第二章　统计学基本知识

§2.1　调查设计与实验设计

资料收集是统计分析的第一步。如第1章所述，其基本方法是抽样和实验。调查设计中，观察者处于被动地位对感兴趣的事物进行研究，如观察吸烟与肺癌是否有关系；实验设计是在严格控制实验条件下，安排实验因素，排除非实验因素的干扰，如物理实验、动物实验和临床试验。实验数据是统计数据的重要来源之一，其处理方法自然是统计软件包处理的典型问题。这里介绍统计实验设计的原则及常见实验设计的概念，更详细的内容可以参阅有关书籍。

§2.1.1　调查设计

1. 普查(mass screening，census)。即全面调查，指对调查范围内全部对象进行调查的方法，其目的是掌握某一时点、一定范围内的研究对象的基本情况，如人口普查、专门人才普查等。人口普查是收集、编制、评价、分析及出版某范围的地区在一个特定时期人口及有关经济和社会资料的全过程，它是重要的国情调查，如1990年全国第四次人口普查。普查的优点是比较有把握地掌握研究对象的基本情况，避免抽样调查中的抽样误差，但需要大量的人、财、物投入，易出现遗漏、由于参与人员多而标准不易统一，大规模全面调查常常只适于做描述研究。

2. 抽样调查(sampling survey)。根据随机的原则，从研究对象的全体中抽取一定数量进行调查的一种调查方法。随机抽样方法有多种，如单纯随机抽样(simple random sampling, SRS)、系统(systmatic) 抽样、分层(stratified) 抽样、整群或集落(cluster) 抽样和多阶段(multistage) 抽样等，在实际应用中可根据具体情况操作和调整。其优点是省时、经济、易于操作，适用范围广，准确性也较好。为保证样本对于总体的代表性，所抽样本应足够大，抽样应随机，调查单位应同分布。

   实际工作中，还可以有其它类型的调查，如个案调查或典型调查，它是在全体研究对象中选取个别研究对象进行的调查。断面调查(cross-sectional study) 是在某一特定时间对调查对象及有关因素进行研究。流行病学研究中还常常用到病例—对照(case-control)研究和队列(cohort)研究(见本章“LOGISTIC 回归”)。

   调查应有计划、有组织、有步骤地进行，调查方案的内容应包括调查目的、调查对象、调查项目、调查表、调查方法、调查人员、调查的组织实施、质量控制与调查进度等。实施时，调查员要经过培训、以保持口径一致，最好进行预调查、复查和数据质量评价等。

   调查数据的分析方法近年来发展很快，其基本思想是考虑这些调查所具有的特征，如基于设计的多阶段抽样调查中的加权，如OSRISIS、SUDAAN、PC CARP ，Stata 5.0 提供许多这一类的方法。考虑数据地域分布的地理信息系统(GIS)方法(如SPLANCS)等。

§2.1.2　实验设计

㈠统计实验设计的原则

1. 对照(control)。为了排除非实验因素的干扰，在进行实验设计时应该设立对照组，并同实验组一样作相同的观察。如观察某种干预措施对儿童缺铁性贫血的影响，可选择一

组儿童给予这种干预，另一组不予干预，间隔一定时间后再对两组儿童的血红蛋白含量进行比较。

2. 随机化(randomization)。是指每一研究单位有均等的机会安排到某个观察组中去，可以消除实验实施时的系统偏差。若无区组，则随机化是随机交换实验的次序以及实际分配因素的水平。出现区组则应对每个区组内的实验进行随机化，然后对区组实验的次序进行随机化。

3. 重复(replication)。指实验组与对照组均要有足够的样本含量，这因为每个观察单位具有变异性，重复观察对于考察响应的平均水平与变动情况是有益的。一种实施办法是在基本的设计中每种情况下的组合做给定的实验数目，另一种是考虑到实验过程须保持一致的环境条件，如温度、湿度，这些条件称做噪声因素。

实验的可靠性是指在同一批对象在不同的时间或等价变量测量时，数据的一致性，也可以指同一总体抽出其它样本时的一致性[12]。

㈡常用实验设计[1]

1. 完全随机化设计(completely randomized design) 是一种最简单的实验设计，没有区组。先根据实验目的选择实验对象，然后用随机化方法将观察单位分别分到实验组和对照组，并给予不同的处理。

2. 随机区组设计(randomized block design) 是将类似的实验对象分到同一组中，称为区组(block)，各区组接受不同的处理，每区组包含的实验单元数为处理数。在每一区组中每个实验对象接受哪一种处理是随机的。配对实验设计是随机区组设计的特例，设计时某些特征相同的两个实验对象配成一对，一个作试验，另一个作对照。在随机区组设计中，若实验的处理数大于每一区组所能容纳的观察单位数时，用平衡不完全区组设计(BIB)，其特点是：所有区组的大小都一样；因子各水平出现次数都一样，每水平在每区组内最多只能出现一次；任一对水平同时出现的次数相同。

3. 交叉实验设计(crossover design) 是把实验对象分为两组，在实验的第一阶段，甲组接受处理，乙组作对照；在实验的第二阶段，乙组接受处理，甲组作为对照。实施时首先将所有对象按某种性状配对，然后随机决定每一对象的试验顺序，经过一个阶段后，处理组与对照组交换。

4. 拉丁方设计(Latin square design) 是使用k 个拉丁字母排成的k x k 方阵的三因素k 水平的设计，在拉丁方中，每个字母在每行每列中只出现一次，实施时先按两因素安排行和列，再按第三个因素的水平随机分到每个拉丁字母。

5. 析因实验设计(factorial experiment design) 是同时检验两种或两种以上因素效应的实验设计。实施方法是首先确定各因素的水平，然后按各因素的各水平的组合确定实验处理，每个受试对象随机地接受处理。其优点是可以了解因素的交互效应。由于随着k 的增大实验次数增加，常常采用部分(fractional) 析因设计。

6. 正交实验设计(orthogonal experimental design) 是在析因设计的基础上发展起来的，其基本思想是用尽可能少的实验次数达到实验的目的。实施时首先确定试验的因素和水平，然后依据正交表安排实验。如$L_4(2)^3$ 表示用4次试验安排一个3个因素，每因素两水平的试验。有时根据需要，可仅使用正交表的部分列来安排试验。

7. 裂区实验设计(split plot design) 是完全随机设计与随机区组设计相结合的设计，同样可做因素和交互效应的研究。在随机区组设计中，只能分析一个处理因素的效应和一个区组效应，若需要分析新的因素但区组对象数无法增加，则应采用裂区实验设计。

8. 嵌套或巢设计(nested design) 是将受试对象成组地分到某试验因素不同水平下的一种设计方法。当试验对象成群出现，不便于对每个对象进行随机化处理，只能成组地随机化分配。

9. 响应曲面设计[6,7] 响应曲面方法(response surface model RSM) 是通过一系列试验来获得响应变量的可靠测量，并且决定一个最合适的模型，最终决定实验因素的最佳配合。响应是测量到的量，其值假设可以通过改变因素的水平而变化。设响应真值可以由因素的某种函数式来表达，对其Taylor 展开式进行一些换算，可把这种关系以多项式的形式表示出来。在几何上，响应与因素的关系式可用超平面表示，也可以等值线图的形式表达。多数响应曲面的研究是一个序贯过程。首先，考虑可能影响响应的因素，然后进行实验，考查这些因素是否真正有影响，再涌现新的想法。

   以下以响应曲面为例进行较详细介绍。

   类似地，分析中的因素是指实验中设定取不同值的变量，一般来说，实验是用因素不同的水平取值来研究感兴趣的响应，直接关心的因素称设计因素，对反应有一定影响但没有直接兴趣的因素称为区组因素。实验的目的之一是避免设计因素与区组因素的混杂。使用正交混杂构造一个设计时，所有因素有相同水平q，q是素数或素数的幂次，一般取值为2，这并不意味着两水平因素的设计不能有两个以上的区组，相反可以用几个两水平的因子来标记两个以上水平的因素，以下的例子是用三个两水平的因素来标记一个8水平的因素。

P_1	P_2	P_3	F
0	0	0	0
0	0	1	1
0	1	0	2
0	1	1	3
1	0	0	4
1	0	1	5
1	1	0	6
1	1	1	7

   因素P_i 仅是用于直接导出因素F 的水平，因而称伪因素(pseudo factors), F 称做导出因素。一般来说，k个q水平的伪因素产生一个q^k 水平的导出因素。区组因素是导出因素，其相关联的P 称作区组伪因素(block pseudo-factor)。在正交混杂设计的构造中，q水平的因素，用q的m 次方个实验，可以区分其前面的m个因素与后面的组合，把前面的m个因素称实验标记因素(run-indexing factors)。设计的分辨(resolution) 能够决定可以独立于其他因素而估计的效应数目。如分辨为5 的设计所有的主效应与两因子交互可以估计。而分辨为4 的实验则某些两因素的交互含有混杂。一般说来，高的分辨需要更大规模的实验。Box 与Draper 列举了响应曲面设计的14 个特性，现列如下：

   1) 在研究区域R 内产生满意的分布。

2) 保证在某点X 处的拟合值$\hat{y}(X)$ 与此处的真值尽量接近。

3) 能方便地看出拟合不当(lack of fit)。

4) 能够进行数据转换。

5) 允许进行区组实验。

6) 允许递增次序的设计能够依次产生。

7) 提供误差的内部估计。

8) 在数据有较大的波动或偏离正态分布假设时不敏感。

9) 只需要最小量的实验单元数。

10) 提供一个简单的数据模式，因而能够地进行一些判断。

11) 计算简便。

12) 当自变量X 发生误差时，设计的行为令人满意。

13) 并不需要自变量不切实际的水平数。

14) 提供一个方差定常(constancy of variance) 假设的检验。

其中较为重要的为1)−3),5)−7),9),11)。

实验正交性(orthogonality) 的含义是拟合的模型的各项相互独立。可旋转性(rotability) 则保证响应的估计仅仅与因素离实验中心的距离有关。响应曲面最基本的问题是估计曲面峰点的坐标，最常用的是一阶设计和二阶设计，其一阶设计模型的形式是：

$$Y = \beta_0 + \beta X + \varepsilon$$

若ε的均值为零，则均值真值为：$\beta_0 + \beta X$

二阶设计包含因素的最高幂次为2，$Y = \beta_0 + X\beta + X'BX$，如：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \varepsilon$$

这时，

$$B = \begin{pmatrix} \beta_{11} & \beta_{12}/2 \\ \beta_{12}/2 & \beta_{22} \end{pmatrix}$$

经过原点位置的移动和坐标的旋转，上式可表达为：

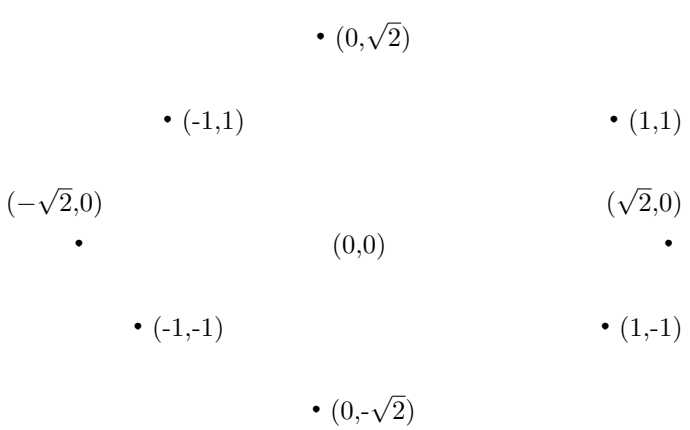
$$y - c_3 = \lambda_1 (x_1 - c_1)^2 + \lambda_2 (x_2 - c_2)^2, \lambda_1, \lambda_2 > 0$$

那么(c_1, c_2, c_3) 即是峰点的一个估计。因为估计二阶函数需要每个因素至少有3个水平，我们可以使用3^k的析因设计来实现，每个因素在-1, 0 和+1 三个水平，k=2 时有九个设计点。当k 增加时，实验的次数猛增，这时可用中心复合设计(central composite design, CCD)来解决，即从通常的析因设计开始，增加2k 个轴点(axial points) 并且在中心点多做几次实验。这种做法的实验次数一般较3^k 要少。

Box-Wilson 设计是一个中心复合设计，由三部分组成：一个完全的或部分的2^k析因设计，因素水平以-1和+1编码，称做析因部分。它有$n_0 \geq 1$ 个中心点，离设计中心α长度的两个轴点，称做设计的轴点。总的设计有$n = 2^k + 2k + n_0$个实验点。如一

个$n_0=1,\alpha=\sqrt{2},k=2$ 的设计为：

$$D=\begin{array}{c}\begin{array}{cc}x_1 & x_2\end{array}\\ \begin{pmatrix}-1&-1\\1&-1\\-1&1\\1&1\\\sqrt{2}&0\\-\sqrt{2}&0\\0&\sqrt{2}\\0&-\sqrt{2}\\0&0\end{pmatrix}\end{array}$$



取$\alpha=\sqrt{2}$设计是可旋转的，因为所有实验点在半径为$\sqrt{2}$ 的球上。在k=3时，α常取为$2^{3/4}=1.682$, $n_0=1$ 时，仅有15次实验，而$3^3=27$次。

Plackett-Burman 引入$n=k+1$ 的k 变量部分2水平析因设计，仅当n 为4 的倍数时可以实现。设计的目的是使设计在此时能以最可能的精度估计所有的主效应。若n 是2 的幂次，并且$n>k+1$，因子间的某些交互影响也能很好地估计。n 是2 的幂次时，Plackett-Burman 设计与标准的部分2水平析因设计等价。它们得到了$n=4,8,12,...,100$ 次实验下$k=3,7,11,...,99$ 个因素的安排方法。构造这种设计，可以这样做：择一由+1 和−1 构成的行，使其+1和−1 的数目分别为$(k+1)/2$ 和$(k-1)/2$，注意这儿由于$k+1$ 是4 的倍数而能够整除。以后列的构造可以从第一列移动一个位置，共移$k-1$ 次。最后，再追加一行皆为−1 的行而得到$n=k+1$ 的设计。现以$n=12,16,20,24$ 为例，设计矩阵的第一行可以是：

n=12 ＋＋－＋＋＋－－－＋－
n=16 ＋＋＋＋－＋－＋＋－－＋－－－
n=20 ＋＋－－＋＋＋＋－＋－＋－－－－－＋＋－
n=24 ＋＋＋＋＋－＋－＋＋－－＋＋－－＋－＋－－－－

要得到完整的设计，循环移动$(n-2)$次，再增加一行符号均为“－”号的行。所有这些设计均是具有复杂alias结构的分解Ⅲ型设计[2]。

Box-Behnken 设计是通过组合2水平析因设计和平衡不完全区组设计(BIB) 而获。如现有一个4 种处理因素6 个区组的BIB，每个处理在实验中出现3 次：

$$\begin{array}{c}\begin{array}{cccc}x_1&x_2&x_3&x_4\end{array}\\ \begin{pmatrix}\star&\star&&\\&&\star&\star\\\star&&&\star\\&\star&\star&\\&\star&&\star\\\star&&\star&\end{pmatrix}\end{array}\quad\begin{array}{c}\begin{array}{cc}x_i&x_j\end{array}\\ \begin{pmatrix}-1&-1\\1&-1\\-1&1\\1&1\end{pmatrix}\end{array}$$

现用右边的2^2 的析因设计去组合，在左边的星号用x_i 代替，右边的星号用x_j代替，没有星号的列填充零，最后再增加几个中点，现增加三个，即可得到一个27个点的设计。一般地，这种设计不是可旋转和区组正交的。

最优设计的准则。线性模型$Y = X\beta_{p\times 1}+\varepsilon$中$\beta$ 的p维可信区域$\beta : (\beta - b)'X'X(\beta - b) \le c$是一个椭圆。$p$个主轴长度的平方是$(X'X)^{-1}$的特征值。一些最有名的选择X的方法就可以从这些值获得。如A、D、E最优。如2^2 设计是正交的，将两因素x_1, x_2 数据进行变换，并取值为± 1，则设计是D最优的。为验证其正交性，记x_1, x_2 两个水平分别是(1,a)、(1,b)，则所有可能的组合是(1,1)，(1,a)，(1,b)，(a,b) ，现记相应的反应观察值是Y_1, Y_a, Y_b, Y_{ab}，关于这些点的响应曲面方程形式为[8]：

$$\begin{pmatrix} Y_1 \\ Y_a \\ Y_b \\ Y_{ab} \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}$$

$$\hat{\beta} = (X`X)^{-1}X`Y = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_a \\ Y_b \\ Y_{ab} \end{pmatrix}$$

$\hat{\beta}_0, \hat{\beta}_1$ 可作为x_1, x_2 主效应的估计。

㈢实验研究中的误差控制与样本含量

误差有几种，即：抽样误差、系统误差、随机测量误差和过失误差。减小抽样误差的方法是使样本在性质上和数量上能够对总体具有代表性，样本含量的估计方法可见有关专著。系统误差往往偏向一个方向，从不同程度上干扰研究结果，应尽力加以避免。偶然造成的随机误差可通过重复测量来消除。过失误差不应出现。

本节最后一部分的内容在SAS/STAT ADX宏与SAS/QC涉及较多，Minitab 10也纳入上述实验设计方法。Epi Info 软件提供了流行病学研究估算样本大小的方法。

§2.2 基础统计分析方法

§2.2.1 基础统计方法

假设$X_1, X_2, ..., X_n$是来自分布$F(x;\theta)$的一个样本，其中分布函数$F(x;\theta)$的形式已知，参数θ未知。参数统计分析包括对参数θ的估计和检验。

记样本统计量$T_n(X_1, ..., X_n)$ 为参数θ的函数$g(\theta)$的一个估计量，它的优良性有一些评价标准，常用的有无偏性(unbiasness) 、优效性(efficiency)、相合性(consistency)与不变性(invariance)等。若$E_\theta[T_n] = g(\theta)$对所有$\theta$和$n$成立，则$T_n$是关于$g(\theta)$ 的无偏估计量，E_θ表示关于θ的期望。T_n是无偏估计量且达到Rao-Cram ér 不等式的下界，则为$g(\theta)$的优效估计量。T_n是无偏估计量且使得估计量方差最小，则称最小方差无偏估计量(MVUE)。统计量T_n是$g(\theta)$的(弱)相合估计量，通常是指当n 增大以概率收敛于$g(\theta)$。若一个估计量的任何函数关于θ的期望值为零，则该估计量是充分的。如当任何分布的期望值与方差是有限的，则样本均值$\sum X_{i=1}^n/n$为无偏估计量。样本方差公式

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2)$$

右端是$(X_1,...,X_n)$ 的二次型，方阵为$I-\frac{1}{n}(1,...,1)'(1,...,1)$，其秩为$n-1$, 故样本方差以$n-1$为分母，且是总体方差的无偏估计。

显然，“样本均值+某个常量”不相合而$[(n-1)/n]\sigma^2$ 则是相合估计量。

假设θ可以表为总体r阶矩$\alpha_1,...,\alpha_r$的函数$\theta=\theta(\alpha_1,...,\alpha_n),\alpha_i=E[X_1^i]$，记$a_1,...,a_r$为相应的样本矩，$a_i=\Sigma_{j=1}^n X_j^i/n$，则矩法是用$a_i$代替$\alpha_i,i=1,...,r$。因由大数定理，若总体的r 阶矩存在，则样本的r 阶矩是依概率收敛于总体的r 阶矩的，这启发我们用样本矩代替总体矩。设$F(x;\theta)$有概率密度函数$f(x;\theta)$，其似然函数为$\Pi_{i=1}^n f(X_i;\theta)$，为样本的联合分布，使其取值最大的估计称做极大似然估计。如正态分布均值与方差的极大似然估计也是其矩估计。矩估计可能不唯一，不同模型的极大似然函数表达式不同，求解时采用的数值方法也就不同，最常用的是牛顿－拉弗森(Newton-Raphson)方法。其他的参数估计方法有贝叶斯估计及最小二乘估计等。

检验统计量服从正态分布、t分布、χ^2分布、F分布时分别称作z检验(u 检验)、t检验、χ^2检验、F检验。

㈠单变量方法

常用描述分布位置的指标有：均值(算术均值、几何均值、调和均值等)、中位数(M)、众数。将样本变量值求和除以样本例数，即是算术均值(AM)。n 个样本变量值的乘积再开n次方就得到几何均值(GM)，变量的调和均值(HM) 是样本各变量值取倒数后求均值，三种均值间有关系：$HM\le GM\le AM$，几何均值与调和均值常用于描述偏态分布数据。一组变量中，50%分位点的数为中位数，众数是样本中出现频数最多的变量值，即频数分布图上对应峰值的变量值，对称分布如正态分布的算术均值、中位数与众数相同。

表示离散的指标有：方差(VAR)、标准差(STD)、绝对偏差中位数(MAD)、极差(R)、四分位差(IQR)。四分位差是样本75%分位点与25%分位点的差，半个四分位差与标准差大致相当。

设$X_1,...,X_n$ 是来自分布$F(x;\theta)$的样本，把它们按升序排列，并记$X_{(1)},...,X_{(n)}$ 就得到了顺序统计量(order statistic)。$X_{(1)}$与$X_{(n)}$称为极值，$X_{(n)}-X_{(1)}$称为极差或样本全距(sample range)。

上述两类指标之间存在着一定的关系，如正态资料：n 约小于12 ，$STD\approx R/\sqrt{n}$; $20<n<40, STD\approx R/4$; n 在100 左右时$STD\approx R/5; n>400$ 则$STD\approx R/6$; 变异系数是样本标准差和算术均值的比值，反映了样本关于均值变异的大小。

位置的一种稳健估计是把数据偏大和偏小的数据去掉的一定比例，这样就得到了截尾均值(α-trimmmed mean)，如20%的截尾均值是两端均去掉20% 后观测的均值。由于不使用仅仅一个数估计，因而一般要较中位数好，中位数是$0.5-\frac{1}{2n}$ 的截尾均值，近似为50% 的截尾均值。截尾均值可以看成一种加权平均，被去掉的数据权重是0，剩下参与估计的权重为1。同样我们可以对不同的观测构造不同的权，这就是M-估计的思想。

