.71    35.29   100.00

       STATISTICS FOR TABLE OF GRADE BY VERBAL
                          McNemar's Test
           Statistic = 5.000        DF = 1        Prob = 0.025
                       Simple Kappa Coefficient
                            95% Confidence Bounds
  Kappa =  0.459     ASE = 0.171        0.123     0.794

  Sample Size = 17
```

希望电子出版社版权所有

第 16 章　类别数据的线性模型：统计程序 PROC CATMOD

16.1　PROC CATMOD 程序概述

CATMOD 程序的全名是 CATegorical data MODeling (或作类别数据的线性模型) 。所采用的线性模型包括变异数分析模型、回归分析模型、对数线性分析、逻辑斯谛回归分析，以及重复观察的实验设计 (亦是变异数分析的一种) 等等。各式模型的参数估计采用下列的方法：

- 与对数有关的线性模型是用最大可能率 (Maximum Likelihood Ratio)的方法来估计参数：
- 一般的线性模型采用加权最小误差平方 (Weighted Least Squares) 的解法来估计参数。

适用于 CATMOD 程序的数据一般可套入行列式的型态：列代表样本，行代表因变量下所分的各类别。下面是一个典型的例子 (类别变量是「高考放榜结果」)：

样本	上榜人数	落榜人数
1.非重点学校，实验班	30	20
2.非重点学校，普通班	23	27
3.重点学校，实验班	29	21
4.重点学校，普通班	24	26

除此之外，CATMOD 程序也接受正方形的共变异数矩阵以及长方形的原始数据 (以个人为单位) 为合法的输入文件形态。下节中就 CATMOD 程序分析时所用到的统计模型作简单的说明。

16.2　CATMOD 程序在分析中所用到的统计模型

■ 线性模型

这个模型适用于含因变量 (如 Y1，Y2) 与自变量 (如 X1，X2) 的实验设计。下面举一例说明 CATMOD 程序的语法：

```
PROC CATMOD;
     WEIGHT WT;
     RESPONSE MARGINALS;
     MODEL Y1*Y2=X1 X2;
QUIT;
```

根据这个程序的指令，因变量 Y1 与 Y2 的边际概率可用自变量 X1 与 X2 的主效果来解释。

■ 对数线性模型

此模型适用于只含因变量的实验设计。下面的程序示范这种模型的撰写方式：

```
PROC CATMOD;
     WEIGHT WT;
     MODEL R1*R2*R3=_RESPONSE_/PRED=FREQ;
     LOGLIN R1|R2 R3;
QUIT;
```

根据这个程序的设定，R1，R2，R3 均为因变量，R1 | R2 R3 的意思是模型应考虑三个反应变量的主效果以及 R1*R2 的二元交互作用。CATMOD 程序执行这种模型时，是用最大可能率法来估计参数的值以及上述程序中所要求的细格频率之估计值。

■ 逻辑斯谛回归模型

适用于含一个名义因变量 (如 R) 以及数个自变量 (如 X1，X2) 的实验设计。下面是一个示范的程序语法：

```
PROC CATMOD;
     WEIGHT WT;
     DIRECT X1 X2;
     MODEL R=X1 X2;
QUIT;
```

根据上面的程序，CATMOD 程序会以最大可能率法估计这个回归模型的参数，即：R 的截距以及 X1，X2 的回归系数。若 X1 与 X2 可当作连续变量来处理，则读者最好用 LOGISTIC (第 20 章) 或 PROBIT (第 19 章) 程序来分析数据。

■ 重复观察的模型

顾名思义，这一类的模型适用于重复观察的实验设计。比方说，R1，R2，R3 代表三个不同时间点 (TIME) 上所收集到的资料，它们与 TIME 以及另一个自变量 (A) 之间的线性关系可用下面的程序来分析：

```
PROC CATMOD;
     WEIGHT WT;
     RESPONSE MARGINALS;
     MODEL R1*R2*R3=_RESPONSE_ |A;
     REPEATED TIME 3 / _RESPONSE_=TIME;
QUIT;
```

希望电子出版社版权所有

■ 变异数分析的模型

这一类的模型适合用来解释数个自变量 (如 A，B) 之主效果以及交互效果与因变量 (如 Y) 的平均数之间的关系 (参见下面的程序)：

```
PROC CATMOD;
     WEIGHT WT;
     RESPONSE MEAN;
     MODEL R=A|B;
QUIT;
```

■ 线性回归模型

这一类模型适合用来解释因变量 (如 Y1，Y2) 的边际概率与一组数值自变量 (如 X1，X2)之间的线性关系。下面的程序示范这种模型的应用：

```
PROC CATMOD;
     WEIGHT WT;
     DIRECT X1 X2;
     RESPONSE MARGINALS;
     MODEL R1*R2=X1 X2;
QUIT;
```

■ 次序变量的逻辑斯谛模型

这种模型适合用来分析次序变量(如 R)与一个连续变量(如 A)之间的线性关系。下面是一个程序的示范：

```
PROC CATMOD;
     WEIGHT WT;
     RESPONSE CLOGITS;
     MODEL R=_RESPONSE_ A;
QUIT;
```

根据这个程序的指令，次序变量(R)的值首先转为累积的对数奇数比，然后 CATMOD 程序利用加权最小平方误差法来估计模型中所有的参数。

■ 条件式逻辑斯谛模型

这一类模型适用于实验组与控制组的配对实验设计。在下面的程序里，我们考虑两个二分的变量：EXP1 与 EXP2；它们分别代表实验组与控制组对两种疾病产生反应的频率差异。CATMOD 程序利用最大可能率法可估计出这两种疾病发病的奇数比：

```
DATA;
  INPUT E1CASE E1CTRL E2CASE E2CTRL CASE WEIGHT;
```

东南大学电子出版社版权所有

```
  EXP1=E1CASE - E1CTRL;
  EXP2=E2CASE - E2CTRL;
CARDS;
[含频率次数的数据]
PROC CATMOD;
    WEIGHT WEIGHT;
    POPULATION E1CASE E1CTRL E2CASE E2CTRL EXP1 EXP2;
    DIRECT EXP1 EXP2;
    MODEL CASE=EXP1 EXP2 / NOINT;
QUIT;
```

■ **样本调查的分析**

这一类的分析方式适用于分析数个 (如 10 个) 反应函数与连续自变量 (如 A，B) 之间的关系。请看下面程序的示范语法：

```
PROC CATMOD;
     RESPONSE READ B1-B10;
     MODEL _F_=_RESPONSE_;
     FACTORS A 2 , B 5 / _RESPONSE_=A B;
QUIT;
```

■ **线性模型与对数线性模型的比较**

这两种模型各有不同的功能和适用范围：

＊线性模型

- 将自变量，因变量划分得非常清楚。
- 主要的功能在于估计参数及统计假设的检定。
- 最适合用来探讨概率间差异的显著程度，执行重复观察的分析，以及检验边际概率是否同等，但不适合用来探讨变量间一般线性的独立关系。
- 由 Grizzel，Starmer 及 Koch 等提倡，相当于将变异数分析模型 (ANOVA Model) 应用在类别数据上。

＊对数线性模型

- 并不事先区分何者为自变量，何者为因变量。
- 主要的功能在于建立模型，执行适合度检验 (Goodness-of-Fit)，且估计列联表中各细格的期待次数。
- 最适合用来探讨变量间一般线性独立的关系，但不适用于概率的比较、执行重复观察的分析或检验边际概率是否相等。
- 由 Bishop，Fienberg，Holland 提倡。

综合以上对 CATMOD 程序的简介，我们可下结论说：CATMOD 程序适合处理类别数据有关：

甲、线性独立的关系
乙、模型适合度检定
丙、同等边际概率的假设
丁、回归模型的有效度
戊、主效果与交互效果在变异数分析法中的统计显著度等
随着以上这几种分析数据的取向，CATMOD 程序会同时估计以下几类参数的值：
共变异数矩阵
累积或通用对数比
边际概率或边际平均数
每一细格内的预测频率或概率
反应函数的预测值

16.3　如何撰写 PROC CATMOD 程序

PROC CATMOD 含十二道指令，它们的格式如下：

```
PROC CATMOD        选项串;
     DIRECT        变量名称串;
     MODEL         反应变量=线性模型 / 选项串;
     LOGLIN        效果名称串 / 选项;
     FACTORS       自变量名称 类别... / 选项串;
     REPEATED      重复变量名称 类别... / 选项串;
     CONTRAST      '标名' 列的解释，列的解释，...;
     POPULATION    变量名称串;
     RESPONSE      函数 / 选项串;
     RESTRICT      参数=定值...;
     WEIGHT        变量名称;
     BY            变量名称串;
```

请读者注意：

- PROC CATMOD 以及 MODEL 这两个指令是必须的，不可省略。
- 如果采用 DIRECT 指令，则 DIRECT 必须在 MODEL 指令之前。
- MODEL 必须在 CONTRAST 指令之前。
- 此程序可在交谈式环境下执行。
- 在两个 RUN；指令中间，可多次使用 RESPONSE 和 CONTRAST 指令；其他指令则只可界定一次。

首先，让我们简单地说明除 PROC CATMOD 指令外，其他各指令的功能：

DIRECT　　要求 CATMOD 程序将某些自变量视为连续变量，其数值是有意义的。
MODEL　　界定实验设计的矩阵以及列联表 (Contingency Table) 的行。

数学出版社版权所有

LOGLIN　　界定对数线性模型的效果。
FACTORS　　界定实验设计矩阵中的效果以及决定反应函数的自变量。
REPEATED　界定重复观察的实验设计矩阵及其自变量。
CONTRAST　界定一个统计假设。
POPULATION 界定列联表的列。
RESPONSE　界定数据中的反应函数。
RESTRICT　将参数的估计值固定。
WEIGHT　　界定一个变量，代表观察体重复出现的次数或加权值。
BY　　用来划分输入文件成数个小文件，然后在各小文件内进行分析。

指令 #1　PROC CATMOD 选项串；

有两个选项可供选择：

(1) DATA= 输入文件名称

指出输入文件之文件名。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的文件，对其数据执行分析。

(2) ORDER=DATA

界定变量下各类别的次序就是它们在输入文件内先后出现的次序。若省略此选项，则各类别的次序以其内在的排列顺序 (如字母的先后次序或类别编号的次序) 为准。

指令 #2　DIRECT 变量名称串；

此指令指示 PROC CATMOD 直接对 DIRECT 变量的数值(而非它们所代表的类别或组别) 加以处理。比方说变量 X 有五个类别，它们之间的差别是以数量来表示，而且这些差别对统计的分析有影响，则读者可用下列的程序来表达：

```
PROC CATMOD;
     DIRECT X;
     MODEL Y=X;
```

这样的写法将导致实验设计矩阵里自变量 X 只有一行，而非四行。并且分析时，CATMOD 程序直接利用 X 的数值，而非 1，2，3，4，5 的类别次序。

若读者在交谈式的环境下执行 CATMOD 程序，则可重复界定 DIRECT 指令，每次设定不同的变量值 (请参考下页的例子)：

```
PROC CATMOD;
     DIRECT X;
     MODEL Y=X;
RUN;
     DIRECT ;
     MODEL Y=X;
RUN;
```

当 CATMOD 程序执行这个程序时，第一个 DIRECT 指令直接接受 X 变量的值，然后用在 MODEL 的指令上。第二个 DIRECT 指令并没有提到 X 变量的名称，因此 CATMOD 程序在执行第二个 MODEL 指令时，会根据 X 的自由度 (如 4) 而自动产生等数目的 (如 4 个) 虚无变量 (Dummy Variable)。

指令 #3　MODEL 反应变量＝线性模型 / 选项串；

PROC CATMOD 程序要求你必须界定一个 MODEL 指令，所以不可省略，但也不可在同一个 CATMOD 程序里界定一个以上的 MODEL 指令。

反应变量

反应变量可以是一个、多个 (以星号 '*' 相联) 或 _F_。_F_ 是 SAS 系统内内设的变量名称，它代表反应函数以及其共变异数矩阵；二者都必须经由文件输入。除 _F_ 之外，一般的反应变量也就是实验设计中的因变量，其组别数决定反应的类别数或列联表中的行数。

线性模型

线性模型的部分说明模型中的主效果 (即自变量) 或主效果间的交互效果，下面是几个线性模型的例子：

MODEL R=A B;	(主效果)
MODEL R= A B A*B;	(主效果 + 交互效果)
MODEL R= A B(A);	(镶嵌效果)
MODEL R= AB;	(二元交叉效果，亦即 A B A*B)
MODEL R= A B(A=1) B(A=2)	(定值镶嵌效果)
MODEL R*S=_RESPONSE_;	(对数线性模型)
MODEL R*S=A _RESPONSE_(A);	(镶嵌重复观察实验)
MODEL _F_=_RESPONSE_;	(直接自文件输入的反应函数)

请读者注意：反应变量＝线型模型，这一部分的写法与读者心目中的实验设计矩阵是息息关系的。若欲深刻了解其理论基础，请参阅本章第 16.4 节「如何产生实验设计的矩阵?」。现在我们说明实验设计矩阵的直接输入。当读者直接输入实验设计的矩阵时，相邻的两列应以逗号分开。整个矩阵必须有 q*s 列 (在此，q 代表每一样本所产生的反应函数，s 代表样本的个数)。第一组 q 列是第一个样本的反应函数，第二组 q 列是第二个样本的反应函数；依此类推，直到第 s 个 q 列为止。当读者直接输入实验设计矩阵时，你可以选择是否要检定某些组的参数为零的假设。若选择执行这个检定，则你必须在矩阵输入之后，分号 (;) 之前，立刻告诉 PROC CATMOD 你要检定那些组的参数。在指令的撰写上，请指明矩阵中的行编号，后接一个等号，再接一个标名 (限制在二十四个字母之内，用单引号括住)。如果你要求检定一组以上的参数，则应以逗号将各组的参数检定分开。最后，用小括号括住每一次的检定，并加上分号。

现在，让我们举一个例子来说明如何撰写这种假设检定的语句：

```
MODEL R=(1   1    0    0,
         1   1    0    1,
```

```
            1   1   0    2,
            1   0   1    0,
            1   0   1    1,
            1   0   1    2,
            1  -1  -1    0,
            1  -1  -1    1,
            1  -1  -1    2) (1='INTERCEPT',
                               2 3='GROUP MAIN EFFECT',
                               4='LINEAR EFFECT OF TIME');
```

如果你直接输入实验设计矩阵，但并不想检定任何参数，则 CATMOD 程序会自动检定 MODELMEAN 的效果 (MODELMEAN 表示此模型和平均数相比，到底能额外解释多少数据?)。若使用上例的矩阵，只要在矩阵后，分号前加上

(2 3 4='MODELMEAN')

即可自动要求 CATMOD 程序检定 MODELMEAN 的效果。

删除号 (/) 后的选项串可分三大类来讨论；第一类选项与分析结果的呈现方式有关；第二类选项与计算过程有关；第三类选项与报表打印的抑止有关。

第一类选项　与分析结果的呈现方式有关，有十个选项：

(1) ONEWAY

为每一个参与分析的变量建立一个单元变量的次数分配表。此表可以告诉我们此变量下各类别出现的次数。

(2) FREQ

为每一对参与分析的变量，建立一个二元变量的次数分配表。

(3) PROB

印出上述二元分配表的期待概率。在同一个样本内，这些概率值的总和应该等于 1。

(4) XPX

印出 $X'S^{-1}X$ 矩阵，也就是正规方程序 (Normal Equations) 所产生的向量内乘积 (Cross Product)。

(5) COV

印出 S_i 矩阵，也就是每一个样本所产生的反应函数值的共变异数矩阵。

(6) COVB

印出参数估计值的共变异数矩阵。

(7) CORRB

印出参数估计值的关系系数矩阵。

下列（8）与（9）选项中，读者只可选用一项：

(8) PREDICT (或 PRED)

印出每一样本中反应函数的实值和期待值，它们之间的误差（误差＝实值－期待值）及标准误。

(9) PRED=PROB 或

东华书局股份有限公司版权所有

PRED=FREQ

当反应函数值等于通用对数时，选项 PRED=PROB 印出各细格（即行与列的交集）内的期待概率；而 PRED=FREQ 则印出各细格内的期待次数。二者不可同时使用。

(10) TITLE='与模型相对应的标题'

这个选项界定一个与模型相对应的标题（前后用单引号括住），如此可帮助读者辨认报表上打印的结果。

第二类选项　与计算过程有关，下分八个选项：

(1) ADDCELL= 正实数 (如 5)

每一细格内的次数会自动增加 5（或正实数的值）。此选项不会影响最大可能比(ML)的方法，但可帮助加权最小误差平方的解法（GLS）如：当某细格内的次数等于 0 时，这个选项可用来避免 0 的对数无解的问题。

(2) AVERAGED

界定两件事：(甲）因变量的结果是可以用模型表示的；（乙）自变量的主要效果是某一样本中所有反应函数的平均数。若读者直接输入实验设计矩阵或在 MODEL 指令中使用_RESPONSE_ 选项，则 AVERAGED 选项自动生效。

(3) MAXITER= 正整数

此选项决定最大可能比的方法 (ML) 中循环 (Iterations) 分析的次数，内设值是 20。

(4) EPSILON= 正实数

此选项规定最大可能比的方法 (ML) 中预定的循环分析停止的标准。在循环分析的过程中，当对数可能比 (Log Likelihood Ratio) 的改变率小于这个选项的值，或当循环分析的次数已超过 MAXITER= 内订的值时，循环分析会停止。内设值是 10 的 -8 次方。

(5) ML

要求以最大可能率 (Maximum Likelihood Ratio) 的方法来估计参数，此选项只可与通用式对数值 (Generalized Logits) 联用。

(6) GLS 或 (7) WLS

这两个关键字分别代表通用最小平方误差估计法或加权最小平方误差估计法，在本程序中 GLS 与 WLS 是同义词。WLS (或 GLS) 法应用在除通用的对数奇数比 (Generalized Logit) 以外的反应函数之估计。

(8) NOINT

使模型中不包括截距。

第三类选项　下列五选项抑止部分统计值的计算 (或打印)：

(1) NODESIGN

抑止印出实验设计矩阵。

(2) NOPARM

抑止印出参数估计值及用来检定参数是否为零之假设的统计数值。

(3) NOPROFILE

抑止印出整个母群的剖面图 (Profile) 或反应剖面图 (Response Profile)。

希望图书出版社版权所有

(4) NORESPONSE

抑止 _RESPONSE_ 矩阵的打印 (只适用于对数线性模型)。

(5) NOITER

抑止循环估计过程里，每一步骤的分析结果 (只适用于最大可能率的参数估计法)。

指令 #4　LOGLIN 效果名称串／选项；

这个指令界定对数线性模型里各式的效果。撰写效果的语法与指令 #3 MODEL 中线性模型的效果部分 (亦即等号=右边的部分) 完全相同，故不赘述。读者可回头参阅 MODEL 指令的解释或本章第 16.4 节的详细说明。

下面举一个例子示范效果名称串的语法：

