SAS软件实习三:方差分析

实验要求：

1. 学会使用SAS软件进行不同的随机化排列——常用试验设计；

2. 常用试验设计的方差分析。

**第一部分 常用试验设计随机化的SAS实现**

基于试验设计原则，设计科学合理的试验是研究人员的基本功。而将试验处理进行随机化是关键。使用随机数字表、随机排列表是很好的方法，因为表中的数字是经过随机产生后定下来的，因此可以重复，但较为麻烦，对于样本量较大的试验，也很费时。在目前计算机应用非常普及的情况下，用统计软件中的随机化功能，事先给出种子数，进行随机化，既简单，又可重复，符合随机化要求。

SAS中的Proc plan过程和随机数函数可用于试验设计。

**Proc plan [seed=随机整数];**

**Factors 行变量名=整数 [ordered] 列变量名=整数 [ordered];/\*默认为随机random\*/**

**[Treatments 处理名=整数（不小于列变量数值）cyclic 2 ] /\*按行安排处理，默认采用随机法random \*/**

**Run;**

SAS系统提供的两个均匀分布随机数函数：

**UNIFORM(seed）：**产生服从均匀分布的随机数。参数seed必须是常数，为0,或5位、6位、7位的奇数。

**RANUNI (seed）：**产生服从均匀分布的随机数。参数seed为小于2\*\*31-1的任意常数。在同一个数据步中对同一个随机数函数的多次调用将得到不同的结果，但不同数据步中从同一种子出发将得到相同的随机数序列。随机数种子如果取0或者负数，则种子采用系统日期时间。

实例如下：

1. 试设计单因素完全随机试验，其中水平数=4，重复数=3。

/\*完全随机排列设计*To generate the completely random design*\*/

DATA RD1;

DO treatment=1 to 4;

DO replication=1 to 3;

r=UNIFORM(2016); OUTPUT;/随机数也可以改为r=ranuni(0)\*/

END; END;

PROC SORT; BY r; RUN;

PROC PRINT noobs;

RUN;

上述程序使用时只需要修改处理数，重复数和seed取值，就可以产生不同的完全随机设计的排列顺序。

医学研究中经常将实验对象按一定数目随机分入不同的实验组和对照组，程序可以略作变化。

实例:对120例病人随机分成两个等比例组，使每组为120例。可以通过plan语句和uniform函数实现病人随机分组。

用PROC PLAN过程进行随机化的SAS程序：

PROC PLAN SEED=210002;

FACTORS n=240; OUTPUT OUT=aaa;

DATA bbb; SET aaa; number=\_n\_;

IF n<=120 THEN group='A'; ELSE group='B';

PROC PRINT NOOBS;

VAR number group;

RUN;

程序先在PROC PLAN语句中给定种子数SEED，本例为210002，通过FACTORS产生1~240之间随机排列的数列，并将结果输出至aaa数据集中，然后用条件语句产生以n为条件的分组结果A和B。用PROC PRINT语句输出VAR后指定的变量值结果，本例的number和group分别对应为序号1~240和组别A、B。事实上，通过改动上述的IF、THEN、ELSE条件语句可很容易实现非等比例或多组的随机化。

用UNIFORM函数实现两组等比例随机化的SAS程序：

DATA ccc;

DO number=1 to 240;

r=UNIFORM(210002);OUTPUT;

END;

PROC RANK data=ccc OUT=ddd;

RANKS r\_rank; VAR r;

DATA eee; SET ddd;

IF r\_rank<=120 THEN group='A';

ELSE group='B';

PROC PRINT NOOBS;

VAR number group;

RUN;

程序中语句r=UNIFORM(210002)根据DO循环确定的次数产生240个0至1之间均匀分布的伪随机数，用PROC RANK过程求出各数对应的秩次，再用条件语句产生以秩次为条件的分组结果A和B。

