**方**

**差**

**分**

**析**

1. **方差分析概述**

1、 方差分析的概念

在科学实验中常常要探讨不同实验条件或处理方法对实验结果的影响。通常是比较不同实验条件下样本均值的差异。方差分析是检验两个或多个样本均值间差异是有具有统计意义的一种方法。

具体的说，所谓方差分析就是利用试验观测值得总偏差的可分解性，将不同条件所引起的偏差和试验误差分解开来，按照一定的规则进行比较，以确定条件偏差的影响程度及其相对大小。当已经确认某几种因素对实验结果又影响时，可应用方差分析检验确定哪种因素对实验结果的影响最为显著，以及估计影响程度。

1. 方差分析中的常用术语
2. 因素与处理

因素是影响因变量变化的客观条件；处理是影响因变量变化的人为条件，也可以通称为因素或因子。例如影响农作物产量的因素有气温、降雨量、日照时间等因素。方差分析中所说的因素是指，从对实验产生影响的众多因素中挑选出的，可通过不同的条件或处理以供考察的因素。

1. 水平与水平数

因素在实验中所取的不同条件（或状态），称为该因素的水平。例如，性别因素在一般实验下只研究两个水平：男和女。因素选定的不同条件的个数称为该因素的水平数。例如上面的水平数为2。

1. 重复与重复数

在相同的条件下进行2次及2次以上的实验，称重复实验。重复的次数称为重复数。

(4) 因素的主效应与因素间的交互效应

某因素单独对实验结果所产生的影响或作用，称该因素的主效应。

在多因素实验中两个及两个以上的因素相互作用，联合对实验结果产生的影响或作用，称为交互效应。例如，有A、B量中药物治疗缺铁性贫血，患者12人，分为4组。实验方案是：第一组用一般疗法；第二组在一般疗法基础上加用A药；第三组在一般疗法基础上加用B药；第四组在一般疗法基础上A、B两药同时使用。则有：

① 比较第二组的均值与第一组的均值是否有显著性差异

② 比较第三组的均值与第一组的均值是否有显著性差异

前两项研究的是A与B两因素的主效应。

③ 比较第四组的均值与第一组的均值是否有显著性差异，研究A药对B药的疗效是否有影响。此为对交互效应的影响。

1. 方差分析的基本类型
2. 纯方差分析

其特点是试验中的因素均可以控制，各因素所确定的所有水平都可以通过努力而达到，也就是说，在进行纯方差分析时，要求除研究对象外应该保证其他条件的一致。

1. 协方差分析

它针对一类具有少量的不可控制或难以控制的因素的试验而言，试验时为了能确定这种因素，并努力消除其影响。例如研究身高与体重的关系时要求按性别进行分析，这样消除性别因素的影响，被测对象往往是不同年龄的。要消除年龄的影响，就应该采用协方差分析。

4 、方差分析的应用条件

 应用方差分析对资料进行统计推断之前应注意其使用条件，包括：

（1）可比性，若资料中各组均数本身不具可比性则不适用方差分析。

（2）正态性，即偏态分布资料不适用方差分析。对偏态分布的资料应考虑用对数变换、平方根变换、倒数变换、平方根反正弦变换等变量变换方法变为正态或接近正态后再进行方差分析。

（3）方差齐性，即若组间方差不齐则不适用方差分析。多个方差的齐性检验可用Bartlett法，它用卡方值作为检验统计量，结果判断需查阅卡方界值表。

**二、方差分析的ANOVA和GLM过程**

**1、 ANOVA过程**

ANOVA过程主要用于处理均衡设计（即：对于每个因素、每个水平的观测数是相等的，另外还可以处理拉丁方设计、正交设计等）的一元、多元方差分析和重复测量的方差分析，也可用于多个变量的对比检验。ANOVA过程最后需要用QUIT语句结束运行。

ANOVA过程语句格式：

PROC ANOVA 选择项1；

CLASS 变量；

MODEL 因变量=效应变量/选择项2；

MANOVA H=效应变量 E=效应变量；

BY 变量；

MEANS 效应变量/选择项3；

RUN;

语句说明：

1.选择项1

（1）DATA=SAS数据集名。

（2）MANOVA：要求PROC ANOVA按多元方式删除那些含有丢失值的观测，即只要在因变量中有丢失值，就在分析中删除这个观测。

（3）OUTSTAT=SAS数据集：定义一个输出数据集，其中含有平方和、F统计量以及模型中每个效应的概率水平。如在MANOVA语句中指定了CANONICAL选择项，且没有M=选择项，数据集中还含有典型分析的一些结果。

2.CLASS语句

在ANOVA过程中要使用的分类变量必须在CLASS语句中说明。在方差分析中，区分分类水平的变量叫分类变量，分类变量必须在CLASS语句中说明。CLASS语句中的变量取值个数叫水平数。分类变量的取值可以是数值型，也可以是字符型，但所表示的意义都是不同水平，而因变量Y则必须是连续型的数值型变量。

3.MODEL语句

MODEL语句用来指明效应变量和自变量的效应。如果没有指明效应，则ANOVA只拟合截距项，也只检验效应变量是否为0。

选项2

（1）INT|INTERCEPT：要求ANOVA过程把截距作为一个效应进行处理，打印出与其有关的假设检验结果。ANOVA过程在模型拟合时总是含有截距，但是，当这个选项缺失时，不输出与其有关的假设检验结果。当截距项在模型中不显著时，可以用NOINT选项去掉它。

（2）NOUNI：不输出单变量分析结果。

4.MANOVA语句

如果MODEL语句中含有多于一个的效应变量，就可以使用MANOVA语句要求进行多元方差分析。使用了MANOVA语句之后，如果某一个观测自变量或效应变量中有丢失值，则在方程分析中删除这条观测。