描述分类数据的指标常用率(rate)、构成比(percentage)及比(ratio)等。率是一种相对指标，用于某事件发生的频度，如出生率、发病率等。构成比用于描述具有某种特征的对象占所有对象的数目，如某小学的一个班级男生占56%，女生占44%。比例是两相关事物发生或出现次数的比值，如人口统计中的性比。

一元统计图主要有直方图(histogram)和条图(bar chart)、圆图(pie chart)、茎叶图(stem-and-leaf plot)、箱尾图(Box-and-whisker plot)。直方图和茎叶图、箱尾图常用于定量数据的描述，条图和圆图常用于描述定性数据。

茎叶图做法是依全距把变量分成不重叠的区间，其大小一般取作$k10^p, k = 0.2, 0.5, 1$，p的值可正可负。茎叶图的茎由这些区间构成，每区间的观察构成了叶，在整个数据范围内观察数的变化反映了分布的形状。区间有许多种分法。

箱尾图依数据的中位数画一条竖线，图中方框的位置相当于四分位差的位置，横线(Whisker)延至方框两边的半个四分差，更远的点即为异常值。

在SPSS/PC+中的箱尾图画一个方框代表四分极差，它用星号表示中位数。方框两端发出的须一直延伸到不是异常值的点。方框越大，观测越分散。若一组观测的最大最小值到盒子的两端距离小于一个四分极差，则从方框两端发出的尾延伸到最大和最小的观测值(用X表示)，若大于此距离但小于1.5个四分极差，则用记号O (outlier) 标记这个点，更远的点如超过3个四分极差用记号E(Extreme)表示。

正态分布是最常用的连续性分布，常用$N(\mu, \sigma^2)$表示，其中μ和σ^2分别为正态分布的均值和方差。据正态分布的样本$(X_1, ..., X_n)$ 可以得到均值和方差的极大似然估计$\hat{\mu} = \sum_i x_i/n$和$\hat{\sigma}^2 = (n-1)S^2/n$，$\sqrt{n}(\overline{X} - \mu)/\sigma$ 服从正态分布$N(0,1)$, 在σ未知时$\sqrt{n}(\overline{X} - \mu)/\mathrm{S}$ 服从自由度为$(n-1)$ 的t分布。同时，$(n-1)s^2/\sigma^2 \sim \chi^2(n-1)$。总体方差95%可信区间为$[(n-1)S^2/\chi^2_{0.05;(n-1)}, (n-1)S^2/\chi^2_{0.95;(n-1)}]$ 内。

对于非正态分布数据，据中心极限定理，在样本数较大时，$\sqrt{n}(\overline{X} - \mu)/\sigma$ 服从标准正态分布，其中μ为总体均值，σ^2为总体方差，$\overline{X}$为样本均值。对总体均值μ可进行u 检验或t检验。方差具有正态分布$N(\sigma^2, 2\sigma^4/(n-1))$。因此可以使用检验$\sqrt{n}(S^2 - \sigma^2)/\sqrt{2S^4} \sim N(0,1)$。

离散性变量常用二项分布和泊松分布描述。考虑n次独立试验，每次试验中事件E以概率p发生，用随机变量X表示E发生的试验次数x，概率为：

$$P[X = x] = \binom{n}{x} p^x (1-p)^{n-x} \equiv b(x; n, p), x = 0, 1, ..., n$$

$\binom{n}{x} \equiv \frac{n!}{x!(n-x)!}, n! = n(n-1)...1, 0! = 1$，期望为$np$，方差为$np(1-p)$。总体率为$\pi, n\pi =$常数$\lambda$时则用泊松分布近似二项分布。函数形式为：

$$P[X = x] = \frac{\lambda^x}{x!} e^{-\lambda}; \lambda > 0, x = 0, 1, 2, ...,$$

其期望与方差均为λ。

在大样本下可以使用近似$(\hat{p} - p)/\sqrt{p(1-p)/n} \sim N(0,1)$。在$p$值近于1 或0 时，$n$ 应至少为100。

【例2.1 】随机抽取某医院400 份病例，有60% 书写合格，则可信区间为：

$$\hat{p} \pm z_{0.025}\sqrt{\hat{p}(1-\hat{p})/n} = 0.60 \pm 1.96\sqrt{(0.60)(0.40)/400} = (0.552, 0.648)$$

检验H: p=0.50 即合格与不合格各占50%

$$z = (\hat{p} - p)/\sqrt{p(1-p)/n} = (0.60 - 0.50)/\sqrt{(0.50)(0.50)/400} = 4.000$$

$p < 0.001$ 对原假设予以拒绝。

对于分组变量，随机变量具有g个类或格子，共观察 n 个对象，格子或分类的频率为$n_i, i = 1, ..., g, n = \Sigma_i n_i$，则描述$n_i$ 的分布为多项分布，n_i 的均值为np_i，方差为$np_i(1-p_i)$，协方差为：$-np_ip_j$。

检验统计量为Pearson χ^2 适合度检验：

$$\sum_{i=1}^{g} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2_{(g-1)}$$

似然比统计量为：$2\sum_{i=1}^{g} n_i \ln[n_i/np_i] \sim \chi^2_{(g-1)}$

【例2.2 】100 个献血者的血型的分布为：A：35，B：25，AB：5，O：35，问是否符合30:30:10:40的比例？

Pearson χ^2为：

$$(35-30)^2/30 + ... + (35-40)^2/30 = 5.208 < \chi^2_{0.10;3} = 6.251$$

似然比统计量为：2[35ln(35/30)+..+35ln(35/40)]=5.668

当n_i与N_i相比是小值时，分布为超几何分布。具有均值$n[N_i/N]$，方差$n[(N-n)/(N-1)][N_i/N][1-N_i/N]$，协方差$-n[(N-n)/(N-1)][N_i/N][N_j/N]$。

【例2.3 】一种遗传学指标的基因突变率为4/1,000,000，试看25,000 次中小于两个的概率？

这是一个二项分布资料，p很小，使用泊松分布近似

$$P(\leq 1) = P(0;0.1) + p(1;0.1) = e^{-0.1}(0.1)^0/0! + e^{-0.1}(0.1)^1/1! = 0.995321$$

探索性数据分析包括一些统计指标和图形表示，还包括寻找异常点、数据转换、研究模型拟合后的残差。

【例2.4 】表2.1为Hoaglin, D.C.(1983) 描述20 个数据的茎叶图和截尾均值：

表 2.1 20 个分析数据

分析数据		数据的茎叶图	
28	29	-4	4
26	22	-3	
33	24	-2	
24	21	-1	
34	25	-0	2
-44	30	0	
27	23	1	69
16	29	2	1234567899
40	31	3	0134
-2	19	4	0

这里茎的单位为10位数，叶为个位数。截尾均值如下：

$$\begin{aligned}
T_{(0.00)} &= (1/20)\Sigma_{i=1}^{20} x_{(i)} = 435/20 = 21.75 \\
T_{(0.05)} &= (1/18)\Sigma_{i=2}^{19} x_{(i)} = 407/18 = 24.39 \\
T_{(0.10)} &= (1/12)\Sigma_{i=3}^{16} x_{(i)} = 407/16 = 25.44 \\
T_{(0.20)} &= (1/12)\Sigma_{i=5}^{16} x_{(i)} = 308/12 = 25.67 \\
T_{(0.30)} &= (1/8)\Sigma_{i=7}^{14} x_{(i)} = 206/8 = 25.75 \\
T_{(0.40)} &= (1/4)\Sigma_{i=9}^{12} x_{(i)} = 102/4 = 25.50
\end{aligned}$$

频数分布(frequency distribution) 也表示了变量在取值范围内不同区间上绝对数目或相对频率或累积频率的变化。

偏度(skewness) 的测量，常用偏度系数，在箱尾图中Whisker 左右长度表示了偏度。$\gamma = E(X-\mu)^3/\sigma^3$。大的负值常表示左偏，否则为右偏。偏度系数用样本计算时公式为：

$$g = \frac{n\sum_i (x_i - \overline{x})^3}{(n-1)(n-2)S^3}$$

正态大样本时g的均值为0，方差为6/n。

为了去掉偏性，可采用Box-Cox 转换：

$$y = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0; \\ \ln(x) & \lambda = 0 \end{cases}, x > 0$$

Hoaglin, D.C.(1989) 等建议对称性转换的使用方法：a . 在尾部数据不重要时，使用对数转换；b . 尾部的对称较重要时，使用平方根转换；c . 对主要部分分布的偏性和极值的合理偏性间做平衡时，用4 次方根。

有时直接写出其极大似然形式，第 4 章第5节有一个在实验设计中的用例。

对称图(symmetry plot) 是用上下两端的第i 个观察绘图。对称的条件是$X_M - X_{(i)} = X_{(n-i+1)} - X_M$。图的斜率为1。中位数$X_M$ 的位置在数据数目为奇数时为$(n-1)/2$，为偶数时为$n/2$。

峰度(kurtosis) 指标。在对称分布中，指示分布中间部分的频率对分布的形状有意义。有

$$\delta = \frac{E(X-\mu)^4}{\sigma^4} - 3$$

此值为正时分布为突起的(leptokurtic)，否则为扁平的(platykurtic)。样本测量：

$$d = \frac{n(n+1)\sum (x_i - \overline{x})^4}{(n-1)(n-2)(n-3)S^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

具渐近分布$N(0, 24/n)$。

去掉峰度的方法通常使用修正的指数转换，其形式类似于Box-Cox 转换：

$$y = SIGN(x - X_M)\frac{(|x - X_M| + 1)^\lambda - 1}{\lambda}$$

$SIGN(.)$ 是符号函数，据函数参量的负值、零和正值分别取值-1, 0 和1。X_M 可以取做均值或中位数。

异常值的检测可以利用$|(x-\overline{x})/s|$并且使用2.70 做界值，其上界为$(n-1)/\sqrt{n}$，在$n < 9$ 时无异常值，修正界值为4 则$n < 18$ 无异常值，在小样本下，利用$X < (Q_1 - 1.5Q)$ 或$X > (Q_3 + 1.5Q)$ 进行比较，$Q = Q_3 - Q_1 \equiv IQR$。由Dixon, W. J.(1950) 的方法是基于顺序统计量，单一的异常值用公式$r_{10} = (X_{(n)} - X_{(n-1)})/R$，R是全距。在n=30时，$p = 0.01, 0.05, 0.10$分别对应$0.341, 0.260, 0.215$，其界值表见Dixon, W.J.(1965). Ratios involving extreme values, Ann. Math. Stat. 22，67-78。

正态性检验，有几种方法，常用的是正态图示、回归方法如Shapiro-Wilk 统计量、Filliben 统计量和D'Agostino 统计量、矩法检查如使用偏度峰度的检验。标准的拟合优度检验是卡方检验、Kolmogorov-Smirnov(K-S) 检验等。K-S 检验使用经验分布函数$F_n(x, \theta) = [X_i \leq x$的数目]/n，$\theta$ 是未知参数。检验统计量

$$D = |\frac{i}{n} - F_{(i)}|, i = 1, 2, ..., n, Z_{(i)} = \frac{(X_{(i)} - \overline{X})}{S}$$

是标化顺序统计量，$F_{(i)} = \Phi(Z_{(i)})$ 的最大值。

图的表示可用Q-Q 图，使用数据分布分位点$x_{(i)}$ 做横坐标，它也是经验分布函数的分位点，$\Phi^{-1}((i-3/8)/(n+0.25))$ 为纵坐标，正态分布时应是一条直线。

Wilks-Shapiro 检验或W-检验. 是基于顺序统计量的方差最优估计量与通常的方差估计的比值，设n个观察排成$x_{(i)} > x_{(i-1)}, i = 2, ..., n$，计算$b = \Sigma_i[x_{(n-i+1)} - x_{(i)}]a_{in}, i = 1, ..., [n/2]$，计算$W_n = b^2/[(n-1)S^2]$，其中S 是$x_i$ 的样本方差，W_n 的百分位点可查表，其值应近于1，否则当W_n 很小时应予拒绝，系数a_{in} 由查表而来，此检验计算较K-S 检验复杂。在样本含量小于50时，W 统计量是Shapiro-Francia W'统计量的良好逼近。

SAS 在样本数小于2000 时计算W-统计量，若样本数大于2000 ，打印Kolmogorov D-统计量，较大值的概率由$(\sqrt{n} - 0.01 + (0.85/\sqrt{n}))D$ 给出，其常用的界值为0.775(0.15), 0.819(0.10), 0.895(0.05), 0.955(0.025) 和1.035(0.01)。

一种简便的情况是在大样本时，考虑偏度与峰度两个指标的正态性，构造检验：$W = n[g/6 + d^2/24]$，它服从自由度为2 的χ^2分布。利用样本均值与方差的独立性，采用刀切法的方法，估计时去掉一个观察，算出均值和方差，每个观察依次轮换，形成了两个新的变量，据数理统计，两个量的相关应为0。因为方差非正态分布，求相关前先对方差开立方根，末了算得的相关采用Fisher 的正态转换。

这里以SPSS/PC+为例，对例2.4的数据计算有关的统计量。其程序为：

```
set length 300.
data list free /x.
begin data.
 28          29        -44          30
 26          22         27          23
 33          24         16          29
 24          21         40          31
 34          25         -2          19
end data.
SET SCREEN OFF.
EXAMINE X /PLOT ALL /MESTIMATOR ALL
          /STATISTICS DESCRIPTIVES EXTREME.
```

其结果如下，按列读取为：均数(21.75)、中位数(25.5)、5%截尾均值(24.39)、标准误(3.94)、方差(310.72)、标准差(17.63)、最小值(-44)、最大值(40)、极差(84)、四分极差(8.5)、偏度(-3.05)及其标准误(.51)、峰度(10.81)及其标准误(.99)。

Mean 21.75	Std Err 3.94	Min -44.00	Skewness -3.05
Median 25.50	Variance 310.72	Max 40.00	S E Skew .51
5% Trim 24.39	Std Dev 17.63	Range 84.00	Kurtosis 10.81
		IQR 8.50	S E Kurt .99

几种M-估计量如下：

Huber (1.34)	25.62	Tukey (4.69)	26.34
Hampel (1.70, 3.40, 8.50)	26.43	Andrew (1.34 * pi)	26.34

设$X_1, ..., X_n$是独立同分布随机变量，通过使$\sum_{i=1}^n \rho(X_i; \theta)$极小而得到参数$\theta$的估计称作M-估计。它包含一大类估计量，如选择$\rho(x_i, \theta) = -\ln f(x_i, \theta)$, $f(x_i, \theta)$为概率密度函数，我们就得

到了极大似然估计。对样本均值有$\rho(x;\theta)=(x-\theta)^2$，由$\min\sum(x-\theta)^2$ 推出$\theta=\sum_i x_i/n$。$\rho(x;\theta)=|x-\theta|$ 的解是样本中位数。R- 估计和L- 估计分别是秩次统计量(rank statistics)和顺序统计量的线性组合或函数，截尾均值($\alpha-$trimmed mean) 就是一种L-估计量。SAS PROC MEANS 中的L_1 是最小绝对偏差(least absolute deviation, LAD)，L_p估计与此相仿。与标准差相应，尺度的估计常用绝对偏差中位数(MAD)来表示，MAD=median$\{|x_i-M|\}$, M=median$\{x_i\}$。这里将以上几种M-估计量解释如下：

Hampel 估计是一种"redescending" M-估计，用三个常数(a,b,c) 来表征。标化观测值绝对值大于c时赋权重为零，0—a 之间的值赋权为1，a—b 和b—c 之间的权随离零的距离而定, 大于c的观测权为0，此处a=1.7，b=3.4，c=8.5。Andrew 估计量也是一种redescending M-估计量。它对于各记录赋的权重没有急剧的变化，而是用一个平滑的正弦曲线来决定各记录的权，标化值绝对值大于c=1.34 π 的记录赋权重为0。Tukey 的biweight 估计量对于标化值大于c=4.685的观测为零，其它权重与离开中心点的距离成反比例。Huber 估计量对标化值小于c=1.339的记录赋权为1，具有较大绝对值记录随离开零的距离增大权重减少。Tukey 的hinges 是每一半数据中点上的值，用于计算盒式图中的四分极差。

本例正态性图示如图 2.1：

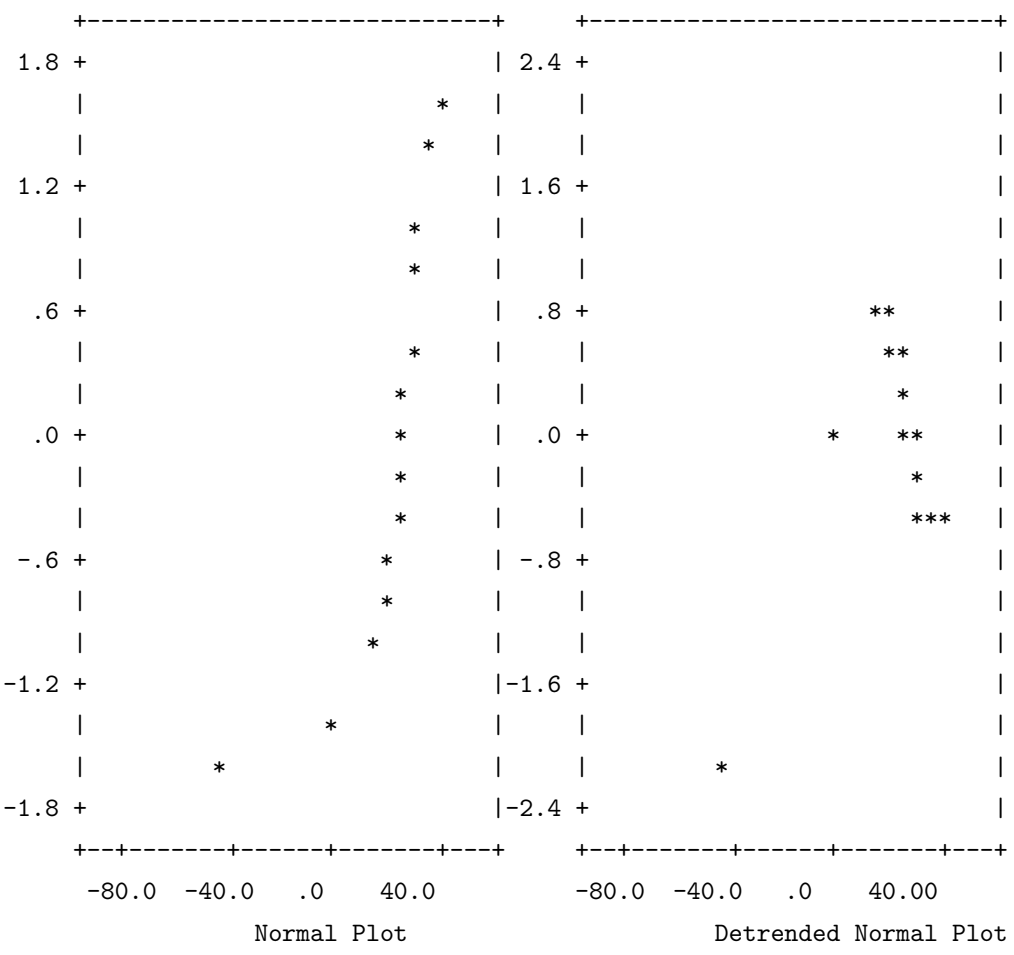


图 2.1 例2.4正态图示

正态性检验统计量：

	Statistic	df	Significance
Shapiro-Wilks	.6460	20	$< .0100$
K-S (Lilliefors)	.2380	20	.0042

箱尾图为图 2.2：

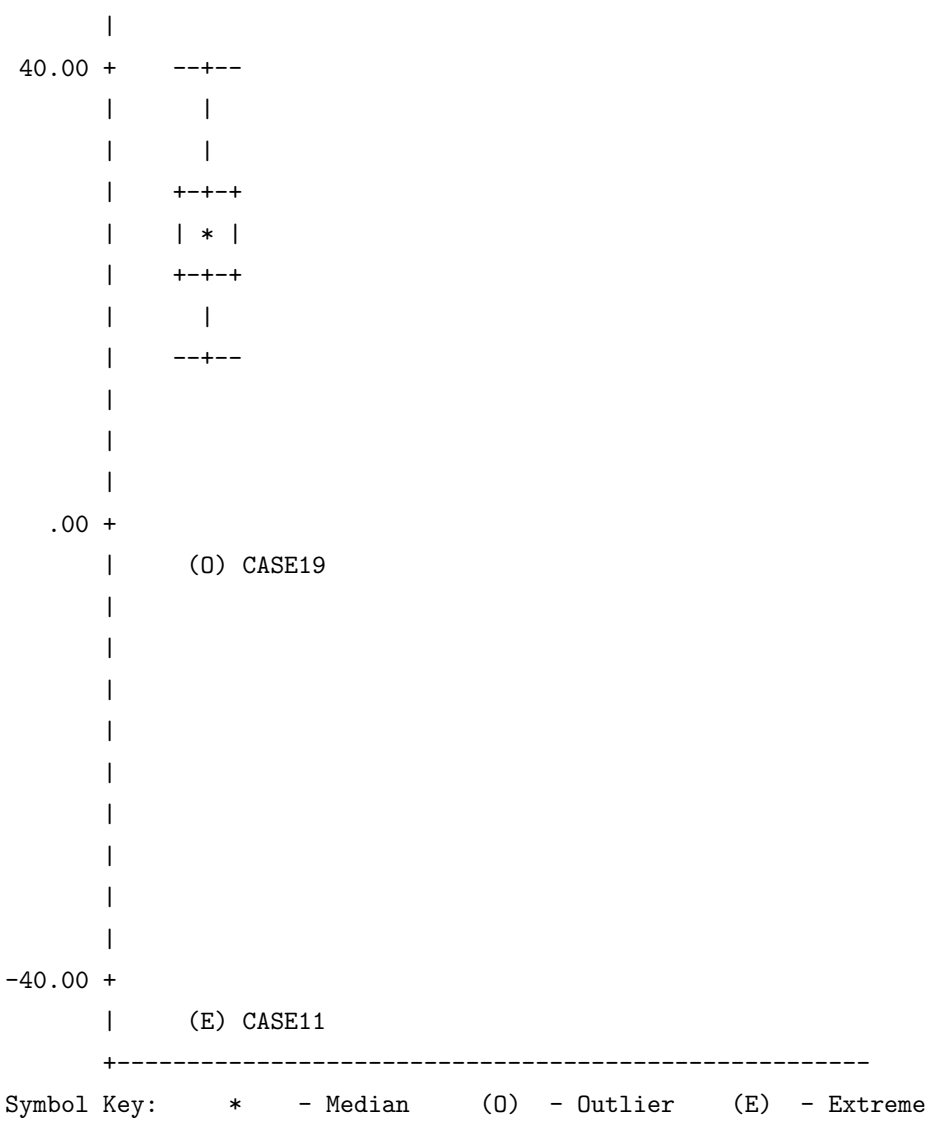


图 2.2 例2.4的箱尾图

正态概率图(P-P 图) 用于检查资料与正态的偏离情况，各点的累积比例与标准正态分布的累积比例绘图。若资料来自于正态分布，则点应近于直线。

去趋势正态图(detrended normal plot) 是观察值为期望值之差而做的图。若样本来自于正态分布，则所有点应聚集在零周围的水平带中，不应该有模式存在。