2. 试设计单因素随机区组试验，其中处理数=8，重复数=4。

/\*随机区组排列设计Random block design\*/

Data new;

Proc plan seed=123456;

Factors blocks=4 ordered plot=8;

Run;

注：修改seed取值可以产生不同的设计。如果不想在output窗口输出结果，而保留到某个数据集中，则在factors语句后面加“/noprint; output out=aaa;”即可。

医学实例：对240例病人按区组随机化方法分成两个等比例组，使每组为120例。

用PROC RANK过程实现区组等比例随机化的SAS程序：

PROC PLAN SEED=210002;

FACTORS block=40 length=6; OUTPUT OUT=fff;

DATA ggg; SET fff; number=\_n\_;

IF length <=3 THEN group='A';

ELSE group='B';

PROC PRINT NOOBS;

VAR number group;

RUN;

用UNIFORM随机化函数实现区组等比例随机化的SAS程序：

DATA hhh;

DO block=1 to 40; DO length=1 to 6;

r=UNIFORM(210002); OUTPUT;

END; END;

DATA iii; SET hhh; number=\_n\_;

PROC RANK OUT = jjj;

RANKS r\_rank; VAR r;

BY block;

DATA kkk; SET lll;

IF r\_rank <=3 THEN group='A';

ELSE group='B';

PROC PRINT NOOBS;

VAR number group;

RUN;

3. 试设计二因素裂区试验，其中主因素水平数=5，副因素水平数=6，重复数=3。

/\*裂区排列设计Split-plot design\*/

Data new;

Proc plan seed=123456;

Factors blocks=3 ordered plot=5 subplot=6;

Run;

4. 试设计5x5拉丁方试验。

/\*拉丁方排列设计5x5 Latin-square design \* /

Data new;

Proc plan;

Factors row=5 ordered col=5 ordered;

Treatments A=5 cyclic B=5 cyclic 2;

Run;

注：通过设置行列数和循环数目可以得到不同的标准拉丁方。具体使用中还需要对行和列进行随机。注意有些结果并不是拉丁方，还需要进行判断。

5. 试设计单因素系统分组设计试验，其中处理（组数）数=5，亚组数=4，重复数=8。

/\*系统分组排列设计Nested blocking design\*/

Data new;

Proc plan;

Factors greenhouse=5 pots=4 plant=8;

Run;

**第二部分 方差分析**

对于某个指标进行方差分析，以分清不同变异来源的差异显著性，进而估计方差各分量或者对不同处理进行多重比较。这是最常用的统计分析方法。SAS提供了ANOVA（Analysis of variance）和GLM（General linear model）两个过程包，用于方差分析。其中ANOVA适用于平衡数据的分析，而GLM适用于非平衡数据，且后者比较耗费时间，当然GLM还有其他的重要分析功能，如回归分析和析因分析等。SAS过程步基本格式如下：

proc anova 选项; /\*或将ANOVA换成GLM\*/

class 分类变量名;

model 依变量=效应项列表; /\*dependent variable, effect factor\*/

[test H=效应项 E=误差项];

[means 效应项/选项];

run;

方差分析各分量进行F测验时，如果效应项的比较项不是误差项，可以用test语句定义。需要进行多重比较时，增加means语句。多重比较方法选项有：LSD或t（最小显著差数法，t测验）；Dunnett('A1')仅与对照处理进行比较；Duncan（新复极差法比较SSR法）；Tukey（Tukey学生氏极差测验HSD法）。SNK（Student-Newman-Keuls Test）；Bonferroni等。

ALPHA, BON, CLDIFF, CLM, DEPONLY,

DUNCAN, DUNNETT, DUNNETTL, DUNNETTU, E, ERROR, GABRIEL, GT2, HOVOUT, HOVTEST,

KRATIO, LINES, LSD, NOSORT, PERITZF, PERITZQ, PLOT, REGWF, REGWQ, SCHEFFE, SIDAK,

SINGULAR, SMM, SNK, T, TUKEY, WALLER, WELCH.