```
PROC CATMOD;
    MODEL A*B*C=_RESPONSE_;
    LOGLIN A|B|C @2;
RUN;
```

根据这个程序的指令，对数线性模型里应含所有的主效果 (即 A，B，C 的效果)以及二元的交互作用 (即 A*B，A*C，B*C)。

删除号 (/) 后的选项有一个，介绍如下：

(1) TITLE='与对数线性模型相对应的标题'

这个选项界定一个与对数线性模型相对应的标题 (前后用单引号括住)，如此可帮助读者辨认报表上打印的结果。

指令 #5　FACTORS 自变量名称 类别…／选项串；

这个指令界定实验设计中的各式效果以及决定反应函数的自变量。若自变量是一个文字变量，则自变量名称部分应含 "$" 符号作为宣告。若自变量下只有一个类别，则可不必提到类别数。否则，必须提到自变量下类别的数目。

删除号 (/) 后的选项有三个，分别介绍如下：

(1) TITLE='与 FACTORS 指令相对应的标题'

这个选项界定一个与 FACTORS 指令相对应的标题 (前后用单引号括住)，如此可帮助读者辨认报表上打印的结果。

(2) _RESPONSE_=实验设计的效果

这个选项所界定的实验设计的效果必须与删除号 (/) 前，FACTORS 指令所提的效果完全相同。若省略此选项，则 CATMOD 程序会自动考虑实验设计中所有可能的主效果以及交互效果。

(3) PROFILE=(一个矩阵)；

这个选项界定各自变量在反应函数上的值。若删除号 (/) 之前 FACTORS 指令提到 p 个自变量，而反应函数的数目是 q 个，则 PROFILE=所界定的矩阵必须是一个 i×p 的矩阵。在此，i 是 q 的分数 (或说 q 是 i 的倍数)。矩阵的两列应

以逗号隔开。矩阵中的元素可以是文字 (十六个字母以内) 或数值，视 FACTORS 的性质而定。不论 FACTORS 的性质如何，PROFILE= 的设定应与上述 _RESPONSE_ 的设定完全相同。下面举几个例子说明这个指令的撰写方式：

```
PROC CATMOD;
 RESPONSE READ Y1 - Y4;
 MODEL _F_=_RESPONSE_;
 FACTORS AGE 2，SEX 2/_RESPONSE_=AGE SEX;
RUN;
```

根据上述的程序，反应函数 (Y1-Y4) 分别来自四个不同的母群；这四个母群的产生是年龄 (下分两组) 与性别 (下分两组) 的所有排列组合。线性模型的设定仅含这两个自变量的主效果。

若读者认为实验的效果应属于定值镶嵌效果，亦即将性别镶嵌在年龄变量内，而且以三十岁作为分组的标准，则上述的 FACTORS 指令可改写为：

```
FACTORS AGE $ 2，SEX $ 2/
        _RESPONSE_=AGE SEX(AGE='under 30')
                       SEX(AGE='30 & over')
        PROFILE=('under 30' Male，
                 'under 30' Female，
                 '30 & over' Male，
                 '30 & over' Female);
```

若再进一步假设，实际收集来的数据中不含三十岁以下的男性，因此反应函数的数目由四减为三，CATMOD 的程序可修改如下：

```
PROC CATMOD;
     RESPONSE READ Y2 - Y4;
     MODEL _F_=_RESPONSE_;
     FACTORS AGE $ 2，SEX $ 2/_RESPONSE_=AGE SEX
     PROFILE=('under 30' Female，
              '30 & over' Male，
              '30 & over' Female);
```

指令 #6　REPEATED 重复变量名称 类别…/选项串；

这个指令的语法与前述 FACTORS 指令的语法完全相同，选项的定义也一致。唯一不同的是，REPEATED 所界定的变量是与重复的观察或测量有关的。在同一个 REPEATED 指令里，读者可同时界定好几个重复变量及其类别个数，它们之间以逗号相隔。

在下列两种情况下，读者应使用此指令：

一、当因变量不只一个，而且 MODEL 指令中又包括 _RESPONSE_ 这个关键字。

二、当你采用重复观察的实验设计，而且因变量就是重复观察的变量。

东海图书出版社版权所有

值得读者注意的是：不可同时界定指令 #4，#5，#6。这是因为它们的功用完全一样，都是用来解释每一样本中反应函数之间的变异数。

下页对这个指令各部分做详细介绍：

重复变量名称

为各重复观察的变量 (其反应函数必须是两个或两个以上) 命名。这个名字必须在八个字母之内，而且必须与其他变量有别。若重复变量是一个文字变量，则在其名称之外，应添加 "$" 的符号。

类别

界定上 N 述重复观察变量的类别，内设值是该变量每一样本的反应函数之个数。

删除号 (/) 后的选项有三个，分别介绍如下：

(1) TITLE='与 REPEATED 指令相对应的标题'

这个选项界定一个与 REPEATED 指令相对应的标题 (前后用单引号括住)，如此可帮助读者辨认报表上打印的结果。

(2) _RESPONSE_=实验设计的效果

这个选项所界定的实验设计的效果必须与删除号 (/) 前，REPEATED 指令所提的效果完全相同。若省略此选项，则 CATMOD 程序会自动考虑实验中所有有关重复变量的主效果以及交互效果。

请看下面两个例子。

例一

```
REPEATED TIME 2，TRTMENT 2;
```

例二

```
REPEATED TIME 2，TRTMENT 2 /_RESPONSE_=TIME|TRTMENT;
```

根据例一的程序，CATMOD 程序考虑两个主效果 (即 TIME 与 TRTMENT) 加上一个二元的交互效果的综合效果对解释 _RESPONSE_ 之变异数的多少。然而例二的程序则考虑这三个效果分别对_RESPONSE_ 的解释量。

(3) PROFILE=(一个矩阵)；

这个选项界定各重复变量在反应函数上的值，若删除号 (/) 之前 REPEATED 部分提到 p 个重复变量，而反应函数的数目是 q 个，则 PROFILE=所界定的矩阵必须是一个 i×p 的矩阵；在此，i 是 q 的分数 (或说 q 是 i 的倍数)。矩阵的两列应以逗号隔开。矩阵中的元素可以是文字 (十六个字母以内) 或数值，视 REPEATED 变量的性质而定。 不论 REPEATED 变量的性质如何，PROFILE= 的设定应与上述 _RESPONSE_ 的设定完全相同。下面举几个例子说明这个指令的撰写方式：

假设有一个重复观察的实验，每一被试连续受测三次 (以 T1，T2，T3 表示)，并且引用 RESPONSE MARGINAL 指令 (见指令 #9)。让我们进一步假设每一因变量含 k 个类别，则 CATMOD 程序会在每一个时间点上计算 (k-1) 个反应函数。由于这些反应函数之间的差异可能与时间有关，我们可用下面的程序来验证这个关系的可能性：

```
PROC CATMOD;
     RESPONSE    MARGINALS;
     MODEL       T1*T2*T3=_RESPONSE_;
     REPEATED    TIME3/_RESPONSE_=TIME;
```

让我们再进一步假设，在每一个时间点 (TIME) 上，每一位受试者照一次 X 光透视；其结果 (或影像) 由两位 X 光专家 (READER) 来解释。如此就产生了一个 3 (TIME) 乘 2 (READER) 的重复观察之实验设计。下面的程序可用来分析这个饱和的模型：

```
PROC CATMOD; RESPONSE MARGINALS;
     MODEL T1R1*T2R1*T1R2*T2R2*T3R1*T3R2=_RESPONSE_;
     REPEATED TIME 3,READER 2/_RESPONSE_
              =TIME READER TIME*READER;
```

请读者注意：MODEL 指令上交互作用的写法 (如：T1R1) 必须与 REPEATED 指令上变量前后出现的次序一致。

根据上面的程序，CATMOD 程序假定六个反应函数与它们对应的交互作用是这样的：

反应函数	交互作用	TIME 类别	READER 类别
1	T1R1	1	1
2	T1R2	1	2
3	T2R1	2	1
4	T2R2	2	2
5	T3R1	3	1
6	T3R2	3	2

如果前述的实验改用对数线性模型 (Log Linear Model) 来分析，则 REPEATED 指令中_RESPONSE_=的写法以 MODEL 指令中提到的因变量为　：

```
PROC CATMOD;
     MODEL TIME*READER=_RESPONSE_;
     REPEATED/_RESPONSE_=TIME|READER;
```

指令 #7　CONTRAST'标名'　列的解释，列的解释，…；

此指令的作用在于建立并检定 MODEL 指令中参数间的线性函数或 LOGLIN 指令中效果是否为零的假设。在此，每一个列的解释包括三部分，即：@n、效果名称和系数。现在，让我们对这几个部分及标名分别加以说明：

'标名'

可长达二十个字母，必须用单引号括住，是用来注解关系的线性函数，不可省略。

列的解释=@n 效果名称 系数...

@n

界定不同之反应函数所代表的不同参数组。比方说：

```
CONTRAST 'Average over Functions' @1 A 1 0 -1
                                  @2 A 1 1 -2;
```

则就第一个反应函数而言，CATMOD 程序检定 A 的第一组与第三组的差异。就第二个反应函数而言，CATMOD 程序检定 A 的第一、二组之和与第三组之两倍的差异。

效果名称

是 MODEL 指令中所列的效果之一。例如，读者可用 INTERCEPT 为截距的参数；读者亦可采用 ALL_PARMS 来代替模型中所有可能的参数。

系数

指用来形成参数间线性联合的系数。若系数的数目少于参数的数目，则不足的那些系数自动以 0 补齐。

下面示范一些例子：

假设有一个自变量 A，下有四个类别。它们相对的参数为 α_1，α_2，α_3 及 α_4。在实际的运算过程中，PROC CATMOD 只利用前三个参数。这是因为第四个参数 α_4，可用前三个表示如下：

$\alpha_4=-\alpha_1-\alpha_2-\alpha_3$

在这种情况下，统计假设 $\alpha_1=\alpha_4$ 可改写成：

$\alpha_1=-\alpha_1-\alpha_2-\alpha_3$ 或

$2\alpha_1+\alpha_2+\alpha_3=0$

所以将上式改写成 CATMOD 的 CONTRAST 指令就演变成：

```
CONTRAST '1 vs.4' A  2  1  1;
```

其中 '1 vs.4' 是标名，@n 省略。A 是效果名称(即 A 的主效果)；系数串是 2，1，1。

用同样的实验设计，让我们另外假设 $(\alpha_1+\alpha_2)/2=\alpha_3$，或者

$\alpha_1+\alpha_2-2\alpha_3=0$

则上式可改写成如下的 CONTRAST 指令：

```
CONTRAST '1&2 vs.3' A  1  1  -2;
```

同理，我们可以试著写下面的几行指令：

```
CONTRAST '1 vs.2'          A   1   -1  0;
CONTRAST '1&2 vs.4'     A  3   3   2;
CONTRAST '1&2 vs. 3&4'     A   2   2   0;
CONTRAST 'MAIN EFFECT'  A  0   1   0,
                        A  1   0   0,
                        A  0   0   1;
```

现在读者对 CONTRAST 指令应该有一些概念了。让我们进一步研究这个指令背后所影射的实验设计矩阵。回头看第一个统计假设 (即 $\alpha_1=\alpha_4$)：

```
PROC CATMOD;
      MODEL Y=A
      CONTRAST '1 vs. 4' A  2  1  1;
```

其对应的实验设计矩阵 (以下简称 C 矩阵) 是这样安排的：

C=[0　2　1　1]

在此，第一个值 (即 0) 代表因变量的截距。如果实验设计中有 B 和 A 两个自变量，而且 B 之下分三个类别，则统计模型是：

MODEL Y=B　A;

故同样的假设 $\alpha_1=\alpha_4$ 在 C 矩阵中的表示是：

C=[0　0　0　2　1　1]

在此，第一个值 (即 0) 仍然代表因变量的截距，第二与第三个 0 则代表 B 的参数。

接下来，让我们学习利用 CONTRAST 指令来表示二元 (或多元) 变量之间的交互效果。让我们再用上例中的两个变量：A 与 B。假如，A 有四个类别，B 有三个，则它们之间所可能产生的主要效果有五个 (或说自由度=5)。它们的写法如下：

```
CONTRAST 'JOINT EFFECT OF A&B' A  1  0  0,
                               A  0  1  0,
                               A  0  0  1,
                               B  1  0  ,
                               B  0  1  ;
```

值得注意的是，上述 0，1 的形态与类别是互相配合的。类别则由选项 ORDER= (见 PROCCATMOD 指令) 来决定。

最后，让我们学习使用 ALL_PARMS 这一个效果名称。它规定变量的参数个数与设计矩阵 (C) 的行数相等；它最有效的应用就是当读者把 C 矩阵直接地输入如下：

```
MODEL Y=(1  0  0  0,
         1  0  1  0,
         1  1  0  0,
         1  1  1  1);
CONTRAST 'MAIN EFFECT OF B' ALL_PARMS 0 1 0 0;
CONTRAST 'MAIN EFFECT OF C' ALL_PARMS 0 0 1 0;
CONTRAST 'B*C  INTERACTION'  ALL_PARMS 0 0 0 1;
```

指令 #8　POPULATION 变量名称串；

这些变量的类别乘积决定样本所代表的母群是如何形成的。如果略去此指令不用，则母群的数目以 MODEL 指令上所列的自变量之类别乘积 (亦即所有可能的类别排列组合)为准。这一个指令有下列两种功能：

甲、 当实验设计矩阵是读者自行输入，而且 MODEL 指令内又不含任何自变量时，POPULATION 指令是唯一可用来指认多个样本的。

乙、 当读者所用的模型是简化模型 (而非饱和模型) 时，POPULATION 指令可帮助读者导出与饱和模型相等的母群个数。

下面这个例子示范第一种功能：

```
DATA ONE;
     INPUT A $ B $ WT @@;
     CARDS;
YES YES 23 YES NO 31 NO YES 47 NO NO 50
;
PROC CATMOD;
     WEIGHT WT;
     POPULATION B;
     MODEL A=(1 0,
              1 1);
```

因为 MODEL 指令中的因变量 A 有两个类别 (即：YES 或 NO)，每一个样本应只有一个反应函数。而 B 变量也有两个类别，所以应该有两个样本。请读者注意实验设计矩阵的列数 (2) 等于反应函数的总数，故这个 MODEL 指令是正确的。如果我们在上述程序中省略 POPULATION 指令，则我们只有一个样本及一个反应函数，如此 MODEL 指令便与上述程序不相配了。

现另举一例说明 POPULATION 指令的第二种功能：

```
DATA TWO;
     INPUT A $ B $ Y WT @@;
     CARDS;
YES     YES     1   23    YES   YES   2   63
YES     NO      1   31    YES   NO    2   70
NO      YES     1   47    NO    YES   2   80
NO      NO      1   50    NO    NO    2   84
;
PROC  CATMOD;
      WEIGHT WT;
      MODEL Y = A  B  A*B;
```

上述的程序导致四个样本 (由于自变量 A 与 B 各有两个类别) 及下列的实验设计矩阵和变异数分析表：

$$X=\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

效果	自由度	χ^2 值	概率
截距	1	48.10	.0001
A	1	3.47	.0625
B	1	.25	.6186
A*B	1	.19	.6638
误差	0	.00	1.000

希望电子出版社版权所有

根据前述的分析，效果 B 及 A*B 均未达到显著水准 (概率均大于 .10)，所以下一步的分析可以试著简化 MODEL 程序，使其只包括 A 的主效果。如果现在读者改写原程序如下(原文件不变)：

```
PROC CATMOD;
     WEIGHT WT;
     MODELY=A;
```

那么我们只剩下两个样本 (因自变量 A 含两个类别)。其分析结果如下：

$$X=\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}$$

效果	自由度	$\chi 2$ 值	概率
截距	1	47.94	.0001
A	1	3.33	.0678
误差	0	0.00	1.0000

相对地，如果读者改写原程序如下 (原文件不变)：