（1）H=效应变量：指定模型中一些效应作为假设检验的矩阵。对于每个H矩阵，H=选择项打印出的特征根及特征向量（其中E是和误差效应相关的矩阵）等。

（2）E=效应变量：规定该变量为误差效应，如果缺省，则使用误差SSCP（残差）矩阵。

5.MEANS语句

用来计算MEANS语句中列出的每个效应所对应的因变量各水平的均值。

选择项3：

（1）BON：对于MEANS语句中的主效应均值之差进行Bonferroni的t检验，控制试验错误率MEER。就是把每个检验的水平取小一些进行多重t检验。对整个试验选一个水平，然后除以检验次数。如要进行10次检验，而将MEER控制在0.05水平，则每次检验需控制在0.005水平。

（2）DUNCAN：对MEANS语句中指定的全部效应执行组间多重比较的Duncanny’s检验。

（3）DUNNETT：对MEANS语句中指定的全部效应执行各组与对照组间多重比较的双尾Dunnett’s t检验。如果要指定对照组所对应的（分组）变量水平，则将该水平所对应的格式化变量值以单引号括起至于选项后的圆括号中。如果要为多个主效应指定对照组，则将各效应之对照组所对应的变量值（先以单引号括起）以空格分隔置于选项后的圆括号中。默认情况下，各效应的第一个水平将被作为对照组使用。例如：MEANS A/DUNNETT(‘A1’)；其中‘A1’是A效应的对照组取值。MEANS A B C/DUNNETT(‘A1’ ‘B1’ ‘C1’)；‘A1’、‘B1’、‘C1’分别是A、B、C各效应的对照组取值。

（4）DUNNETTL：进行DUNNETT单尾t检验（DUNNETT’s One-Tailed T-test）。它检验是否一处理显著小于对照组。

（5）DUNNETTU：进行DUNNETT单尾t检验（DUNNETT’s One-Tailed T-test）。它检验是否一处理显著大于对照组。

（6）REGWQ：对应MEANS语句中的所有主效应进行多重F检验。

REGWQ检验控制MEER。BON和t检验都用到了比较每一对均值的检验，而REGWQ检验是一种多级检验。它先检验所有N歌均值间的差异，再对N-1歌均值的各集合进行检验，再对N-2个各集合进行检验，依此类推。

（7）SNK：对MEANS语句中指定的全部主效应执行Student-Newman-Keuls组间多重比较。

（8）T|LSD：对应MEANS语句中所有主效应进行两两t检验，相当于在单元观测数相等时的Fisher’s检验。用重复t检验控制CER（比较错误率），每对做一次，控制每对检验的错误率来控制CER。

（9）GABRIEI：对于MEANS语句中的所有主效应进行Gabriel多重对比检验。

（10）CLM：要求ANOVA过程对MEANS语句中变量的每个水平的均值的结果按置信区间形式输出。

（11）LINES：当设置“BON”、“DUNCAN”、“REGWQ”、“SNK”、“T”等选项时，将多重比较结果以如下方式显示：将各组均值以降序方式排列，用线段将差异不显著的均数相连接。此选项不能与“DUNNETT”、“DUNNETTL”、“DUNNETTU”选项一起使用。

（12）ALPHA=p：给出均值之间对比检验的显著水平。一般p值可以取0.0001到0.9999之间的一切值。

***注：常用SNK法（Q检验）、LSD|T法、DUNCAN法（新复极差法）、BON法、REGWQ法。其中REGWQ法和BON法控制实验错误率（EER），REGWQ的检验功效比BON好。***

***LSD|T法（最小显著差法）、DUNCAN法（新复极差法）控制的是比较错误率（CER），无法控制实验错误率(EER)。***

6.输出说明

ANOVA过程中都要输出两个基本方差分析表，其中各项含义分别为：

（1）CLASS语句指名的变量名字。

（2）CLASS变量的水平个数及变量的各水平取值。

（3）数据集中的观测数及由于丢失值而不使用的观察数。

（4）SUM:因变量的总平方和。

（5）SS：模型的平方和。

（6）误差平方和

（7）误差均方（MSE），它是的估计值

（8）F值，它等于模型均方除以误差均方，用以检验模型的有效性，即检验除截距之外的任何其他参数均为0。同时用于检验效应的各个组均值是否相等及模型的显著性。

（9）Prob>F：F统计量的显著概率

（10）R Square：复回归系数，它等于模型均方除以总平方和；用于度量在因变量的变差中能够由模型决定的比例有多少。

7.方差分析中的效应表示方式（也适用于GLM过程）

（1）主效应就用变量本身表示，如A、B、C等。

（2）交互效应可在效应变量之间加上\*（星号）表示，如A\*B,A\*C,B\*C。

（3）嵌套效应是在主效应或者交叉效应之后用一圆括号括起来的一个变量或一串变量来表示，如B(A)、C\*D(A B)。

举例说明：

如果模型中A、B、C是主效应，Y是响应变量，则如下方式分别表示;

Proc anova;

Class A B C;

Model Y=A B C; 主效应模型：如果模型中A、B、C是主效应，Y是响应变量.