本例正态图示和检验统计量均提示该数据不符合正态性分布。

进行检验时，所采用的方法与均值与方差是已知还是未知以及观察的数目有关，可用

下面的思路[15]：首先，进行的检验是关于率、均值还是方差？

关于率的检验：是一个率还是两个率？

一个率的检验，$z=(p-\pi)/\sqrt{\pi(1-\pi)/n}$

两个率的检验，使用$\chi^2=\Sigma(o-e)^2/e$

关于方差的检验：是一个还是两个？

一个方差的检验，使用$\chi^2_{(n-1)}=(n-1)S^2/\sigma^2$

两个方差的检验，使用$F_{v_1,v_2}=S_1^2/S_2^2$

关于均值的检验：是一个或是两个？

一个均值检验：样本数≥30？

是，方差已知时使用$z=(\overline{X}-\mu)/[\sigma/\sqrt{n}]$

方差未知时使用样本标准差代替总体标准差用上式进行 z 检验。

否，变量是正态时使用$t_{n-1}=(\overline{X}-\mu)/[S/\sqrt{n}]$

变量非正态时使用非参检验

两个均值检验：样本数≥30？

是，方差已知时使用$z=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$

方差未知时用样本方差代替总体方差继续使用上式。

否，两个变量是正态的吗？

是，方差相等时使用，$t_{n_1+n_2-2}=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{S^2/n_1+S^2/n_2}}$

其中$S=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

方差不齐时使用t'检验。

$$t_f=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{S^2/n_1+S^2/n_2}$$

其自由度为(Satterthwaite's approximation)：

$$f=\frac{|S_1^2/n_1+S_2^2/n_2|^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}}$$

否，使用Mann-Whitney-wilcoxon 检验法。

以上方法，在各统计软件包中实现方式不一，如SAS PROC TTEST 提供两种t检验，方差不齐时采用t'检验。

多个样本均值的检验使用方差分析，方差分析中两两比较时有Fisher(LSD)、Duncan、Student-Newman-Keuls(SNK)、Tukey(Honestly Significant Different , HSD) 和Scheffë 法，这些检验的效率越来越低，而一类错误的机会也逐渐减小。Fisher 法逐个控制比较(comparisonwise)的一类误差；Tukey 的HSD 法控制整个比较过程(experimentwise)的 I 类误差，所以进行所有对子的比较，而出错机会是5%；Duncan 法处于两者之间，其界值取决于均值在排列中距离的远近，相邻均值处理同LSD，否则界值增加，但总小于Tukey法；Scheffë 检验最保守，但却可用于其它方式的比较。如：$H:\mu_1=(\mu_2+\mu_3)/2$ 可以等价地写成，$\mu_1-(\mu_2+\mu_3)/2=0$，检验误差为$1+(-1/2)^2+(-1/2)^2=3/2$，在SAS 中使用CONTRAST 语句进行此类检验，SAS 共有十七种比较的方法而在SPSS 中有七种。比较方法的选用取决于具体问题，如在条件较严格的物理实验中常用Fisher 法而在一过性事件和社会科学中Tukey 方法反而常用。

在SPSS/PC+中的Tukey-B 方法也是一种两两比较方法，把均值由小到大排列，然后对排列中每种比较求得一个距离，该法使用Tukey HSD 和SNK 的均值计算每步差的取值范围。

对分类数据分析，若仅对一个分类变量，则数据通常用频数表表示，它列出变量的取值和发生频数；若有两个或以上分类变量，一个对象的profile 定义为它在各分类上的取值，这样的数据可以把对象的profile 连同其频数一起列成频数表；若恰好有两个分类变量，则常用两维列联表(contingency table)的形式，其行列由每个分类变量不同取值构成。行列的交叉成为格点(cell) ，格点记录了相应profile的频数；对于多个分类变量，用多维列联表表示，在SAS FREQ和CATMOD 用不同的方法表示。

㈡两变量方法

1. 两变量的图示常用散点图、分组的箱尾图、茎叶图(back-to-back stem- and-leaf plot)等。

为了检验两个变量是否正态分布，可采用以下统计量：

$$\begin{pmatrix} X-\overline{X} \\ Y-\overline{Y} \end{pmatrix}' \begin{pmatrix} S_2 & S_{XY} \\ S_{XY} & S_Y^2 \end{pmatrix}^{-1} \begin{pmatrix} X-\overline{X} \\ Y-\overline{Y} \end{pmatrix}$$

应服从$\chi^2(2)$分布。

分类数据中，双向分类数据的χ^2检验假设：①有N次相同的实验；②每次试验有k种可能的结果；③K 个结果的概率保持不变；④实验是独立的；⑤K个格子的预计反应数应至少为5。有关列联表稳健性的讨论可见Hoaglin, D.C .(1983)，多维列联表使用对数线性模型来处理，见第13章。

一个行数为R 列数为C 的$R\times C$ 列联表独立性使用Pearson χ^2来检验，表示$R\times C$列联表行列因素相关的程度有许多统计量，其中之一是ϕ，其公式是：$\phi=\sqrt{\chi^2/N}$，其下界为0，当观测值与期望值相同时为0，而其上界是$\sqrt{\min(r-1,c-1)}$，两行或两列时，上界为1，故最常用。对于观察格子为A，B，C，D时，四格表$\phi=\frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$。

Cramer's $V=\frac{\phi}{\sqrt{\min(r-1,c-1)}}$，其范围是$0\sim 1$。

列联表系数$CC=\sqrt{\frac{\chi^2}{\chi^2+N}}$，$\chi^2$值最小时该量最大，上界随表的增大而趋于1。

现在看一个吸烟与与肺癌关系的的例子。x:1=不吸烟，0=吸烟；Y:1=死于其它疾病，0=死于肺癌。原始数据如下：

X：0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1

Y：1 1 0 0 0 0 0 0 0 1 1 0 1 1 1 1 0 1 1 1 0

可以算得其积矩(Pearson)相关系数r 和列联表系数 ϕ 为0.302。按列联表形式算得$\chi^2=1.818$。显然，χ^2是x与y两变量关系的显著性而列联表系数是其大小。$\phi=\sqrt{\chi^2/N}$，但只用于2x2表，对于更大的表使用V 统计量。SAS 程序如下：

```
data phi;
input x y @@;
cards;
0 1   0 0   1 0   1 0
0 1   0 0   1 1   1 1
0 0   0 0   1 1   1 1
0 0   0 1   1 1   1 1
0 0   0 1   1 1   1 0
proc print;
proc corr;
```

```
run;
proc freq;
   table x*y/chisq expected nopercent nocol norow;
run;
```

```
 X        Y
 Frequency|
 Expected |       0|       1|  Total
 ---------+--------+--------+
0 |      6 |      4 |     10
  |    4.5 |    5.5 |
 ---------+--------+--------+
1 |      3 |      7 |     10
  |    4.5 |    5.5 |
 ---------+--------+--------+
 Total      9      11       20

STATISTICS FOR TABLE OF X BY Y

Statistic   DF  Value       Prob
------------------------------------------------------
Chi-Square                          1     1.818       0.178
Likelihood Ratio Chi-Square         1     1.848       0.174
Continuity Adj. Chi-Square          1     0.808       0.369
Mantel-Haenszel Chi-Square          1     1.727       0.189
Fisher's Exact Test (Left)                              0.965
(Right)                0.185
(2-Tail)                              0.370
Phi Coefficient                    0.302
Contingency Coefficient            0.289
Cramer's V                         0.302

Sample Size = 20
WARNING:  50% of the cells have expected counts less
          than 5. Chi-Square may not be a valid test.
```

两率比较的功效：

设$p_A = p_B \approx 60\%$，期望$p_A - p_B = 6\%$，试问样本量不同取值下的检验功效？

$p_A - p_B$的方差为$\frac{\pi_A(1-\pi_A)}{(n/2)} + \frac{\pi_B(1-\pi_B)}{(n/2)} \approx 0.6 \times 0.4 \times (4/n) = 0.96/n$

$$Z = \frac{(p_A - p_B) - (\pi_A - \pi_B)}{\sqrt{0.96/n}} \sim N(0,1)$$

当$p_A - p_B = 0.06, H_0 : \pi_A - \pi_B = 0$　功效为：

$$p\left[\frac{|p_A - p_B| - 0.06}{\sqrt{0.96/n}} \geq z_{\alpha/2}\right]$$

$$\alpha = 0.05, Z_{0.025} = 1.96$$

有

$$p\left[\frac{|p_A - p_B| - 0.06}{\sqrt{0.96/n}} > 1.96 - 0.06\sqrt{n/0.96}\right] + p\left[\frac{|p_A - p_B| - 0.06}{\sqrt{0.96/n}} < -1.96 - 0.06\sqrt{n/0.96}\right]$$

$$\begin{aligned} &= P[Z > 1.96 - 0.06\sqrt{n/0.96}] + p[z < -1.96 - 0.06\sqrt{n/0.96}] \\ &= \Phi[1.96 - 0.06\sqrt{n/0.96}] + \Phi[-1.96 - 0.06\sqrt{n/0.96}] \end{aligned}$$

当$n = 50, \approx 0.07$; $n = 200, \approx 0.14$。

作为列联表探索性数据分析的用例，这里给出一个中位数平滑的例子(V. A. Sposito, On Median Polish and L1 Estimators, Comp. Stat. and Data Anal., V5. N3.，1989)。

列联表为$\begin{pmatrix} 1 & 8 & 3 \\ 5 & 9 & 2 \\ 6 & 4 & 7 \end{pmatrix}$　每行中位数$3, 5, 6$

以行开始进行第一个半步，每行减去行中位数，结果为：

```
 -2   5   0  | 3          -2    1     0  | -2
  0   4  -3  | 5           0    0    -3  |  0
  0  -2   1  | 6           0   -6     1  |  1
 ------------+---        ----------------+---
  0   4   0  |            0    4     0  |  5
```

5 是前一行效应估计的中位数，由于第二半步时每行、列的中位数为0，过程停止，得：行效应$\overline{\alpha} = (-2, 0, 1)$，列效应$\overline{\beta} = (0, 4, 0)$。对应于模型：

$$Y_{ij} = \mu_i + \beta_j + \varepsilon_{ij}\ \varepsilon_{ij} = \begin{pmatrix} -2 & 1 & 0 \\ 0 & 0 & -3 \\ 0 & -6 & 1 \end{pmatrix}$$

软件包Statgraphics 能够进行上述计算。

2. 两变量分析最常用的手段是变量的线性相关。设随机向量(ξ, η) 服从二维正态分布，参数为$(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，其中$\rho$ 是ξ 和η 的相关系数，现在$(X_1, X_1), ..., (X_n, Y_n)$是来自$(\xi, \eta)$ 的简单随机样本，相关系数的公式是：

$$r = \frac{\sum_{i=1}^n (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum_{i=1}^n (X - \overline{X})^2 \sum_{i=1}^n (Y - \overline{Y})^2}}$$

据Schwartz 不等式，变量的相关系数应在$[-1, 1]$ 内。相关系数是否有显著性意义，可利用r-检验，查其界值表判断其显著性；或者使用t-检验，其公式是$t = r\sqrt{(n-2)/(1-r^2)}$；最后，还可以使用Fisher 的z-转换进行检验。根据相关为组内相关或组间相关的不同，对子数的校正方法也不相同。

在线性回归分析中，应该注意的是回归诊断技术。其主要内容为：

误差项是否满足独立性、等方差性、正态性。

选择线性模型是否合适，是否存在曲线等关系？

是否有异常样品存在，即异常点？

回归模型是否过多依赖于某些样品，即模型稳健性如何？

自变量之间是否高度相关，即多重共线性？

【例2.5】表 2.2是Anscomb构造的数据(Anscomb, F.J., 1973, Graphs in statistical analysis. Am. Statist.,27,17-21)，前四列可分为三组，第一列是共用的自变量x，最后两列是第四组的x,y。

表 2.2 Anscomb(1973)设计的四个数据例子

x1	y1	y2	y3	x4	y4
10.0	8.04	9.14	7.46	8.0	6.58
8.0	6.95	8.14	6.77	8.0	5.76
13.0	7.58	8.74	12.74	8.0	7.71
9.0	8.81	8.77	7.11	8.0	8.84
11.0	8.33	9.26	7.81	8.0	8.47
14.0	9.86	8.10	8.84	8.0	7.04
6.0	7.24	6.13	6.08	8.0	5.25
4.0	4.26	3.10	5.39	19.0	12.50
12.0	10.84	9.13	8.15	8.0	5.56
7.0	4.82	7.26	6.42	8.0	7.91
5.0	5.68	4.74	5.73	8.0	6.89

相应的SAS回归程序为：

```
data anscomb;
input x1 y1 y2 y3 x4 y4;
cards;
......
proc reg data=anscomb;
L1:model y1-y3=x1;
L4:model y4=x4;
run;
```

结果算得四组回归结果是相同的, y=3+0.5x, $R^2 = 0.667$，可见通常的检验方法对四组数据的区分是无能为力的，而由残差图可清楚地看出残差所隐含的模式。

从SAS、SPSS/PC+、Stata 等可以获得回归诊断统计量。如曲线关系利用标化残差与x的点图来发现；方差不一致利用标化残差对x 或y 预测值点图的方法；残差间的相关，使用Durbin-Watson 统计量。这些诊断统计量在线性回归有最简单的形式，现将模型$y = \beta_0 + \beta_1 x + \varepsilon$有关的诊断量列如下，

1. 杠杆(leverage)：

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{\sum (x_j - \overline{x})^2}$$

可见x离均值较远时h_{ij}就大，提示一个距离的解释。

$$\hat{y}_i = b_0 + b_1 x_i = \sum h_{ij} y_j$$

它反映了对y估计值的影响，$\frac{(n-2)(h_{ii}-1/n)}{1-h_{ii}}$ 服从$t(n-2)$ 分布，因而实算值大于界值时可认为是高杠杆点。

2. 残差(residual)：普通残差为$e_i = y_i - \hat{y}_i$，删除残差为

$$e_{(-i)i} = y_i - \hat{y}_{(-i)i} = \frac{e_i}{1 - h_{ii}}$$

反映了预测值与实际值的差，故也称预测残差，下标$-i$ 表示去掉第i 个观察后的结果。

预测残差平方和$PRESS = \sum e^2_{(-i)i}$ 与$\sum e^2$ 的比反映了缺失观察的影响，可用于变量的筛选。

标准化残差和学生化残差(studentized residual)：由于$E(e_i) = 0, V(e_i) = \sigma^2(1 - h_{ii})$，直接比较残差是不合适的，用标准化残差$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$。记其最大的一个为$R_n = MAX|r_i|$，其上界值为：

$$\frac{(n-2)F_{1,(n-2)}(\alpha)}{n - 3 + F_{1,(n-2)}(\alpha)}$$

另一种学生化残差是$t_i = \frac{e_i}{s_{(-i)}\sqrt{1-h_{ii}}}$

$$s^2_{(-i)} = \frac{(n-2)s^2 - \frac{e_i^2}{1-h_{ii}}}{n-3}$$

因s^2并不与e_i独立，(x_i, y_i) 的统计量就可以写成下式：

$$t = \frac{e_i\sqrt{(n-3)}}{\sqrt{(n-2)s^2(1-h_{ii}) - e_i^2}}$$

应注意的是，小的值仍可以是高影响点，用相同的界值进行所有的比较时就有可能使Ⅰ类误差增大，考虑用Bonferroni 修正。要保证所有单侧检验水平是α，则界值是α，对于双侧检验则是$\alpha/2n$。

3. 库克距离(Cooks' D)：

$$D_i = \sum \frac{(\hat{y}_{(-i)j} - \hat{y}_j)^2}{2s^2}, i, j = 1, ..., n$$

或：$D_i = \frac{r_i^2 h_{ii}}{2(1-h_{ii})}, i = 1, ..., n$ 可见其受r_i与h_{ii} 的影响。

Belsey, Kuh 和Welsch (1980) 提出了DF 簇方法：

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(-i)i}}{s_{(-i)}/\sqrt{h_{ii}}} = \sqrt{\frac{h_{ii}}{(1-h_{ii}}}\frac{e_i}{s_{(-i)}\sqrt{1-h_{ii}}}$$

可见它与D_i是可比的，仅仅以$s_{(-i)}$代替了$\sqrt{2}s$。与此相仿，有$DFBETA_{1i} = \frac{b_1 - b_{(-i)1}}{C_1 s_{(-i)}}, i = 1, ..., n, C_1^2 = \frac{n}{n\sum x^2 - (\sum x)^2}$，是$b_1$ 的方差除以σ^2 。$DFBETA_{0i}$ 亦与此相仿。Welsch 建议用加权最小二乘求取回归系数，使用的权w_i当$DFFITS_i \leq 0.34$ 时为1，否则为$0.34/|DFFITS_i|$。

考虑观察值对于方差协方差阵的影响，原来估计值与估计值的比值COVRATIO 为：

$$\frac{n-1}{(n-2)+t_i}(1-h_{ii})$$

在$1/n \le h_{ii} \le 2/n$ 及$|t_i| \le 2$ 时在$[1-3/n, 1+3/n]$ 内。否则当$|COVRATIO-1| > 3/n$ 时，(x_i, y_i) 点应予以研究。另外，$h_{ii} \to 0$ 时，COVRATIO 比值近于1，较大较负的t_i 可导致更大的COVRATIO。

4. 自相关的检验(Durbin-Watson test)

检验统计量

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

在n比较大时近似有：

$$\sum_{i=1}^{n} e_i^2 \approx \sum_{i=2}^{n} e_{i-1}^2 \approx \sum_{i=2}^{n} e_i^2$$

故$d \approx 2-2r = 2(1-r)$, r 是一阶自回归模型AR(1) 的参数，AR(1) 的模型是：

$$e_i = \rho e_{i-1} + u_i, u_i \sim N(0, \sigma^2), V(e) = \frac{\sigma^2}{1-\rho^2}$$

$\rho_s = \rho^s$ 是自相关系数，d可能的取值为(0,4)，上述结论便于记忆相关的符号。

$H: \rho = 0 \leftrightarrow A: \rho > 0$, 在$d < d_L^\star$ 时拒绝，在$d > d_U^\star$ 时接受；

$H: \rho = 0 \leftrightarrow A: \rho < 0$, 在$d > 4 - d_L^\star$ 时拒绝，在$d < 4 - d_U^\star$时接受。

一般地，由于Box 和Jenkins 自相关和移动平均模型的形式是：

$$ARMA(p,q): e_t = \phi_1 e_{t-1} + ... + \phi_p e_{t-p} + v_t + \theta_1 v_{t-1} + ... + \theta_q v_{t-q}$$

它由两部分组成：

自相关AR(p)：$e_t = \phi_1 e_{t-1} + ... + \phi_p e_{t-p}$

移动平均MA(q)：$e_t = v_t + \theta_1 v_{t-1} + ... + \theta_q v_{t-q}$

它们属于时间序列分析的内容，SAS/ETS、SPSS/PC+ Trend、SYSTAT Series 及TSP 软件均可进行时间序列分析。

序列间的相关还可用游程检验(Wald-Wolfowitz) 来进行。检验是针对按一定顺序排列的二分类变量，连续出现一种结果(符号)为一个游程，全部游程数记为r。记出现正号的数目为$n_1, n_2, n = n_1 + n_2$，据界值表，r 过大或过少均表示符号的变化是不随机的。

$$\mu = \frac{2n_1 n_2}{n} + 1, \quad \sigma^2 = \frac{2(n_1 n_2)(2n_1 n_2 - n)}{n^2(n-1)}$$

$$z = \frac{|r-\mu| - 0.5}{\sigma}$$

游程检验可由SAS/QC 的SHEWHART 过程完成；在SYSTAT 中，命令RUNS 计算游程检验。

多数回归诊断技术提供的最好方法还是删除观察量，较好的方法是采用稳健回归的方法。这些方法的研究近年来很活跃。除此以外，非线性回归与相关、校准(calibration) 方法亦经常采用。

§2.2.2 非参统计方法

属于单样本方法有符号检验(sign test)、Wilcoxon 符号秩次检验、Kendall 检验、Spearman 检验、K-S 检验，两样本方法有Wilcoxon-Mann-Whitney 检验、Kolmogorov-Smirnov 检验，多样本检验用Kruskal-Wallis等方法[36]。

符号检验是最简单的一种，它只考虑符号，而Wilcoxon 符号秩次检验考虑了秩的绝对值大小Spearman检验与Kendall检验是关于相关系数的。K-S检验可用于分布的拟合，与χ^2检验的区别是它仅适于连续型分布，其精确的界值是已知的，对于任意样本大小均更有效。

Wald-Wolfowitz 游程检验也是一种非参检验，其思想是：两个样本来自于同一个总体，则将两样本的观测值混合按小到大排列，用一个游程表示同一组数据的数列，如果游程数很少，表明两样本来自不同的总体。