```
PROC CATMOD;
     WEIGHT WT;
     POPULATION A B;
     MODEL Y=A;
```

则我们仍会保有四个样本 (它们由指令 POPULATION 中的 A，B 变量导出)。其分析结果如下：

$$X=\begin{pmatrix}1 & 1\\ 1 & 1\\ 1 & -1\\ 1 & -1\end{pmatrix}$$

效果	自由度	χ^2 值	概率
截距	1	47.76	.0001
A	1	3.30	.0694
误差	2	0.35	.8374

上述两个改写程序的差别在于后者 (也就是利用 POPULATION 指令的那个程序) 导致四个样本，而前者只产生两个。由于这两个改写程序所用的参数估计法十分接近，所以两者分析的结果亦应十分类似。

指令 #9　RESPONSE　函数／选项串；

此指令的目的在于界定反应概率的函数以便进行参数的检定。读者可同时使用多个 RESPONSE 指令，每一个指令会导致一个不同的分析法。下面就此指令的各部分做详细的介绍：

函　数　有八种设定的方法

(1)　MARGINAL (或 MARGINALS)

规定反应函数以 MODEL 指令中各因变量的边际概率来代表。

(2)　MEAN (或 MEANS)

规定反应函数以 MODEL 指令中各因变量的平均数为代表。请注意：此选项要求所有的因变量必须都是数值变量。

(3)　LOGIT (或 LOGITS)

是函数的内设值，规定反应函数以 MODEL 指令中各因变量之边际概率的通用对数为代表，通用对数就是取两个概率比例的对数值。这个比例的分母是因变量最后一个类别的边际概率，而分子则分别是该因变量所有其他类别的边际概率，因此，LOGIT 所引出的反应函数的数目等于各因变量的自由度。

(4)　CLOGIT (或 CLOGITS)

规定反应函数以 MODEL 指令中各因变量之累积边际概率的通用对数为代表。C_k 通用对数仍是取两个概率比例的对数值。这个比例的分子是第 k 组的累积概率 (以 C_k 表示)，分母则是 $(1-C_k)$ 。只有当因变量的组别是循序渐进时 (如教育程序从小学毕业，中学毕业，而至大学毕业)，此函数值的界定才有意义。

(5)　ALOGIT (或 ALOGITS)

规定反应函数以 MODEL 指令中各因变量之相邻两类别的 (边际概率之) 对数奇数比为代表。对数奇数比的分母是第 k 组的边际概率，分子则是第 (k+1) 组的边际概率。

(6)　JOINT

规定反应函数以联合反应概率为代表。

(7)　READ 变量名称串

这种设定的方式是藉著一个输入文件，内含所有的反应函数以及它们的共变异数矩阵。反应函数的个数就是 READ 变量串的个数。有关这个语法的示范，请参考指令 #5FACTORS 的选项 PROFILE= 的说明。

(8)　矩阵的转换

包括 LOG、EXP、矩阵运算及矩阵累加运算等。它们的定义简介如下：

运　　算	定　　义
LOG	对数
EXP	指数
矩阵运算	线性组合
矩阵累加运算	矩阵运算加上常数

LOG，EXP 的运算是显而易见的，在此不赘述。但请读者注意：转换的过程中，若牵涉到多重运算，则 CATMOD 程序由左到右逐一执行。若有两个矩阵运算相邻，且以星号 (*) 相连，则 CATMOD 程序会自动计算两者的乘积。

矩阵运算以及矩阵累加运算

矩阵运算是一系列的数字；它将矩阵的横列写在同一列上，以逗号分开。如下：

```
RESPONSE 1 0 0,0 1 0;
```

这个示范表示原数据的反应函数有三组 (以 P1，P2，P3 表之)。经过上述的矩阵运算后，这三个组就转换成两个线性独立的值 (以 F1，F2 表之)，如下所示：

$$\begin{pmatrix} F1 \\ F2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} P1 \\ P2 \\ P3 \end{pmatrix}$$

在此，P1，P2，与 P3 代表三组的反应比例，而 F1 与 F2 代表转换后的函数值。假设有一种情况：每一位学生的美工成品由两位老师来评分；这两位老师都用“上”，“中”，“下”三级来审核。则这两位评审老师之间的一致性可用下式求出：

```
RESPONSE 1  00  010   001  0 1;
```

上述语法所影射的矩阵运算是这样的：

$$F=P11+P22+P33=[1\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1]*\begin{pmatrix} P11 \\ P12 \\ P13 \\ P21 \\ P22 \\ P23 \\ P31 \\ P32 \\ P33 \end{pmatrix}$$

在此，P_{ij} 表示一位学生由第一位老师处得到 i 的成绩，由第二位老师处得到 j 的成绩。

如果牵涉到多重运算，CATMOD 程序由右到左逐一执行。比方说：

```
RESPONSE 1 -1 LOG;
```

$$\text{等于}\ F=[1\ -1]*\begin{pmatrix} LOG\ (P1) \\ \\ LOG\ (P2) \end{pmatrix}$$

也等于 F1=LOG(P1)-LOG(P2) 或

=LOG[P1/(1-P1)];

最后一种表示法也就是通用对数 (LOGIT) 的定义。由于只有两个类别，所以 P2=1-P1。

下面再举一个包含多重运算的例子：

```
RESPONSE 1 -1 EXP 1 0 0 1, 0 1 1 0 LOG;
```

等于下列的矩阵运算：

F=A*EXP[B*LOG(P)];

在此，P 是概率向量，A=[1 -1] ，而 $B=\begin{pmatrix} 1\ 0\ 0\ 1 \\ \\ 0\ 1\ 1\ 0 \end{pmatrix}$ 。

如果上例中的四个反应函数是由两个因变量分别导出的 (每个因变量各有两个类别)，则上述的写法还可以再简化成：

希望电子出版社版权所有

```
F=P11*P22 - P12*P21;
```

此式代表一个 2*2 列联表中细格人数 (或百分比) 内乘积的比例。

选项串　删除号 (/) 后的选项有三个：

(1)　OUT= 输出文件名称

这个文件含每一个样本内反应函数的实际及预测值，它们的标准误差，以及预测的误差。下列是输出文件内含的 SAS 内设的变量名称及定义：

变量	定义
SAMPLE	样本的编号
TYPE=FUNCTION 或	反应函数的预测 / 实际值
TYPE=PROB 或	每一细格的概率
TYPE=FREQ	每一细格的次数
NUMBER	反应函数 (或概率或次数) 的编号
OBS	原数据值
SEOBS	原数据值的标准误差(Standard Error)
PRED	预测值
SEPRED	预测值的标准误差
RESID	原数据值减预测值的误差。
BY	输入文件中原含的 BY 变量
因变量值	若模型通用对数线性模型，则因变量值就是反应函数值。然当 _TYPE_='FUNCTION' 时，这些因变量值自动成为遗漏数据。
自变量值	代表母群矩阵的列。

(2)　OUTSTAT= 参数的输出文件名称

这个输出文件与前一个不同，因其包含参数的预测值，以及它们的预测共变异数矩阵。它含下列的变量：

变量	定义
BY	输入文件中原含的 BY 变量
TYPE=EST 或	参数的估计
TYPE=COV	参数之间的共变异数矩阵的估计
NAME=(空白) 或	当 _TYPE_=EST
NAME=B1，B2 等	代表 _TYPE_=COV 矩阵中的参数
B1，B2 等等	估计的各参数值
METHOD=WLS 或	若参数估计的方法是加权最小平方误差法
METHOD=ML	若参数估计的方法是最大可能率法

现在举几个常见的 RESPONSE 指令，分别就其写法及结果说明之：

希望电子出版社版权所有

写法	结果
RESPONSE MARGINAL;	各因变量的边际概率
RESPONSE MEAN;	各因变量的边际平均数
RESPONSE LOGITS;	边际概率的通用式对数值
RESPONSE JOINT;	联合概率或细格概率
RESPONSE 1 -1 LOG;	1 与 -1 的对数值
RESPONSE OUT=PRED1;	通用对数及一个输出文件
RESPONSE 1 2 3;	求出三个反应类别 (以 1，2，3 表示) 的平均次数 (或平均概率)
RESPONSE READ B1-B4;	将四个反应函数及其共变异数矩阵直接自输入文件中读进来

(3) TITLE='与 RESPONSE 指令相对应的标题'

这个选项界定一个与 RESPONSE 指令相对应的标题 (前后用单引号括住)，如此可帮助读者辨认报表上打印的结果。

指令 #10　RESTRICT 参数=定值…;

这个指令限制某些参数的估计值。其语法示范如下：

```
RESTRICT B2=1 B4=0 B6 = -1;
```

根据这个程序的限制，第二个参数 (即 B2) 的估计值必须是 1，第四个 (B4) 必须是 0，第六个 (B6) 必须是 -1。其余参数值的估计也受到这三个条件的牵制。

值得读者注意的是，若你重复界定 RESTRICT 的指令，则在后的参数值取代在前的参数值。而且，如果你在交谈式环境下执行 CATMOD 程序，则新的 MODEL，LOGLIN，FACTORS，REPEATED 等指令有自动取代旧的 RESTRICT 指令之效力。这个指令只适用于主机上 6.06 或 6.07 版。

指令 #11　WEIGHT 变量名称;

此变量的值代表各细格的次数。比方说一个 2*2 的次数分配表如下所示：

	A=1	A=2
B=1	2	1
B=2	3	0

这个表的原数据可能是这样安排的

观察体	A	B
1	1	1
2	1	1
3	1	2
4	1	2
5	1	2
6	2	1

若将这组数据先经过 PROC FREQ 的处理：

```
PROC FREQ DATA=RAW;
     TABLES A*B/OUT=SUMMARY;
```

这一步所产生的输出文件含次数的资料。文件 SUMMARY 的内容如下：

OBS	A	B	COUNT
1	1	1	2
2	1	2	3
3	2	1	1

再将这个文件进一步输入 PROC CATMOD 作分析而且指明 COUNT 的值就代表各细格内观察体的次数：

```
PROC CATMOD DATA=SUMMARY;
     WEIGHT COUNT;
     MODEL A=B;
```

如此，读者当可明白 WEIGHT 变量的值就代表观察体的次数或加权值。

指令 #12　BY 变量名称串；

PROC CATMOD 依据此指令所列举的变量将文件分成几个小的文件，然后对每一个小的文件分别执行分析。当读者选用此指令时，文件内的数据必须先按 BY 变量串的值做由小到大的重新排列，这个步骤可借 PROC SORT 达成。

16.4　如何产生实验设计的矩阵

实验设计矩阵的每一横列代表一个母群，横列也是所有自变量值的排列组合。矩阵的每一直行是由 MODEL，LOGLIN，FACTORS 和 REPEATED 四道指令的效果决定。现在以四个段落讨论如何产生实验设计矩阵。

一、线性模型中的效果

MODEL，LOGLIN，FACTORS 与 REPEATED 四道指令中所含的「效果」部分，可大致分为下列六类：

六类效果

a. 交互的效果
   以一个或多个 (以星号*连接) 类别变量表之。如：A 或 A*B 或 A*B*C。
b. 镶嵌的效果
   类似上述交互的效果，但以小括号括住一部分的交互效果，如：B(A) 表示 B 被镶嵌在 A 效果里；或 C(A*B) 或 A*B(C*D)。
c. 定值镶嵌的效果

类似上述镶嵌的效果，但小括号内的交互效果后增加一个等号及数值。如：B(A=1) 或 C(A*B=1) 或 A*B(C=1*D=1)。

d. 直接的效果

在 DIRECT 指令中表示，如：DIRECT X Y;。

e. 主效果加上部分的交互效果

如 LOGLIN A|B|C @2; 表示对数线性模型中含 A，B，C 三个主效果以及所有二元的交互作用 (亦即 A*B，A*C，B*C)。

f. 直接效果与其他直接效果或交互效果的混合

如：X*Y 或 X*X*X 或 X*A*B(C*D=1)。在此，X*Y 表示直接效果和交互效果的联合。

X*A*B(C*D=1) 则表示直接效果，交互效果和定值镶嵌效果的联合。

变量次序的解释

当 PROC CATMOD 在诠释交互效果与镶嵌效果的变量时，CATMOD 程序以各变量第一次出现的次序为准。比方说，

```
MODEL R= B A A*B C(A B);
```

在此，因为 B 出现在 A 之前，故 CATMOD 程序视交互效果 A*B 为 B*A。另外 C(A B) 就是 C(A*B)，故 CATMOD 程序会将这个镶嵌效果解释成 C(B*A)。

@ 符号的使用

界定模型中交互作用的最高次方，请参见上述 (e) 的示范语法。

竖号的使用

竖号 (|) 是用来简化交互及镶嵌效果的撰写过程。一个冗长的指令，如：

```
MODEL Y=A  B  C  A*B  B*C  A*C  A*B*C;
```

可用竖号精简成　　MODEL Y=A | B | C;

竖号使用的原则如下：

a. 竖号的左、右两方各是一个效果，它们的交互效果也是一个效果。所以 A | B 表示三个效果：A，B 和 A*B。

b. 可以在一个程序中使用多个竖号。CATMOD 程序以由左至右的顺序来解释。如：CATMOD 程序视 A | B | C 为 [A | B |] | C，即 [A B A*B] | C。也就是 A B A*B C A*C A*B*C。

c. 交互与镶嵌的变量可以混合着使用 。比方说：A(B) | C(D) 代表 A*C(B D) 以及其他的效果。

d. 重复的变量会被取消。比方说：A(C) | B(C) 表示 A*B(C)，而非 A*B(C C)。

e. 互相抵触的效果，如：一个变量若同时造成交互和镶嵌的效果，则会从模型中剔除。比方说：A(B) | B(D E) 会产生一个效果是 A*B(B D E)，但这个效果是自相矛盾的，所以会从模型中剔除。

下面我们再举几个使用竖号的例子：

A | C(B)	等于	A	C(B)		A*C(B)
A(B) | C(B)	等于	A(B)	C(B)		A*C(B)
A(B) | B(D E)	等于	A(B)	B(D E)		
A | B(A) | C	等于	A	B(A)	C	A*C B*C(A)

二、一个样本产生一个反应函数

(1) 截距

当每一个样本产生一个反应函数时，除非读者在 MODEL 指令中选用 NOINT 选项，或直接输入实验设计矩阵，否则实验设计矩阵的第一行应该都是 '1'，它的对应参数是截距。

(2) 主要效果

如果一个自变量 A 有 k 个类别，则其主要效果有 (k-1) 个自由度，而且它的实验设计矩阵有 (k-1) 行。这些行所对应的参数 α_1，α_2，...，α_{k-1} 与变量的前 (k-1) 个类别相对照。最后一个类别 (第 k 个类别) 的参数则不必写出。因为它与其他所有参数的值加起来必须等于 0。此外实验设计矩阵中第 i 行在第 i 列的值必须是 1；在最尾列的值必须是 -1 。有关实验设计矩阵与自变量参数之间的关系，请看下面的例子：

自变量的类别	实验设计矩阵的行	
A	A	
1	1	0
2	0	1
3	-1	-1
B	B	
1	1	
2	-1	

(3) 交互效果 (又称互动效果)

交互效果 (如：A*B)是主要效果的横向相乘积。其自由度 (就是此交互效果的实验设计矩阵的行数) 是这两个 (或多个) 主效果的自由度相乘积。现举一例说明 (见下页)：

自变量		实验设计矩阵的行				
A 与	B 的类别	A		B	A*B	
1	1	1	0	1	1	0
1	2	1	0	-1	-1	0
2	1	0	1	1	0	1
2	2	0	1	-1	0	-1
3	1	-1	-1	1	-1	-1
3	2	-1	-1	-1	1	1

希望电子出版社版权所有

(4) 镶嵌效果

A(B)念成 "A 镶嵌在 B 效果里"，表示在 B 的每一个类别里都有 A 的主要效果。若以 n_a 与 n_b 分别代表 A 与 B 变量的类别，则 A(B) 的实验设计矩阵应有 $(n_a-1)*n_b$ 个行。利用上面二元变量的设计，让我们看它的矩阵是如何排列的：

自变量		实验设计矩阵的行			
B 与 A 的类别		A(B)			
1	1	1	0	0	0
1	2	0	1	0	0
1	3	−1	−1	0	0
2	1	0	0	1	0
2	2	0	0	0	1
2	3	0	0	−1	−1

上面实验设计矩阵代表四个线性组合：第一个组合是当 B=1 时，A1 与 A3 相比 (见矩阵的第一行)，第二个组合是当 B=1 时，A2 与 A3 相比 (见矩阵的第二行)，第三与第四组合则相当于第一与第二组合，只是 B=2 (见矩阵的第三与第四行)。

(5) 定值镶嵌效果

一般而言，定值镶嵌效果的表示法，较镶嵌效果精细。以上述的二元变量为例：我们可用 A(B=1) 和 A(B=2) 定值镶嵌效果的表示法将 A(B) 分为两部分来分别说明。这两个定值镶嵌效果的实验设计矩阵各有 (n_a-1) 个自由度。比方说：

自变量		实验设计矩阵的行			
B 与 A 的类别		A(B=1)		A(B=2)	
1	1	1	0	0	0
1	2	0	1	0	0
1	3	-1	-1	0	0
2	1	0	0	1	0
2	2	0	0	0	1
2	3	0	0	-1	-1

(6) 直接效果

直接效果的值是根据原变量的值，而非 0，+1 或 -1 等值。假如 X1 与 X2 这两个变量已在 CATMOD 程序的 DIRECT 指令中宣告成直接效果，则其实验设计的行还是保留原变量值。试看下例：