Model Y= A B C A\*B A\*C B\*C A\*B\*C;模型中带有交互效应

Model Y=A B(A) C(A) B\*C(A)；模型中含有嵌套、交叉及主效应

**2、 GLM过程**

GLM过程是一个非常通用的方差分析方法，GLM过程用到的统计方法 ：回归分析、方差分析、协方差分析、多元方差分析和偏相关，GLM过程同时还提供了多种诊断方法：随机效应检验、常用的假设检验对比估计和多变量的对比检验等。相对于ANOVA过程，GLM过程在处理不均衡设计时更有效。GLM采用最小二乘估计法拟合一般线性模型，在此基础上进行其他的分析。最后需要QUIT语句从过程中退出。

GLM过程语句格式：

PROC GLM 选择项1；

CLASS 变量；（以上两个语句必须出现在MODEL语句之前）

MODEL 因变量=自变量/选择项2；

CONTRAST ‘LABEL’EFFECT(效应名) VALUES(取值)<效应 取值>.../选项3；

LSMEANS 效应/选择项4；

MEANS 效应/选择项5；

OUTPUT OUT=SAS数据集 关键字=新变量名；

RUN;

语句说明：

1.选择项1

（1）DATA=SAS数据集。

（2）MANOVA:要求删去具有缺失值的观测的多元模型，同ANOVA过程中的一样。

2.CLASS语句

说明分类变量，同ANOVA过程一样；它必须在MODEL语句之前。

3.MODEL语句

MODEL语句用来指明因变量和自变量效应。如果没有指明效应变量，则GLM过程只拟合截距，也只检验因变量是否为0。

选择项2：

（1）INT|INTERCEPT：要求GLM过程把截距作为一个效应进行处理，打印出与其有关的假设检验结果。PROC GLM在模型拟合时总是含有截距，如果缺省，则不打印出与其有关的假设检验结果。

（2）NOINT：当截距项在模型中不显著时，可以用NOINT选项去掉它。

（3）ALPHA=p：指定置信区间的水平。p值通常取0.01、0.05、0.10。如果缺省此选择项，则用0.05水平。

（4）CLI：对每个观测的预测值打印出置信限，CLI不能同CLM一起使用。

（5）CLM：对每个观测的平均预测值打印出置信限。

（6）P：对自变量没有丢失值的观测打印出观察值、预测值、残差值等。

（7）SOLUTION：要求打印出正规方程的解（即参数估计值），当用CLASS语句时，SOLUTION是有用的。当没有使用CLASS语句时，如果缺省该选项，GLM也会打印出一个解。

4.CONTRAST语句

CONTRAST ‘LABEL’EFFECT VALUES;

CONTRAST ‘LABEL’INTERCEPT VALUES;

CONTRAST ‘LABEL’EFFECT VALUES<效应 取值></选项3>；

CONTRAST语句使我们可以用自定义的方式进行假设检验，一个GLM过程中可以同时使用多条CONTRAST语句，但CONTRAST语句必须出现在MODEL语句后。如果同时使用CONTRAST、MANOVA、REPEATED、TESE语句，则相应的对比检验，将作为MANOVA、REPEATED、TESE分析的一部分而实现。

CONTRAST语句中的各项说明：

（1）‘LABEL’：为CONTRAST语句中的标签，是必须的，作为标签的字符串需要用单引号括起来。

（2）EFFECT|INTERCEPT:为所有检验的效应，它必须是MODEL语句所定义的效应之一。当模型中拟合截距项时，则INTERCEPT可作为一个效应。

（3）VALUES：它是和效应相关的L向量的元素，用以指定相应效应各水平的值，从而进行假设检验。

选择项3：

（1）E：输出整个L向量。

（2）E=EFFECT：指明模型中的某个效应作为误差项。GLM过程将把这个效应作为单变量F检验的分子。

5.LSMEANS语句

计算LSMEANS语句中的每一个效应的最小二乘均值。

边缘均值的最小二乘估计（LSM）是针对非均衡设计的，而类和子类的算术平均值是对均衡设计的。LSM指数假设设计是均衡时，类或子类边缘均值所期望的简单估计。

对于模型中涉及分类变量的效应都能计算出最小二乘均值。

6.MEANS语句

对于模型中不含有连续变量的任何效应，都可以出现在MEANS语句中，并且MEANS语句对于这些效应的各个水平计算对应的因变量之均值及其标准误差等。

假设a和b两个效应每个均有两个水平，则语句：

Proc glm；

Class a b；

Model y=a a\*b；

Means a\*b/bon；

对于a\*b的四个水平组合分别打印出相应的均值和标准差。但对于下面的模型：

Model y=x a a\*x；

其中x是连续型变量，进行协方差分析。效应x和a\*x就不能用于MEANS语句中。

选择项5：

参见ANOVA过程中的MEANS语句中的选项说明。

7.OUTPUT语句

格式：OUTPUT OUT=SAS数据集

PREDICTED|P=NAMES RESIDUAL=NAMES

L95M=NAMES U95M=NAMES

L95=NAMES U95=NAMES

STDP=NAMES STDR=NAMES

STUDENT=NAMES;

说明：

STDP=NAMES：平均预测值的标准误。

STDR=NAMES：残差的标准差。

STUDENT=NAMES：学生化残差。

8.GLM模型的参数说明

GLM过程按照MODEL语句中的说明构造一个线性模型。每个效应都在设计矩阵X中产生一列或几列。

**三、单因素方差分析**

**1、实例**

某厂进行合成反应试验，与考察某种触媒用量对合成物产出量的影响。先选取三种触媒用量A1，A2，A3。各做4次试验，实验数据见表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| A1 | 74 | 69 | 73 | 78 |
| A2 | 79 | 81 | 75 | 78 |
| A3 | 82 | 85 | 80 | 79 |

**data** sasuser.examine;

\*list方式导入数据;

input level $ x @@;

cards;

A1 74 A1 69 A1 73 A1 67

A2 79 A2 81 A2 75 A2 78

A3 82 A3 85 A3 80 A3 79

;

**run**;

**proc** **univariate** data=sasuser.examine normal;

\*利用univariate作正态性检验;

class level;

**run**;

**proc** **glm** data=sasuser.examine;

class level;

\*level是分类变量，class语句必须在model语句前;

model x=level;

\*指明效应变量WSC和自变量variety的效应;

means level/duncan hovtest;

\*指定用DUNCAN进行多重比较，但还要用hovtest做方差齐性检验;

means level/dunnett ('A1');

**run**;

**quit**;/\*当数据经检验不满足方差分析的前提条件时，增加此过程进行非参数检验\*/

**proc** **npar1way** wilcoxon;

class level;

var x;