这些检验结果的判断一般据查表和正态分布、χ^2分布近似方法，软件包多基于近似方法，此处作主要介绍。

㈠ Mann-Whitney 秩和检验

有时称Mann-Whitney U 或Wilcoxon Rank Sum W Test，是一种非参检验，检验两个容量分别为m和n的独立样本是否来自同一总体，它使用了观察的秩次，故较中位数检验(median test)更为有效。

H:总体1与总体2的相对频数分布是相同的。

A:总体1的频数分布相对于总体2向左右移动。

检验统计量为U，在$U \leq U_c$时拒绝原假设。

做法是先把$m+n$ 个观察自小到大排序，每个样本的观察值序号之和称为秩和，记为T_1与T_2。编秩时相同的秩次取其平均值。小样本时两者较小一个$\geq T_U$ 或$\leq T_L$ 时拒绝，T_U 和T_L可由专门的表查出；大样本时，m 与n 的值均应至少为10，可用Z-检验。

【例2.6】测得铅作业与非铅作业两组工人的血铅值($\gamma/100g$) 如表2.3，判断两组工人血铅值之间是否存在差别？

秩和(期望值)为：T_2=59.5 (63.0), T_1=93.5 (90.0)。

H: 铅作业工人与非铅作业工人血铅的分布是相同的

A: 两组工人血铅分布是不同的。

计算统计量为：$U = nm + \frac{m(m+1)}{2} - T_1 = (10)(7) + \frac{7(8)}{2} - 93.5 = 4.5$

U 的期望值是mn/2，方差是nm(n+m+1)/12，所以采用正态近似，

$$z = \frac{U - nm/2}{\sqrt{nm(n+m+1)/12}} = \frac{4.5 - (10)(7)/2}{(10)(7)(10+7+1)/12} = -2.97$$

计算结果亦列于上表。Wilcoxon 秩和检验与Mann-Whitney 检验是等价的，有时又称Mann-Whitney-Wilcoxon 检验。结果可由SAS PROC NPAR1WAY 指定Wilcoxon 选项而得，检验又是Kruskal-Wallis 两组时的情况，χ^2=8.8813, 自由度1, P=0.0029。使用SPSS/PC+，结果类似，校正相同秩次后，Z值为-2.9801, P=0.0029，均表明两组工人血铅值存在显著差异。

㈡Wilcoxon 符号秩次检验: 配对检验。

即Wilcoxon matched-pairs signed-ranks test，是一种两样本非参方法，检验两个变量的分布是否相同。它不对分布的形状作任何假设，算出变量差值的绝对值，并由小到大排列，检验根据正差和负差之和进行。

H: 总体1与总体2的相对频数分布相同。

A: 总体1的频数分布相对于总体2向左右移动。

表 2.3 两组工人血铅值的秩和检验

工人号	组号	血铅	秩次
1	1	5	1.5
2	1	5	1.5
3	1	6	3
4	1	7	4
5	1	9	5
6	1	12	6
7	1	13	7
8	1	15	8
9	1	18	10.5
10	1	21	13
11	2	17	9
12	2	18	10.5
13	2	20	12
14	2	25	14
15	2	34	15
16	2	43	16
17	2	44	17

差值为零者忽略不计，使用差为负的秩和或差为正的秩和，在秩和少于界值时拒绝原假设。在组数足够大如$N \geq 25$时采用正态逼近的办法。

【例2.7 】用两种饲料喂8只大白鼠后，测定其肝中维生素A的含量(国际单位/mg)，正常组(normal)与维生素缺乏组(deff)结果见表2.4，问不同饲料的效果有无差别？

大样本的算式如下：

$$Z = \frac{T_{-} - [n(n+1)/4]}{\sqrt{[n(n+1)(2n+1)/24]}} = \frac{1-(8)(9)/4}{\sqrt{(8)(9)(17)/24}} = -4.20$$

在H 下，$Z > Z.05 = 1.96$，故应拒绝原假设。

应用SPSS/PC+中语句NPAR TESTS /WILCOXON normal deff，有Z = -2.3805，P =.0173。使用语句T-TEST /PAIRS normal deff. 进行配对t 检验，均值分别为3.3188和2.5063，差的均值为0.8123，标准误为0.193，t=4.21，P=0.004 。在SAS中可以使用PROC MEANS进行这个检验，语句为proc means t prt var std stderr; var d;run;。

㈢Kruskal-Wallis H 检验: 比较K 个总体相对频数分布。

H:K个总体的相对频数分布是相同的

A:至少有两个总体的相对频数分布是不同的

检验统计量：

$$H = \frac{12}{n(n+1)} \Sigma_i \frac{T_i^2}{n_i} - 3(n+1)$$

其中n_i= 第i个样本的观察数, T_i=第i个样本的秩和, n=Σn_i为总的样本数目

表 2.4 配对资料符号秩和检验两种鼠肝中维生素A含量

OBS	正常饲料	维生素E缺乏组	差值(d)	秩次
1	3.55	2.45	1.10	6
2	2.00	2.40	-0.40	-1
3	3.00	1.80	1.20	7
4	3.95	3.20	0.75	3
5	3.80	3.25	0.55	2
6	3.75	2.70	1.05	5
7	3.45	2.50	0.95	4
8	3.05	1.75	1.30	8

$T_+ = 35, T_- = 1$。由界值表，$\alpha < 0.05$，应拒绝原假设。

假设: 1. K 个样本独立并随机地由各自的总体抽出。2. 为使卡方逼近较为稳妥，每个样本应至少有 5 个或更多的观察。3. 秩次重复时，它们的秩次排列好象它们没有重复时的秩次求和取平均而得。

【例2.8 】研究社会经济状况与在校成绩的关系，将社会经济状况分为三等。看入校新生绩点成绩间的差别。原始数据如表2.5：

表 2.5 社会经济状况与在校成绩的关系

下　等	中　等	上　等
2.87 (10)	3.23 (16)	2.25 (5)
2.16 (3.5)	3.45 (18)	3.13 (14)
3.14 (15)	2.76 (8)	2.44 (6)
2.51 (7)	3.77 (20)	3.27 (17)
1.80 (2)	2.97 (11)	2.81 (9)
3.01 (12.5)	3.53 (19)	1.36 (1)
2.16 (3.5)	3.01 (12.5)	
$T_1 = 53.5$	$T_2 = 104.5$	$T_3 = 52$

$H = 6.13 > \chi^2_{.05;(3-1)} = \chi^2_{2;0.05} = 5.99, P < 0.05$ 拒绝原假设。

㈣ Friedman 检验: 随机区组设计

H: K个总体的相对频数分布是相同的

A: 至少有两个总体的相对频数分布是不同的

检验统计量：

$$F = \frac{12}{bk(k+1)} \Sigma_i \frac{T^2}{n_i} - 3b(k+1)$$

其中b=实验中使用的区组数, k=处理数, T_i=第i种处理的秩和。在$F > \chi^2_{\alpha;(k-1)}$界值时拒绝原假设。

假设：1. K个处理随机分配给每个区组内的K个实验单元。2. 为保证χ^2分布的适度，区组数或处理数均应超过 5 。3. 秩次重复时，它们的秩次排列好象它们没有重复时的秩次求

和取平均而得。

【例2.9 】一个食品公司进行了一个单盲试验，让 6 个受试者随机地品尝三种不同的咖啡 A 、B 、C ，并把三种咖啡的优劣排序，得表2.6结果：

表 2.6 6 个受试者评判三种咖啡的结果

受试者	A	B	C
1	1	3	2
2	2	3	1
3	1	2	3
4	1	3	2
5	2	3	1
6	1	3	2
	$T_a = 8$	$T_b = 17$	$T_c = 11$

$$F = \frac{12}{(6)(3)(3+1)}[(8)^2 + (17)^2 + (11)^2] - 3(6)(3+1) = 7.0 > 5.99, P < 0.05$$

利用SPSS 和Minitab 可以进行这个检验，SAS 的示范程序中含有这个检验的例子。

㈤ Spearman 和Kindall 秩和相关检验

Spearman 相关是利用变量对间秩次的相关来说明两个变量的关系，由于编秩时信息的损失，它仅是一个单调性检验，并不是一个真正的线性关联。计算时分别对两个变量进行排秩，把编成的秩次仿Pearson 积矩相关进行计算，就得到Spearman 相关。在相同秩次较多时，建议用秩次代入积矩相关的公式中进行计算。利用两各观察秩次的差(d) 时，可用下面的公式，在相同秩次较多时应予校正。

H: 样本X与样Y是独立的

A: 较大的X趋于同较大的Y配对

记X与Y秩次为R_i, T_i，其均值皆为$\frac{n+1}{2}$, 检验统计量为：

$$r_s = 1 - \frac{2(2+1)}{n-1} + \frac{2}{n(n^2-1)}\Sigma_i R_i T_i$$

也用下面的形式：检验统计量: $r_s = 1 - \frac{6\Sigma_i d_i^2}{n(n^2-1)}$，大样本时$r_s\sqrt{n-1}$ 服从标准正态分布

肯德尔τ系数$\tau = 2\frac{\sum_{i<j} i_{ij}}{n(n-1)}$

$i_{ij} = SIGN[(x_i - x_j)(y_i - y_j)], SIGN(.)$ 是符号函数。对于(x_i, y_i)和(x_j, y_j)，若当$x_i > x_j$时$y_i > y_j$，则是一致的(concordance)，否则是不一致的(discordance)。肯德尔相关系数反映了数据这种一致与不一致的情况，该统计量在观测重复数较大时除以下式来校正：

$$(\frac{n(n-1)}{2} - n_x)^{0.5}(\frac{n(n-1)}{2} - n_y)^{0.5}$$

其中n_x与n_y是x与y的重复数。

大样本时用$3\tau\sqrt{\frac{n(n-1)}{2(2n+5)}} \sim N(0,1)$

【例2.10 】肝癌病因研究中，某地调查十个乡的肝癌死亡率(1/20万)与某食物中黄曲霉毒素相对含量的关系，数据列于表2.7，试分析两者是否存在相关？

表 2.7 黄曲霉素相对含量与肝癌死亡率

乡编号	黄曲霉素相对含量		肝癌死亡率		
	X	d	Y	d	秩次差
1	3.7	4	46.6	7	-3
2	1.0	2	18.9	2	0
3	1.7	3	14.4	1	2
4	0.7	1	21.5	3	-2
5	4.0	5	27.3	4	1
6	5.1	6	64.6	9	-3
7	5.5	7	46.3	6	1
8	5.7	8	34.2	5	3
9	5.9	9	77.6	10	-1
10	10.0	10	55.1	8	2

Spearman 相关分析结果

检验统计量:

$$r_s = 1 - \frac{6\Sigma_i d_i^2}{n(n^2-1)} = 1 - \frac{6(42)}{10(10^2-1)} = 0.7545$$

P=0.0133, 相关有显著意义。

以上几个非参检验在软件包中实现很方便，例2.10的SAS程序为：

```
data list;
   input x y ;
cards;
 3.7   46.6
 1.0   18.9
 1.7   14.4
 0.7   21.5
 4.0   27.3
 5.1   64.6
 5.5   46.3
 5.7   34.2
 5.9   77.6
 10.0  55.1
proc corr pearson spearman kendall nosimple;
   var x y;
run;
```

结果如下：Pearson 相关系数0.69754, P=0.0249，Spearman 相关系数0.74545, P=0.0133, Kendall Tau b 0.51111, P=0.0397。

也可以利用PROC RANK进行变量x、y的排序，生成的排序变量直接用于Pearson 相关公式得出Spearman相关系数。

§2.3 多元分析

§2.3.1 均值的检验

1. 单样本检验

设$X_1, ..., X_n$ 是独立同分布、来自p维正态分布$N_p(\mu, \Sigma)$的样本，其中均值向量μ和协方差矩阵Σ未知，现要检验假设H: $\mu = \mu_0$，A:$\mu \neq \mu_0$，

检验统计量为$T^2 = n(\overline{X} - \mu_0)' S^{-1}(\overline{X} - \mu_0)$ ，其中样本均值向量$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, 样本协方差矩阵$S = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})'}{n-1}$。在假设H成立的条件下，$\frac{n-p}{p(n-1)} T^2 \sim F_{p,n-p}$。因为在显著水平为$\alpha$时，$\frac{n-p}{p(n-1)} T^2 \geq F_{p,n-p;\alpha}$则拒绝假设H，其中$F_{p,n-p;\alpha}$表示自由度为$p, n-p$的F分布的右侧分位点。

【例2.11】8 个人服用某种药物后，将其血糖和血压记录于表 2.8，研究是否由于药物作用造成的特性改变为某一定量零。

表 2.8 八个人对某种药物的反应结果[21]

编号	1	2	3	4	5	6	7	8
血糖	30	90	-10	10	30	60	0	40
收缩压	-8	7	-2	0	-2	0	-2	1
舒张压	-1	6	4	2	5	3	4	2

现在，样本均值向量$\overline{X} = \begin{pmatrix} 31.25 \\ -0.75 \\ 3.125 \end{pmatrix}$ 方差矩阵$S = \begin{pmatrix} 1069.64 & 82.5 & 16.964 \\ 82.5 & 17.357 & 6.393 \\ 16.964 & 6.393 & 4.694 \end{pmatrix}$

检验$H: \mu = 0$ 对$A: \mu \neq 0$

$$T^2 = n\overline{X}' S^{-1} \overline{X} = 79.064$$

$\frac{n-p}{p(n-1)} T^2 = [\frac{5}{3(7)} 79.064 = 18.825 > F_{3,5;0.05} = 5.41$

则拒绝H,即药物作用造成的改变量不为零。

总体均值μ可用$\overline{X}$来估计。相应于单变量的区间估计，均值向量μ的$100(1-\alpha)\%$可信区域为：

$$\{\mu : n(\overline{X} - \mu)' S^{-1}(\overline{X} - \mu) \leq \frac{(n-1)p}{n-p} F_{p,n-p;\alpha}\}$$

现仅就本例中的收缩压和舒张压改变来考虑，95%可信区域为：

$$0.116(\mu_1 + 0.75)^2 - 0.314(\mu_1 + 0.75)(\mu_2 - 3.125) + 0.427(\mu_2 - 3.125)^2 \leq 1.50$$

在多变量时，也可对μ 的所有线性函数的同时可信区间感兴趣，使用Bonferroni 法。本例为$(-5.18 \leq \mu_1 \leq 3.68)$ 和$(0.825 \leq \mu_2 \leq 5.425)$

上述过程在BMDP 3D 程序为：

```
/PROBLEM   TITLE IS 'BLOOD DATA'.
           VARIABLES ARE 3.
/VARIABLES NAMES ARE SUGAR,SYST,DIAS.
/TEST      VARIABLES ARE SUGAR,SYST,DIAS.
           HOTELLING.
```

```
  /PRINT     DATA.
             COVARIANCE.
             CORRELATION.
  /END
```

上述问题在SAS GLM 中处理，程序为：

```
* Hotelling T square;
data;
input sugar syst dias @@;
a=1;
cards;
30  -8  -1   30  -2   5
90   7   6   60   0   3
-10 -2   4   0   -2   4
10   0   2   40   1   2
proc corr cov;
   var sugar syst dias;
run;
proc glm;
   class a;
   model sugar syst dias=a/noint;
   manova h=a;
quit;
```

PROC CORR 印出变量间的协方差阵(选项COV)，MANOVA 语句据因素A对变量行检验。

其均值与协方差矩阵分别为：

$$\overline{X} = \begin{pmatrix} 31.250 \\ -0.750 \\ 3.125 \end{pmatrix}$$

及

$$S = \begin{pmatrix} 1069.642857 & 82.500000 & 16.964286 \\ 82.500000 & 17.357143 & 6.392857 \\ 16.964286 & 6.392857 & 4.696429 \end{pmatrix}$$

过程默认打印单变量的检验，变量SUGAR：F 值为7.30，P 值0.0305；变量SYST：F=0.26, P=0.6263；变量DIAS：F=16.63，P=0.0047；自由度均为1,7。

多元方差分析(MANOVA)的结果：

```
    Characteristic Roots and Vectors of: E Inverse * H, where
     H = Type III SS&CP Matrix for A   E = Error SS&CP Matrix
Characteristic Percent      Characteristic Vector  V'EV=1
   Root
                              SUGAR           SYST          DIAS
```

```
11.294801079    100.00    0.01087150    -0.14412966    0.23692200
0.0000000000      0.00   -0.01076657     0.02062595    0.11261594
0.0000000000      0.00    0.00193692     0.08070481    0.00000000
```

```
            Manova Test Criteria and Exact F Statistics for
                 the Hypothesis of no Overall A Effect
       H = Type III SS&CP Matrix for A   E = Error SS&CP Matrix
                          S=1    M=0.5    N=1.5
Statistic                   Value        F   Num DF Den DF  Pr > F
Wilks' Lambda           0.08133519  18.8247    3        5  0.0037
Pillai's Trace          0.91866481  18.8247    3        5  0.0037
Hotelling-Lawley Trace 11.29480108  18.8247    3        5  0.0037
Roy's Greatest Root    11.29480108  18.8247    3        5  0.0037
```

后面两部分结果指明是矩阵$E^{-1}H$ 的特征值和特征向量，Wilks 的似然比为0.081335，对应的F 值为18.8247，对比F(3,5) 界值，概率为0.0037 ，表明药物作用的存在。S, M, N 的意义在SAS 说明书中有说细的说明，如：SAS/STAT User's Guide, Release 6.03 Edtion, pp 16-17。Wilks' Lambda 是一种多元显著性检验。取值范围为0—1，其较大的值提示均值不存在差异，当所有均值相同时取值为1。有时也称U统计量。Hotelling 迹是根据特征值之和进行的多元显著性检验。

相应的SPSS/PC+程序如下：

```
SET MORE OFF.
data list free /sugar syst dias.
begin data.
30  -8  -1   30  -2   5
90   7   6   60   0   3
-10 -2   4   0   -2   4
10   0   2   40   1   2
end data.
manova sugar syst dias.
```

系统自动按析因分析处理，结果与SAS相同。单变量F检验自由度为1,7，F值和P值与SAS相同。

2. 两样本配对T^2检验

记$d_i, i = 1, .., .n$是两配对样本的差，设$d_i \sim N_p(\delta, \Sigma)$

检验$H : \delta = 0$, 对$A : \delta \neq 0$

若$\frac{n(n-p)}{p(n-1)}\overline{d}' S_d \overline{d} \geq F_{p,n-p,\alpha}$ 则拒绝H。

其中$\overline{d} = \frac{\sum d_i}{n}$, $(n-1)S_d = \sum_{i=1}^{n}(d_i - \overline{d})(d_i - \overline{d})'$

以下数据是Maindonald,JH[16]的一个例子。

```
 -0.2  1.6  1.3
  8.2 11.1  1.1
```

```
-1.9 -2.2  0.9
 4.4  6.2  2.5
 1.5  4.6  2.0
 2.1  2.7  0.3
 1.7  1.6  1.8
-1.5 -0.2  3.0
 2.3  6.9  3.4
```

使用SAS分析语句与前面例子相同。

```
data;
input d1-d3;cards;
.... 数据行....
proc glm;
class a;
model d1 d2 d3=a/noint;
manova h=a;
run;
```

结果Hotelling-Lawley Trace=4.27933866，F=8.5587，自由度为3，6，P= 0.0138，可以认为各差值在0.05水平上有差异。类似地，SPSS/PC+语句是：

```
data list free /d1 d2 d3.
begin data.
.... 数据行....
end data.
manova d1 to d3/print cellinfo(means).
```

结果如下：

```
EFFECT .. CONSTANT
Multivariate Tests of Significance (S = 1, M = 1/2, N = 2 )
Test Name         Value  Approx. F Hypoth. DF   Error DF  Sig. of F
Pillais          .81058    8.55868       3.00       6.00       .014
Hotellings      4.27934    8.55868       3.00       6.00       .014
Wilks            .18942    8.55868       3.00       6.00       .014
Roys             .81058
```

3. 两样本均值检验

设$X \sim N_p(\mu_1, \Sigma)$, $Y \sim N_p(\mu_2, \Sigma)$，$X$ 与Y 的独立样本分别为$X_1, ..., X_{n_1}$及$Y_1, ..., Y_{n_2}$。

检验$H : \mu_1 = \mu_2$ 对$A : \mu_1 \neq \mu_2$

现有μ_1, μ_2 及Σ 的估计量

$$\overline{X} = \frac{\sum_{i=1}^{n_1} X_i}{n_1}, \quad S_1 = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})(X_i - \overline{X})'}{n_1 - 1}$$

$$\overline{Y} = \frac{\sum_{i=1}^{n_2} Y_i}{n_2}, \quad S_2 = \frac{\sum_{i=1}^{n_2}(Y_i - \overline{Y})(Y_i - \overline{Y})'}{n_2 - 1}$$

$$S_p = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$$

$$T^2 = \frac{(\overline{X}-\overline{Y})' S_p^{-1} (\overline{X}-\overline{Y})}{(1/n_1 + 1/n_2)}$$

若由样本观察值计算得到$\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}T^2 \geq F_{p,n_1+n_2-p-1;\alpha}$ 则以显著水平α拒绝H。

对任一$a \neq 0, a'(\mu_1-\mu_2)$ 的$(1-\alpha)100\%$ 可信区间为：

$$a'(\overline{X}-\overline{Y}) - [T_\alpha^2(\frac{1}{n_1}+\frac{1}{n_2})a'S_p a]^{0.5} \leq a'(\mu_1-\mu_2) \leq a'(\overline{X}-\overline{Y}) + [T_\alpha^2(\frac{1}{n_1}+\frac{1}{n_2})a'S_p a]^{0.5}$$

其中

$$T_\alpha^2 = \frac{p(n_1+n_2-2)}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1;\alpha}$$

【例2.12】8 只狗分至两组接受不同的实验处理，第一组是控制组，第二组每只狗腿上置一金属盘，然后测量狗腿的张力与压力，其数据见表 2.9。

表 2.9 两组狗对某种处理的实验结果[21]