```
DIRECT X1 X2;
```

自变量		实验设计矩阵的行	
X1 与 X2 的类别		X1	X2
1	1	1	1
2	4	2	4
3	9	3	9

希望出版社版权所有

三、一个样本产生两个或两个以上的反应函数

在这种情况下，实验设计矩阵的构造依 MODEL 指令中是否选用 AVERAGED 选项而定。下面我们分别讨论这两种情况：

(1) MODEL 指令中无 AVERAGED 选项

假设自变量 A 有两个类别，而且程序如下：

```
PROC CATMOD;
    MODEL Y=A;
```

如果因变量 Y 的数据来自两个不同的样本，而且每一个样本 (以及其母群) 只有一个反应函数，则实验设计的矩阵是这样安排的：

实验设计矩阵

样　本	截　距	A
1	1	1
2	1	-1

当因变量 Y 的数据来自三个不同的样本时，每一样本产生两个反应函数。因此，前述的实验设计矩阵应改写如下：

样　本	反应函数的编号	实验设计矩阵 截　距	A
1	1	1　0	1　0
1	2	0　1	0　1
2	1	1　0	-1　0
2	2	0　1	0　-1

请读者注意：每一样本内的反应函数串总是被编在一起的。

(2) MODEL 程序中含 AVERAGED 选项

假设自变量 A 有两个类别，因变量 Y 有三个类别，而且程序如下：

```
PROC CATMOD;
RESPONSE 1 0 0 / 0 1 0;
MODEL Y=A / AVERAGED;
```

则每个母群将产生两个反应函数，所以上述的程序将导出下列的实验设计矩阵 (模型含两个自由度)：

样　本	反应函数的编号	实验设计矩阵 截　距	A
1	1	1	1
1	2	1	1
2	1	1	-1
2	2	1	-1

希望电子出版社版权所有

请读者注意：只要 MODEL 指令里含 _RESPONSE_ 关键字，则 CATMOD 程序自动假设 MODEL 指令同时也含了 AVERAGED 选项。这是因为关键字 _RESPONSE_ 本身即表示一个母群内 q 种反应函数值之间的差异。若读者的实验设计不含重复的变量 (所以，程序中无 REPEATED 的指令)，则每一个主效果的自由度将是 q-1。因此，若 q=3，则实验设计的矩阵如下所示：

实验设计矩阵

反应函数	_RESPONSE_ 的值		
1	1	0	
2	0	1	(自由度=2)
3	-1	-1	

若母群不止一个，则 _RESPONSE_ 的效果是所有母群的平均。此外，_RESPONSE_ 的效果可与其他的效果相互交叉，甚至镶嵌在另一个效果之内。若实验设计含重复的变量 (所以，程序中有 REPEATED 指令)，则关键字 _RESPONSE_ 不再出现在实验设计矩阵的报表上，取而代之的是变量的名称。比方说，一个实验设计含一个组间变量 (下称 A)，一个重复变量 (下称 TIME，分 TIME1 与 TIME2 两个时段)，而且 MODEL 指令的撰写如下：

```
PROC CATMOD;
     RESPONSE    MARGINALS;
     MODEL       TIME1*TIME2=A _RESPONSE_ A*_RESPONSE_;
     REPEATED    TIME 2 / _RESPONSE_=TIME;
```

则实验设计的矩阵是这样安排的：

实验设计矩阵的行

样本	反应函数	截距	A	TIME	A*TIME
1	1	1	1	1	1
2	2	1	1	-1	-1
3	1	1	-1	1	-1
4	2	1	-1	-1	1

然而，若前述 TIME1 与 TIME2 重复变量下，各分三个时段，则它们各自拥有两个自由度 (=3-1)。当这两个自由度与上述的两个样本相配合时，我们会得到一个 8*8 的实验设计矩阵如下：

实验设计矩阵的行

样本	反应函数	截距		A		TIME		A*TIME	
1	P(TIME1=1)	1	0	1	0	1	0	1	0
1	P(TIME1=2)	0	1	0	1	0	1	0	1
1	P(TIME2=1)	1	0	1	0	-1	0	-1	0
1	P(TIME2=2)	0	1	0	1	0	-1	0	-1

2	P(TIME1=1)	1	0	-1	0	1	0	-1	0
2	P(TIME1=2)	0	1	0	-1	0	1	0	-1
2	P(TIME2=1)	1	0	-1	0	-1	0	1	0
2	P(TIME2=2)	0	1	0	-1	0	-1	0	1

进一步而言，若设计中含两个重复变量 (如，TIME 与 PLACE)，则矩阵的设计按照 REPEATED 指令中 _RESPONSE_ 的撰写方式而来：

```
PROC CATMOD;
    RESPONSE MARGINALS;
    MODEL R11*R12*R21*R22=_RESPONSE_;
    REPEATED TIME 2，PLACE 2 / _RESPONSE_=TIME PLACE;
```

由于 TIME 和 PLACE 两变量下各分两组，因此，CATMOD 程序会导出四个反应函数；其中两个是与 TIME 的主效果有关的，另两个则与 PLACE 的主效果有关。以下是上述程序所导出的实验设计矩阵：

				实验设计矩阵的行		
反应函数	因变量	TIME	PLACE	截距	_RESPONSE_	
1	R11	1	1	1	1	1
2	R12	1	2	1	1	-1
3	R21	2	1	1	-1	1
4	R22	2	2	1	-1	-1

四、对数线性模型

一般而言，关键字 _RESPONSE_ 自动引发对数的线性模型。假设自变量 X 与 Y 各有两类别，其线性模型含主效果与交互效果。则下列的程序：

```
PROC CATMOD;
    MODEL X*Y=_RESPONSE_;
    REPEATED/_RESPONSE_=X Y X*Y;
```

自动导出四个反应的概率：P11，P12，P21 以及 P22。从这四个概率值可再导出三个对数反应函数，其定义如下：

F1=Log(P11/P22)，

F2=Log(P12/P22)，

F3=Log(P21/P22)。

从线性模型看来，原概率的对数值与自变量之间的关系如下：

$$\begin{pmatrix} \text{Log P11} \\ \text{Log P12} \\ \text{Log P21} \\ \text{Log P22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \gamma - \lambda \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

希望电子出版社版权所有

$$= \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \beta - \delta \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

在此，γ 与 β 代表参数的向量，λ 与 δ 代表标准化常数以便使概率值加起来等于 1。由于上述概率值与 F1，F2 与 F3 之间的关系，同样的线性模型可改写成：

$$\begin{matrix} F1 \\ F2 = \\ F3 \end{matrix} \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \text{Log P11} \\ \text{Log P12} \\ \text{Log P21} \\ \text{Log P22} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \beta$$

$$= \begin{pmatrix} 2 & 2 & 0 \\ 2 & 0 & -2 \\ 0 & 2 & -2 \end{pmatrix} \beta$$

因此，实验设计矩阵的排列如下：

样　本	反应函数	实验设计矩阵的行 X	Y	X*Y
1	1	2	2	0
1	2	2	0	-2
1	3	0	2	-2

让我们再讨论另一个比较简单的模型 (请比较此程序与上一个程序的 REPEATED 指令)：

```
PROC CATMOD;
    MODEL X*Y=_RESPONSE_;
    REPEATED/_RESPONSE_=X Y;
```

根据原线性模型，对数据率与参数之间的关系如下：

$$\begin{pmatrix} \text{Log P11} \\ \text{Log P12} \\ \text{Log P21} \\ \text{Log P22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{pmatrix} \gamma - \lambda \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \end{pmatrix} \beta - \delta \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

希望电子出版社版权所有

若以 F1，F2，F3 等反应函数表示，则前式可改写成：

$$\begin{pmatrix} F1 \\ F2 \\ F3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \text{Log P11} \\ \text{Log P12} \\ \text{Log P21} \\ \text{Log P22} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \end{pmatrix} \beta$$

$$= \begin{pmatrix} 2 & 2 \\ 2 & 0 \\ 0 & 2 \end{pmatrix} \beta$$

因此，实验设计矩阵的排列如下：

样　本	反应函数	实验设计矩阵的行 X	Y
1	1	2	2
1	2	2	0
1	3	0	2

为了使读者确定 β 的转换矩阵是正确的，SAS 在报表上会打印出这个转换矩阵如下：

反应函数	_RESPONSE_ 矩阵 1	2
1	1	1
2	1	-1
3	-1	1
4	-1	-1

若读者不须检验这个矩阵，则可在 MODEL 指令的选项部分添加 NORESPONSE 的关键字。这样，报表上就不会打印出这个转换矩阵。

16.5　十四个范例

例一：线性反应函数，反应类别 r 等于 2

下面数据用来研究消费者对两种清洁剂 (以 M 及 X 表示) 的偏好。因变量是消费者对牌子的选择 (M 牌或 X 牌)。自变量有三，均是类别变量：水质 (以 SOFTNESS 表示)，是否有用过 M 牌清洁剂的经验 (PREV) 及水温 (TEMP)。分析的模型也有两种：饱和模型 (含三个自变量的主效果及其交互效果) 和简化模型 (只含三个主效果)。下面是

希望电子出版社版权所有

程序和输出文件。

程　序

```
TITLE 'DETERGENT PREFERENCE STUDY';
DATA DETERG;
    INPUT SOFTNESS $ BRAND $ PREV $ TEMP $ COUNT @@;
    CARDS;
SOFT X YES HIGH 19 SOFT X YES LOW 57 SOFT X NO HIGH 29 SOFT X NO LOW 63
SOFT M YES HIGH 29 SOFT M YES LOW 49 SOFT M NO HIGH 27 SOFT M NO LOW 53
MED  X YES HIGH 23 MED  X YES LOW 47 MED  X NO HIGH 33 MED  X NO LOW 66
MED  M YES HIGH 47 MED  M YES LOW 55 MED  M NO HIGH 23 MED  M NO LOW 50
HARD X YES HIGH 24 HARD X YES LOW 37 HARD X NO HIGH 42 HARD X NO LOW 68
HARD M YES HIGH 43 HARD M YES LOW 52 HARD M NO HIGH 30 HARD M NO LOW 42
;
PROC CATMOD;
    RESPONSE 1 0;
    WEIGHT COUNT;
    MODEL BRAND=SOFTNESS|PREV|TEMP /FREQ PROB NODESIGN;
    TITLE2 'SATURATED MODEL';
PROC CATMOD;
    RESPONSE 1 0;
    WEIGHT COUNT;
    MODEL BRAND=SOFTNESS PREV TEMP / NOPROFILE;
    TITLE2 'MAIN-EFFECTS MODEL';
RUN;
```

结　果

经两种模型分析后，结果皆显示：消费者对清洁剂的偏好受过去使用经验 (Chi-Square=20.96，P=0.0000) 及水温 (Chi-Square=3.95，P=0.0468) 的影响，然而水质这个因素 (Chi-Square=0.01 或 0.23，P=0.9066 或 0.6342) 则与使用清洁剂的偏好无关。

报表 16.1　线性反应函数，反应类别 r 等于 2

```
                DETERGENT PREFERENCE STUDY
                   SATURATED MODEL
                   CATMOD PROCEDURE                    RESPONSE PROFILES
Response: BRAND            Response Levels (R)=    2      Response  BRAND
Weight Variable: COUNT      Populations     (S)=   12      ---------------
Data Set: DETERG           Total Frequency (N)=  1008          1    M
Frequency Missing: 0        Observations  (Obs)=   24          2    X
```

POPULATION PROFILES

Sample	SOFTNESS	PREV	TEMP	Sample Size
1	HARD	NO	HIGH	72
2	HARD	NO	LOW	110
3	HARD	YES	HIGH	67
4	HARD	YES	LOW	89
5	MED	NO	HIGH	56
6	MED	NO	LOW	116
7	MED	YES	HIGH	70
8	MED	YES	LOW	102
9	SOFT	NO	HIGH	56
10	SOFT	NO	LOW	116
11	SOFT	YES	HIGH	48
12	SOFT	YES	LOW	106

RESPONSE FREQUENCIES

Sample	Response 1	Number 2
1	30	42
2	42	68
3	43	24
4	52	37
5	23	33
6	50	66
7	47	23
8	55	47
9	27	29
10	53	63
11	29	19
12	49	57

RESPONSE PROBABILITIES

Sample	Response 1	Number 2
1	0.41667	0.58333
2	0.38182	0.61818
3	0.64179	0.35821
4	0.58427	0.41573
5	0.41071	0.58929
6	0.43103	0.56897
7	0.67143	0.32857
8	0.53922	0.46078
9	0.48214	0.51786
10	0.4569	0.5431
11	0.60417	0.39583
12	0.46226	0.53774

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	983.13	0.0000
SOFTNESS	2	0.09	0.9575
PREV	1	22.68	0.0000
SOFTNESS*PREV	2	3.85	0.1457
TEMP	1	3.67	0.0555
SOFTNESS*TEMP	2	0.23	0.8914
PREV*TEMP	1	2.26	0.1324
SOFTNESS*PREV*TEMP	2	0.76	0.6850
RESIDUAL	0	.	.

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.5069	0.0162	983.13	0.0000
SOFTNESS	2	-0.00073	0.0225	0.00	0.9740
	3	0.00623	0.0226	0.08	0.7830
PREV	4	-0.0770	0.0162	22.68	0.0000
SOFTNESS*PREV	5	-0.0299	0.0225	1.77	0.1831
	6	-0.0152	0.0226	0.45	0.5007
TEMP	7	0.0310	0.0162	3.67	0.0555
SOFTNESS*TEMP	8	-0.00786	0.0225	0.12	0.7265
	9	-0.00298	0.0226	0.02	0.8953

希望电子出版社版权所有

PREV*TEMP	10	-0.0243	0.0162	2.26	0.1324
SOFTNESS*PREV*TEMP	11	0.0187	0.0225	0.69	0.4064
	12	-0.0138	0.0226	0.37	0.5415

DETERGENT PREFERENCE STUDY

MAIN-EFFECTS MODEL

CATMOD PROCEDURE

Response: BRAND Response Levels (R)= 2

Weight Variable: COUNT Populations (S)= 12

Data Set: DETERG Total Frequency (N)= 1008

Frequency Missing: 0 Observations (Obs)= 24

Sample	Response Function	DESIGN MATRIX 1	2	3	4	5
1	0.41667	1	1	0	1	1
2	0.38182	1	1	0	1	-1
3	0.64179	1	1	0	-1	1
4	0.58427	1	1	0	-1	-1
5	0.41071	1	0	1	1	1
6	0.43103	1	0	1	1	-1
7	0.67143	1	0	1	-1	1
8	0.53922	1	0	1	-1	-1
9	0.48214	1	-1	-1	1	1
10	0.45690	1	-1	-1	1	-1
11	0.60417	1	-1	-1	-1	1
12	0.46226	1	-1	-1	-1	-1

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	1004.93	0.0000
SOFTNESS	2	0.24	0.8859
PREV	1	20.96	0.0000
TEMP	1	3.95	0.0468
RESIDUAL	7	8.26	0.3100

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.5080	0.0160	1004.93	0.0000
SOFTNESS	2	-0.00256	0.0218	0.01	0.9066
	3	0.0104	0.0218	0.23	0.6342
PREV	4	-0.0711	0.0155	20.96	0.0000
TEMP	5	0.0319	0.0161	3.95	0.0468

科学出版社版权所有

例二：平均数反应函数，r=3

下列的数据是用来比较四家医院对同一病症采用不同手术 (以 TRT 表示，下分 A，B，C，D 四类) 的效果。因变量是手术后病人出现并发症的严重性。这个严重性分三级：0 表示完全没有并发症，0.5 表示轻度并发症，1 表示中度并发症。由于这三级有轻重之分，所采用的反应函数自动成为平均数函数。模型采用主效果模型 (即医院、手术种类两个主效果)。

程　序