**run**;

输出结果：

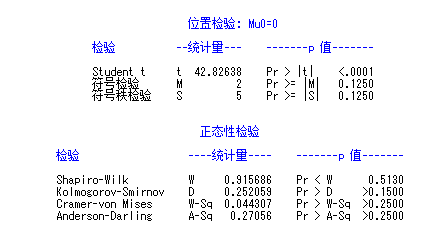
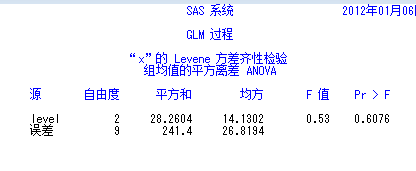


图1 正态性检验

由于样本数小，所以应采用Shapiro-Wilk法进行正态性检验，W=0.915686，Pr<W=0.5130，大于0.05，可以接受数据服从正态分布的假设条件。在符合正态性的情况下，可以采用Student’s（学生 t）检验进行总体均值为0的近似t检验。此处t=42.82638，Pr>|t|<0.0001，说明总体均值极显著不为0，因此可以使用方差分析法进一步对数据进行分析。

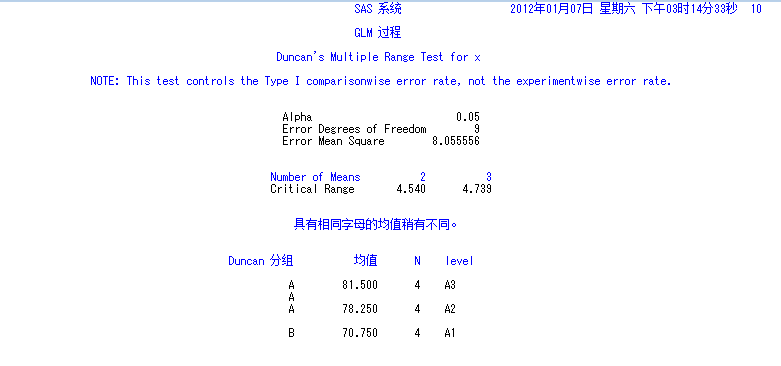


**图2 方差齐性检验**

在比较之前一般要考虑方差齐性的问题。这里可以用GLM过程中的HOVTEST选项实现，实际上这也是在SASA的GLM过程中做方差齐性检验的唯一途径。此处的F值是0.53，Pr>F为0.6076，大于0.05，说明不同组的x方差齐。

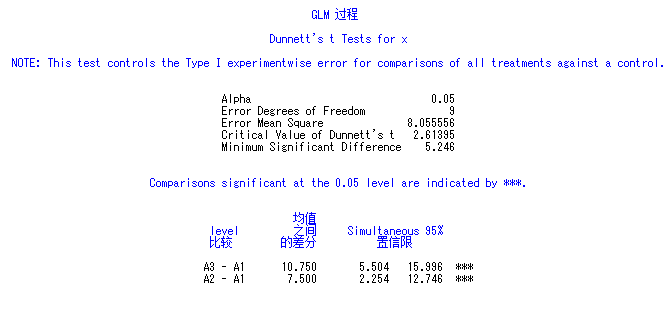
**图3 总体图**

Pr>F为0.0013，小于显著性水平0.05，则可推断因素A是显著的，即三种触媒用量水平对合成物的影响时有差异的。



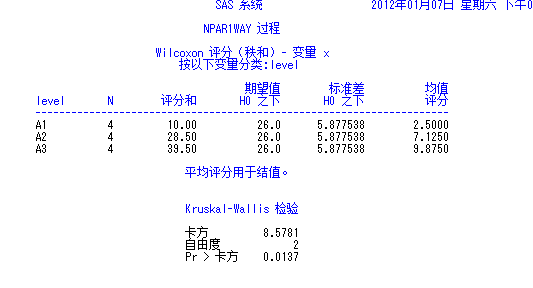
**图4 DUNCAN多重比较**

两组之间的比较看第一列的字母有相同者，表明无统计学意义的差异，不同者有差异。由此可知A1与A2、A3有显著差异，而A2与A3无显著差异。



**图5 ——处理组与对照组的比较**

输出中标有“\*\*\*”的2行表明各组均值与对照组a1均值间的比较在0.05水平上是显著的。具体地说，两均值之差的最小显著界限值是5.246，所给出的上、下是均值之差的95%置信限，从这里也可得出差异显著的结论。 A3均值高于A1均值1.750，A2均值高于A1均值7..500，说明A3、A2两个用量水平下的合成物产出量比对照组A1都高，是较好的水平。



**图6——非参数检验**

1. **双因素试验方差分析**

**1、实例分析（双因素不重复）**

为研究三聚磷酸纯度与聚合时间及温度的关系，经试验得表4-1的纯度数据，试在显著性水平为0.05下分析聚合时间及温度对纯度的影响。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 温度tp  纯度  时间time | 320 | 340 | 360 | 380 |
| 1 | 82.7 | 82.0 | 88.5 | 86.0 |
| 2 | 92.1 | 94.5 | 94.0 | 95.0 |
| 3 | 90.2 | 94.0 | 96.0 | 94.5 |

**data** a;

Do time=**1** to **3**;

Do tp=**320** to **380** by **20**;

Input Pu @@;

output;

end;

end;

Cards;

82.7 82.0 88.5 86.0

92.1 94.5 94.0 95.0

90.2 94.0 96.0 94.5

;

**run**;

**proc** **anova** data=a;

class time tp;

model pu=time tp;

means time tp/regwq snk;