Means语句后面加Hovtest，可进行方差同质性检验（Homogeneity of Variance）。可选的方差齐性检验方法有：hovtest=levene 输出以离差的平方为离差变量的 levene's检验（默认）； hovtest=levene (type=ABS) 输出以绝对离差为离差变量的 levene's检验；hovtest=bartlett 输出 Bartlett 检验（常用）；hovtest=obrien 输出O'Brien's 检验（O'Brien's Spread Variable）；hovtest= bf 输出以绝对离差为离差变量的bf检验（Brown and Forsythe's Test）

方差分析过程：一般先进行方差同质性分析，判定是否可以进行联合方差分析，再根据方差显著性情况增加means语句选项，进行多重比较。

**补充PROC NOVA后缀选项说明：**

DATA=数据集名; 指明进行方差分析的数据集。

MANOVA：需要多变量模型（清除缺值观测）。如果依变量出现缺值，程序将从分析中清除该观测值。当利用互作模型运行PROC ANOVA，期望进行多变量分析时，MANOVA是一个非常有用的选项。

MULTIPASS：重新导入数据集，如有必要，而不是写入依变量取值。该选项将会耗费执行时间和磁盘空间，故很少使用。

NAMELINE=n：指定效应名称长度为n个字符，n取值范围为20~200字符。缺省默认20个字符。

NOPRINT：限制结果正常输出。如果仅仅创造输出数据集，可用此选项。

ORDER=DATA | FORMATTED | FREQ | INTERNAL：指明分类变量水平的顺序。该顺序决定模型中的哪个参数对应与数据的哪个水平。注意该选型作用于所有的分类变量。具体含义参见SAS help。

OUTSTAT=输出数据集名；指明输出结果的数据集，输出项目包括模型中每个效应的平方和，自由度，F统计量和概率水平。

**实例3-1-1 四种药剂对水稻苗高的影响**

**单因素完全随机设计（单向分组资料组内观察值数相等，平衡数据）**

（资料来源：袁志发、贠海燕主编《试验设计与分析》page66）

Data new;

input a$@@;

do i=1 to 6;

Input y@@; output;

end;

Cards;

A1 19 23 21 13 76 19

A2 21 24 27 20 92 23

A3 20 18 19 15 72 18

A4 22 25 27 22 96 24

run;

proc anova data=new;

class a;

model y=a;

means a/t; /\*Dunnett('A1')\*/

run;

还可用另一个过程步进行相似的分析：

proc glm;

class a;

model y=a;

means a/Duncan;

lsmeans a/pdiff e;

run;

**附：解读GLM输出结果中不同类型平方和**

当你使用GLM过程对资料进行方差分析时,在不另加特殊选择的条件下, 系统将自动按２种计算方法求出各因素所引起的离均差平和(Type I ss,　Type Ⅲ SS), 如果在MODEL语句后面加上特殊选择项"/SS1 SS2 SS3 SS4;",可按4种计算方法求出各因素所引起的离差平和(Type Ⅰ SS, ... , Type Ⅳ SS), 也可以任选其中某些项。

当所分析的资料属于平衡资料(即因素的各水平组合下观测的例数相等)时,4种计算方法的结果是一致的,此时,用ANOVA过程取代GLM过程进行方差分析, 速度快、需要的内存少。当所分析的资料属于不平衡资料时,必须用GLM过程,4种计算方法的结果不会完全一致,此时,如何正确地选择其中之一来下统计结论呢?下面给出几条参考性建议:

(1)选择Type Ⅰ SS 算法的结果应满足下列条件之一

①各因素以适当的顺序排列在平衡的方差分析模型中 　　这话的意思是各因素在模型语句中排列的先后顺序与计算结果有密切关系, 并且资料属于平衡资料。一般来说, 单个因素(称为主效应)按其在专业上的作用由大到小的顺序排在模型语句中,交互作用项应放在主效应之后。如: MODEL Y=A B C　A\*B A\*C B\*C A\*B\*C;