```
TITLE 'DUMPING SYNDROME DATA';
DATA OPERATE;
    INPUT HOSPITAL TRT $ SEVERITY $ WT @@;
    CARDS;
1 A NONE 23   1 A SLIGHT  7   1 A MODERATE 2
1 B NONE 23   1 B SLIGHT 10   1 B MODERATE 5
1 C NONE 20   1 C SLIGHT 13   1 C MODERATE 5
1 D NONE 24   1 D SLIGHT 10   1 D MODERATE 6
2 A NONE 18   2 A SLIGHT  6   2 A MODERATE 1
2 B NONE 18   2 B SLIGHT  6   2 B MODERATE 2
2 C NONE 13   2 C SLIGHT 13   2 C MODERATE 2
2 D NONE  9   2 D SLIGHT 15   2 D MODERATE 2
3 A NONE  8   3 A SLIGHT  6   3 A MODERATE 3
3 B NONE 12   3 B SLIGHT  4   3 B MODERATE 4
3 C NONE 11   3 C SLIGHT  6   3 C MODERATE 2
3 D NONE  7   3 D SLIGHT  7   3 D MODERATE 4
4 A NONE 12   4 A SLIGHT  9   4 A MODERATE 1
4 B NONE 15   4 B SLIGHT  3   4 B MODERATE 2
4 C NONE 14   4 C SLIGHT  8   4 C MODERATE 3
4 D NONE 13   4 D SLIGHT  6   4 D MODERATE 4
;
PROC CATMOD ORDER=DATA;
    WEIGHT WT;
    RESPONSE 0 0.5 1;
    MODEL SEVERITY=TRT HOSPITAL / FREQ ONEWAY;
    TITLE2 'MAIN-EFFECTS MODEL';
RUN;
```

结　果

分析结果显示不同的手术 (特别是前面两个) 最易导致不同程度的并发症 (Chi-Square=4.17，P=0.0411)，至于病人在哪家医院接受手术则与并发症的严重性无关 (Chi-Square=0.60，0.44，或 2.28)。

希望电子出版社版权所有

报表 16.2　平均数反应函数，r=3

DUMPING SYNDROME DATA

MAIN-EFFECTS MODEL

CATMOD PROCEDURE

Response: SEVERITY	Response Levels (R)= 3	
Weight Variable: WT	Populations (S)= 16	
Data Set: OPERATE	Total Frequency (N)= 417	
Frequency Missing: 0	Observations (Obs)= 48	

ONE-WAY FREQUENCIES

Variable	Value	Frequency
SEVERITY	NONE	240
	SLIGHT	129
	MODERATE	48
TRT	A	96
	B	104
	C	110
	D	107
HOSPITAL	1	148
	2	105
	3	74
	4	90

POPULATION PROFILES

Sample	TRT	HOSPITAL	Sample Size
1	A	1	32
2	A	2	25
3	A	3	17
4	A	4	22
5	B	1	38
6	B	2	26
7	B	3	20
8	B	4	20
9	C	1	38
10	C	2	28
11	C	3	19
12	C	4	25
13	D	1	40
14	D	2	26
15	D	3	18
16	D	4	23

RESPONSE PROFILES

Response	SEVERITY
1	NONE
2	SLIGHT
3	MODERATE

DUMPING SYNDROME DATA

MAIN-EFFECTS MODEL

Sample	Response Function	DESIGN MATRIX 1	2	3	4	5	6	7

RESPONSE FREQUENCIES

Sample	Response Number 1	2	3

教育电子出版社版权所有

1	0.17188	1	1	0	0	1	0	0	1	23	7	2
2	0.16000	1	1	0	0	0	1	0	2	18	6	1
3	0.35294	1	1	0	0	0	0	1	3	8	6	3
4	0.25000	1	1	0	0	-1	-1	-1	4	12	9	1
5	0.26316	1	0	1	0	1	0	0	5	23	10	5
6	0.19231	1	0	1	0	0	1	0	6	18	6	2
7	0.30000	1	0	1	0	0	0	1	7	12	4	4
8	0.17500	1	0	1	0	-1	-1	-1	8	15	3	2
9	0.30263	1	0	0	1	1	0	0	9	20	13	5
10	0.30357	1	0	0	1	0	1	0	10	13	13	2
11	0.26316	1	0	0	1	0	0	1	11	11	6	2
12	0.28000	1	0	0	1	-1	-1	-1	12	14	8	3
13	0.27500	1	-1	-1	-1	1	0	0	13	24	10	6
14	0.36538	1	-1	-1	-1	0	1	0	14	9	15	2
15	0.41667	1	-1	-1	-1	0	0	1	15	7	7	4
16	0.30435	1	-1	-1	-1	-1	-1	-1	16	13	6	4

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	248.77	0.0000
TRT	3	8.90	0.0307
HOSPITAL	3	2.33	0.5065
RESIDUAL	9	6.33	0.7069

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.2724	0.0173	248.77	0.0000
TRT	2	-0.0552	0.0270	4.17	0.0411
	3	-0.0365	0.0289	1.59	0.2073
	4	0.0248	0.0280	0.78	0.3757
HOSPITAL	5	-0.0204	0.0264	0.60	0.4388
	6	-0.0178	0.0268	0.44	0.5055
	7	0.0531	0.0352	2.28	0.1312

例三：对数回归分析，标准的反应函数

下列数据是有关铸铁的资料。铁块首先以不同的温度去锻链 (HEAT)，又以不同的时间浸在溶液中(SOAK)。因变量是铁块能否被卷曲的程度 (Y)。当 Y=1 时，表示铁块能

东南大学出版社版权所有

被卷曲；其个数以 COUNT 表之。当 Y=0 时，则表示铁块尚不能被卷曲，其个数也以 COUNT 代表。在下列的程序中，CATMOD 程序采用主效果 (HEAT 和 SOAK) 模型来分析这组数据。

程　序

```
TITLE 'MAXIMUM LIKELIHOOD LOGISTIC REGRESSION';
DATA INGOTS;
    INPUT HEAT SOAK NREADY NTOTAL @@;
    COUNT=NREADY; Y=1; OUTPUT;
    COUNT=NTOTAL-NREADY; Y=0; OUTPUT;
    DROP NREADY NTOTAL;
    CARDS;
 7 1.0 0 10  7 1.7 0 17  7 2.2 0  7   7 2.8 0  12
 7 4.0 0  9 14 1.0 0 31 14 1.7 0 43  14 2.2 2  33
14 2.8 0 31 14 4.0 0 19 27 1.0 1 56  27 1.7 4  44
27 2.2 0 21 27 2.8 1 22 27 4.0 1 16  51 1.0 3  13
51 1.7 0  1 51 2.2 0  1 51 4.0 0  1
;
PROC CATMOD;
    WEIGHT COUNT;
    DIRECT HEAT SOAK;
    MODEL Y=HEAT SOAK / FREQ ML NOGLS COVB CORRB;
RUN;
```

结　果

分析结果显示：铁块是否能被卷曲受锻链时温度的影响 (Chi-Square=11.95，P=0.0005)；与浸在溶液内时间的长短无关 (Chi-Square=0.03，P=0.0005)。

所求得的对数据归模型如下：

Y 的自然对数之预测值=5.5592+(-0.0820)*温度+(-0.0568)*时间。

报表 16.3　对数据归分析，标准的反应函数

```
                    MAXIMUM LIKELIHOOD LOGISTIC REGRESSION

                          CATMOD PROCEDURE        ESPONSE PROFILES

Response: Y             Response Levels (R)=  2     Response  Y

Weight Variable: COUNT  Populations      (S)= 19    -----------

Data Set: INGOTS        Total Frequency (N)=387          1    0

Frequency Missing: 0    Observations   (Obs)= 25         2    1
```

希望电子出版社版权所有

POPULATION PROFILES

Sample	HEAT	SOAK	Sample Size
1	7	1	10
2	7	1.7	17
3	7	2.2	7
4	7	2.8	12
5	7	4	9
6	14	1	31
7	14	1.7	43
8	14	2.2	33
9	14	2.8	31
10	14	4	19
11	27	1	56
12	27	1.7	44
13	27	2.2	21
14	27	2.8	22
15	27	4	16
16	51	1	13
17	51	1.7	1
18	51	2.2	1
19	51	4	1

RESPONSE FREQUENCIES

Sample	Response Number 1	2
1	10	0
2	17	0
3	7	0
4	12	0
5	9	0
6	31	0
7	43	0
8	31	2
9	31	0
10	19	0
11	55	1
12	40	4
13	21	0
14	21	1
15	15	1
16	10	3
17	1	0
18	1	0
19	1	0

MAXIMUM LIKELIHOOD LOGISTIC REGRESSION

MAXIMUM LIKELIHOOD ANALYSIS

Iteration	Sub Iteration	-2 Log Likelihood	Convergence Criterion	Parameter Estimates 1	2	3
0	0	536.49592	1.0000	0	0	0
1	0	152.58961	0.7156	2.1594	-0.0139	-0.003733
2	0	106.76066	0.3003	3.5334	-0.0363	-0.0120
3	0	96.692171	0.0943	4.7489	-0.0640	-0.0299
4	0	95.383825	0.0135	5.4138	-0.0790	-0.0498
5	0	95.345659	0.000400	5.5539	-0.0819	-0.0564
6	0	95.345613	4.8289E-7	5.5592	-0.0820	-0.0568
7	0	95.345613	7.731E-13	5.5592	-0.0820	-0.0568

希望电子出版社版权所有

```
MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE
Source              DF   Chi-Square      Prob
-------------------------------------------
INTERCEPT            1        24.65    0.0000
HEAT                 1        11.95    0.0005
SOAK                 1         0.03    0.8639
LIKELIHOOD RATIO    16        13.75    0.6171

  ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES
                              Standard    Chi-
Effect    Parameter  Estimate    Error    Square   Prob
-------------------------------------------------------
INTERCEPT       1     5.5592    1.1197     24.65  0.0000
HEAT            2    -0.0820    0.0237     11.95  0.0005
SOAK            3    -0.0568    0.3312      0.03  0.8639

          COVARIANCE MATRIX OF
     THE MAXIMUM LIKELIHOOD ESTIMATES
                1           2           3
    ---------------------------------------
     1    1.2537133   -0.0215664   -0.2817648
     2   -0.0215664    0.0005633    0.0026243
     3   -0.2817648    0.0026243    0.1097020
          CORRELATION MATRIX OF
     THE MAXIMUM LIKELIHOOD ESTIMATES
                1           2           3
    ---------------------------------------
     1    1.0000000   -0.8115214   -0.7597670
     2   -0.8115214    1.0000000    0.3338261
     3   -0.7597670    0.3338261    1.0000000
```

例四：对数线性模型，一个因变量

利用对数线性模型将例一的资料重新分析。在本例中，第一个模型仍是饱和模型（三个主效果加交互效果），第二个模型是简化模型（只含三个主效果）。

程　序

```
TITLE 'DETERGENT PREFERENCE STUDY';
DATA DETERG;
    INPUT SOFTNESS $ BRAND $ PREV $ TEMP $ COUNT @@;
    CARDS;
```

希望电子出版社版权所有

```
SOFT X YES HIGH 19 SOFT X YES LOW 57 SOFT X NO HIGH 29 SOFT X NO LOW 63
SOFT M YES HIGH 29 SOFT M YES LOW 49 SOFT M NO HIGH 27 SOFT M NO LOW 53
MED  X YES HIGH 23  MED  X YES LOW 47 MED  X NO HIGH 33 MED   X NO LOW 66
MED  M YES HIGH 47  MED  M YES LOW 55 MED  M NO HIGH 23 MED   M NO LOW 50
HARD X YES HIGH 24 HARD X YES LOW 37 HARD X NO HIGH 42 HARD X NO LOW 68
HARD M YES HIGH 43 HARD M YES LOW 52 HARD M NO HIGH 30 HARD M NO LOW 42
;
PROC CATMOD;
    RESPONSE 1 0;
    WEIGHT COUNT;
    MODEL BRAND=SOFTNESS|PREV|TEMP /FREQ PROB NODESIGN;
    TITLE2 'SATURATED MODEL';
PROC CATMOD;
    WEIGHT COUNT;
    MODEL BRAND=SOFTNESS PREV TEMP / NOPROFILE;
    TITLE2 'MAIN-EFFECTS MODEL';
RUN;
```

结　果

虽将原数据转换成对数，且用两种模型来分析，分析的结果仍与例一相同。也就是说：消费者对清洁剂的偏好受过去使用某品牌清洁剂之经验 (Chi-Square=19.28，P=0.0000) 及水温的影响 (Chi-Square=3.63，P=0.0566)，水质的因素与消费者对清洁剂的偏好无关 (Chi-Square=0.01 或 0.19)。

报表 16.4　对数线性模型，一个因变量

```
                    DETERGENT PREFERENCE STUDY

                         SATURATED MODEL

                        CATMOD PROCEDURE

Response：BRAND              Response Levels      (R)=        2
Weight Variable：COUNT       Populations          (S)=       12
Data Set：DETERG             Total Frequency      (N)=     1008
Frequency Missing：0         Observations       (Obs)=       24

     POPULATION PROFILES

         (与例一同)

     RESPONSE PROFILES

         (与例一同)

    RESPONSE FREQUENCIES

         (与例一同)

    RESPONSE PROBABILITIES

         (与例一同)

                    ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES
```

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	0.21	0.6505
SOFTNESS	2	0.10	0.9522
PREV	1	21.80	0.0000
SOFTNESS*PREV	2	3.80	0.1493
TEMP	1	3.63	0.0567
SOFTNESS*TEMP	2	0.20	0.9066
PREV*TEMP	1	2.25	0.1335
SOFTNESS*PREV*TEMP	2	0.74	0.6916
RESIDUAL	0	.	.

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.0305	0.0673	0.21	0.6505
SOFTNESS	2	-0.00418	0.0940	0.00	0.9645
	3	0.0278	0.0947	0.09	0.7688
PREV	4	-0.3140	0.0673	21.80	0.0000
SOFTNESS*PREV	5	-0.1214	0.0940	1.67	0.1963
	6	-0.0636	0.0947	0.45	0.5017
TEMP	7	0.1281	0.0673	3.63	0.0567
SOFTNESS*TEMP	8	-0.0311	0.0940	0.11	0.7407
	9	-0.0096	0.0947	0.01	0.9190
PREV*TEMP	10	-0.1009	0.0673	2.25	0.1335
SOFTNESS*PREV*TEMP	11	0.0766	0.0940	0.66	0.5311

DETERGENT PREFERENCE STU

MAIN-EFFECTS MODEL

CATMOD PROCEDURE

Response：BRAND		Response Levels	(R)=	2
Weight Variable：COUNT		Populations	(S)=	12
Data Set：DETERG		Total Frequency	(N)=	1008
Frequency Missing：0		Observations	(Obs)=	24

DETERGENT PREFERENCE STUDY

MAIN-EFFECTS MODEL

Sample	Response Function	DESIGN MATRIX 1	2	3	4	5
1	-0.33647	1	1	0	1	1
2	-0.48184	1	1	0	1	-1
3	0.58315	1	1	0	-1	1
4	0.34033	1	1	0	-1	-1
5	-0.36101	1	0	1	1	1
6	-0.27763	1	0	1	1	-1
7	0.71465	1	0	1	-1	1
8	0.15719	1	0	1	-1	-1
9	-0.07146	1	-1	-1	1	1
10	-0.17284	1	-1	-1	1	-1
11	0.42286	1	-1	-1	-1	1
12	-0.15123	1	-1	-1	-1	-1

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	0.20	0.6519
SOFTNESS	2	0.20	0.9033
PREV	1	19.28	0.0000
TEMP	1	3.63	0.0566
RESIDUAL	7	8.18	0.3169

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.0302	0.0669	0.20	0.6519
SOFTNESS	2	-0.0099	0.0911	0.01	0.9133
	3	0.0391	0.0902	0.19	0.6643
PREV	4	-0.2817	0.0642	19.28	0.0000
TEMP	5	0.1277	0.0670	3.63	0.0566

例五：对数线性模型，三个因变量

下面的数据与植物的插枝移植有关。三个因变量都是二分的，其值以 1 或 2 表示：生存状态 V1（1=死、2=活），插枝的时间 V2（1=春天、2= 即剪即插枝），及剪下来的枝干长短 V3（1=长、2=短）。我们对这三者间的交互效果最感兴趣。

希望电子出版社版权所有

3	0	3800.12	0.0000255
4	0	3800.12	3.6909E-9

Parameter Estimates

Iteration	1	2	3	4	5	6
0	0	0	0	0	0	0
1	-3.7E-18	-1.85E-18	1.85E-18	-0.2125	0.2125	0.3083
2	0.0494	0.0752	-0.0752	-0.2486	0.2486	0.3502
3	0.0555	0.0809	-0.0809	-0.2543	0.2543	0.3568
4	0.0556	0.0810	-0.0810	-0.2544	0.2544	0.3569

MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
V3	1	2.64	0.1041
V2	1	5.25	0.0220
V3*V2	1	5.25	0.0220
V1	1	48.94	0.0000
V3*V1	1	48.94	0.0000
V2*V1	1	95.01	0.0000
LIKELIHOOD RATIO	1	2.29	0.1299

MAXIMUM LIKELIHOOD PREDICTED VALUES FOR RESPONSE FUNCTIONS AND FREQUENCIES

	-------Observed-------			-------Predicted------		
Sample	Function Number	Function	Standard Error	Function	Standard Error	Residual
1	1	-0.2924782	0.10580617	-0.2356473	0.09848616	-0.056831
	2	-0.9115175	0.12918766	-0.9494184	0.1299476	0.03790099
	3	-0.9115175	0.12918766	-0.9494184	0.1299476	0.03790099
	4	-0.2924782	0.10580617	-0.2356473	0.09848616	-0.056831
	5	-0.6695054	0.11887171	-0.6936188	0.12017169	0.02411336
	6	-0.4519851	0.11092108	-0.3896985	0.1022668	-0.0622866
	7	-1.908347	0.19246494	-1.7314626	0.14296911	-0.1768845
	F1	156	11.4302231	161.096138	11.0737946	-5.0961381
	F2	84	8.75499857	78.9038609	7.8086133	5.09613909
	F3	84	8.75499857	78.9038609	7.8086133	5.09613909
	F4	156	11.4302231	161.096138	11.0737946	-5.0961381
	F5	107	9.75058759	101.903861	8.9243041	5.09613941
	F6	133	10.7039226	138.096139	10.3343404	-5.0961386
	F7	31	5.47713048	36.0961431	4.82631486	-5.0961431
	F8	209	12.7866711	203.90386	12.2128493	5.09614031

例六：对数线性模型，含零频率

下面的数据是有关六对松鼠猴的性行为以及它们性角色 (主动或被动) 之间的关系。性行为的分类以 R，S，T，U，V，W 表示。由于一对猴子不可能同时采用主动以及被动的性角色，所以有些变量的组合必定造成零的频率。这种情况称为结构性的零频率。CATMOD 程序会自动剔除这些组合。若是由于随机取样产生的零频率 (称为随机零频率)，CATMOD 程序会以极小的正实数 (如：10 的-20 次方) 来取代它们。

程　序