**run**;

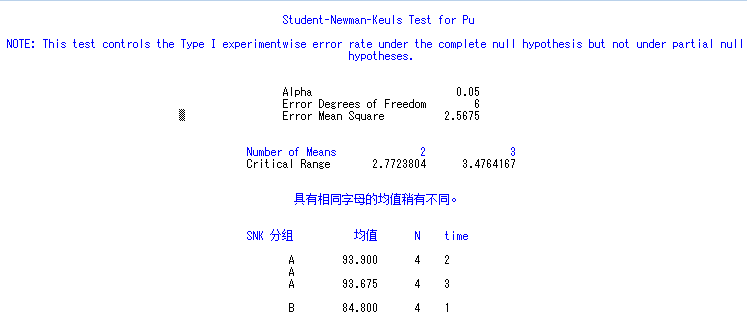
**quit**;

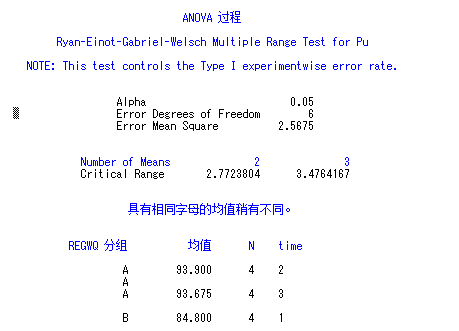
输出结果：



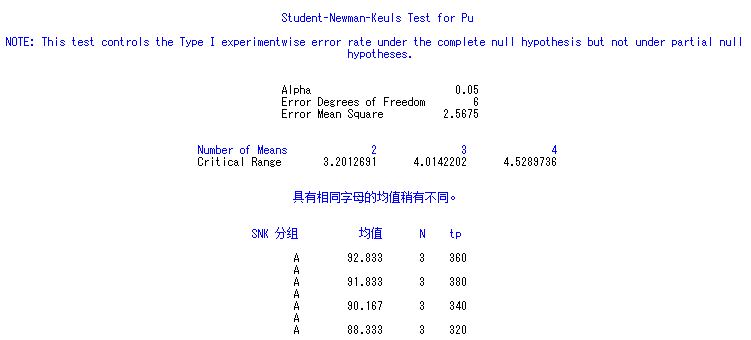
**图4-1 总体的分析表**

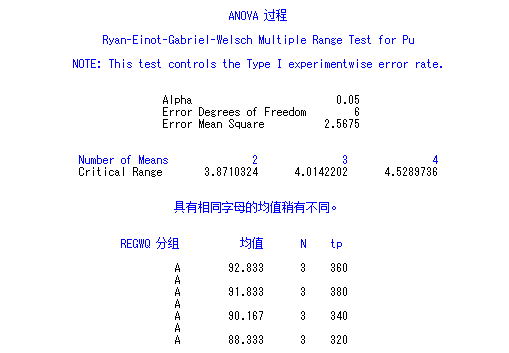
从结果看，tp因素之间均无显著差异， 但time因素之间有十分显著的差异。





REGWQ和SNK两种方法均检验的time为1的组与time为2和3的组有显著差异，time为2的组与time为3的组无显著差异。





REGWQ和SNK对因素tp的多重检验结果，表明tp即温度对纯度的影响不大。总体均值在时间time的影响下有显著的变化。

**实例（双因素重复）**

一位工程师设计研制一种用在某种装置内的电池。工程师决定考察三种极板材料和三种温度（-10℃、20℃、50℃）下对电池有效使用寿命的影响。在每种极板材料与温度搭配条件下检测四个电池。试验观察所得电池寿命数据如表4-2所示。

试在显著水平α=0.05下，对试验数据进行方差分析。

|  |  |  |  |
| --- | --- | --- | --- |
| 温度B  材料A | -10 | 20 | 50 |
| A1 | 130 155  74 180 | 34 40  80 75 | 20 70  82 58 |
| A2 | 150 188  159 126 | 126 122  106 115 | 25 70  58 45 |
| A3 | 138 110  168 160 | 174 120  150 139 | 96 104  82 60 |

**表4-2**

**data** test;

\*利用循环输入数据集;

Do A=**1** to **3**;

Do B=-**10** to **50** by **30**;

Do i=**1** to **4**;

Input y @@;

Output;

End;

End;

End;

Cards;

130 155 74 180 34 40 80 75 20 70 82 58

150 188 159 126 126 122 106 115 25 70 58 45

138 110 168 160 174 120 150 139 96 104 82 60

;

**data** test(drop= i);

set test;

**run**;

\*正态性检验;

**proc** **univariate** data=test normal;

class A B;

**proc** **glm**;

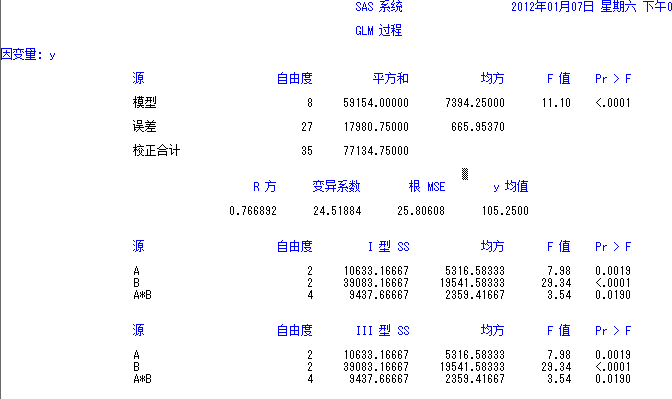
class A B;

model y= A B A\*B;

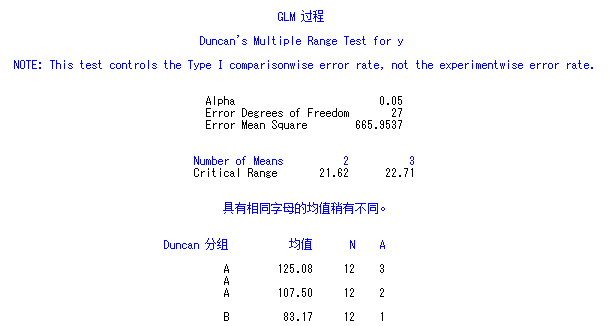
means A B A\*B/duncan snk;