②纯嵌套模型 　　因素按其在专业上作用大小,从主要到次要的顺序排列在纯嵌套模型中。如: 　　MODEL Y=A B(A) C(B A); 　　意思是A因素的各水平中包含着B因素,B因素的各水平中又包含着C因素。

③多项式回归模型 　　在模型语句中,低次项必须在前,高次项必须在后。如: MODEL Y=X1 X1\*X1;

(2)选择Type Ⅱ SS 算法的结果应满足下列条件之一

①任何平衡的模型;

②任何主效应模型;

③任何纯回归模型;

④一个效应不包含在任何其他效应之中。

(3)选择Type Ⅲ SS 和Type Ⅳ SS应满足下列条件之一

这２种方法的计算结果在绝大多数场合下是完全一致的, 当资料中有缺失网格(Missing Cells,即因素的某些水平组合下根本没有试验数据),且某效应包含在另一个效应之中时, 两法有些差别,需借助专业知识帮助选定。

①一个效应不涉及其他不包含该效应的参数;

②因素的效应与因素在模型中排列的先后顺序无关;

③算法所依赖的各种假设与资料中有、无缺失网格无关,它们不是网格中数据个数的函数;

④各因素的离差平和之和通常不等于由整个模型所算得的离差平和,Type Ⅱ SS有时也有这一特性,只有Type Ⅰ SS不具有这一特性。

总之,当使用GLM过程分析资料后, 只是在不平衡资料(包括平衡不完全配伍组设计资料)时才会面临从4种算法中选择一种,通常只需从Type Ⅰ SS和Type Ⅲ SS中选择一种。

**单因素完全随机设计（单向分组资料组内观察值数不相等，非平衡数据）**

**实例3-2-4 三头母猪所产仔猪的断奶体重（公猪相同）**

（资料来源：袁志发、贠海燕主编《试验设计与分析》page78）

Data new;

do a=1 to 3;

Input n;

do i=1 to n;

input y@@; output;

end; end; drop n i;

cards;

8

21 29 24 22 25 30 27 26

10

20 25 25 23 29 31 24 26 20 21

6

24 22 28 25 21 26

;

run;

proc glm;

class a;

model y=a;

means a/Duncan;

lsmeans a/pdiff e;

run;

尝试调用ANOVA过程：

proc anova;

class a;

model y=a;

means a/Duncan;

run;

**说明：资料为不平衡数据，宜选用GLM过程分析结果。**

**单因素系统分组设计（单向分组资料组内又分亚组的数据）**

**实例3-2-5 培养液处理对盆栽作物株高的影响（4液，3盆，4株）**

（资料来源：袁志发、贠海燕主编《试验设计与分析》page79）

data daten;

do i=1 to 4;

do trt=”A”, “B”, ”C”, ”D”;

do pot=1 to 3;

input y@@;

output;

end;

end; end;

cards;

50 35 45 50 55 55 85 65 70 60 60 65

55 35 40 45 60 45 60 70 70 55 85 65

40 30 40 50 50 65 90 80 70 35 45 85

35 40 50 45 50 55 85 65 70 70 75 75

;

run;

proc anova;

class trt pot i;

model y= trt pot(trt); /\* 定义嵌套效应模型 \*/

test H=trt e=pot(trt); /\* 指明亚组效应pot(trt）为处理效应误差；若无特别指明，则剩余误差为F测验误差 \*/

means trt/Duncan e=pot(trt);

run;

试比较有、无误差定义的分析结果！

**实例3-3-2小麦品比试验**

（资料来源：袁志发、贠海燕主编《试验设计与分析》page100）

单因素随机区组设计试验的方差分析（双向资料组内单个观察值）

data WC;

do a=1 to 8;

do b=1 to 3;

input x @@; output;

end; end;

cards;