```
OPTIONS NODATE;
TITLE 'BEHAVIOR OF SQUIRREL MONKEYS';
DATA DISPLAY;
    INPUT ACTIVE $ PASSIVE $ WT @@;
    IF ACTIVE NE 'T';
    IF ACTIVE  NE PASSIVE THEN IF WT=0 THEN WT=1E-20;
    CARDS;
```

```
RR 0  RS 1  RT 5  RU 8  RV 9  RW0
SR 29 SS 0  ST14  SU 46 SV 4  SW0
TR 0  TS 0  TT 0  TU 0  TV 0  TW0
UR 2  US 3  UT 1  UU 0  UV38  UW2
VR 0  VS 0  VT 0  VU 0  VV 0  VW1
WR 9  WS25  WT 4  WU 6  WV13  WW0
;
PROC CATMOD;
    WEIGHT WT;
    MODEL ACTIVE*PASSIVE=_RESPONSE_ / ML NOGLS FREQ PRED=FREQ NOPARM;
    REPEATED / _RESPONSE_ =ACTIVE PASSIVE;
    TITLE2 'TEST QUASI-INDEPENDENCE FOR THE INCOMPLETE TABLE';
RUN;
```

结　果

经过七次的循环估计后，对数线性模型中的参数估计即趋向稳定。分析结果显示：松鼠猴的性行为和它们的两性角色 (主动或被动) 息息关系 (主动之 Chi-Square=56.58，P=0.0000；被动之 Chi-Square=47.94，P=0.0000)。

报表 16.6　对数线性模型，含零频率

```
                BEHAVIOR OF SQUIRREL MONKEYS
        TEST QUASI-INDEPENDENCE FOR THE INCOMPLETE TABLE
                     CATMOD PROCEDURE
Response：ACTIVE*PASSIVE                      RESPONSE PROFILES
Weight Variable:WT
Data Set：DISPLAY                        Response  ACTIVE  PASSIVE
Frequency Missing：0                     -------------------------
Response Levels (R)=    25                   1     R       S
Populations      (S)=    1                   2     R       T
Total Frequency (N)=   220                   3     R       U
Observations   (Obs)=   25                   4     R       V
                                             5     R       W
                                             6     S       R
                                             7     S       T
                                             8     S       U
                                             9     S       V
          Sample                            10     S       W
Sample     Size                             11     U       R
----------------                           12     U       S
    1       220                            13     U       T
```

14	U	V
15	U	W
16	V	R
17	V	S
18	V	T
19	V	U
20	V	W
21	W	R
22	W	S
23	W	T
24	W	U
25	W	V

RESPONSE FREQUENCIES

Response Number

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	5	8	9	1E-20	29	14	46	4	1E-20	2	3	1	38

Response Number

Sample	15	16	17	18	19	20	21	22	23	24	25
1	2	1E-20	1E-20	1E-20	1E-20	1	9	25	4	6	13

BEHAVIOR OF SQUIRREL MONKEYS

TEST QUASI-INDEPENDENCE FOR THE INCOMPLETE TABLE

RESPONSE MATRIX

	1	2	3	4	5	6	7	8	9
1	1	0	0	0	0	1	0	0	0
2	1	0	0	0	0	0	1	0	0
3	1	0	0	0	0	0	0	1	0
4	1	0	0	0	0	0	0	0	1
5	1	0	0	0	-1	-1	-1	-1	-1
6	0	1	0	0	1	0	0	0	0
7	0	1	0	0	0	0	1	0	0
8	0	1	0	0	0	0	0	1	0
9	0	1	0	0	0	0	0	0	1
10	0	1	0	0	-1	-1	-1	-1	-1
11	0	0	1	0	1	0	0	0	0
12	0	0	1	0	0	1	0	0	0
13	0	0	1	0	0	0	1	0	0
14	0	0	1	0	0	0	0	0	1
15	0	0	1	0	-1	-1	-1	-1	-1
16	0	0	0	1	1	0	0	0	0
17	0	0	0	1	0	1	0	0	0
18	0	0	0	1	0	0	1	0	0
19	0	0	0	1	0	0	0	1	0
20	0	0	0	1	-1	-1	-1	-1	-1
21	-1	-1	-1	-1	1	0	0	0	0
22	-1	-1	-1	-1	0	1	0	0	0
23	-1	-1	-1	-1	0	0	1	0	0

东南大学出版社版权所有

24	-1	-1	-1	-1	0	0	0	1	0
25	-1	-1	-1	-1	0	0	0	0	1

MAXIMUM LIKELIHOOD ANALYSIS

Iteration	Sub Iteration	-2 Log Likelihood	Convergence Criterion
0	0	1416.3054	1.0000
1	0	1238.2417	0.1257
2	0	1205.1264	0.0267
3	0	1199.5068	0.004663
4	0	1198.6271	0.000733
5	0	1198.5611	0.0000551
6	0	1198.5603	6.5351E-7
7	0	1198.5603	1.217E-10

Parameter Estimates

Iteration	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0
1	-0.4976	1.1112	0.1722	-0.8804	-0.006978	0.0827	-0.4735	0.7287	0.5791
2	-0.3420	1.0962	0.5612	-1.7549	0.2233	0.3899	-0.4086	0.7875	0.5728
3	-0.1570	1.2687	0.7058	-2.3992	0.3034	0.4360	-0.3162	0.8812	0.6703
4	-0.0466	1.3791	0.8170	-2.8422	0.3309	0.4625	-0.2890	0.9085	0.6968
5	-0.002748	1.4230	0.8609	-3.0176	0.3334	0.4649	-0.2866	0.9110	0.6992
6	0.002760	1.4285	0.8664	-3.0396	0.3334	0.4649	-0.2865	0.9110	0.6992
7	0.002837	1.4285	0.8665	-3.0399	0.3334	0.4649	-0.2865	0.9110	0.6992

MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
ACTIVE	4	56.58	0.0000
PASSIVE	5	47.94	0.0000
LIKELIHOOD RATIO	15	135.17	0.0000

MAXIMUM LIKELIHOOD PREDICTED VALUES FOR RESPONSE FUNCTIONS AND FREQUENCIES

		-------Observed-------		-------Predicted------		
Sample	Function Number	Function	Standard Error	Function	Standard Error	Residual
1	1	-2.5649494	1.03774904	-0.973554	0.33901898	-1.5913953
	2	-0.9555114	0.52623481	-1.7250404	0.34543788	0.76952896
	3	-0.4855078	0.44935852	-0.5275144	0.30925387	0.0420066
	4	-0.3677248	0.43362909	-0.7392682	0.24900568	0.37154345
	5	-48.616651	1E10	-3.560517	0.63410407	-45.056134
	6	0.80234647	0.33377513	0.32058886	0.2662902	0.48175761

希望电子出版社版权所有

7	0.07410797	0.38516444	-0.2993416	0.29563358	0.37344956
8	1.26369204	0.31410541	0.89818441	0.25085737	0.36550763
9	-1.178655	0.57177187	0.6864306	0.17339604	-1.8650856
10	-48.616651	1E10	-2.1348182	0.60807083	-46.481833
11	-1.8718022	0.75955453	-0.2414953	0.28721789	-1.6303069
12	-1.4663371	0.64051262	-0.1099394	0.30356781	-1.3563977
13	-2.5649494	1.03774904	-0.8614257	0.31479379	-1.7035236
14	1.0726368	0.32130806	0.12434644	0.20434511	0.94829036
15	-1.8718022	0.75955453	-2.6969023	0.61743258	0.82510014
16	-48.616651	1E10	-4.1478747	1.02450813	-44.468777
17	-48.616651	1E10	-4.0163187	1.03006239	-44.600332
18	-48.616651	1E10	-4.7678051	1.03245707	-43.848846
19	-48.616651	1E10	-3.5702791	1.02079389	-45.046372
20	-2.5649494	1.03774904	-6.6032817	1.16128927	4.03833233
21	-0.3677248	0.43362909	-0.3658417	0.20295917	-0.001883
22	0.65392647	0.34194017	-0.2342858	0.23279368	0.88821229
23	-1.178655	0.57177187	-0.9857722	0.23940797	-0.1928828
24	-0.7731899	0.49354812	0.21175381	0.18500696	-0.9849437

BEHAVIOR OF SQUIRREL MONKEYS

TEST QUASI-INDEPENDENCE FOR THE INCOMPLETE TABLE

MAXIMUM LIKELIHOOD PREDICTED VALUES FOR RESPONSE FUNCTIONS AND FREQUENCIES

		-------Observed-------		-------Predicted------		
	Function		Standard		Standard	
Sample	Number	Function	Error	Function	Error	Residual
1	F1	1	0.99772468	5.25950838	1.36156002	-4.2595084
	F2	5	2.21051208	2.48072585	0.6910659	2.51927415
	F3	8	2.77652497	8.21594841	1.85514611	-0.2159484
	F4	9	2.93799561	6.64804868	1.50931986	2.35195132
	F5	1E-20	1E-10	0.39576868	0.2402678	-0.3957687
	F6	29	5.01769596	19.1859928	3.14791495	9.81400723
	F7	14	3.62064786	10.321716	2.16959874	3.67828404
	F8	46	6.03173426	34.1846262	4.42870591	11.8153738
	F9	4	1.98173478	27.6609647	3.72278813	-23.660965
	F10	1E-20	1E-10	1.64670026	0.95271227	-1.6467003
	F11	2	1.40777064	10.936396	2.12321968	-8.936396
	F12	3	1.72020083	12.4740717	2.55433555	-9.4740717
	F13	1	0.99772468	5.8835826	1.3806555	-4.8835826

东南大学出版社版权所有

F14	38	5.60681404	15.7672979	2.68469221	22.2327021
F15	2	1.40777064	0.93865177	0.55164479	1.06134823
F16	1E-20	1E-10	0.21996583	0.22177911	-0.2199658
F17	1E-20	1E-10	0.2508934	0.25370612	-0.2508934
F18	1E-20	1E-10	0.11833763	0.12031391	-0.1183376
F19	1E-20	1E-10	0.39192393	0.39325479	-0.3919239
F20	1	0.99772468	0.01887928	0.02172759	0.98112072
F21	9	2.93799561	9.6576454	1.80865595	-0.6576454
F22	25	4.70734436	11.0155266	2.27501884	13.9844734
F23	4	1.98173478	5.19563797	1.18445235	-1.195638
F24	6	2.41585671	17.2075014	2.77209793	-11.207501
F25	13	3.49740163	13.9236886	2.24158038	-0.9236886

例七：重复观察的实验，一个样本，两种反应

每一受试者分别服用三种药物 (以 A，B，C 表之)后，记录他们的反应，并将反应分成两类 (F 表示好的反应，U 表示不好的反应)。每一位受试者需被重复地观察三次。分析的步骤有二：首先，测受试者对这三种药物的两种反应频率 (或作边际概率) 是否相等。初步分析的结果否定这个边际概率相等的假设。所以，第二步骤分析只针对药物 A 与 B，以探讨它们的边际概率是否真的相等。

程　序

```
TITLE 'ONE-POPULATION DRUG STUDY';
DATA DRUGS;
    INPUT DRUGA $ DRUGB $ DRUGC $ COUNT @@;
    CARDS;
F F F 6 F F U 16 F U F 2 F U U 4 U F F 2 U F U 4 U U F 6 U U U 6
;
PROC CATMOD;
    WEIGHT COUNT;
    RESPONSE MARGINALS;
    MODEL DRUGA*DRUGB*DRUGC=_RESPONSE_;
    REPEATED DRUG 3;
    TITLE2 'TEST MARGINAL HOMOGENITY';
PROC CATMOD;
    WEIGHT COUNT;
    RESPONSE MARGINALS;
    TITLE2 'MODEL FOR EQUALITY OF A AND B MARGINAL PROBABILITIES';
    MODEL DRUGA*DRUGB*DRUGC=(1 0,
                             1 0,
```

```
                                        0 1 ) /NOPROFILE;
RUN;
```

结　果

如前所述，三种药物的两个边际概率在第一步骤的分析中已证实是不等的。因此第二步骤的分析重点是放在 A 与 B 两种药物上，而分析的结果同样地证实它们的两个边际概率不等。所以，我们可以间接推论说 C 药物的边际概率也不等于 A 或 B 的边际概率。因此，三种药物在病人身上造成的效果不尽相同。

报表 16.7　重复观察的实验，一个样本，两种反应

```
                          ONE-POPULATION DRUG STUDY
                          TEST MARGINAL HOMOGENITY
                              CATMOD PROCEDURE
                                          RESPONSE PROFILES
Response：DRUGA*DRUGB*DRUGC            Response  DRUGA  DRUGB  DRUGC
Weight Variable：COUNT                 ----------------------------
Data Set：DRUGS                            1     F      F      F
Frequency Missing：0                       2     F      F      U
Response Levels (R)=     8                 3     F      U      F
Populations      (S)=    1                 4     F      U      U
Total Frequency (N)=    46                 5     U      F      F
Observations  (Obs)=     8                 6     U      F      U
                                           7     U      U      F
            Sample                         8     U      U      U
   Sample     Size
   ----------------
       1        46
       Function  Response  DESIGN MATRIX        Source       DF   Chi-Square     Prob
Sample  Number   Function   1    2    3          ----------------------------------------
----------------------------------------         INTERCEPT     1       146.84    0.0000
    1        1    0.60870   1    1    0          DRUG          2         6.58    0.0372
             2    0.60870   1    0    1
             3    0.34783   1   -1   -1          RESIDUAL      0          .        .
                 ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES
                                              Standard    Chi-
       Effect            Parameter  Estimate    Error    Square   Prob
       ------------------------------------------------------------------
       INTERCEPT                1    0.5217    0.0431    146.84  0.0000
       DRUG                     2    0.0870    0.0507      2.95  0.0861
                                3    0.0870    0.0507      2.95  0.0861
```

希望电子出版社版权所有

```
                    ONE-POPULATION DRUG STUDY
          MODEL FOR EQUALITY OF A AND B MARGINAL PROBABILITIES
                        CATMOD PROCEDURE
      Response：DRUGA*DRUGB*DRUGC    Response Levels (R)=     8
      Weight Variable：COUNT         Populations     (S)=     1
      Data Set：DRUGS                Total Frequency (N)=    46
      Frequency Missing：0           Observations  (Obs)=     8
                              DESIGN              ANALYSIS OF VARIANCE TABLE
    Function    Response      MATRIX
Sample  Number    Function     1      2     Source       DF   Chi-Square      Prob
-----------------------------------------   ------------------------------------
   1       1      0.60870     1      0     MODEL|MEAN    1         6.58    0.0103
           2      0.60870     1      0
           3      0.34783     0      1     RESIDUAL      1         0.00    1.0000
              ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES
                                     Standard    Chi-
          Effect            Parameter  Estimate    Error    Square   Prob
     ------------------------------------------------------------
     MODEL                     1     0.6087     0.0613    98.54  0.0000
                               2     0.3478     0.0702    24.53  0.0000
```

例八：重复观察的实验，三个样本，两种反应

下列数据是根据三组受试 (以 1，2，3 表示) 在四次实验 (以 A，B，C，D 表示) 中的反应收集来的。反应是二分的，以 0 或 1 表示。分析的步骤有二：首先，以传统的重复观察的变异数分析法处理。在这个步骤里两个主效果 [即实验效果 (TRIAL) 及组别 (GROUP)] 及两者间的交互作用被用来解释反应频率的不同。分析的结果显示交互作用达显著水准。所以在第二个步骤里，分析的重点放在第三组受试者上，看看他们在四次实验中有没有不同的反应频率。

程　序

```
TITLE 'TWO-POPULATION REPEATED MEASURES';
DATA GROUP;
    INPUT A B C D GROUP WT @@;
    CARDS;
1 1 1 1 2 2   0 0 0 0 2 2   0 0 1 0 1 2   0 0 1 0 2 2
0 0 0 1 1 4   0 0 0 1 2 1   0 0 0 1 3 3   1 0 0 1 2 1
0 0 1 1 1 1   0 0 1 1 2 2   0 0 1 1 3 5   0 1 0 0 1 4
0 1 0 0 2 1   0 1 0 1 2 1   0 1 0 1 3 2   0 1 1 0 3 1
1 0 0 0 1 3   1 0 0 0 2 1   0 1 1 1 2 1   0 1 1 1 3 2
1 0 1 0 1 1   1 0 1 1 2 1   1 0 1 1 3 2
```

```
;
PROC CATMOD;
    WEIGHT WT;
    RESPONSE MARGINALS;
    MODEL A*B*C*D=GROUP _RESPONSE_ GROUP*_RESPONSE_/FREQ NODESIGN;
    REPEATED TRIAL 4;
    TITLE2  'SATURATED MODEL';
PROC CATMOD;
    WEIGHT WT;
    RESPONSE MARGINALS;
    MODEL A*B*C*D=GROUP _RESPONSE_ (GROUP=3) /NOPROFILE NOPARM;
    REPEATED TRIAL 4;
    TITLE2 'TRIAL NESTED WITHIN GROUP 3';
RUN;
```

结　果

第二步分析的结果显示：第三组受试者在四次实验里的反应频率有显著的不同(Chi-Square=75.07，显著度 < 0.001)。

报表 16.8　重复观察的实验，三个样本，两种反应