	控制组(X)				实验组(Y)			
编号	1	2	3	4	1	2	3	4
张力	131.5	145	191	150	40.5	80	50	90
压力	9	12	30	36	54	74.5	64.5	60.5

控制组与实验组的均值分别是$\overline{X} = \begin{pmatrix} 141.875 \\ 21.75 \end{pmatrix}$和$\overline{Y} = \begin{pmatrix} 65.125 \\ 63.375 \end{pmatrix}$

合差协方差阵的估计值为$S = \begin{pmatrix} 309.90 & 86.36 \\ 86.36 & 124.99 \end{pmatrix}$

现在，

$$T^2 = \frac{(\overline{X}-\overline{Y})' S^{-1} (\overline{X}-\overline{Y})}{\frac{1}{n_1}+\frac{1}{n_2}} = 116.7$$

$$T_{0.05}^2 = \frac{p(n_1+n_2-2)}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1;0.05} = 13.9$$

拒绝两实验结果相同的假设。

可信区间为$(\overline{x}_1-\overline{y}_1) \pm (13.9 \times 0.5 \times 309.9)^{0.5} = 76.75 \pm 46.41$ 和$(\overline{x}_2-\overline{y}_2) \pm (13.9 \times 0.5 \times 124.99)^{0.5} = -41.625 \pm 29.47$

使用Bonferroni 可信区间为

$$(\overline{x}_1-\overline{y}_1) \pm 3.03\sqrt{(0.5 \times 309.9)} = 76.75 \pm 37.7 \quad (\overline{x}_2-\overline{y}_2) \pm 3.03\sqrt{(0.5 \times 124.99)} = -41.625 \pm 24.0$$

其中$t_{0.0125,7} = 3.03$

BMDP 程序为：

```
/PROBLEM   TITLE IS 'DOG DATA'.
           VARIABLES ARE 3.
/VARIABLES NAMES ARE STRIDE, STRAIN, TREAT.
           GROUPS IS TREAT.
```

```
  /GROUP     CODE(3) ARE 1,2.
             NAMES(3) ARE CONTROL, TREAT.
  /TEST      VARIABLES ARE STRIDE,STRAIN.
             GROUPS ARE 1,2.
             HOTELLING.
  /PRINT     DATA.
             COVARIANCE.
             CORRELATION.
  /END
```

SAS GLM 相应的程序如下：

```
* tests difference between two populations;
data dogs;
input stride strain treat @@;
cards;
131.5  9.01 1  40.5  54.02 2
145.0 12.01 1  80.0  74.52 2
141.0 30.01 1  50.0  64.52 2
150.0 36.01 1  90.0  60.52 2
proc glm;
   class treat;
   model stride strain=treat;
   manova h=treat;
quit;
```

一元方差分析结果：STRIDE：均值103.5，F= 38.2, P=0.0008；STRAIN：均值42.57，F=27.74，P=0.0019，F的自由度均是1,6。

多元方差分析结果：

```
  Characteristic   Percent       Characteristic Vector  V'EV=1
       Root
                                          STRIDE         STRAIN
    19.455518487    100.00           0.02253459    -0.03337111
    0.0000000000      0.00           0.01258064     0.02319117
         Manova Test Criteria and Exact F Statistics for
            the Hypothesis of no Overall TREAT Effect
 H = Type III SS&CP Matrix for TREAT   E = Error SS&CP Matrix
                        S=1    M=0    N=1.5
Statistic                  Value        F   Num DF Den DF  Pr > F
Wilks' Lambda           0.04888656  48.6388      2      5  0.0005
Pillai's Trace          0.95111344  48.6388      2      5  0.0005
Hotelling-Lawley Trace 19.45551849  48.6388      2      5  0.0005
Roy's Greatest Root    19.45551849  48.6388      2      5  0.0005
```

拒绝处理结果相同的假设。

相应的SPSS/PC+程序如下：

```
data list free/ stride strain treat.
begin data.
131.5  9.01 1 40.5  54.02 2
145.0 12.01 1 80.0  74.52 2
141.0 30.01 1 50.0  64.52 2
150.0 36.01 1 90.0  60.52 2
end data.
manova stride strain by treat(1,2)
   /DESIGN treat /print homogeneity(all) ERROR(COVARIANCES SSCP).
```

Cochran和Bartlett-Box检验均显示齐性。多元Box M 检验、F检验和χ^2 检验的结果如下：

Boxs M =	5.06310	
F WITH (3,6479) DF =	1.07931,	P = .357 (近似)
Chi-Square with 3 DF =	3.23476,	P = .357 (近似)

其余结果同上。

在SAS中，PROC DISCRIM 可用于方差阵齐性检验。对于协方差阵不等的两正态总体均值检验，是多元的Behrens-Fisher 问题，可采用Scheffë 或Yao 方法处理，见文献[21]。

4. 推广的t检验

单变量t检验的一个推广是p种处理与其中单一响应变量行比较，每个对象或实验单元在连续的时间接受这p个处理，第j个观察是$X_j = (x_{1j}, ..., x_{pj})', j = 1, ..., n, x_{ij}$是第$i$种处理对第$j$个对象的反应。为了比较，考虑：

$$\begin{pmatrix} \mu_1 - \mu_2 \\ ... \\ \mu_1 - \mu_p \end{pmatrix} = \begin{pmatrix} 1 & -1 & ... & 0 \\ ... & ... & ... & ... \\ 1 & ... & ... & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ ... \\ \mu_p \end{pmatrix} = C_1\mu$$

或

$$\begin{pmatrix} \mu_2 - \mu_1 \\ ... \\ \mu_p - \mu_{p-1} \end{pmatrix} = \begin{pmatrix} -1 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & -1 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ ... \\ \mu_p \end{pmatrix} = C_2\mu$$

C_1, C_2 称为对比矩阵，有p-1个行线性无关，每个都是一个对比向量，对比向量各元素之和为0。处理的均值相同时$C_1\mu = C_2\mu = 0$。事实上在处理间无差别时的假设就成了C μ =0，C是对比矩阵的任何一种择取。这时有均值$C\overline{X}$，方差CSC'，T^2统计量就是：

$$T^2 = n(C\overline{X})'(CSC')^{-1}(C\overline{X}) \sim \frac{(n-1)(p-1)}{n-p+1} F_{p-1,n-p+1;\alpha}$$

【例2.13】下面是J. Atlee 关于19条狗的试验数据。19 只狗开始给予苯巴比妥，然后每条狗在两种不同压力的CO_2上加上卤烷(Halothane)，最终测到狗的心跳间隔(毫秒)。用C表示CO_2，C+和C-表示其高低两个水平；用H表示卤烷，用H+和H-表示其高低两水平，原始数据列于表 2.10：

表 2.10 J. Atlee 19 条狗的实验数据

狗号	C+H-	C-H-	C+H+	C-H+	狗号	C+H-	C-H-	C+H+	C-H+
1	426	609	556	600	11	349	382	473	497
2	253	236	392	395	12	429	410	488	547
3	359	433	349	357	13	348	377	447	514
4	432	431	522	600	14	412	473	472	446
5	405	426	513	513	15	347	326	455	468
6	324	438	507	539	16	434	458	637	524
7	310	312	410	456	17	364	367	432	469
8	326	326	350	504	18	420	395	508	531
9	375	447	547	548	19	397	556	645	625
10	286	286	403	422					

其均值与协方差矩阵分别为：

$$\overline{X}=\begin{pmatrix}368.21053\\404.63158\\479.26316\\502.89474\end{pmatrix}\quad S=\begin{pmatrix}2819.29 & 3568.42 & 2943.50 & 2295.36\\3568.42 & 7963.13 & 5303.99 & 4065.46\\2943.50 & 5303.99 & 6851.32 & 4499.64\\2295.36 & 4065.46 & 4499.64 & 4878.99\end{pmatrix}$$

记其理论均值为μ_1,μ_2,μ_3,μ_4。考虑以下几种效应：

卤烷H $(\mu_3+\mu_4)-(\mu_1+\mu_2)$

CO_2 $(\mu_1+\mu_3)-(\mu_2+\mu_4)$

H-C 的交互 $(\mu_1+\mu_4)-(\mu_2+\mu_3)$

即

$$C=\begin{pmatrix}-1 & -1 & 1 & 1\\1 & -1 & 1 & -1\\1 & -1 & -1 & 1\end{pmatrix}=\begin{pmatrix}C_1'\\C_2'\\C_3'\end{pmatrix}$$

$T^2=n(C\overline{X})'(CSC')^{-1}(C\overline{X})=116$

在$\alpha=0.05$时，$\frac{(n-1)(p-1)}{n-p+1}F_{p-1,n-p+1;\alpha}=[18(3)/16]F_{3,16;0.05}$ =10.94。因此拒绝处理相同的假设，卤烷效应的95%可信区为：

$$(\overline{X}_3+\overline{X}_4)-(\overline{X}_1+\overline{X}_2)\pm\sqrt{10.94}\sqrt{C_1'SC_1/19}=209.31\pm73.70$$

其它两个可信区间分别是-60.05 ± 54.70 和-12.79 ± 65.97。

5.重复数据的检验

当对同一实验单元进行多次测量，测量的结果彼此相关。若这些测量性质不同，如重量、长度、宽度，则用多元方差分析等方法；若测量是在实验因素如时间、药物不同剂量等不同水平上进行，则由重复测量(repeated measures)方差分析处理。

随机顺序下可假设其协方差阵是$\Sigma=\sigma^2[I(1-\rho)+\rho ee']$, $e=(1,...,1)'$。即组内相关(intraclass correlation)模型。

重复测量分析有许多特点：因为误差中除去了个体间差异的影响，所以效率要高，精度提高，实验所需对象数目减少。分析技术可以是一元或者多元，文献建议观察数少于处理数＋10时使用一元分析。它假设数据的分布符合多元正态、各观察独立并且是球形(sphericity，这在多元分析方法中不需要)。球形要求所有重复测量对的方差相同，尽管在SPSS/PC+中有这样的检验，却不推荐。若不满足球形，则Ⅰ类误差增大。

Greenhouse 和Geisser(1959) 利用 ε 对重复测量结果进行调整。当球形假设成立时，$\varepsilon = 1$，最差时为1/(k-1)，k为处理数。SAS 输出它的Huynh 和Feldt(1976) 修正，当重复测量设计中偏离球形假设时，分子和分母自由度均乘以该值，使用调整后的自由度计算观察显著性水平。Huynh 和Feldt 证明Greenhouse-Geisser ε比较保守，特别对于小样本更是如此。常用的轮廓分析或形象分析(profile analysis)，特色是构造合适的对比矩阵，SYSTAT 的MGLH 模块和SAS、SPSS/PC+都能进行，这要求数据被很好地标化，以两组做为例：

$$\mu_1 = (\mu_{11}, \mu_{12}, ..., \mu_{1p})', \mu_2 = [\mu_{21}, \mu_{22}, ..., \mu_{2p}]'$$

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2}$$

要检验$H : \mu_1 = \mu_2$即两总体具有相同的均值，有三个检验：H_1. 轮廓是相似的吗?H_2.若轮廓相似，是重合的吗? H_3.若轮廓重合，轮廓的各水平是相同的吗?

$H_1 : \mu_{1i} - \mu_{1,i-1} = \mu_{2i} - \mu_{2,i-1}, i = 2, ..., p$ 或$C(\mu_1 - \mu_2) = 0$ 相对于$A_1 : C(\mu_1 - \mu_2) \neq 0$

$$C_{(p-1)\times p} = \begin{pmatrix} -1 & 1 & 0 & ... & 0 & 0 \\ 0 & -1 & 1 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & -1 & 1 \end{pmatrix}$$

从而$Ce = 0, e = (1, ..., 1)'$，以上检验即$H_1 : C(\mu_1 - \mu_2) = \gamma e$相对于$A_1 : C(\mu_1 - C\mu_2) \neq \gamma e$，$\gamma$为轮廓间的平均差异。检验统计量为：

$$\frac{n_1 + n_2 - p}{(n_1 + n_2 - 2)(p-1)}(\overline{X}_1 - \overline{X}_2)'C'[(\frac{1}{n_1} + \frac{1}{n_2})CSC']^{-1}C(\overline{X}_1 - \overline{X}_2)$$

与$F_{p-1,n_1+n_2-p;\alpha}$ 相比较

H_2与A_2即$H_2 : \gamma = 0$相对于$A_2 : \gamma \neq 0$

$$T^2 = (\overline{X}_1 - \overline{X}_2)'C'[(\frac{1}{n_1} + \frac{1}{n_2})CSC']^{-1}C(\overline{X}_1 - \overline{X}_2)$$

检验统计量$(\frac{1}{n_1}+\frac{1}{n_2})^{-1}(e'S^{-1}(\overline{X}-\overline{X}_2))^2(e'S^{-1}e)^{-1}(1+\frac{T^2}{n_1+n_2-2})^{-1}$ 与$F_{1,n_1+n_2-p-1;\alpha}$ 相比较，γ的极大似然估计为: $\hat{\gamma} = \frac{e'S^{-1}(\overline{X}_1-\overline{X}_2)}{e'S^{-1}e}$

H_3与A_3即$H_3 : \mu_1 = \delta e, \mu_2 = \xi e$，$\delta, \xi$未知，或$H_3 : C(\mu_1+\mu_2) = 0$ 相对于$A_3 : \mu_1 \neq \delta e, \mu_2 \neq \xi e$，有

$$\frac{(n_1 + n_2 - p)(n_1 + n_2)}{(n_1 + n_2 - 2)(p-1)}\overline{X}'C'[CSC']^{-1}C\overline{X}$$

与$F_{p-1,n_1+n_2-p;\alpha}$比较，$\overline{X} = \frac{n_1\overline{X}_1+n_2\overline{X}_2}{n_1+n_2}$。

第 4 章给出了一个利用PROC GLM 进行三组轮廓分析的例子，第 5 章也给出了相应的SPSS/PC+程序。有关协方差阵的其它检验可见如[21]。

§2.3.2 回归分析

1. 回归分析是研究应用最为广泛的多元分析技术。

因变量Y 和p 个自变量$X_1, ..., X_p$ 线性回归模型是

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon, \varepsilon \sim N(0, \sigma^2 I)$$

n 次观测$(y_i, x_{i1}, ..., x_{ip}), i = 1, ..., n$，具有如下线性关系：

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, i = 1, .., n$$

这里假定$\varepsilon_i \sim N(0, \sigma^2)$, 其回归系数的最小二乘估计为：$\hat{\beta} = (X'X)^{-1}X'Y \equiv HY$，$X = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix}$，$Y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}$。

2. 有关回归方程的统计量

⑴R^2和校正后的R^2

R^2是度量一个线性模型拟合优度的常用统计量，它不仅是自变量X和因变量Y复相关系数的平方，还是因变量Y与其预测值相关系数的平方。由于样本R^2 对模型拟合好坏趋于做出一个乐观估计，使用校正R^2 以更好地刻划模型拟合情况。校正后的R^2为：

$R_a^2 = R^2 - \frac{(1-R^2)p}{n-p-1}$，其中$p$ 为自变量个数，n 为观测个数。

⑵方差分析表

其F检验用于检验线性回归方程的零假设：$H : \beta_1 = ... = \beta_p = 0$,即因变量与所有自变量无线性关系。F统计量可以表为：F=回归均方/残差均方$\sim F_{p,n-p-1}$

因此，可对R^2 给予另一种解释：R^2是能被模型所解释的因变量的那一部分比例，即：R^2=1-残差平方和/总平方和，R_a^2=1-[残差平方和/(n-p-1)]/[总平方和/(n-1)]

⑶R^2 改变量

R^2 改变量是评价自变量相对重要性的一个常用指标，即考察当一个变量进入方程时R^2 的增量$R^2 - R(i)^2$，其中$R(i)^2$ 是当除第i个自变量外其它自变量均包括在方程中时的复相关系数的平方。显然R^2 的改变量大则意味着第i个变量提供其它已在方程中的变量所不能提供的信息量最大。

⑷条件数(condition number)

方阵$X'X$ 的条件数定义为$k = \lambda_1/\lambda_p$，其中λ_1 和λ_P 分别是$X'X$ 的最大和最小特征根，常用来刻划多重共线性是否存在以及严重程度。经验上。若k 取值在范围$100 \sim 1000$，则认为存在中等或较强的多重共线性；若$K > 1000$，则认为存在严重的多重共线性。

3. 关于自变量的统计量

⑴t统计量

对零假设$H : \beta_i = 0$，可用具有自由度为$1, (n - p - 1)$ 的F 统计量进行检验，但由于具有k个自由度的t值平方等于具有$1, k$ 自由度的F 值，故可用t统计量检验上述假设。

⑵标准化回归系数

一般来说，回归方程中各自变量测量单位不同, 因此不能将其系数大小视为变量重要性标志。标准化回归系数使系数在一定程度上具有可比性, 它是当所有变量用其标准化形式时自变量的系数，可以直接从回归系数计算。

$\tilde{\beta}_i = \beta_i S_i / S_y, S_i$ 是第i 个自变量的标准差，S_y 是因变量Y 的标准差。

⑶方差估计

仅仅按照每个自变量系数的t值的显著性判断对预测的重要性是危险的，因为那些具有较大方差的系数估计是不可靠的。所以，我们常常关心回归系数方差的估计量。

⑷变量容许性

记R_i^2 为当第i 个自变量被视为因变量且它与其它自变量计算产生回归方程时的复相关系数，R_i^2 较大同时表明第i 个自变量几乎是其它自变量的线性组合。$1 - R_i^2$ 反映了其它自变量未解释的变异比例，称为第i 个自变量的容许性。容许性是描述自变量之间相互依赖的常用指标。一个容许性小的变量进入方程，不仅导致估计的方差增大，还会引起一些计算上的问题。此外，即使某一备选变量具有可接受的容许性从而可以进入方程，但由此可能导致原来已在方程中的那些变量的容许性变得不可接受的低。因此，在逐步回归中的每一步，都应当重新计算方程中全部变量的容许性。在SPSS REGRESSION中，通过TOLERANCE 设置变量容许性，其默认值为0.01。

⑸部分相关系数与偏相关系数

对于R^2 的改变量$R^2 - R(i)^2$ 带有正负号的平方根称为部分相关系数，部分相关系数是当其它自变量的线性效应从X_i 中消除之后Y 和X_i 的相关系数。另一个重要的系数是$PR_i^2 = [R^2 - R(i)^2]/[1 - R(i)^2]$ 其分子是部分相关系数的平方，分母是当除第i 个自变量以外所有其它自变量包含在方程中时解释变异的比例。带有正负号的PR_i^2 平方根称为偏相关系数，可解释为当其它自变量的线性效应从X_i 和Y 两者中消除后第i个自变量与因变量的相关系数。注意在绝对意义上部分相关系数不大于偏相关系数。

⑹关于未入选变量的统计量

对于尚未进入方程的自变量，可通过如下统计量考察其性质：如果该变量进入方程，其回归系数；对该系数为零的假设的t检验及概率水平；与因变量的偏相关系数与容许性等。由这些统计量可帮助判断下一步应该进入的变量。

通常，对于一个具体问题，我们事先并不知道上述模型是否合适。因此，有必要利用观测数据对模型假设的合理性给予考察，这种考察主要是通过残差分析进行的。若对数据而言模型是合适的，则残差E_i 作为e_i 的估计具有与e_i 类似的特征。

残差分析的内容很丰富，现只就SPSS REGRESSION 中有关残差分析的三个主要方面作一简要介绍，其一是利用残差分析考察上述模型假设的合理性，其二是探查对回归分析产生较大作用的异常点的强影响点。最后是多元共线性问题。设$e = (I - H)Y$, 它服从正态分布$N(0, (I-H)\sigma^2)$，方差估计用$(I-H)s^2$代替，就有标准化残差，$e_i/s\sqrt{1-h_{ii}}, h_{ii}$是H 的第$i$ 个对角元，在SAS 中称为STUDENT。据Belsley, Kuh and Welsch (1980) 的建议，使用$s_{(-i)}$代替s_i，在SAS 中称为RSTUDENT。

4. 模型假设合理性考察

⑴线性假设。方法一：作“残差图——预测值”图(即以预测值为横轴，残差为纵轴，将各观测个体相应的点绘于图上)。如果直线性和方差齐性得以满足，则预测值与残差值之间应不存在任何关系，即在假设满足时，残差应随机分布在通过0 的水平直线所展开的带状区域内(通常是|残差| ≤ 2)，如果从图上可看出任何变化模式，就应怀疑上述假设是否满足。方法二：作“残差——某自变量图”，同样若假设得以满足，残差应随机分布于水平带内。特别地，可考虑以未入选方程的那些自变量作残差图，若发现残差不是随机分布，可考虑将该变量包括进方程内。

⑵等方差性假设。如上所述，可利用作图的方法，若方差随自变量或预测值的增长而增长(或减小)，则应当怀疑Y 对所有X 值均为等方差这一假设。在SPSS REGRESSION 中，利用SCATTERPLOT 子命令可以对指定的一对变量作其散点图，以考察其直线性和等方差性假设。

⑶误差独立性假设。只要数据是按顺序(如时间) 收集，就应当作“残差—顺序变量”图，这是因为即使时间并未作为模型中的一个变量，它也可能影响残差。

如果残差与收集顺序无关，在上述图形上不应该发现任何变化模式，当误差项是正相关时，残差在一段上为正，另一段则为负；当误差呈负相关时，残差符号变化很频繁。利用Durbin-Watson 统计量可以检验相邻残差项是否为序列相关。残差自相关可用e_i与$e_{(i-1),i}$ 为下标的图来表现，用Durbin-Watson 检验。

⑷正态性假设。方法一，作残差直方图。方法二，作观测残差累积分布—期望残差累积分布“图(P-P 图)”。显然，当两者一致时，应产生一条直线，其中期望残差作横轴，观测残差为纵轴。正态性检验的Q-Q 图正常时为一条过原点的线，其斜率依赖于残差的标准差。另外可以使用W-统计量及Shapiro-Francia 统计量，它是观察次序残差与正态次序统计量的相关系数的平方。