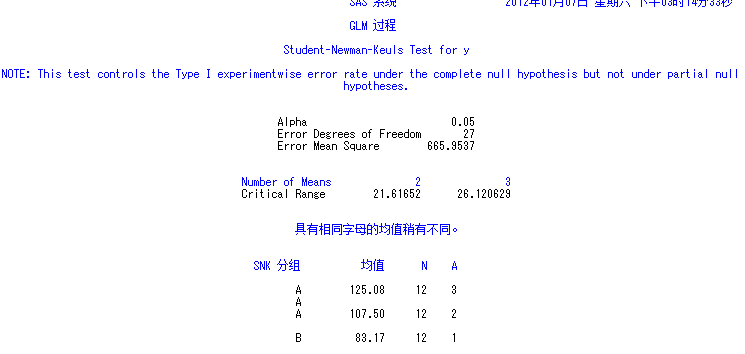
**run**;**quit**;

输出结果：

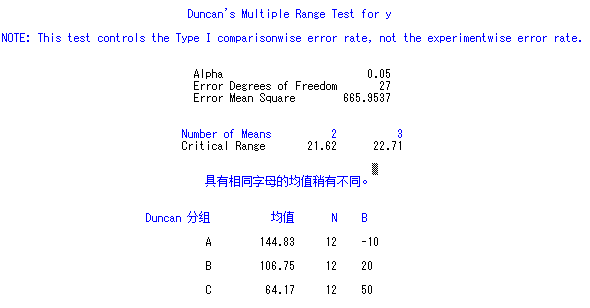


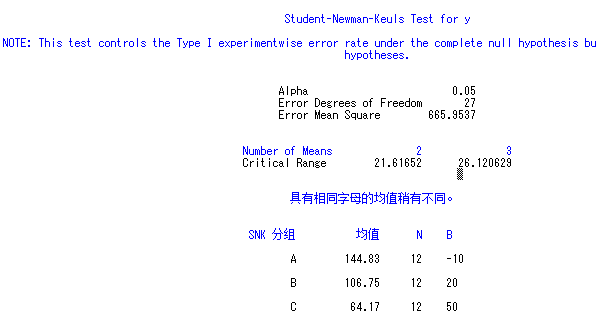
因为因素A、B以及他们的交互作用A\*B的Pr>F均小于显著性水平0.05，所以两个因素及它们的交互作用都对电池寿命有显著的影响。



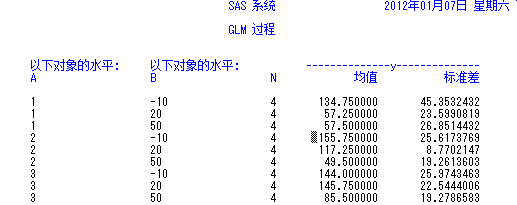
在DUNCAN法对A因素多重比较得出，A2和A3的DUNCAN分组为同一组，则无显著差异。A1与A2、A3有显著差异。

SNK法的检验结果与DUNCAN的检验结果一致。



用DUNCAN进行多重比较，结果说明在B因素的影响下，温度为-10℃、20℃、50℃的均值有显著性差异。

用SNK多重比较得出的结果与DUNCAN的一致。温度为10℃时，电池寿命相对较长。



以上是A\*B交互作用的均值比较，由Mean列可以看出A=2、B=-10时，电池的寿命最长，为155.75。

由此例分析得知A=2 B=-10时，电池寿命最长，为155.75，同时与A=3、B=-10和A=3、B=20较为接近，若考虑经济成本，我们需要比较这两组配比之间是否存在差异。可以使用CONTRAST语句进行比较。

**proc** **glm** data=test;

class A B;

model y=A B A\*B;

contrast 'B(-10) vs B(20)/A3' B **1** -**1** **0** A\*B **0** **0** **0** **0** **0** **0** **1** -**1** **0**;

**run**;

**proc** **glm** data=test;

class B A;

model y=B A A\*B;

contrast 'A2 vs A3/B(-10)' A **0** **1** -**1** B\*A **0** **1** -**1**;

**run**;**quit**;



其他输出内容同上，这里仅给出CONTRAST语句的输出。由输出得到：在A=2，B=-10，A=3、B=-10，A=3、B=20这三种情况下的电池寿命都较长，根据专业知识及考虑经济成本，这三种都可以推荐使用。

小结：当满足正态性和方差齐性时才能进行方差分析，其中方差齐性检验不必单独使用一个T检验，直接在方差分析过程中加入HOVTEST即可。满足条件：分析先看总体模型的Pr>F,再用不同的比较方法做多重比较，最好是同具有分别控制CER和EER错误的方法，如分别控制CER和EER的T和REGWQ同时使用。当有较为接近而又需要继续区分的组别时，可以使用CONTRAST语句进行比较。当不满足方差齐性条件时，则只能用非参数检验。

1. **多因素试验方差分析**

**1、实例分析：**

在某研究所，要研究某一化工产品生产过程中催化剂用量、原料产地、反应时间对产品生成量的影响。催化剂用量取三个水平A1，A2，A3；原料产地取两个水平B1,B2；反应时间取3个水平C1，C2，C3，试验结果见表5-1，试在α=0.01下进行方差分析。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **C**  **A B** | | **C1** | **C2** | **C3** |
| **A1** | **B1**  **B2** | **73.2 73.9** | **74.6 74.7** | **75.1 75.4** |
| **A2** | **B1**  **B2** | **76.2 78.4** | **77.2 78.4** | **77.7 79.2** |
| **A3** | **B1**  **B2** | **75.2 75.9** | **76.3 77.1** | **76.9 77.8** |

**data** test;

do a=**1** to **3**;

do b=**1** to **2**;

do c=**1** to **3**;

do i=**1** to **2**;