10.9 11.3 12.2

10.8 12.3 14.0

11.1 12.5 10.5

9.1 10.7 11.1

11.8 13.9 14.8

10.1 10.6 11.8

10.0 11.5 14.1

9.3 10.4 12.4

; run;

proc anova;

class a b;

model x = a b; /\*没有互作项。比较添加a\*b后的结果 \*/

means a /LSD;

run;

方差分析结果可与教材表3-3-9（page106）比较。

**二因素随机设计试验的方差分析（组内有等重复观察值的双向分组资料）**

**实例3-2-10 不同土质、施肥条件下树苗高度资料**

（资料来源：袁志发、贠海燕主编《试验设计与分析》page90）

data treeheight;

do a=1 to 3;

do i=1 to 3;

do b=1 to 4;

input y @@;

output;

end; end; end; drop i;

cards;

|  |  |  |  |
| --- | --- | --- | --- |
| 31 | 45 | 76 | 60 |
| 35 | 43 | 71 | 64 |
| 30 | 47 | 75 | 62 |
| 50 | 61 | 88 | 74 |
| 52 | 63 | 90 | 75 |
| 54 | 65 | 92 | 70 |
| 40 | 55 | 86 | 70 |
| 45 | 52 | 84 | 71 |
| 44 | 56 | 84 | 69 |

;

run;

proc anova;

class a b;

model y = a b a\*b;

means a b/Duncan; /\*可设置alpha=0.01，默认为0.05\*/

run;

**单个拉丁方试验的方差分析**

**实例3-4-1水稻品比拉丁方试验产量结果**

（资料来源：袁志发、贠海燕主编《试验设计与分析》page114）

data Riceyield;

Input r$ c$ v$ y;

cards;

r1 c1 D 37

r1 c2 A 38

r1 c3 C 38

r1 c4 B 44

r1 c5 E 38

r2 c1 B 48

r2 c2 E 40

r2 c3 D 36

r2 c4 C 32

r2 c5 A 35

r3 c1 C 27

r3 c2 B 32

r3 c3 A 32

r3 c4 E 30

r3 c5 D 26

r4 c1 E 28

r4 c2 D 37

r4 c3 B 43

r4 c4 A 38

r4 c5 C 41

r5 c1 A 34

r5 c2 C 30

r5 c3 E 27

r5 c4 D 30

r5 c5 B 41

;

run;

proc anova;

class r c v;

model y = r c v;

means v / lsd;

run;

**二因素随机区组设计**

**实例3-3-5 小麦不育系与恢复系杂交组合产量资料**

（资料来源：袁志发、贠海燕主编《试验设计与分析》page 109）

Data wheat\_FXM;

do r = 1 to 3;

do b = 1 to 4;

do a = 1 to 3;

input y @@; output;

end;

end;

end;

cards ;

|  |  |  |
| --- | --- | --- |
| 5.0 | 4.6 | 4.6 |
| 5.3 | 5.6 | 5.9 |
| 5.6 | 5.8 | 7.4 |
| 5.3 | 5.4 | 5.4 |
| 5.1 | 4.6 | 4.4 |
| 5.4 | 5.4 | 5.2 |
| 5.7 | 5.9 | 6.2 |
| 5.2 | 5.1 | 5.4 |
| 4.9 | 4.8 | 4.8 |
| 5.2 | 5.2 | 6.0 |
| 5.4 | 5.9 | 7.0 |
| 5.6 | 5.0 | 4.6 |

;

run;

proc anova ;

class a b r;

model y = a b a\*b r;

means a b a\*b/ t;

run;

**二因素裂区设计**

**实例3-5-2 中耕施肥处理小麦产量资料**

（资料来源：袁志发、贠海燕主编《试验设计与分析》page 126）

注意：裂区设计各效应误差项不尽相同

Data wheatff;

do a = 1 to 3; do b = 1 to 4; do r = 1 to 3;

input y @@; output;

end; end; end;

cards ;