⑸偏回归图。偏回归图是考察合理性的另一重要工具。对第j个自变量的偏回归图由两个残差构成，第一个是删除了第j个自变量后，其余自变量对因变量所做回归的残差(常称为偏回归残差)，第二个是自变量j与其余自变量回归所做回归的残差。在SPSS REGRESSION 中，利用PARTIALPLOT 子命令可对指定自变量作偏回归图。

⑹数据修正。当发现数据不符合模型假设时，可考虑对数据作适当变换。当直线性假设不符时，可根据描点和实际知识背景知识对数据作某些变换。当等方差假设不符时，可考虑对因变量作变换，使变换后的数据的误差方差相等。当误差独立性不成立时，可采用“两步估计法”。

5. 检查异常点和强影响点

⑴对某些观测个体，若其残差明显比其它观测个体的残差大很多，则称为异常点。由于异常点具有绝对值较大的残差，故可以直接使用残差图探查异常点。其次，可以通过直方图检查异常点。SPSS REGRESSION 对学生化残差的绝对值大于3.16 的观测个体直方图中都用”Out” 标记出来，最后还可以用逐点残差异常图来获得详细信息。

⑵马氏距离和中心化杠杆。马氏距离反映了某个观测点到观测中心的距离。第i 个观测个体的中心化杠杆值定义为：$L_i = D_i/(n-1)$，D_i 是第i 个观测点到中心的马氏距离。L_i 的取值范围从$-1/n$ 到$(n-1)/n$，均值为n/p。L_i 的取值越大则对回归的影响越大。

⑶强影响点。在一组数据中，对参数的估计具有特别大影响的观测个体称为强影

响点。这样的点被删除后回归直线与删除前相比有很大不同。识别强影响点的方法之一是当怀疑某个点是强影响点时删除该点，重新计算残差，考察其变化情况。一个点被删除后所计算的残差称为删除残差，删除残差除以其标准误则产生学生化删除残差。对回归效果有潜在影响的点是在空间上离$\overline{X}$ 较远的点，此距离可用h_{ii}来表示，它是投影阵H 的第i 个元，因为$\Sigma h_{ii}=p$，故其均值为p/n，Belsley, Kuh & Welsch (1980) 建议用$2p/n$做为它的界值(CUTOFF)，其上界是$3p/n$。有四种反映点的影响的统计量，它们是通过去掉该点来反映的。Cook's D 反映对回归系数估计值的影响，DFFITS 反映对预测值的影响，DFBETAS 反映对特定回归系数的影响，COVRATIO 反映对参数估计量方差—协方差阵的影响。前面三个可以想象成对去掉观察i 后对k 个线性无关的回归系数的影响，即$k(\beta-\beta_{-i})$ 可以写成与一般线性模型类似的二次型。三种度量对应不同的k值。建议用$2\sqrt{p/n}$做DFFITS 界值，其上界为$\sqrt{p}$，对Cook 距离近似用$4/n$ 做界值，其上界为1 ，对DEBETAS 建议用$2/\sqrt{n}$做界值，其上界为1，对于COVRATIO，建议值为$1\pm 3p/n$做界值。Cook 距离反映了当第i 个点被删除后在所有残差中的变化，常用于识别强影响点，其定义为：$C_i=\Sigma_{k=1}^{N}(\hat{y}_{k(i)}-\hat{y}_k^2)/[ps^2]$ 其中p 为自变量个数，s^2 是残差方差的估计量。很明显，强影响点时库克距离较大。

⑷异常点和强影响点的处理。发现异常点，应当根据专业知识的数据收集情况对其进行慎重分析处理。若发现是由于数据失误导致异常或强影响点，应当删除观测数据。若发现数据确系系统身产生，则应予以保留。考虑用一些稳健方法进行参数估计。

6. 多元共线性问题

若$|X'X|$不满秩，即多元共线性(multicollinearity) 问题，正规方程将有非正常解，为此出现了岭回归及主成分回归等相关方法。共线性诊断中的条件指标(condition index) 是矩阵条件数(condition number) 的推广，第i 个条件指标是最大特征值与第i 个特征值之比，Belsley 建议在10 附近为有相关的影响，大小100 为严重共线性，看一下有几个条件指标指示共线性的存在，从每个共线性中去掉一变量，若去掉变量后拟合效果太差，则应在拟合和共线性之间进行折衷。共线性诊断也可采用有方差扩张因子，若其值大于10时有共线性存在。

共线性处理常用方法：1. 适当的变量转换，增加一些观察，但这往往并不现实；2.从模型中去掉一些变量，两个自变量高度相关，模型中通常包括一个就够了；3.进行主成分回归。许多回归诊断结果提示的最好方法是删除记录，另外可采用稳键回归(robust regression) 等方法。岭回归(ridge regression)是一种有偏估计。对于回归方程$y=\beta_0+X\beta+\varepsilon,\beta_0$ 的估计是y的均值，而β的岭估计是：

$$\beta(k)=(X'X+kI)^{-1}X'y$$

$$Var[\beta(k)]=(X'X+kI)^{-1}(X'X)(X'X+kI)^{-1}\sigma^2$$

据Hoerl 与Kennard，存在$k>0$使$E[\hat{\beta}(k)-\beta]^2<E[\beta-\beta]^2$ k=0 时即通常的最小二乘估计，在k 增大时，方差扩张因子减少而$\beta_{(k)}$ 的偏倚增大，k 通常可取如$0.005\sim 0.2$，每个回归系数对k 做图给出岭迹(ridge trace)并由此确定k的取值。不同的作者有不同的做法。也可使用方差扩张因子VIF进行k 的选择，此时VIF 应于范围$1\sim 10$。

回归诊断在SAS(REG)、SPSS(REGRESSION)、Stata(regress) 等软件包均可以得到，SAS有专用过程RIDGE 进行岭回归分析，RIDGE 也可做为PROC REG 的选项进行岭回归分析。在SPSS 中使用RIDGEREG 宏定义进行岭回归分析。启用方法：

RIGDGEREG DEP=因变量/enter自变量/start= /stop= /inc= /k=后面选项设定k 值。

描述、分析和研究因素间的相互关联和影响，可以通过一些统计指标如相关系数和典型相关，还可以通过考查其相关的结构。据Karlin, S. (1983)，常用的方法有：通径分析、结构方程模型以及方差分量分析，它们都基于线性的假设。结构方程模型在第14章介绍，第4章有CALIS的例子。结构方程模型与通径分析方法有密切的联系。

多元线性模型包含多元回归模型、多元方差分析模型用协方差分析模型等。在SAS 中有专门的过程MIXED 处理混合模型数据。方差分量分析是基于一般线性模型，估计各变异来源的方差组分大小，可用于遗传学分析等。由平衡资料估计方差分量即为方差分析法(ANOVA)，即把方差分析表中的均方作为其期望值的估计。对于非平衡资料估计方差分量采用Henderson 法Ⅰ、Ⅱ、Ⅲ及极大似然法(ML)、约束极大似然法(REML)、MINQUE、MIVQUE和I-MINQUE 等，在SAS 中方差分量分析使用过程VARCOMP 实现。

§2.3.3 方差分析

许多研究，要归结到几个样本均数的比较。设有k个样本均数，要比较它们的差别，若取检验水准$\alpha = 0.05$，对这些均数用t—检验两两比较，共有$k(k-1)/2$个比较，总结论的检验水准就成为$1-(1-\alpha)^k$，可见这样做既不经济，效率也低。方差分析正是处理这一类问题的统计学方法，方差分析有时也称变异数分析。

方差分析的应用条件是：各个样本来自正态总体；各个样本是相互独立的随机样本；各总体方差相等。

方差分析的基本思想是把所有数据的总变异(离均差来方和)分解成几个部分，然后对各部分的变异进行比较。完全随机设计或单因素设计，是把受试对象完全随机地分配到各个处理组中去。处理组可以为两组或多组，各组样本含量可以相等，也可以不等。完全随机设计方差分析把总变异区分为处理组间变异和组内变异。配伍组设计，也称随机区组设计，是扩展的配对设计。配伍组设计的方差分析可以把总变异分为处理组间变异、配伍间变异和误差三个部分，较完全随机设计提高了效率。若研究的因素很多，可以使用析因设计或正交设计。

使用MANOVA可控制整个实验水平上的误差，同时考虑了因变量间的关联。其基本假设是独立性、方差协方差阵相等、正态性。MANOVA需要更多的样本量，受异常值的影响也更大，其假设因变量间的线性组合。显著性检验准则常用的有四种，即Roy最大特征根、Wilks λ、Hotelling迹、Pillai准则。它们之间最基本的不同是对“不同维”上因变量的差异的评定方法。Roy准则利用第一个特征根来评价，这样其功效和特异度比较好，最适于因变量在某一准上存在强相关的情形，同时也是违背准则时受影响最大的。

对于方差分析假设的检验常用的如Bartlett检验、进行必要的数据转换等。

所谓平衡是指在分类变量的交叉下的记录个数相同；表2.11是一个不平衡设计的例子。

表 2.11 2×2 设计中的格子均值

数目	第一列	第二列	数值	第一列	第二列
第一行	2	200	第一行	10	20
第二行	200	2	第二行	30	40

第一行总均值=19.90。在行均值之间约差10，行效应存在；列均值之间亦约差10，列效应也是有效的，可以化出交互项，效应也存在，原因在于模型受了大样本的影响。表2.12也是一个不平衡设计的例子，负值表示缺失，括号内是每个格子中观察值的数目。

表 2.12 不平衡设计的例子

		因	子	B
数据(n_{ij})		1	2	3
因子 A	1	2,4,6	4,6	5
		(3)	(2)	(1)
	2	12,8	11,7	-1
		(2)	(2)	(0)

方差分析模型是$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$

表 2.12 的SAS运算结果列如表2.13：

表 2.13 表 2.12的运算结果

来源	自由度	I	II	III	IV
A	1	60.00	57.02	54.55	54.55
B	2	0.32	0.32	0.21	1.50
A×B	1	2.18	2.18	2.18	2.18

对线性模型$Y = X\beta + \varepsilon, E(Y) = X\beta$，分析的基本目的是在可能的情况下估计或检验$\beta$或其线性组合，这可以由观测Y的线性组合来做到。又设β的线性组合是$L\beta$，则应有Y的线性组合，使其期望为$L\beta$，这就要求这样的组合存在，也就是说能找到X的相应的线性组合，因而X的行也就成为L的发生集，而且因为$X = X(X'X)^{-}(X'X), X'X$的行也成为L的发生集。根据设计是否平衡等因素，L取为不同的形式。如 I 型平方和是用修正扫描算子计算$X'X$的g2逆求解正规方程组的副产品，可用许多方法得到，一种方法是对$X'X$ 向前的Doolittle分解，跳过任何对角元为零的情形。

SAS 能输出四种平方和，I 型为顺序平方和，使用Searle的记号，$SS(A) = R(\alpha|\mu), SS(B) = R(\beta|\alpha,\mu), SS(AB) = R(\gamma|\alpha,\beta,\mu)$ 等等，其各SS的大小依赖于效应进入模型的顺序，在效应的排列很好时是有用的，如用于多项式回归可以看出是否需要继续引入高次项。在多项式回归中，相应于正交多项式检验，项的贡献可以很清楚。其它的特点如：所有效应的SS之和与模型SS 相同，若剩余是独立正态分布则各SS是独立的。对于不平衡资料，其假设是格点的函数，结果一般与平衡资料不同。II 、III、IV 型称为偏平方和，“偏”的含义是指进行了其它效应的调整，即每一种均调整了其它的分类效应，调整准则不同，检验与效应进入模型的顺序无关。如第 II 种$SS(A) = R(\alpha|\beta,\mu), SS(B) = R(\beta|\alpha,\mu), SS(AB) = R(\gamma|\alpha,\beta,\mu)$，一般不具有平衡设计中的那种相等分布(equitable distribution) 或正交(orthogonality)特性。III、IV 与 II 的区别是高阶交互或嵌套效应的系数也进行调整以满足正交性条件(III) 或等分布(IV)，这些效应的系数不再依赖于格子数n_{ij}，只要$n_{ij} \neq 0$时III、IV 是相同的。出现$n_{ij} = 0$时，III具有正交

特性，检验好象是针对“效应和为零”，而Ⅳ具有平衡特性，用非零格子的子集作为平衡数据集，结果不唯一。对于平衡资料，四种平方和相同；在模型没有交互时，Ⅱ=Ⅲ；在所有格子非空(all-cell-filled data)时Ⅲ=Ⅳ。在PROC GLM中，可估计函数由选择项E1-E4给出，几种平方和由S1-S4给出。

SPSS用MANOVA和ANOVA进行方差分析，并区分独立(UNIQUE)、顺序(SEQUENTIAL)等离均差平方和。

§2.3.4 主成分分析

在实际应用中，为了全面分析问题，提出的指标(或变量)往往很多，每个指标都在不同程度上反映了所要研究的课题的某些信息，由于指标之间常常具有一定的相关性。因此希望找到较少的几个彼此不相关的指标，来代替原来的指标并且尽可能地反映原来指标的信息，这就是主成分分析的思想方法。

1. 主成分的定义及求法

设$X=(X_1,X_2,...,X_p)$为p维随机变量，有二阶矩存在，记$\mu=E(X),\Sigma=Var(X)$，考虑它的线性变换$Z_i=l_i'X,i=1,...,p$, 其中$l_1'=(l_{11},l_{12},...,l_{1p})$，易见

$$Var(Z_i)=l_i'\Sigma l_i,\quad Cov(Z_i,Z_j)=l_i'\Sigma l_j,i\neq j$$

因此$Var(Z_1)$越大，表明Z_1包含的信息就越多，若$Z_1=l_1'X$满足

$$l_1'l_1=1,Var(Z_1)=\max Var(l'X)$$

则称Z_1是X的第一主成分，Z_1是X的所有线性变换$l'X$ 中最能综合原p个变量信息的一个线性变换，其中，使得方差达到最大的向量即为主成分系数。如果第一个主成分不足以代表原p个变量的信息，考虑第二主成分Z_2，为了有效地代表原变量的信息，第一主成分Z_1 已有的信息就不需要出现在第二主成分Z_2中，即$Cov(Z_1,Z_2)=0$。由此可得$l_1'l_2=0$或$l_2'l_1=0$。因此，若满足

$$l_2'l_2=1,l_2'l_1=0,Var(Z_2)=\max_l Var(l'X)$$

则称Z_2是X的第二主成分，一般地，如果$Z_i=l_i'X$ 满足

$$l_i'l_i=1,l_i'l_j=0,j=1,...,i-1,Var(Z_i)=\max_l Var(l'X)$$

则称Z_i是X的第i个主成分。

假定$\lambda_i,i=1,2,...,p$为方差矩阵$Var(X)=\Sigma$ 的p个特征根并且它们由大到小的排列为$\lambda_1\geq\lambda_2\geq...\geq\lambda_p\geq0$，则$X$的第$i$个主成分的系数向量$l_i$就是第$i$个特征根$\lambda_i$ 所对应的正则化特征向量，其中$l_i'l_i=1,l_i'l_j=0,i\neq j,i,j=1,...,p$

若记所有主成分构成的向量为Z, 相应的主成分系数矩阵为L,则上述线性变换即为$Z=L'X$, 而且可以得到以下结论：

(1). $L'L=I_p$, 即L 是正交阵，因此，主成分代表原变量空间中的垂直向量，或者说，主成分是对原变量进行了一次正交变换。

(2). Z的分量间互不相关，即相关系数矩阵$Cov(Z_i,Z_j)=0$。

(3). Z的p个分量是按方差大小，由大到小排列的。

(4). $Var(Z) = L'\Sigma L = diag(\lambda_1, \lambda_2, ..., \lambda_p)$。因此，由特征根$\lambda_i, i = 1, 2, ..., p$的大小可知它所包含信息的多少。一般，称$\lambda_k/\Sigma_{i=1}^{p}\lambda_i$为第k个主成分的方差贡献率，$\Sigma_{i=1}^{k}\lambda_k/\Sigma_{i=1}^{p}\lambda_i$称为前$k(k \leq p)$个主成分的累积方差贡献率。通常在实际应用中，当前个主成分的累积方差贡献率超过85%时，则用这前k 个主成分就可以描述原来p个变量的信息了。

(5). 相关系数矩阵$\rho(Z_k, X_i) = \sqrt{\lambda_k}l_{ki}/\sqrt{\sigma_{ii}}$其中$\sigma_{ii}$为$\Sigma$的第$i$个主对角元素，表示原变量在主成分中的负荷量，在实际应用中，为了消除量纲的影响，往往把原变量标准化，标准化后的协差阵即为原变量的相关阵。此时，要观察原变量的负荷量，只需观察l_{ki}的系数，也即Z_k的系数。

2. 样本主成分

在实际问题中，$X' = (X_1, ..., X_n)$的协差阵或相关阵常常是未知的，于是抽取随机样本，得到观察数据阵X，从X可得样本协差阵或样本相关阵，并求得它们的特征根以及对应的正则化特征向量，得到的主成分，即为样本主成分。若记$Z_i = l_i'X$为的第i个样本主成分，$Z_{ki} = l_i'X_k, i = 1, ..., p, k = 1, ..., n$, 称$Z_k = (Z_{k1}, Z_{k2}, ..., Z_{kp})', k = 1, ..., n$ 为主成分得分。

3. 主成分回归

在回归分析中，当自变量$X_1, X_2, ..., X_p$之间存在多重共线性关系时，$|X'X|$接近于0，此时用通常的最小二乘估计求得的回归方程就可能出现一些不符合实际的情况，主成分回归是一种可以选择的方法。设自变量$X_1, ..., X_p$ 和因变量Y有对应关系，首先对自变量进行中心化，仍记作$X = \{x_{ij}\}$，则有线性回归方程：

$$\hat{y} = \hat{c}_1x_1 + ... + \hat{c}_px_p$$

其中回归系数$\hat{c}_j, j = 1, ..., p$是通常的最小二乘解。

记矩阵X的奇异值分解$X'X = U\Lambda U'$ 中的正交阵为U, X 的主成分为$w_1, ..., w_p$, 现求Y对这些主成分变量的回归，得：

$$\hat{y} = \hat{b}_1w_1 + ... + \hat{b}_pw_p$$

则$\hat{b}$与$\hat{c}$有关系$\hat{b} = U'\hat{c}$，且$\Sigma\hat{y}^2 = \Sigma\hat{\lambda}b^2$

主成分回归的回归系数只与其相应的主成分有关，与其它主成分无关；主成分回归平方和与原来的回归平方和相等，且等于各主成分对 y 的回归平方贡献之和，此性质可用于变量的筛选。若主成分有明确的实际意义，则把主成分看成单个自变量，要减少参加回归的主成分，仅去掉若干个主成分即可；若主成分没有明确的意义，或仍希望用原始变量来求回归方程，则首先找出贡献最小者，再据其组合系数U值的大小进行取舍，因为该主成分对回归贡献小，则其对应的起主要作用的原始变量贡献也应较小，应予以舍弃。SAS 用PRINCOMP过程进行主成分分析。

§2.3.5 因子分析

1. 因子分析是识别代表大量相关变量相互关系的一组少量(通常是不可测的) 因子的统计技术。因子分析试图用最小个数的不可测的所谓公因子的线性函数与特殊因子来对原

来观测的变量进行描述，这样做的目的，是尽可能合理地解释存在于原变量间的相关性，并且简化变量的维数与结构。

例如: 教师想从各门课程考试的成绩中了解学生的“理解能力”,“计算能力”,“记忆能力”等等。成绩是可以测定的，而这些“能力”是不可直接测定的，习惯上称此为公共因子(common factor)，显然，成绩的好坏受这些公共因子的影响，而每门课都有其特殊性，因此,它还受到一个不可测的特殊因子(unique factor)的影响，因子分析就是根据一些变量$X_1...X_p$(相当于各门课程的成绩) 来得到公共因子$f_1...f_q(q<p)$(相当于“理解能力”,“记忆能力”等等)的统计方法。

2. 因子模型

(1). 初始因子模型

设$X=(X_1,X_2,...,X_p)'$为p维可观测随机变量，而且X已中心化并仍记作X,由q个公共因子$f=(f_1,...,f_q)'$所支配，其相应的因子模型为：

$$X_{p\times 1}=A_{p\times q}f_{q\times 1}+\varepsilon_{p\times 1}$$

这里假定：

$$E(X)=0,Var(X)=\Sigma>0,E(f)=0,Var(f)=I_q,q<p$$

$$E(\varepsilon)=0,Cov(f,\varepsilon)=0$$

$$Var(\varepsilon)=D=diag(e_1^2,e_2^2,...,e_p^2)$$

特殊因子$\varepsilon_i,i=1,...,p$彼此不相关且具有单位方差，每个公共因子至少对两个变量有贡献，否则它将成为特殊因子。

据模型有: $\Sigma=AA'+D\equiv H+D$

记$\Sigma=(\sigma_{ij})_{p\times p},A=(a_{ij})_{p\times q},H=(h_{ij})_{p\times p}$ 且$h_{ij}=\Sigma a_{ik}a_{kj},h_{ii}=h_i^2$ 则

$$\sigma_{ii}=h_i^2+e_i^2,\sigma_{ij}=h_{ij},i\neq j,i,j=1,...,p$$

由此可知：

1) a_{ij} 反映了X_i 与f_j 之间的相关。

$a_{ij}=Cov(X_i,f_i)$，且当$Var(X_i)=1$ 时$a_{ij}=\rho(X_i,f_j)$，因此通常称a_{ij} 为第i个变量X_i在第j个公共因子f_j上的载荷量，A为公因子$f_1,...,f_q$的载荷矩阵。

2) $h_i^2,i=1,...,p$ 反映了公因子对X_i的影响作用，称此为公因子方差或称共性估计值。

3) $g_k^2=\Sigma_{i=1}^p a_{ik}^2,k=1,...,q$ 反映了第k个公共因子f_k对X的各分量X_i的方差贡献之和，是衡量每个公因子相对重要的一个尺度。