Input y @@;

Output;

end;

end;

end;

end;

cards;

73.2 73.9 74.6 74.7 75.1 75.4

76.2 78.4 77.2 78.4 77.7 79.2

75.2 75.9 76.3 77.1 76.9 77.8

;

**data** test(drop=i);

set test;

**run**;

**proc** **glm**;

class a b c;

model y=a b c;

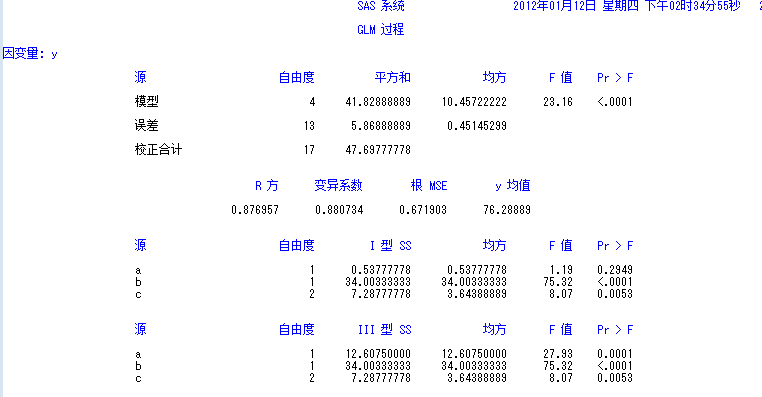
manova h=a b c; \*进行多元方差分析;

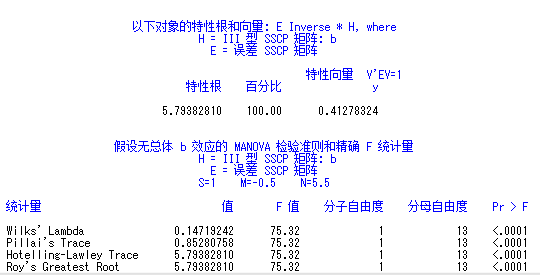
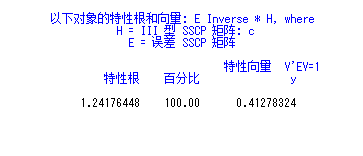
lsmeans a b c/pdiff; \*用最小二乘法检验a b c各因素在不同水平下的两两进行比较;

**run**;

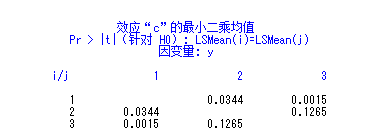
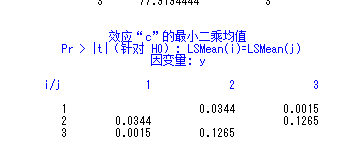
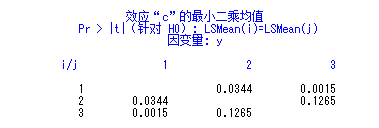
**quit**;

输出结果：



方差分析的总体模型十分显著，因为pr>F为<0.0001。三个因素除了a在第一型离差平方和的Pr>F等于0.2949大于0.1，不显著之外，其他因素均显著。  

以上是MANOVA语句进行的多元方差分析，我们选择Wilk’Lambda检验方法，三因素a、b、c在模型中也不显著，p值均小于0.1。

以上是LSMEANS语句用最小二乘法分别检验a、b、c各因素在不同水平下两两均值相等的概率。所有的概率值均小于0.1水平，说明a、b、c三因素对稻草粉酶解得糖率没有显著影响。

**小结：关键在于用LSMEANS做多因素两两比较**

1. **协方差分析**

协方差分析是有时用来提高实验精确度的一种方法。设在实验中，要考察的指标的响应变量是Y，存在另一变量，例如X，而且Y和X的关系是线性的。还有，设X不能被实验者控制，但可以随着Y一起被观察到。变量X叫做协变量。协方差涉及到如何来调节协变量对被考察的响应变量的效应。如果不能进行这样的调节，则伴随变量会提高误差均方值并使由于处理不同引起的响应变量值之间的真正差异更难检测。这样一来，协方差分析是调节不可控的多余变量的效应的一种方法。正如我们将会见到的，这一方法是方差分析和回归分析的综合。

**实例：**

三台不同的机器为纺织公司生产单丝纤维。研究者感兴趣与决定三台机器生产的纤维的抗断强度是否有差异。但是，纤维的强度亦受它的纤度或粗细所影响；较粗的纤维一般强于较细的纤维。从每台机器上选取五根纤维样品的随机样本，每一样本的纤维强度（y）和对应的直径(x)如表8-1所示。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 机器一 | | 机器二 | | 机器三 | |
| Y | x | y | x | y | x |
| 36 | 20 | 40 | 22 | 35 | 21 |
| 41 | 25 | 48 | 28 | 37 | 23 |
| 39 | 24 | 39 | 22 | 42 | 26 |
| 42 | 25 | 45 | 30 | 34 | 21 |
| 49 | 32 | 44 | 28 | 32 | 15 |

**图8-1**

**data** test;

do n=**1** to **3**;

do i=**1** to **5**;

input y x @@;

output;

end;

end;

cards;

36 20 41 25 39 24 42 25 49 32

40 22 48 28 39 22 45 30 44 28

35 21 37 23 42 26 34 21 32 15

;

**Data** test(drop= i);

set test;

**run**;

**proc** **univariate** normal;

class n;

var y;

**run**;

**proc** **anova**;

class n;

model y=n;

means n/duncan snk;

**run**;

**proc** **glm**;

class n;

model y=n x n\*x/ss1;

**run**;

**proc** **glm**;

class n;

model y=x n x\*n/ss1;