```
                    TWO-POPULATION REPEATED MEASURES
                          SATURATED MODEL
                          CATMOD PROCEDURE

                                          RESPONSE PROFILES
 Response：A*B*C*D                       Response  A  B  C  D
 Weight Variable：WT                     ---------------------
 Data Set：GROUP                             1     0  0  0  0
 Frequency Missing：0                        2     0  0  0  1
 Response Levels (R)=    13                  3     0  0  1  0
 Populations     (S)=     3                  4     0  0  1  1
 Total Frequency (N)=    45                  5     0  1  0  0
 Observations  (Obs)=    23                  6     0  1  0  1
                                             7     0  1  1  0
                                             8     0  1  1  1
   POPULATION PROFILES                       9     1  0  0  0
                  Sample                    10     1  0  0  1
 Sample  GROUP     Size                     11     1  0  1  0
 -----------------------                    12     1  0  1  1
     1       1        15                    13     1  1  1  1
     2       2        15
```

```
        3       3       15
                    TWO-POPULATION REPEATED MEASURES
                          SATURATED MODEL
                        RESPONSE FREQUENCIES
                          Response Number
  Sample    1    2    3    4    5    6    7    8    9   10   11   12   13
  ----------------------------------------------------------------------
       1    0    4    2    1    4    0    0    0    3    0    1    0    0
       2    2    1    2    2    1    1    0    1    1    1    0    1    2
       3    0    3    0    5    0    2    1    2    0    0    0    2    0
     ANALYSIS OF VARIANCE TABLE                ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES
Source            DF  Chi-Square     Prob                                   Standard   Chi-
-----------------------------------------  Effect       Parameter  Estimate   Error   Square   Prob
INTERCEPT          1      354.88    0.0000  ----------------------------------------------------------
GROUP              2       24.79    0.0000  INTERCEPT          1    0.5833   0.0310   354.88   0.0000
TRIAL              3       21.45    0.0001  GROUP              2    0.1333   0.0335    15.88   0.0001
GROUP*TRIAL        6       18.71    0.0047                     3   -0.0333   0.0551     0.37   0.5450
                                             TRIAL              4    0.1722   0.0557     9.57   0.0020
RESIDUAL           0          .       .                        5    0.1056   0.0647     2.66   0.1028
                                                                6   -0.0722   0.0577     1.57   0.2107
                                             GROUP*TRIAL        7   -0.1556   0.0852     3.33   0.0679
                                                                8   -0.0889   0.0953     0.87   0.3511
                                                                9    0.0889   0.0822     1.17   0.2793
                                                               10   -0.0556   0.0800     0.48   0.4877
                                                               11    0.0111   0.0866     0.02   0.8979
                                                               12   -0.0111   0.0824     0.02   0.8927
                    TWO-POPULATION REPEATED MEASURES
                    TRIAL NESTED WITHIN GROUP 3
                          CATMOD PROCEDURE
     Response：A*B*C*D                    Response Levels (R)=    13
     Weight Variable：WT                  Populations     (S)=     3
     Data Set：GROUP                      Total Frequency (N)=    45
     Frequency Missing：0                 Observations  (Obs)=    23
       Function    Response                  DESIGN MATRIX
Sample   Number     Function      1       2       3       4       5       6
-----------------------------------------------------------------------------
   1        1      0.73333       1       1       0       0       0       0
            2      0.73333       1       1       0       0       0       0
            3      0.73333       1       1       0       0       0       0
```

	4	0.66667	1	1	0	0	0	0
2	1	0.66667	1	0	1	0	0	0
	2	0.66667	1	0	1	0	0	0
	3	0.46667	1	0	1	0	0	0
	4	0.40000	1	0	1	0	0	0
3	1	0.86667	1	-1	-1	1	0	0
	2	0.66667	1	-1	-1	0	1	0
	3	0.33333	1	-1	-1	0	0	1
	4	0.06667	1	-1	-1	-1	-1	-1

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	386.94	0.0000
GROUP	2	25.42	0.0000
TRIAL(GROUP=3)	3	75.07	0.0000
RESIDUAL	6	5.09	0.5319

例九：重复观察的实验，一个样本，四种反应

下列的数据是七千四百七十七位女士左眼 (以 LEFT 表示) 及右眼 (以 RIGHT 表示) 的视力资料 (导致重复观察的变量是 SIDE)。每一双眼睛可能有四种反应 (以 1，2，3，4 表示)，因此双眼可能产生的反应型态有十六种。对每一双眼睛而言，四种可能的反应导出三个彼此独立的边际概率，因此 SIDE 的自由度是 3。

程　序

```
TITLE 'VISION SYMMETRY';
DATA VISION;
    INPUT RIGHT LEFT COUNT @@;
    CARDS;
1 1 1520   1 2  266   1 3  124   1 4  66
2 1  234   2 2 1512   2 3  432   2 4  78
3 1  117   3 2  362   3 3 1772   3 4 205
4 1   36   4 2   82   4 3  179   4 4 492
;
PROC CATMOD;
    WEIGHT COUNT;
    RESPONSE MARGINALS;
    MODEL RIGHT*LEFT=_RESPONSE_/FREQ;
    REPEATED SIDE 2;
    TITLE2 'TEST OF MARGINAL HOMOGENITY';
```

希望电子出版社版权所有

```
RUN;
```

结　果

分析结果显示：双眼对四种反应的频率不同，其显著度达 0.0075 (Chi-Square =11.98)。

报表 16.9　重复观察的实验，一个样本，四种反应

```
                    VISION SYMMETRY
              TEST OF MARGINAL HOMOGENITY
                    CATMOD PROCEDURE
Response：RIGHT*LEFT                RESPONSE PROFILES
Weight Variable：COUNT
Data Set：VISION                   Response  RIGHT  LEFT
Frequency Missing：0                ---------------------
Response Levels (R)=    16             1        1     1
Populations     (S)=    1              2        1     2
Total Frequency (N)=  7477             3        1     3
Observations  (Obs)=    16             4        1     4
                                       5        2     1
                                       6        2     2
          Sample                       7        2     3
  Sample    Size                       8        2     4
  ---------------                    9        3     1
      1      7477                     10       3     2
                                      11       3     3
                                      12       3     4
                                      13       4     1
                                      14       4     2
                                      15       4     3
                                      16       4     4
                    RESPONSE FREQUENCIES
                          Response Number
Sample    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
-------------------------------------------------------------------------------------------
   1   1520  266  124   66  234 1512  432   78  117  362 1772  205   36   82  179  492

                                        ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES
     Function Response     DESIGN MATRIX                                Standard    Chi-
Sample Number Function  1   2   3   4   5   6   Effect  Parameter  Estimate  Error   Square   Prob
-------------------------------------------------  ---------------------------------------------
```

希望电子出版社版权所有

1	1	0.26428	1	0	0	1	0	0	INTERCEPT	1	0.2597	0.00468	3073.03	0.0000
	2	0.30173	0	1	0	0	1	0		2	0.2995	0.00464	4160.17	0.0000
	3	0.32847	0	0	1	0	0	1		3	0.3319	0.00483	4725.25	0.0000
	4	0.25505	1	0	0	-1	0	0	SIDE	4	0.00461	0.00194	5.65	0.0174
	5	0.29718	0	1	0	0	-1	0		5	0.00227	0.00255	0.80	0.3726
	6	0.33529	0	0	1	0	0	-1		6	-0.00341	0.00252	1.83	0.1757

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	3	78744.17	0.0000
SIDE	3	11.98	0.0075
RESIDUAL	0	.	.

例十：重复观察的实验，两个样本，四种反应

两组神经系统有障碍的病人 (GROUP=1，2)，经过两位专家 (OBS1，OBS2) 分别诊断后，被归入下列四种病况之下：

1. 已经陷入中枢神经系统麻痹的状态，
2. 极有可能是中枢神经系统的麻痹，
3. 有 50% 的可能是中枢神经系统麻痹，
4. 以上皆非。

如例九的处理方法，这四种类别最后导出三个互相独立的反应函数，所以每一位专家效果的自由度也等于 3。

程　序

```
TITLE 'MULTIPLE SCLEROSIS DATA ANALYSIS';
DATA ONE;
    INPUT GROUP OBS1 OBS2 WT @@;
    CARDS;
1 4 4 10   1 4 1  3   1 4 2  7   1 4 3  3
1 1 4  1   1 1 1 38   1 1 2  5   1 1 3  0
1 2 4  0   1 2 1 33   1 2 2 11   1 2 3  3
1 3 4  6   1 3 1 10   1 3 2 14   1 3 3  5
2 4 1  1   2 4 2  2   2 4 3  4   2 4 4 14
2 3 1  2   2 3 2 13   2 3 3  3   2 3 4  4
2 2 1  3   2 2 2 11   2 2 3  4   2 2 4  0
2 1 1  5   2 1 2  3   2 1 3  0   2 1 4  0
;
PROC CATMOD;
    TITLE2 'SATURATED MODEL';
```

东南大学出版社版权所有

```
        WEIGHT WT;
        RESPONSE MARGINALS;
        MODEL OBS1*OBS2=GROUP _RESPONSE_ GROUP*_RESPONSE_/FREQ NODESIGN;
        REPEATED OBSERVER 2;
    PROC CATMOD;
        TITLE2 'TEST MARGINAL HOMOGENITY BY OBSERVER NESTED WITHIN GROUP';
        WEIGHT WT;
        RESPONSE MARGINALS;
        MODEL OBS1*OBS2=GROUP _RESPONSE_(GROUP) /NOPROFILE NODESIGN;
        REPEATED OBSEEVER 2;
    RUN;
```

结　果

结果显示：两位专家在病人的分类诊断上有显著的不同 (Chi-Square=47.90，P=0.0000)，并且四种病况出现的次数也有显著的差异 (Chi-Square=37.62，P=0.0000)。

报表 16.10　重复观察的实验，两个样本，四种反应

```
                    MULTIPLE SCLEROSIS DATA ANALYSIS
                         SATURATED MODEL
                           CATMOD PROCEDURE

 Response：OBS1*OBS2
 Weight Variable：WT
 Data Set：ONE
 Frequency Missing：0
 Response Levels (R)=    14
 Populations      (S)=    2
 Total Frequency (N)=   218
 Observations   (Obs)=   27

   POPULATION PROFILES
                  Sample
 Sample  GROUP     Size
 -----------------------
     1       1      149
     2       2       69
```

```
                RESPONSE PROFILES
 Response  OBS1  OBS2
 --------------------
     1       1     1
     2       1     2
     3       1     4
     4       2     1
     5       2     2
     6       2     3
     7       3     1
     8       3     2
     9       3     3
    10       3     4
    11       4     1
    12       4     2
    13       4     3
    14       4     4
```

```
                RESPONSE FREQUENCIES
                Response Number
```

希望电子出版社版权所有

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	38	5	1	33	11	3	10	14	5	6	3	7	3	10
2	5	3	0	3	11	4	2	13	3	4	1	2	4	14

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	3	875.79	0.0000
GROUP	3	37.62	0.0000
OBSERVER	3	47.90	0.0000
GROUP*OBSERVER	3	14.09	0.0028
RESIDUAL	0	.	.

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.2836	0.0240	139.46	0.0000
	2	0.3112	0.0255	148.88	0.0000
	3	0.1967	0.0206	91.63	0.0000
GROUP	4	0.1459	0.0240	36.92	0.0000
	5	-0.0294	0.0255	1.32	0.2499
	6	-0.0424	0.0206	4.25	0.0392
OBSERVER	7	-0.0780	0.0153	26.15	0.0000
	8	-0.0231	0.0219	1.11	0.2911
	9	0.0801	0.0205	15.22	0.0001
GROUP*OBSERVER	10	-0.0562	0.0153	13.60	0.0002
	11	0.0566	0.0219	6.71	0.0096
	12	0.000413	0.0205	0.00	0.9839

MULTIPLE SCLEROSIS DATA ANALYSIS

TEST MARGINAL HOMOGENITY BY OBSERVER NESTED WITHIN GROUP

CATMOD PROCEDURE

Response：OBS1*OBS2		Response Levels (R)=	14
Weight Variable：WT		Populations (S)=	2
Data Set：ONE		Total Frequency (N)=	218
Frequency Missing：0		Observations (Obs)=	27

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	3	875.79	0.0000
GROUP	3	37.62	0.0000
OBSERVER(GROUP)	6	69.01	0.0000
RESIDUAL	0	.	.

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob

清华电子出版社版权所有

INTERCEPT	1	0.2836	0.0240	139.46	0.0000
	2	0.3112	0.0255	148.88	0.0000
	3	0.1967	0.0206	91.63	0.0000
GROUP	4	0.1459	0.0240	36.92	0.0000
	5	-0.0294	0.0255	1.32	0.2499
	6	-0.0424	0.0206	4.25	0.0392
OBSERVER(GROUP)	7	-0.1342	0.0216	38.78	0.0000
	8	0.0336	0.0263	1.63	0.2016
	9	0.0805	0.0190	17.92	0.0000
	10	-0.0217	0.0216	1.01	0.3138
	11	-0.0797	0.0349	5.21	0.0225
	12	0.0797	0.0364	4.79	0.0286

例十一：重复观察的实验，生长曲线的对数分析

本资料来自一个长期的实验，研究对象以四位病人为样本。这四个样本分别含有两种诊断结果（DIAG）：轻微或严重，以及两种药方（TRT）：传统与新药方不同的排列组合。经过一段治疗期后，每一位病人在三个不同的时间（第一、二、四种）回医院作药物反应的测验。反应分成两类：正常（以 N 表之）或不正常（以 A 表之）。时间变量的值（原是 1，2，4），先取以 2 为底的对数将它们转换成 0，1，与 2。同样地，将反应的次数亦转换成对数。然后这些对数值用重复观察的变异数分析法处理之。

程　序

```
TITLE 'GROWTH CURVE ANALYSIS';
DATA GROWTH2;
      INPUT DIAG $ TRT $ WEEK1 $ WEEK2 $ WEEK4 $ COUNT @@; CARDS;
MILD  STANDARD  N N N 16  SEVERE STANDARD N N N  2
MILD  STANDARD  N N A 13  SEVERE STANDARD N N A  2
MILD  STANDARD  N A N  9  SEVERE  STANDARD N A N  8
MILD  STANDARD  N A A  3  SEVERE  STANDARD N A A  9
MILD  STANDARD  A N N 14  SEVERE STANDARD A N N  9
MILD  STANDARD  A N A  4  SEVERE STANDARD A N A 15
MILD  STANDARD  A A N 15  SEVERE STANDARD A A N 27
MILD  STANDARD  A A A  6  SEVERE STANDARD A A A 28
MILD       NEW  N N N 31  SEVERE      NEW N N N  7
MILD       NEW  N N A  0  SEVERE      NEW N N A  2
MILD       NEW  N A N  6  SEVERE      NEW N A N  5
MILD       NEW  N A A  0  SEVERE      NEW N A A  2
MILD       NEW  A N N 22  SEVERE      NEW A N N 31
MILD       NEW  A N A  2  SEVERE      NEW A N A  5
```

```
MILD       NEW  A A N  9  SEVERE      NEW A A N  32
MILD       NEW  A A A  0  SEVERE      NEW A A A  6
;
PROC CATMOD ORDER=DATA;
  TITLE2 'REDUCED LOGISTIC MODEL'; WEIGHT COUNT;
  POPULATION DIAG TRT; RESPONSE LOGIT;
  MODEL WEEK1*WEEK2*WEEK4=( 1 0 0 0 ,
                            1 0 1 0 ,
                            1 0 2 0 ,
                            1 0 0 0 ,
                            1 0 0 1 ,
                            1 0 0 2 ,
                            0 1 0 0 ,
                            0 1 1 0 ,
                            0 1 2 0 ,
                            0 1 0 0 ,
                            0 1 0 1 ,
                            0 1 0 2 )  (1='MILD DIAGNOSIS, WEEK 1',
                                       2='SEVERE DIAGNOSIS, WEEK 1',
                                       3='TIME EFFECT FOR STD TRT',
                                       4='TIME EFFECT FOR NEW TRT')
                                       / FREQ;
   CONTRAST 'DIAGNOSIS EFFECT, WEEK 1' ALL_PARMS 1  -1  0  0;
   CONTRAST 'EQUAL TIME EFFECTS' ALL_PARMS 0  0  1  -1;
 RUN;
```

结　果

除了轻微的病人在第一周的反应测验无显著差异外，严重的病人其反应会随时间而改变。另外，两种药方的治疗效果也随著治疗期的长短而有不同 (P=0.0000)。

报表 16.11　重复观察的实验，生长曲线的对数分析