值得注意的是，以上模型只是对均值为0的随机变量而言，若X为标准化随机变量，则模型中的Σ应是相关矩阵R。

(2). 旋转后的因子模型

初始因子模型建立后，每个变量在公共因子上的载荷量往往没有很明显的差别，因此，不易对公共因子作出解释，这时，需要对载荷矩阵进行进一步简化，使得各列元素

向0和1两极分化，但保持各变量的公因子方差$h_i^2, i = 1, ..., p$ 不变，这种变换方法称为因子的旋转。

旋转的主要思想在于获得一个简单的结构，希望每个因子都对部分变量有非零载荷，以便对因子作出解释。希望每个变量也仅对部分因子有非零载荷，这样使因子之间相互不同(因为若几个因子在某个变量上均有较高载荷，则很难解释清楚这些因子各有何区别、特点)。SPSS提供了几种转换方法，最常用的是方差极大旋转法(VARIMAX)，它试图尽可能减少在一个因子上具有较高载荷的变量个数，使因子便于被解释。另外两种方法是四次极大化(QUARTIMAX) 和等方差极大化(EQUAMAX) 方法。此外，为了便于简化因子载荷矩阵，也可考虑采用斜交旋转。

a. 正交旋转(othogonal rotation)

在初始因子模型中，若在A阵后面乘上$q \times q$的正交阵Γ，f 的前面乘上Γ'，此时，相当于对公共因子进进行正交旋转，正交旋转后的因子模型为：

$$X = Af + \varepsilon = A\Gamma\Gamma' f \equiv A^\star f^\star + \varepsilon$$

$A^\star$ 中的元素向0和1两极分化，$f^\star$为旋转后的公因子便于解释，因为

$$f^\star = \Gamma' f, E(f^\star) = \Gamma' E(f) = 0$$

$$Var(f^\star) = \Gamma' Var(f)\Gamma = I_q$$

$$Cov(f^\star, \varepsilon) = Cov(\Gamma' f, \varepsilon) = \Gamma' Cov(f, \varepsilon) = 0$$

所以，旋转后的公共因子也是不相关的。

b. 斜交旋转(oblique rotation)

若在公共因子之前乘上的非奇异矩阵，此时，相当于对公共因子进行斜交旋转，斜交旋转后的因子模型为$X = Af + \varepsilon = AB^{-1}Bf + \varepsilon = AB^{-1}f^{\star\star} + \varepsilon$

由于

$$E(f^{\star\star}) = E(Bf) = BE(f) = 0$$

$$Var(f^{\star\star}) = Var(Bf) = BB'$$

不一定为单位阵。

因此，公共因子通过斜交旋转后，变为相关的因子，尽管如此，斜交旋转常产生比正交旋转更有用的模型。

(3). 因子得分模型

无论是初始因子模型，还是旋转后的因子模型，它们都是用可观测变量的线性组合来表示公共因子，并计算这些公共因子的估计值，这种估计值叫做因子得分。对第k个观测个体，第j个因子的得分估计为：$f_{jk} = \Sigma_{i=1}^{p} w_{ji} X_{ik}$为第$k$ 个观测个体第i 个变量的标准化值，w_{ji} 是第j 个因子第i 个变量的因子得分系数，$W = \{w_{ij}\}$称为得分矩阵。

SPSS FACTOR 提供的因子得分系数估计的方法有：Anderson-Rubin 方法、回归方法和Bartlett 方法。

3. 参数估计

估计载荷矩阵和特殊因子的协差阵常用的方法常有以下几种：

(1). 主成分分析法。从p维可观测变量$X_{p\times 1}$的观测数据阵，可得到的样本协差阵，假设由大到小排列的特征根所对应的正则化特征向量为$l_1,...,l_p$，则当最后$p-q$个特征根较少时，S可近似地分解为

$$S=\lambda_1 l_1 l_1'+...+\lambda_q l_q l_q'+D=AA'+D$$

上述的A和D即为因子模型中载荷矩阵和特殊因子协方差阵的估计。

由于A中第j列元素与主成分的函数只相差一个常数，$\sqrt{\lambda_j}, j=1,...,q$，故这个估计方法通常称为主成分分析法。当量纲不同时，可把原变量标准化，类似地从相关阵R出发求得A和D的估计，将其进行谱分解而获得特征向量$V_1,V_2,...,V_p$及其特征根$\lambda_1\geq\lambda_2\geq ...\lambda_P\geq 0$，因此对事先设定的临界值$t$，若$q$ 满足：$\sum_{i=1}^{q}\lambda_i/\sum_{i=1}^{p}\lambda_i\geq t \quad \sum_{i=1}^{q-1}\lambda_i/\sum_{i=1}^{p}\lambda_i < t$

则因子载荷矩阵估计为：

$$A=(\sqrt{\lambda_1}V_1,\sqrt{\lambda_2}V_2,...,\sqrt{\lambda_S}V_q)$$

V_i是λ_i 相应的特征向量。特殊因子方差估计为：$\hat{\delta}_i^2=1-\Sigma_{j=1}^{q}a_{ij}^2$，公共因子方差为$\hat{h}^2=\Sigma_{j=1}^{q}a_{ij}^2$

(2). 主因子分析法。也称主轴析因法，这是对主成分法所作的一种修正，从相关阵出发求得A与D的估计。由于标化变量的协差阵即为原变量的相关阵R，因而有$R=AA'+D$，而且若有特殊因子方差的一个初始估计$\hat{\delta}^2$，则由$R-D=AA'$可知先验公因子方差的估计为，$\hat{h}_i^2=1-\hat{\delta}^2$。

记$R^\star=R-D$(称之为约相关阵)，则$R^\star$的对角元是$h_i^2, i=1,...,p.$ 由$R^\star$可得其特征根$\lambda_1^\star\geq\lambda_2^\star\geq ...\geq\lambda_p^\star>0$，取前$q$个较大的特征根，并计算其对应的特征向量，则$R^\star$可近似地分解成$R^\star=AA'$，其中$A=(\sqrt{\lambda_1}l_1,...,\sqrt{\lambda_q}l_q)$，由此可求得

$$\delta_i^2=1-\sum_{j=1}^{q}a_{ij}^2, i=1,...,p, D=diag(\delta_1^2,...,\delta_p^2)$$

在实际应用中，由于D是未知，用D的初始估计来求得A和D的估计也是一个近似解，因此，常用迭代主因子法来获得一个更好的解，即用上面得到的D 作为特殊因子协差阵的估计，重复上述步骤，直到得到稳定解，当特殊因子方差的初始值为0时，主因子分析法即为主成分分析法。

若已知先验公因子方差h_i^2的初始估计值，同样也可推得A和D 的主因子解，常用的初始估计有以下几种方法：

取h_i^2为第i个变量与其它所有变量的多重相关系数的平方

取h_i^2为第i个变量与其他变量相关系数绝对值的最大值

取$h_i^2=1$(此时，主因子法等于主成分法)

(3). 极大似然法。假定$f\sim N_q(0,I_q),\varepsilon\sim N(0,D)$，随机变量$X_1,...,X_n$ 为来自正态总体$N_p(\mu,\Sigma)$的简单随机样本，则样本似然函数为$L(\mu,\Sigma)$，取$\mu=\overline{X}$(样本均值)，$\Sigma=AA'+D$，

则$L(\mu,\Sigma)$为A、D的函数，使似然函数达到最大，A、D 从式$diag(S) = diag(AA' + D)$和$A = S(AA' + D)^{-1}A$求得，其中S为样本协差阵。

(4) 广义最小二乘法。载荷矩阵A和特殊方差矩阵D由极小化下列目标函数产生：$tr(S-\Sigma)'H(S-\Sigma)$，其中$\Sigma = D + AA', H = \Sigma^{-1}$或为其相合估计，S 为样本协方差阵。

(5) 未加权最小二乘法。与最小二乘法类似，但用于极小化的目标函数不作加权。

除了上述因子提取方法外，还有α方法和象因子法。

4. 模型的检验

因子模型建立以后，可用似然比检验的方法检验其是否合适：

$H : \Sigma = AA' + D$ 相对于$A : \Sigma \neq AA' + D$

若$\lambda = |S|^{n/2}/|\Sigma|^{n/2}$ 的值很小，则似然比检验拒绝H，其中的S为样本协差阵，由渐近理论可知-2ln λ 的分布是自由度为f的χ^2分布，其中$f = 0.5[(p-q)^2-(p+q)]$。Bartlett(1951)建议用如下的近似值

$$-[n-1-(1/6)(2p+5)-(2/3)q]\ln(|S|/|\Sigma|) = \chi_f^2$$

若$\chi_f^2 \geq \chi_{f;\alpha}^2$，则拒绝$H$，此处$\chi_{f;\alpha}^2$为显著性水平为$\alpha$，自由度为$f$的$\chi^2$临界值，若拒绝H，则认为所选的模型不合适，必须增加一个公共因子以重新求得载荷矩阵A的估计，然后再次检验模型，重复这个步骤，直到模型合适为止。

由于因子分析的目的之一就是试图获得能够解释变量关联的因子，因此在一个合适的因子模型中，变量必须相关。SPSS FACTOR 提供了下列考察变量间相关的途径。

(1) Bartlett 检验：用以检验关于相关阵为单位阵的假设，若不能拒绝该假设，则不宜使用因子模型。

(2) 偏相关系数：当其余变量的线性影响消除后，两个变量之间的偏相关系数应当较小，这可用反象相关(anti-image correlation) 来刻画。

(3) Kaiser-Meyer-Oklin (KMO) 指数：

$$KMO = \frac{\sum\sum_{i\neq j} r_{ij}^2}{\left[\sum\sum_{i\neq j} r_{ij}^2 + \sum\sum_{i\neq j} b_{ij}^2\right]}$$

其中r_{ij} 是变量i 和j 的样本相关系数，b_{ij} 是度量i 和j 的偏相关系数。若KMO 值较小，说明不宜做因子分析，根据Kaiser 的划分，$KMO > 0.9$ 为优秀，$0.8 < KMO < 0.9$ 为良好，$0.7 < KMO < 0.8$ 为中等，$KMO < 0.5$ 不能接受。

因子分析通常的步骤是：⑴计算并考察所有变量所构成的相关矩阵；⑵提取能代表数据的因子；⑶通过旋转变换使因子更具可解释性；⑷对每个观测个体计算因子得分。

确证型因子分析大致步骤为：设计一个理论模型，是构造因果关系的路径图并将其转化为一系列结构方程模型和度量模型，选择矩阵类型和模型估计，最后是对所识别的模型进行评价，看其适合度如何，进行模型解释和模型修正。

R.A. Johnson [18]建议因子分析用以下步骤， 1.首先进行主成分因子分析，绘出因子得分图并找出可疑观察，计算标准得分；进行最大方差旋转； 2.进行极大似然法因子分析和方

差极大化旋转；3.比较因子分析的结果：因子载荷聚合的模式是相同的吗？绘图比较主成分法和极大似然法的得分； 4.对于其他数目的共因子重复上述步骤，看是否有其它因子对于数据解释有用； 5.把大的数据集分成两部分，分别进行分析。

SAS 的因子分析过程为FACTOR，使用CALIS进行确证型分析。BMDP用4M进行因子分析。

§2.3.6 典型相关分析

1. 是研究两组随机变量间的相关关系的一种方法，其中，每组变量可能包含有多个变量。设$X=(X_1,...,X_p),Y=(Y_1,...,Y_q)$分别是$p$维和$q$ 维随机向量(假定$p\le q$)，典型相关分析就是要研究X与Y之间的相关性，当$p=q=1$时，X与Y之间的关系的大小用相关系数去衡量。当$p=1$，$q=n$ 时，X 与Y之间的关系大小用复相关系数去衡量，当$p\neq 1,q\neq 1$时，X与Y之间的相关大小用典型相关系数去衡量。

2. 总体典型相关

   令
   $$E\begin{pmatrix}X\\Y\end{pmatrix}=0,Cov\begin{pmatrix}X\\Y\end{pmatrix}=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{pmatrix}$$
   其中Σ_{11},Σ_{22}分别为$p\times p,q\times q$ 阶正定矩阵，Σ_{12}为$p\times q$阶矩阵，且$\Sigma_{12}=\Sigma_{21}'$我们用X与Y的线性组合$\alpha'X$与$\beta'Y$之间的相关性来描述X与Y 之间的相关性。

   X与Y之间的第一典型相关(系数)为X的线性组合$\alpha'X$与Y的线性组合$\beta'Y$二者之间的极大相关，也就是
   $$\rho_1=\frac{\alpha_1'\Sigma_{12}\beta_1}{\sqrt{(\alpha_1'\Sigma_{11}\alpha_1)(\beta_1'\Sigma_{22}\beta_1)}}=\max_{\alpha,\beta}\frac{\alpha'\Sigma_{12}\beta}{\sqrt{(\alpha'\Sigma_{11}\alpha)(\beta'\Sigma_{22}\beta)}}$$
   可以验证ρ_1^2即为$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$的最大特征根，$\alpha_1$为相应特征向量，而$\beta_1$为$\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$的最大特征根所对应的特征向量，称$V_1=\alpha'X,W_1=\beta'Y$ 为第一对典型变量，α_1,β_1为典型系数或典型权数，称方差为1的典型变量的系数为正则化典型系数。

   类似地，因$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$与$\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ 非负定，且$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$与$\Sigma_{11}^{-0.5}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-0.5}$有相同特征根，若令$A=\Sigma_{11}^{-0.5}\Sigma_{12}\Sigma_{22}^{-0.5}$, 记$\rho_1^2\ge\rho_2^2\ge,...,\ge\rho_p^2$ (假定$p\le q$)为AA'和$A'A$的有序特征根，$\alpha_1,\alpha_2,...,\alpha_p$为对应于$AA'$的有序特征根的$p$维特征向量，$\beta_1,\beta_2,...,\beta_p$为对应于$A'A$ 的有序特征根的q维特征向量，则称$V_i=\alpha_i'X,W_i=\beta_i'Y,i=1,...,p$ 为第i对典型变量，它们之间的相关为第i个典型相关。注意V_i 与W_i 各自的变量之间不相关并且
   $$Var(\alpha_i'X)=Var(\beta_i'Y)=1,Cov(\alpha_iX,\beta_iY)=\rho_i,i=1,2,...,p$$

3. 样本典型相关

   在实际问题中，Σ往往是未知的，故要通过样本来估计。假定从总体中抽取一个大小为n的样本($n>p+q$)，每个样本有X,Y两组指标，若S为其样本协差阵，则将S分割如$S=\begin{pmatrix}S_{11}&S_{12}\\S_{21}&S_{22}\end{pmatrix}$，其中$S_{11},S_{12}$ 分别是$p\times p$ 和$q\times q$ 阶矩阵，$S_{12}=S_{21}'$。

   令$\lambda_1^2\ge\lambda_2^2\ge...\ge\lambda_p^2$为$S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}$的有序特征根(假定$p\le q$)则称$1\ge\lambda_1\ge\lambda_2...\ge\lambda_p\ge 0$为样本典型相关(系数)。

4. 典型相关的显著性检验

求出典型变量对和典型相关系数后，把具有显著性意义的典型相关系数所对应的典型变量对保留下来，并给予合理的解释，若第i个典型相关系数近似为0，那么,这一对典型变量对于解释原来两组变量间的相关性就没有意义，因此，必须对ρ_i进行显著性检验，若$(X \quad Y)' \sim N_{p+q}(0,\Sigma)$，则可以对典型相关系数作$\chi^2$检验，检验假设为：

$H_{0i}:\rho_i=0, i=1,...,p$

检验以上假设可用Bartlett关于大样本χ^2的统计量，取统计量

$$Q_i=-[n-i-0.5(p+q+1)]\Sigma_{k=i}^{p}\ln(1-\lambda_k^2)$$

其中λ_k是由样本观测数据得到的第k个典型相关系数

对较大的样本量n，在H_{0i}为真时$Q_i \sim \chi^2(f_i)$, 其中$f_i=(p-i+1)(q-i+1)$，对给定的显著性水平α，当$Q_i > \chi_\alpha^2(f_i)$时，拒绝H_{0i}，即认为$\rho_i \neq 0$，于是再对ρ_{i+1}作检验，依次下去，若某一个$Q_j < \chi_\alpha^2(f_j)$, 则接受原假设，认为$\rho_j=0$，则有$\rho_{j+1}=\rho_{j+2}=...=\rho_p=0$。注意应首先对$\Sigma_{12}=0$ 进行检验。

5. 分析

在得到典型相关变量对后，可以用各变量的载荷反映其在相应的典型相关变量对中的作用。

载荷是变量组$X=(X_1,X_2,...,X_p)'$的任一指标$X_i, i=1,...,p$ 与其线性组合$V_j=\alpha_j'X$的相关系数，$\gamma_{X_iV_j}=\frac{Cov(X_i,V_j)}{\sqrt{Var(X_i)}\sqrt{Var(V_j)}}, i,j=1,...,p$ 称为X_i在V_j中的载荷。类似地，Y_i与W_j的相关系数：$\gamma_{Y_iV_j}=\frac{Cov(Y_i,W_j)}{\sqrt{Var(Y_i)}\sqrt{Var(W_j)}}, i,j=1,...,q$ 称为Y_i在W_j中的载荷。

现虑一个$n\times p$阶矩阵X，分块为g个$n_j\times p$阶阵$X^j, j=1,...,g$，X^j 的n_j个行来自群体$\Pi_j, j=1,...,g$。用$n\times(g-1)$ 阶矩阵Y表示X的分群标志：

$$y_{ij}=\begin{cases}1 & x_i\in\Pi_j;\\ 0 & x_i\notin\Pi_j\end{cases} \qquad i=1,...,n, j=1,...,g-1$$

可把判别分析问题化为典型分析问题，首先引进指示变量矩阵Y，然后求X与Y的最大典型相关系数λ_1和相应的典型向量a_1，即可得到判别函数$Y=a_1X$。

SAS 典型相关分析过程为CANCORR。SPSS 有专用的宏定义CANCORR，其使用方法为：CANCORR SET1=变量表1 /SET2=变量表2。BMDP相应的程序是6M。

§2.3.7 判别分析

1. 是根据待判个体的某些特定指标的观测值判断其类别归属的统计分析技术。从数学上看，就是对具有分布函数$F_i(x)$ 的k 个母体$G_i, i=1,...,k$ 判定给定的待判个体x 来自那个母体。常见的判别分析方法有Fisher 线性判别、Bayes 判别、距离判别、核密度法等。在SPSS/PC+ DSCRIMINANT 中的基本假定是参与判别分析的母体具有正态等方差分布，以线性判别函数和Bayes 分类规则进行判别。

以下简要介绍Fisher 线性判别函数、Bayes 分类规则以及在DSCRIMINANT 中使用的一些概念。

2. 普通判别(非逐步判别)

(a) Fisher 线性判别函数

对于p 维观测$X=(X_1,X_2,...,X_p)'$，线性判别函数可表为：

$$D(X)=\beta_0+\beta_1X_1+...+\beta_pX_p$$

Fisher 线性判别函数的系数b_i 的选取原则是使上述函数$D(X)$ 的取值尽可能在不同类别中不同，换言之，是选取使比率λ=组间平方和/ 组内平方和达到最大的系数b_i。

对观测个体X_0，代入上述Fisher 线性判别函数后，可得其判别得分$D(X_0)$。Fisher 准则只提供了确定判别系数的准则，并未涉及对个体的分类规则(allocation rule)。

(b) Bayes 分类规则

利用判别得分，DSCRIMINANT 基于Bayes 分类规则将待判个体判给某一类别。由Bayes 逆概率公式，在已知某个观测个体X_0 具有得分$D(x_0)$ 的条件下其属于母体i 的概率

$$P(G_i|D(X_0))=\frac{P(D(X_0)|G_i)P(G_i)}{\Sigma_{j=1}^{k}P(D(X_0)|G_j)P(G_j)}$$

其中$P(G_i)$ 为母体G_i 的先验概率(prior probability)。

先验概率是关于各母体的一种先验知识，其估计可由许多方法得到。对混合抽样问题，可用所抽到的每个母体中个体的比例作为对先验概率的估计。对于固定抽样问题，则只能考虑采用其他途径估计。最后，若各母体等可能出现或对其先验信息完全未知，则可对所有母体采用等先验概率。在DSCRIMINANT 中，为上述三种形式的先验概率在PRIORS 子命令中设计了相应的输入方式。

根据Bayes 分类规则，基于其判别得分$D(X_0)$，待判个体X_0将按最大后验概率$P(G_i|D(X_0))$确定其属于哪一个母体，即分类为规则：

若

$$P(G_i|D(X_0))=\max_j P(G_j|D(X_0))$$

则将观测个体X_0 (其判别得分为$D(X_0)$) 判为属于母体G_i。

(c) 训练样本、待判样本

一个典型的判别问题通常分为两步，首先根据训练样本(其所包含的观测个体的类别归属已知) 构造判别函数，其次对待判样本(不知其观测个体属于哪个母体) 或对验证样本(通过对其判别结果对错判率进行估计)进行判别。在DSCRIMINANT 中，利用SELECT 子命令可将训练样本和待判样本(或验证样本) 的个体标识出来。

(d) 标准化和典型判别函数

标准化和非标准化判别系数：非标准化系数是在变量采用原有度量单位时求得的判系数，而标准化系数(如同在回归分析中那样) 是当变量按零均值单位方差标准化后的系数。标准化的意义在于标准化矫正了由于不同测量单位所引起的偏差，从而使判别系数的大小反映出其各变量的不同重要程度。

典型判别系数：当线性判别函数按照Fisher 准则(使组间平方和与组内平方和之比最大) 求得时，它恰好等于典型相关分析中的典型相关系数，因此，它又称做典型判别系数。

(e) 多类判别函数个数

对多类判别，可建立多个判别函数，对P 个预测变量，K 类判别问题，通常有m 个函数，m 满足$m \leq \min(k-1, p)$。一般说来，在Fisher 线性判别函数中，以最大待征根对应的判别能力为最大，因此常只用它进行判别，但有时在多类判别中，问题较为复杂，也可建立多个判别函数。在DSCRIMINANT 中，采用FUNCTION 子命令决定判别函数的个数。