**run**;

**proc** **glm**;

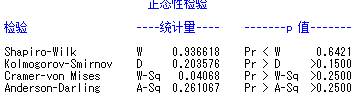
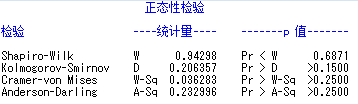
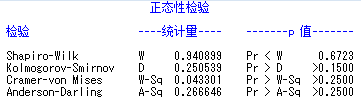
class n;

model y=x n/solution ss3;

lsmeans n/stderr pdiff;

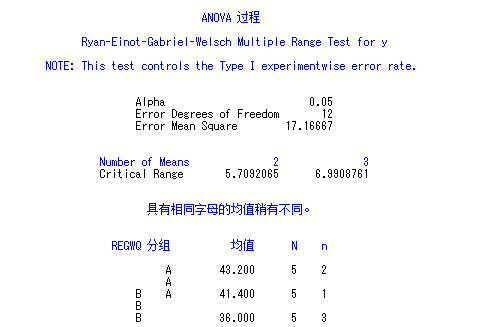
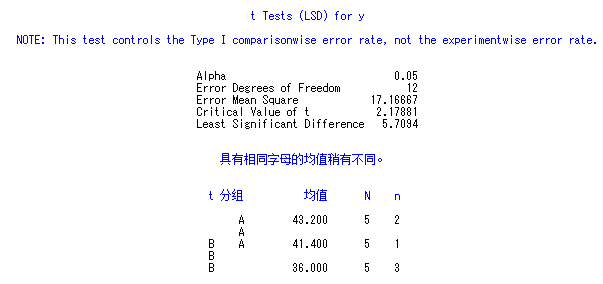
**run**;

**quit**;



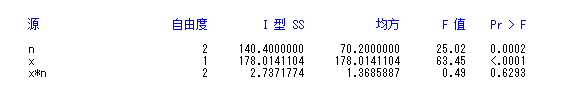
按不同的组做正态性检验。Pr>F值均大于显著性水平0.05，所以可说明服从正态分布。

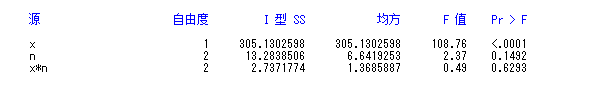


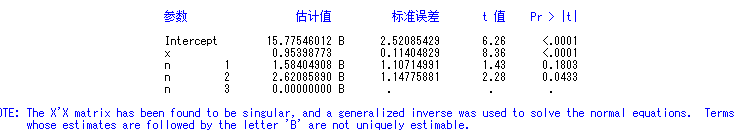


这是ANOVA过程的输出结果，即不考虑x的影响时，机器各水平下样本纤维强度最强的未修正均数之间有显著差异（Pr>F=0.0422,小于0.05）。使用控制试验错误率DUNCAN和比较错误率（T|LSD）进行检验，纤维强度n1与n2、n1与n3之间没有显著差异，都用相同符号标识，但n2与n3之间有差异。

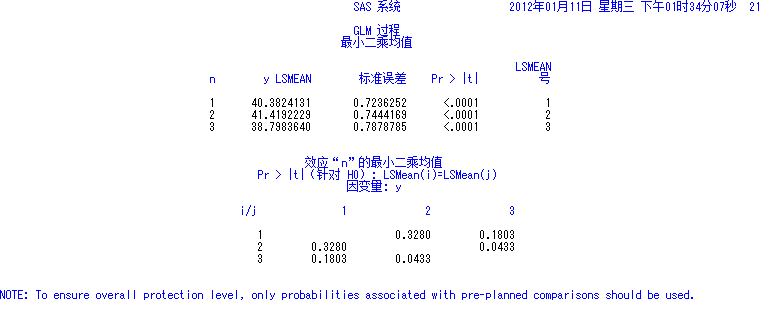
由专业知识可知， 直径x对纤维强度y必存在一定的影响，故选用协方差分析较为合适。





由以上两个图可以得出结论：n和x单独对y的影响是十分显著的，但n\*x对y的影响并不显著，所以三条直线回归方程的总体斜率相等。

这是最后一个GLM过程步输出的第2部分，即对各回归系数检验的结果。对定性变量a而言，是以最后一个水平为基准的，即以第3种机器为对照组。由Pr>|t|一列的输出可以得知：当纤维直径x作为一个因素考察时，机器二与机器三存在显著的差异，因为Pr>|t|=0.0433，小于0.05。机器一与机器三无显著差异，Pr>|t|=0.1803，大于0.05。这一结论与不考虑直径x时的结果相同，说明前面用单因素方差分析法得到的结论正确。然而，有时未考虑与y有关系的因素时，单因素可能会得出不正确的结论，这时则需要用协方差分析。

这是最后一个GLM过程步输出的第4部分，即y的3个修正均数之间的两两比较的结果。设y的修正均数为38.7983640（以n3对照组），因为n1（40.9824131）与n2（41.4192229）均值相等的概率为0.3280，大于0.05；n1（40.9824131）与n3（38.7983640）均值相等的概率为0.1457，大于0.05，所以n1与n2、n1与n3间无显著差异。而n2（41.4192229）与n3（38.7983640）均值相等的概率为0.433，小于0.05，存在显著的差异。扣除了纤维直径x对纤维强度y的影响之后，可看出n2对纤维强度y有较好的影响，修正均值最高（41.4192229）。但由于n1与n2两种机器没有显著差异，所以生产纤维时可以推荐使用n1和n2。具体使用时可进一步分析这两种这两种机器的性能，要考虑生产的经济效益。

小结：先判断自变量是否对响应变量有一定的影响，判断步骤：先不考虑该因变量，对响应变量做方差分析，多重比较，再将该因变量加入到多重比较中，查看结果是否一致，不一致说明因变量对响应变量有影响，这时可做协方差分析。当不考虑与响应变量有影响的因素时，对同一个实验的单因素方差分析和协方差分析得出的结果不一定相同，