```
                    GROWTH CURVE ANALYSIS
                  REDUCED LOGISTIC MODEL
                     CATMOD PROCEDURE
Response：WEEK1*WEEK2*WEEK4
Weight Variable：COUNT
Data Set：GROWTH2
Frequency Missing：0                    RESPONSE PROFILES
Response Levels (R)=     8
Populations     (S)=     4            Response  WEEK1  WEEK2  WEEK4
```

Total Frequency (N)= 340

Observations (Obs)= 29

POPULATION PROFILES

Sample	DIAG	TRT	Sample Size
1	MILD	STANDARD	80
2	MILD	NEW	70
3	SEVERE	STANDARD	100
4	SEVERE	NEW	90

1	N	N	N
2	N	N	A
3	N	A	N
4	N	A	A
5	A	N	N
6	A	N	A
7	A	A	N
8	A	A	A

RESPONSE FREQUENCIES

	Response Number							
Sample	1	2	3	4	5	6	7	8
1	16	13	9	3	14	4	15	6
2	31	0	6	0	22	2	9	0
3	2	2	8	9	9	15	27	28
4	7	2	5	2	31	5	32	6

GROWTH CURVE ANALYSIS

REDUCED LOGISTIC MODEL

Sample	Function Number	Response Function	DESIGN MATRIX 1	2	3	4
1	1	0.05001	1	0	0	0
	2	0.35364	1	0	1	0
	3	0.73089	1	0	2	0
2	1	0.11441	1	0	0	0
	2	1.29928	1	0	0	1
	3	3.52636	1	0	0	2
3	1	-1.32493	0	1	0	0
	2	-0.94446	0	1	1	0
	3	-0.16034	0	1	2	0
4	1	-1.53148	0	1	0	0
	2	0	0	1	0	1
	3	1.60944	0	1	0	2

ANALYSIS OF VARIANCE TABLE

希望电子出版社版权所有

Source	DF	Chi-Square	Prob
MILD DIAGNOSIS, WEEK 1	1	0.28	0.5955
SEVERE DIAGNOSIS, WEEK 1	1	100.48	0.0000
TIME EFFECT FOR STD TRT	1	26.35	0.0000
TIME EFFECT FOR NEW TRT	1	125.09	0.0000
RESIDUAL	8	4.20	0.8387

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
MODEL	1	-0.0716	0.1348	0.28	0.5955
	2	-1.3529	0.1350	100.48	0.0000
	3	0.4944	0.0963	26.35	0.0000
	4	1.4552	0.1301	125.09	0.0000

ANALYSIS OF CONTRASTS

Contrast	DF	Chi-Square	Prob
DIAGNOSIS EFFECT, WEEK 1	1	77.02	0.0000
EQUAL TIME EFFECTS	1	59.12	0.0000

例十二：重复观察的实验，两个重复观察的变量

本数据关系一群病人对两种诊断法 (TRTMENT 下分两类：传统的与创新的) 的反应。反应的结果 (正或负) 分两个时间 (以 1,2 表之) 来观察。如此,诊断法 (TRTMENT) 与时间 (TIME) 构成两个重复观察的变量。

分析的步骤有三：首先，反应次数以饱和模型来处理。其次，将模型简化成主效果模型，以此简化后的模型来处理同样的数据。最后，CATMOD 程序探讨两种诊断法的正确性 (以 ACCURACY 表之)。由于数据不完全，要如何撰写 RESPONSE 指令，请读者仔细研究下面程序的示范：

程　序

```
TITLE 'DIAGNOSTIC PROCEDURE COMPARISON';
DATA A;
    INPUT STD1 $ TEST1 $ STD2 $ TEST2 $ WT @@; CARDS;
NEG NEG NEG NEG 509    NEG NEG NEG POS  4  NEG NEG POS NEG  17
NEG NEG POS POS   3 NEG POS NEG NEG 13  NEG POS NEG POS   8
NEG POS POS POS   8 POS NEG NEG NEG 14  POS NEG NEG POS   1
POS NEG POS NEG  17 POS NEG POS POS  9  POS POS NEG NEG   7
POS POS NEG POS   4 POS POS POS NEG  9  POS POS POS POS 170
;
```

```
PROC CATMOD;
    TITLE2 'MARGINAL SYMMETRY, SATURATED MODEL'; WEIGHT WT;
    RESPONSE MARGINALS; MODEL STD1*TEST1*STD2*TEST2=_RESPONSE_/FREQ NOPA-RM;
    REPEATED TIME 2, TRTMENT 2 / _RESPONSE_=TIME TRTMENT TIME;
PROC CATMOD;
    TITLE2 'MARGINAL SYMMETRY , REDUCED MODEL'; WEIGHT WT;
    RESPONSE MARGINALS; MODEL STD1*TEST1*STD2*TEST2=_RESPONSE_/NOPROFILE
    CORRB;
    REPEATED TIME 2, TRTMENT 2 / _RESPONSE_= TRTMENT ;
PROC CATMOD;
    TITLE2 'SENSITIVITY AND SPECIFICITY ANALYSIS, MAIN-EFFECTS';
    MODEL STD1*TEST1*STD2*TEST2=_RESPONSE_/COVB NOPROFILE;
    REPEATED TIME 2, ACCURACY 2 / _RESPONSE_= TIME ACCURACY; WEIGHT WT;
    RESPONSE EXP  1 -1 0  0 0  0 0  0 ,
                  0  0 1 -1 0  0 0  0 ,
                   0  0 0  0 1 -1 0  0 ,
                   0  0 0  0 0  0 1 -1
           LOG
               0 0 0 0  0 0 0  0 0 0 0  1 1 1 1 ,
               0 0 0 0  0 0 0  1 1 1 1  1 1 1 1 ,
               1 1 1 1  0 0 0  0 0 0 0  0 0 0 0 ,
               1 1 1 1  1 1 1  0 0 0 0  0 0 0 0 ,
               0 0 0 1  0 0 1  0 0 0 1  0 0 0 1 ,
               0 0 1 1  0 0 1  0 0 1 1  0 0 1 1 ,
               1 0 0 0  1 0 0  1 0 0 0  1 0 0 0 ,
               1 1 0 0  1 1 0  1 1 0 0  1 1 0 0 ;
RUN;
```

结　果

结果显示当数据用饱和模型处理时，会对两种诊断法的反应有显著的不同。其余的效果如时间或时间与诊断法之间的交互作用则未达 0.05 的显著程度。以简化的模型解释数据时，也会得到与上述一致的分析结果。

最后，分析诊断的正确性，亦可发现时间的因素与诊断的正确性同时达到 0.05 的显著程度。

报表 16.12　重复观察的实验，两个重复观察的变量

```
                    DIAGNOSTIC PROCEDURE COMPARISON
                  MARGINAL SYMMETRY , SATURATED MODEL
                   CATMOD PROCEDURE
                                        RESPONSE PROFILES
Response：STD1*TEST1*STD2*TEST2          Response  STD1  TEST1  STD2  TEST2
```

Weight Variable：WT

Data Set：A

Frequency Missing：0

Response Levels (R)= 15

Populations (S)= 1

Total Frequency (N)= 793

Observations (Obs)= 15

Sample	Sample Size
1	793

1	NEG	NEG	NEG	NEG
2	NEG	NEG	NEG	POS
3	NEG	NEG	POS	NEG
4	NEG	NEG	POS	POS
5	NEG	POS	NEG	NEG
6	NEG	POS	NEG	POS
7	NEG	POS	POS	POS
8	POS	NEG	NEG	NEG
9	POS	NEG	NEG	POS
10	POS	NEG	POS	NEG
11	POS	NEG	POS	POS
12	POS	POS	NEG	NEG
13	POS	POS	NEG	POS
14	POS	POS	POS	NEG
15	POS	POS	POS	POS

RESPONSE FREQUENCIES

Response Number

Sample	1	2	3	4	5	6	7	9	10	11	12	13	14	15
1	509	4	17	3	13	8	8	1	17	9	7	4	9	170

DIAGNOSTIC PROCEDURE COMPARISON

MARGINAL SYMMETRY , SATURATED MODEL

Sample	Function Number	Response Function	DESIGN MATRIX 1	2	3	4
1	1	0.70870	1	1	1	1
	2	0.72383	1	1	-1	-1
	3	0.70618	1	-1	1	-1
	4	0.73897	1	-1	-1	1

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	2387.12	0.0000
TIME	1	1.11	0.2920
TRTMENT	1	9.08	0.0026
RESIDUAL	1	2.40	0.1215

DIAGNOSTIC PROCEDURE COMPARISON

MARGINAL SYMMETRY , REDUCED MODEL

Response：STD1*TEST1*STD2*TEST2

Weight Variable：WT

Data Set：A

Frequency Missing：0

Response Levels (R)= 15

Populations (S)= 1

Total Frequency (N)= 793

Observations (Obs)= 15

希望电子出版社版权所有

DIAGNOSTIC PROCEDURE COMPARISON

MARGINAL SYMMETRY , REDUCED MODEL

Sample	Function Number	Response Function	DESIGN MATRIX 1	2
1	1	0.70870	1	1
	2	0.72383	1	-1
	3	0.70618	1	1
	4	0.73897	1	-1

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	2386.97	0.0000
TRTMENT	1	9.55	0.0020
RESIDUAL	2	3.51	0.1731

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.7196	0.0147	2386.97	0.0000
TRTMENT	2	-0.0128	0.00416	9.55	0.0020

CORRELATION MATRIX OF THE PARAMETER ESTIMATES

	1	2
1	1.0000000	0.0419376
2	0.0419376	1.0000000

DIAGNOSTIC PROCEDURE COMPARISON

SENSITIVITY AND SPECIFICITY ANALYSIS, MAIN-EFFECTS MODEL

CATMOD PROCEDURE

Response：STD1*TEST1*STD2*TEST2　　Response Levels (R)= 15

Weight Variable：WT　　Populations (S)= 1

Data Set：A　　Total Frequency (N)= 793

Frequency Missing：0　　Observations (Obs)= 15

Sample	Function Number	Response Function	DESIGN MATRIX 1	2	3
1	1	0.82251	1	1	1
	2	0.94840	1	1	-1
	3	0.81545	1	-1	1
	4	0.96964	1	-1	-1

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	6448.79	0.0000
TIME	1	4.10	0.0428
ACCURACY	1	38.81	0.0000
RESIDUAL	1	1.00	0.3178

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.8892	0.0111	6448.79	0.000
TIME	2	-0.00932	0.00460	4.10	0.042

COVARIANCE MATRIX OF THE PARAMETER ESTIMATES

	1	2	3
1	0.00012260	2.292E-06	0.00010137
2	2.292E-06	0.00002116	-5.873E-06
3	0.00010137	-5.873E-06	0.00012697

希望电子出版社版权所有

ACCURACY	3	-0.0702	0.0113	38.81	0.000

例十三：条件式逻辑斯谛分析，1∶4 配对组实验

这个例子旨在示范 RESTRICT 指令的使用。分析的数据来自 Breslow (1982，第 668 页)，是关于酒精对胃肠癌的影响。整个实验设计是采用 1∶4 的配对组结构，亦即针对每一个癌症病人，随机抽取四位配对的健康受试者为控制组。文件 (MATCH) 内含三个变量，分别说明如下：

NC= (1+4)人的小组内酷好喝酒的人数

CASE= 一个二分的反应变量，下分 1 (得了胃肠癌)、2 (尚未得胃肠癌) 两个类别

WT= 各细格内人数的多少

在参数估计的过程中，将第二个参数 (亦即 B2) 的值局限为 1。

(因鉴于 RESTRICT 指令只适用于主机 6.06 或 6.07 版，下列的程序是在 VAX 8*** 机型上执行的。)

程　序

```
DATA MATCH;
    INPUT NC CASE WT @@;
    OFFSET=LOG(NC/(5-NC));
CARDS;
1 1 5 1 2 15 2 1 19 2 2 8 3 1 10 3 2 7 4 1 6 4 2 0
;
PROC CATMOD DATA=MATCH;
    WEIGHT WT;
    POPULATION NC;
    DIRECT OFFSET;
    MODEL CASE=OFFSET;
    RESTRICT B2=1;
RUN;
```

结　果

根据分析的结果，酷爱酒精的人得此癌症的相对风险为：

截距的指数值=$e^{(0.681)}$=1.98

因此，好喝酒的人比滴酒不沾的人多一倍的机会得胃肠癌。

报表 16.13　条件式逻辑斯谛分析，1∶4 配对组实验

```
                    The SAS System
                   CATMOD PROCEDURE
Response：CASE                    Response Levels (R)=     2
Weight Variable：WT                Populations     (S)=     4
Data Set：MATCH                   Total Frequency (N)=    70
Frequency Missing：0               Observations  (Obs)=     7
```

科学电子出版社版权所有

POPULATION PROFILES

Sample	NC	Sample Size
1	1	20
2	2	27
3	3	17
4	4	6

RESPONSE PROFILES

Response	CASE
1	1
2	2

MAXIMUM-LIKELIHOOD ANALYSIS

Iteration	Sub Iteration	-2 Log Likelihood	Convergence Criterion	Parameter Estimates 1	2
0	0	97.040605	1.0000	0	0
1	0	86.887942	0.1046	0.2857	1.0000
2	0	84.545264	0.0270	0.6802	1.0000
3	0	84.545252	1.4287E-7	0.6811	1.0000
4	0	84.545252	2.731E-16	0.6811	1.0000

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	6.89	0.0087
OFFSET	0*	.	.
LIKELIHOOD RATIO	3	6.20	0.1022

NOTE：Effects marked with '*' contain one or more redundant or restricted parameters.

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.6811	0.2595	6.89	0.0087
OFFSET	2	1.0000 *	.	.	.

NOTE：Parameters marked with '*' are restricted.

希望电子出版社版权所有

例十四：交易行为的预测

在这个例子里，我们示范如何预测一位顾客购买某样产品的可能性。文件 (LOAN) 内含四个变量，分别说明如下：

EDUC　　=客户的教育水准，下分高 (HIGH)、低 (LOW) 两组

INCOME　　=客户的经济状况，也分为高 (HIGH)、低 (LOW) 两组

PURCHASE　　=客户是否购买某产品的结果，YES 表示买了，NO 表示始终未买。

WT　　=上述三个变量排列组合后各细格的人数。

实验的假设是认为消费者购买某样产品的可能性受其经济情况 (INCOME) 与教育水准(EDUC)的影响。

下列程序是在 VAX 8*** 主机之 SAS 6.06 版的环境下执行的。

程　序

```
DATA LOAN;
    INPUT EDUC $ INCOME $ PURCHASE $ WT;
CARDS;
HIGH HIGH   YES     54
HIGH HIGH   NO      23
HIGH LOW    YES     41
HIGH LOW    NO      12
LOW  HIGH   YES     35
LOW  HIGH   NO      42
LOW  LOW    YES     19
LOW  LOW    NO       8
;
PROC CATMOD DATA=LOAN ORDER=DATA;
    WEIGHT WT;
    RESPONSE MARGINALS/OUT=PRED;
    MODEL PURCHASE=EDUC INCOME/PRED;
DATA PRED2;
    SET PRED;
    KEEP EDUC INCOME _PRED_;
PROC SORT DATA=PRED2;
    BY DESCENDING _PRED_;
PROC PRINT DATA=PRED2;
RUN;
```

结　果

CATMOD 程序分析的结果再经过 PROC SORT 整理以及 PROC PRINT 打印后，我们不难发现教育程度高的客户一般比教育程度低的客户购买欲强。若将教育程度控制在高

希望电子出版社版权所有

(HIGH) 或低 (LOW)的定值上时，经济状况较差 (LOW) 的顾客，其购买欲比经济情况较优的为强。这两个分析的结果可帮助一间公司决定广告宣传的策略或地区等。

报表 16.14　交易行为的预测

```
                    The SAS System
                   CATMOD PROCEDURE
Response：PURCHASE                 Response Levels (R)=     2
Weight Variable：WT                Populations     (S)=     4
Data Set：LOAN                     Total Frequency (N)=   234
Frequency Missing：0               Observations  (Obs)=     8
                 POPULATION PROFILES
                                   Sample
            Sample  EDUC  INCOME    Size
            ----------------------------
              1   HIGH   HIGH        77
              2   HIGH   LOW         53
              3   LOW    HIGH        77
              4   LOW    LOW         27
                 RESPONSE PROFILES
                Response  PURCHASE
                ------------------
                     1      YES
                     2      NO
          Response       DESIGN MATRIX
    Sample   Function       1       2       3
    ----------------------------------------
        1      0.70130      1       1       1
        2      0.77358      1       1      -1
        3      0.45455      1      -1       1
        4      0.70370      1      -1      -1
            ANALYSIS-OF-VARIANCE TABLE
Source                  DF   Chi-Square      Prob
-------------------------------------------------
INTERCEPT                1       418.36    0.0000
EDUC                     1         8.85    0.0029
INCOME                   1         4.70    0.0302
RESIDUAL                 1         1.84    0.1745
   ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES
                              Standard      Chi-
```

希望电子出版社版权所有

Effect	Parameter	Estimate	Error	Square	Prob
INTERCEPT	1	0.6481	0.0317	418.36	0.0000
EDUC	2	0.0924	0.0311	8.85	0.0029
INCOME	3	-0.0675	0.0312	4.70	0.0302

PREDICTED VALUES FOR RESPONSE FUNCTIONS

		-------Observed-------		-------Predicted------		
Sample	Function Number	Function	Standard Error	Function	Standard Error	Residual
1	1	0.7012987	0.05215845	0.67293982	0.04779446	0.02835888
2	1	0.77358491	0.05748688	0.80803395	0.05158569	-0.034449
3	1	0.45454545	0.05674434	0.48811031	0.05107719	-0.0335649
4	1	0.7037037	0.08787719	0.62320444	0.06486699	0.08049927

OBS	EDUC	INCOME	_PRED_
1	HIGH	LOW	0.80803
2	HIGH	HIGH	0.67294
3	LOW	LOW	0.62320
4	LOW	HIGH	0.48811

希望电子出版社版权所有