29 28 32

37 32 31

18 14 17

17 16 15

28 29 25

31 28 29

13 13 10

13 12 12

30 27 26

31 28 31

15 14 11

16 15 13

; run;

proc glm ;

class a b r;

model y = a b r a\*b a\*r /ss3;

test h= r e=a\*r;

test h= a e=a\*r;

means a / duncan e=a\*r;

means b /duncan;

run;

**二因素条区设计**

**实例3-5-4：水稻移栽期和绿肥二因素试验**

（资料来源：袁志发、贠海燕主编《试验设计与分析》page 136）

DATA Ricetq;

DO blk=1 TO 6;

INPUT a $ b $ y @@;

OUTPUT;

END;

CARDS; /\*以下数据直接按照田间排列方式录入\*/

A1 B2 376 A1 B3 386 A1 B1 355

A3 B2 455 A3 B3 476 A3 B1 433

A2 B2 480 A2 B3 496 A2 B1 446

A2 B1 549 A2 B3 533 A2 B2 540

A1 B1 396 A1 B3 388 A1 B2 406

A3 B1 492 A3 B3 482 A3 B2 512

A2 B2 500 A2 B3 482 A2 B1 513

A1 B2 347 A1 B3 337 A1 B1 387

A3 B2 468 A3 B3 435 A3 B1 476

A2 B3 413 A2 B1 469 A2 B2 436

A3 B3 334 A3 B1 436 A3 B2 398

A1 B3 201 A1 B1 298 A1 B2 280

A3 B1 458 A3 B3 413 A3 B2 434

A1 B1 366 A1 B3 333 A1 B2 356

A2 B1 474 A2 B3 425 A2 B2 465

A2 B3 490 A2 B2 509 A2 B1 502

A3 B3 447 A3 B2 473 A3 B1 487

A1 B3 348 A1 B2 356 A1 B1 397

;

PROC ANOVA;

CLASS blk a b;

MODEL y=blk a a\*blk b b\*blk a\*b;

TEST H=a E=a\*blk;

TEST H=b E=b\*blk;

MEANS a / T E=a\*blk;

MEANS b / T E=b\*blk;

RUN;

如果试验数据经过初步整理，将每个区组内的A和B水平按顺序排列，那么建立数据集的过程就可以简化为：

DATA Ricetq;

DO blk=1 TO 6; DO a=1 TO 3; DO b=1 TO 3;

INPUT y @@; OUTPUT;

END; END; END;

CARDS; /\*以下数据按区组、A、B的顺序初步整理\*/

355 376 386 446 480 496 433 455 476

396 406 388 549 540 533 492 512 482

387 347 337 513 500 482 476 468 435

298 280 201 469 436 413 436 398 334

366 356 333 474 465 425 458 434 413

397 356 348 520 509 490 487 473 447

;

**单一自由度的比较**

data w1;

do group=1 to 5;

do i=1 to 4;

input x @@; output;

end; end;

cards;

24 30 28 26

27 24 21 26

31 28 25 30

32 33 33 28

21 22 16 21

; run;

proc glm data=w1;

class group;

model x=group;

means group/hovtest;

means group/ snk bon dunnett('1');

means group/ snk alpha=0.01;

contrast 'A B C D vs E' group 1 1 1 1 -4;

contrast 'C D vs A B' group 1 1 1 1 -4;

contrast 'D vs C' group 1 1 1 1 -4;

contrast 'A vs B' group 1 -1 0 0 0;

run;

作业三：Page168-172

习题3, 5, 6, 7, 8, 9, 10, 11.12, 13, 14, 15.

建议：数据先输入Excel表格，再转帖到SAS数据步中。每种数据首先说明是哪一种类型的数据，注意数据集建立方法。