3. 逐步判别

逐步判别中有两个关键要素，变量选择准则和变量选择方法。

DSCRIMINANT 中提供了五种变量选择准则，可利用子命令METHOD 选择使用，下面将五个准则加以介绍。

设有k 个母体$G_1, G_2, ..., G_k$，p 个预测变量，记n_i 为第i 个母体的样本含量，$\overline{X}_{ij}$ 为第j 个母体第i 个变量的均值，$\overline{X}_i$ 为所有母体合并后第i 个变量的均值，w_{ij} 为组内协方差阵的逆阵第(i,j) 个元素，常用的变量选择选准则有以下几种：

①Rao 的V 统计量

$$V = (n-k)\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}\sum_{l=1}^{k} n_l(\overline{X}_{il} - \overline{X}_i)(\overline{X}_{jl} - \overline{X}_j)$$

也称Lawley-Hotelling 迹，V 的渐近分布为$\chi^2_{p(k-1)}$。

显然，组内均值的差值越大，V 就越大，因此，考察一个变量贡献大小可看将其加入到模型后V 的增量有多大，其增量显著性检验可基于χ^2分布进行。

②Mahalanobis 距离(马氏距离)

母体a 和母体b 之间的样本距离定义为：

$$D^2_{ab} = (n-k)\Sigma_{i=1}^{p}\Sigma_{j=1}^{p} w_{ij}(\overline{X}_{ia} - \overline{X}_{ib})(\overline{X}_{ja} - \overline{X}_{jb})$$

变量选择时首先计算所有母体之间的两两距离，其次，对具有最小距离的两个母体，选择具有最大D^2 的变量入选模型。

③组间F 统计量

基于马氏距离，可构造两母体均值相等的零假设检验，其对应的统计量为：

$$F = \frac{n-1-p}{p(n-2)}\frac{n_a n_b}{n_a + n_b}D^2_{ab}$$

F 值可用于变量选择,在每一步,选择具有最大F 值的变量入选。

④未解释变异和

由于马氏距离与回归分析中的R^2成比例，即：$R^2 = cD^2$，对于每一对母体a 和b，从回归分析角度来看未解释变异为$1 - R^2_{ab}$，其中R^2_{ab} 为复相关系数的平方。在变量的选择时，选取使“未解释变异和”最小的变量入选。

⑤Wilks 的λ统计量

选择使Wilks λ最小的变量入选模型是最为常用和读者最为熟悉的一种选择准则。

在DSCRIMINANT 中提供三种常用的变量选择的算法：前进法、后退法和逐步选择法。其原理与逐步回归分析一样，只是变量入选的准则不同。三种变量选择法可利用ANALYSIS 子命令任选其一。

下面以逐步选择法为例说明其实施步骤：(i) 在所有变量$x_1, ..., x_p$ 中挑选对选择准则而言最大可能接受值的变量入选；(ii) 当第一个变量入选后，对尚未入选变量，根据准则重新计算评判，选出下一个具有最大可能接受值的变量入选；(iii) 对已入选变量，应重新评价以考察其重要程度是否因其他入选变量的进入而发生变化，及时删除满足删除准则的变量；(iv) 如此继续，直至既不能入选新变量又不能删除已入选变量为止。此时，利用最终选入的变量构成判别函数。

在上述算法过程中，DSCRIMIANT 将在每一步显示一个当前已入选变量表，在最后一步，给出一个综览表，以说明各次入选和删除变量的过程以及统计量的显著水平。

SAS 的判别分析过程为DISCRIM，典型判别分析的过程是CANDISC。

§2.3.8 聚类分析

1. 是在一组观测个体类别归属未知的条件下，根据其观测指标在数值上的特征进行归类的统计分析技术。常见的聚类方法有：系统聚类法，动态聚类法、分解法、有序聚类法等。在进行聚类分析时，应明确：使用哪些变量？变量的距离如何确定？使用哪些准则进行类的归并？SPSS/PC+ CLUSTER 命令提供了系统聚类法，以下略加介绍。

2. 相似系数和距离

观测个体的归类是基于对个体之间关系的某种度量进行的。常用的度量有相似系数和距离，即根据观测个体之间的相似或距离远近进行归类，SPSS/PC+ CLUSTER 中采用的相似系数和距离介绍如下：

设$X' = (x_1, x_2, ..., x_m)$，和$Y' = (y_1, y_2, ..., y_m)$ 为两个具有m 个变量指标的观测个体。以下的求和与取最大的下标范围均是1.,,,m。

①SEUCLID (平方欧氏距离) $d(X,Y) = \Sigma(x_i - y_i)^2$

②EUCLID (欧氏距离) $d(X,Y) = \sqrt{\Sigma(x_i - y_i)^2}$

③COSINE (夹角余弦——相似系数) $C(X,Y) = \Sigma x_i y_i / \sqrt{\Sigma x_i^2 \Sigma y_i^2}$

④BLOCK (绝对值距离) $d(X,Y) = \Sigma|x_i - y_i|$

⑤CHEBYSHEV (切比雪夫距离) $d(X,Y) = \max|x_i - y_i|$

⑥POWER(P,R) (绝对值幂距离) $d(X,Y) = \sqrt[r]{\Sigma|x_i - y_i|^p}$

其中，当r=p 时为明考夫斯基(Minkowski) 距离，r=p=1 时为绝对值距离，r=p=2 时为欧氏距离。

3. 系统聚类法及类间距离

系统聚类法的基本思想是，首先将n个待聚观测个体看作n小类，然后规定个体之间的距离或相似系数(SPSS/PC+ CLUSTER 中选择上述六种之一) 以及规定各类之间的距离，选择距离最小的两类作为新的一类，然后重新计算所有各类之间的距离，再次选择距离最小的两类并为一类，如此继续，直至并为一类为止。这一归类过程可用一张聚类图或谱系图形象地表示出来。

因此，系统聚类法实施的另一个重要前提是定义度量各类之间距离的方法，用$G_1, G_2, ...,$表示它们分别有观测数$n_1, n_2, ...,$的类，d_{ij} 表示观测个体i与j 的距离，D_{pq} 表示G_p 与G_q 的距离，SPSS/PC+ CLUSTER 中所提供的度量类间距离方法有：

①类间平均法(BAVERAGE) $D_{pq}^2 = (1/n_p n_q)\Sigma d_{ij}^2, i \in G_p, j \in G_q$

②类内平均法(WAVERAGE) $D_{pq}^2 = [\Sigma d_{ij}^2 + \Sigma d_{kl}^2]/[C_{np}^2 + C_{nq}^2], i, j \in G_p, k, l \in G_q$

③最短距离法(SINGLE) $D_{pq} = \min d_{ij}, i \in G_p, j \in G_q$

④最长距离法(COMPLETE) $D_{pq} = \max d_{ij}, i \in G_p, j \in G_q$

⑤重心法(CENTROID) $D_{pq} = d_{\overline{x}_p \overline{x}_q}$ 其中$\overline{x}_p$ 和$\overline{x}_q$ 分别为类G_p 和G_q 的均值重心且必须用平方欧氏距离，若将G_p 和G_q 合并为新的一类G_r 时，其新的(均值)重心定义为：$x_r = (1/n_r)(n_p \overline{x}_p + n_q \overline{x}_q), n_r = n_p + n_q$

⑥中间类法(MEDIAN) $\tilde{D}_{pq} = \tilde{d}_{XX}$

应当注意的是，与重心法不同之处在于在中间类法中，当G_p 和G_q 并为一类时，其新的(中间类法) 重心定义为：$\tilde{X} = 0.5(\overline{X}_p + \overline{X}_q)$

⑦WARD 法设将 n 个观测个体分成 k 类，$G_1, G_2, ..., G_k$，用x_{it} 表示G_t 中的第i 个样品(注意x_{it} 为m 维向量)，首先计算类G_t中观测个体的离差平方和：$S_t = \Sigma_{i=1}^{n_t}(\overline{x}_{it} - x_t)'(\overline{x}_{it} - x_t)$. 当$k$固定时，要选择使$S = \Sigma_{t=1}^{k} S_t$极小的分类。

多个样本点聚类时，如果数目很大，宜采用K-means聚类。

SAS 使用CLUSTER和FASTCLUS、ACECLUS进行系统聚类和K-means聚类，在BMDP中相关的模块为1M、2M、3M和KM。

§2.3.9 分类数据分析

与连续数据分析有许多类似之处，这里主要介绍对数线性模型和logistic 回归分析，对数线性模型在第13章也有一些讨论。

1. 对数线性模型

对数线性模型与回归模型有相似之处，在对数线性模型中，所有被用于分类的变量均作为自变量，而因变量为交叉表中各点上的理论频数。

例如，对于变量X 和Y 的$r \times s$ 列联表第i行第j 列格的模型为：

$$\ln(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

其中λ_i^X 和λ_j^Y 称为主效应，λ_{ij}^{XY} 称为交互效应，m_{ij} 为格点(i, j) 的理论频数。

对数线性模型可分为分层(hierarchical) 和不分层两大类，其中分层模型更为常用，它是指如果模型中出现了某一组变量的某种交互项，则必存在这些变量所有可能组合的低阶项。

例如，在$r \times s$列联表中，若分层模型中出现λ_{ij}^{XY}，则必出现λ_i^X 和λ_j^Y，若模型不包括λ_j^Y，则必不包括λ_{ij}^{XY}。

SPSS 的HILOGLINEAR 用于处理分层对数线性模型，以下对它作一介绍。

分层对数线性模型有两个重要特例：饱和模型指包含了所有变量的主效应和这些变量所有可能的交互效应项的对数线性模型；独立变量模型指不包含变量的交互效应

项的对数线性模型称为独立变量模型。例如上例中不包含λ_{ij}^{XY}时则为独立变量模型。显然，饱和模型和独立变量模型是一般分层对数线性模型的两个极端情形。

(a) 模型参数及估计

对数线性模型参数有一个重要性质：在其余变量不变的条件下，任意一变量的所有效应之和为零。如上述$r \times s$列联表中：

$$\Sigma_{i=1}^{r}\lambda_i^X = 0, \Sigma_{i=1}^{s}\lambda_i^Y = 0, \Sigma_{i=1}^{r}\lambda_{ij}^{XY} = 0, \Sigma_{i=1}^{s}\lambda_{ij}^{XY} = 0$$

$i = 1, ..., r, j = 1, ..., s$

上述性质在对数线性模型的参数估计时非常有用。在HILOGLINEAR 中，只提供与参数自由度相等的参数估计，其余参数依上述性质自行推导。

如上所述，对于分层对数线性模型，指定了最高阶交互效应项，实际上就指定了整个模型的结构。在HILOGLINEAR 中，利用DESIGN了命令指定饱和模型生成类中的某模型的最高阶交互项，以指定该模型的结构。

HILOGLINEAR 在给出参数的估计值时，是按照上述最高阶交互项中最左边的变量顺序每取一值，最右边变量依次取其与自由度相应的所有可能值的方式给出参数估计。

例如，在变量为X和Y的3×3表中，若指定最高阶交互项为XY，则顺序给出下列参数的估计。

$\lambda_{11}^{XY}, \lambda_{12}^{XY}, \lambda_{21}^{XY}, \lambda_{22}^{XY}, \lambda_1^X, \lambda_2^X, \lambda_1^Y, \lambda_2^Y$

其余参数均可由约束方程导出。

HILOGLINEAR 采用迭代比例拟合算法为多维列联表拟合分层对数线性模型。对迭代的控制可通过HILOGLINEAR中的CRITERIA 子命令设置估计精度和最大迭步数实现。

(b) 检验

常见的关于拟合的假设检验有以下两种：χ^2 拟合优度检验

$$\chi^2 = \sum\sum \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \sim \chi^2_{(r-1)(s-1)}$$

似然比χ^2 检验

$$2\sum\sum n_{ij} \ln(\frac{n_{ij}}{m_{ij}})$$

式中n_{ij} 是实际观测频数，$\hat{m}_{ij}$ 为其相应的拟合值，下标i, j的求和是对表中所有格点而言，在大样本情况下这两种统计量等价。

(c) 残差与诊断

模型拟合好坏还可通过考察基于模型产生的期望格点与观测格点的差值—即残差进行，易见，当模型拟合较好时，残差应较小。与回归分析类似，考察标准化残差—它由残差除以其标准差构成—要比直接考察残差更为合理。

标准化残差=[观测格点数- 期望格点数]/$\sqrt{\text{期望格点数}}$

若模型是合适的，则标准化残差将有渐近标准正态分布。因此，当标准化残差绝对值大于1.96，表明拟合很不好。残差的诊断类似于回归分析中的情形。(1) 观察“标准化残差—格点观察频数”图和“标准化残差——格点期望频数”图；在HILOGLINEAR 中，通过PLOT 子命令中的RESID 作上述图形显示。(2) 观察正态概率图，该图是以正态概率为纵柚，以标准化残差为横轴构成的图。显然，当模型拟合合适时，图形应为一条直线。在HILOGLINEAR 中，通过PLOT 子命令中的NORMPROB 产生上述图形显示。关于残差诊断的进一步讨论，可参见“回归残差分析”一节。

(d) 模型选择

⑴模型分块选择法

类似于回归分析中利用复相关系数的改变量来刻划新增变量的贡献，在分层对数线性模型中通常利用似然比χ^2统计量检验一个模型在增加某新的效应项时其对应的贡献大小，从而决定模型的选择。

在HILOGLINEAR 中将自变动给出两类假设检验：关于所有k阶及其更高阶效应为零的假设和关于k阶效应为零的假设。

⑵后退删除法

由于在对数线性模型中后退法比前进法对模型选择更合适，所以HILOGLINEAR只提供了后退法。其初始模型可以是任何分层对数线性模型，程序将计算所有最高阶交互项卡方值的观测显著性水平，对于其观测值水平大于保留准则值的效应项，若其删除导致似然比卡方最小显著性改变，则可给予删除。注意在这一步中，为确保是分层模型，只考察对应于生成类的效应项。类似地，在删除某个效应项基础上，再次比较观测显著性水平和似然比卡方以决定其它的效应项是否删除。后退删除法通过METHOD 子命令中的BACKWARD 选择项指定。

2. LOGISTIC 回归

是一种用于结果为二分类数据的多元分析方法。流行病学分析致病因素与结果的关系，有两种方法经常使用，一种称定群或队列研究，在调查开始时，将人群分成两组，分别暴露于某种因素，另一组用做对照，此后随访观察他们在一个时期内的结局，称为前瞻性定群研究；否则若用现有对象，追溯所研究暴露因素的影响，就是历史性定群研究。第二种是调查病人和非病人暴露于于某危险因素的方法，称做病例对照研究，若病例或样本均是人群中的随机样本，则病例与对照使用的样本数目不一定相等，称为成组病例对照研究；否则对每一病例按特定条件如年龄、性别、住址等找出对照，就是配对病例—对照研究，对照可以是一个或多，称做1: 1 配对和1:M 配对。LOGISTIC 回归是分析这一类数据的有力方法，SPSS/PC+ 用LOGISTIC REGRESSION 命令，SAS 和SYSTAT 也都可以做成组或等级分组的LOGISTIC 回归分析，SAS 还可进行条件LOGISTIC 回归。现从一般多元回归分析出发，模型对二分类数据的因变量Y_i 预计取值应是一个$0 \sim 1$ 的概率P_i，通常的回归效果必然不佳，现做LOGIT 变换$\ln(\frac{P}{1-P})$，有：

$$\ln(\frac{P}{1-P}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

或

$$P_i = \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_i x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_i x_{ij})}, \quad i = 1, ..., n$$

即成组的LOGISTIC 模型。其中$(k+1)$ 维向量$(\beta_0,\beta_1,...,\beta_k)$ 是待估计参数，模型对于$\beta'X$ 是单调的。现设某个二分类的反应变量Y 在事件发生时取值为$'A'$，未发生时取值为$'B'$，又设有一个做为回归变量的危险因素X在出现时取值为1，不出现时取值为0，根据LOGISTIC 模型

$$P(Y ='\ A'|X=1)=\frac{\exp(\alpha+\beta)}{1+\exp(\alpha+\beta)}$$

$$P(Y ='\ B'|X=0)=\frac{\exp(\alpha)}{1+\exp(\alpha)}$$

其中α是截距，β是回归系数，对于具有危险因子的个体来说，事件的比数(odds) 定义为$\frac{P(Y='A'|X=1)}{P(Y='B'|X=1)}=\frac{P(Y='A'|X=1)}{1-P(Y='A'|X=1)}=\exp(\alpha+\beta)$；类似地，对于没有危险因素的个体比数为$\exp(\alpha)$，比数比(odds ratio)就是这两个比数的比，即$I=\frac{\exp(\alpha+\beta)}{\exp(\alpha)}=\exp(\beta)$。由此可见回归系数$\beta$表示了因素由0 变到1 的对数比数的变化。当因素的编码是a 与b 时$I=\exp((b-a)\beta),\beta$则是因素变化一个单位时对数比数的变化。在SAS PROC LOGISTIC 中，因素的比数比由$\hat{I}=\exp(\hat{\beta})$ 给出，可信限由$\exp(\hat{\beta}\pm z_{\alpha/2}s(\hat{\beta}))$ 算出。

LOGISTIC 似然函数正是Bernoulli 变量的似然函数，由下式表达：

$$\prod_i P_i^{Y_i}(1-P_i)^{1-Y_i}=\exp(\sum_i Y_i(\beta_0+\sum_{j=1}^k\beta_i x_{ij}))\prod_i(1+\exp(\beta_0+\sum_{j=1}^k\beta_i x_{ij}))^{-1}$$

上式表明，LOGISTIC 模型是$(k+1)$ 个参数的指数簇分布，具有充分统计量$s_0=\sum y_i,s_1=\sum y_i x_{i1},...,s_k=\sum y_i x_{ik}$。对数似然函数关于$\beta_j$的一阶导数为$V_j(\beta)=s_j-\sum x_{ij}\frac{\exp(\beta_0+\sum_{j=1}^k\beta_i x_{ij})}{1+\exp(\beta_0+\sum_{j=1}^k\beta_i x_{ij})}$, 即$s_j$ 的观察值减去期望值；β 的协方差阵由信息矩阵的逆给出，信息矩阵是：

$$M_{jl}(\beta)=\sum x_{ij}\frac{\exp(\beta_0+\sum_{j=1}^k\beta_i x_{ij})}{(1+\exp(\beta_0+\sum_{j=1}^k\beta_i x_{ij})^2},j=0,1,...,k,l=0,1,...,k$$

从$\beta_0=\ln(\frac{s_0}{n-s_0}),\beta_1...,\beta_k=0$ 开始，用$\beta+[M(\beta)]^{-1}V(\beta)$ 不断修正至收敛。

针对回归系数的检验，根据方程估计中的正态近似，常用正态 z 检验(Wald's test)，计分检验(score test) 以及似然比检验(likelihood ratio test)，有关说明可详参Rao, C.R. (1973)。

计分检验统计量由对数似然函的一阶与二阶偏导计算，正态性检验是利用参数估计值与其近似标准误的比值，Wald's 检验则是该量的平方。

设有模型$logit\quad p(Y=1|X)=\beta_0$ 及$logit\quad p(Y=1|X)=\beta_0+\beta_1X_1$，它们的对数似然函数分别为$L(\hat{\beta}_0)$ 和$L(\hat{\beta})$，针对假设$\hat{\beta}_1=0$ 计算$-2(L(\hat{\beta}_0)-L(\beta))$，该统计量具有近似自由度为1 的$\chi^2$分布。

在SAS PROC LOGISTIC 中引入了Hosmer-Lemeshow 统计量，其原理如下：在求得回归系数后，代入原公式得到每个反应的概率，把这些概率从小到大进行排列，然后按如下规则分成大约十组：记N 为观察个体总数，M 是每个亚组的个体数$M=[0.1N+.5],[x]$是x 的整数部分，若原始数据是未分组的，则各组都有不同M 取值的个体；若资料是分组的，则观察个体按观察的组界进行分组，对应第一个观察的个体分到第一组，设第一个观察有n_1 个体，第二个观察有n_2个体，当$n_1<M$ 及$n_1+[0.5n_2]\le M$ 时第二个观察也

放到第一组，一般来说若$(j-1)$ 个观察已放到第k 组，而第k 组有c 个个体，第j 个观察的个体在$c \le M, c+[0.5n_j] \le M$ 时被放到第k 组，否则放到下一组，另外若最后一组的个体数目不超过$[0.5 \times N]$，则把最后两组合并。如此分得g 组，通过观察与计算频数的$2 \times g$表，计算Hosmer-Lemeshow 拟合优度检验，统计量为：

$$\chi^2_{hw} = \sum_{i=1}^{g} \frac{(O_i - N_i\overline{\pi}_i)^2}{N_i\overline{\pi}_i(1-\overline{\pi}_i)}$$

其中N_i 是第i 组中的个体数，O_i 是第i 组中的事件数，$\overline{\pi}_i$ 是第i 组的事件结果的平均估计概率，统计量与$\chi^2(g-2)$分布进行比较。

SAS 对于模型中包含分类变量的情况，可输出多个系数，克服了把分类变量作为连续型的不合理性。如种族变量race 有“1.黑人、2.白人、3.西班牙人、4.其他”样的分类，做为连续量处理则不合理，而用哑变量(dummy variable) 对变量进行编码，不同的软件编码方式不同，如：

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}$$

第一种使用参照组编码(reference cell coding)，第二种是与总均值的编码(deviation from means coding)，第三种为正交编码。在伪变量回归中，常数项表示了对照的均值或者是总观察的均值，或是组均值未加权的值，SAS 把最后一个分类作参照组，而GLIM则取第一组作参照，Hosmer, D.W.和S. Lemeshow (1989) 对此进行了说明。

SAS 分类数据分析的主要过程是FREQ和CATMOD。

3. 其它方法

对应分析(corresponding analysis)是用图形方式表达两维列联表关系的方法。行和列用图上的点来表示，点的位置表示了关系，点的坐标也是典型相关模型分数的一种方式。优点：首先，多个分类变量可以经过列联表数据简明地表达，这一手段使得研究者利用现有数据或收集一般名义的或不容易测量的数据进行分析。因为因子分析只能针对所有分析变量均为区间类型时的情形。其次，因为对应分析不仅勾画了行与列的关系，也表达了它们自身分类间的关系，比如：列的许多特征比较接近，则它们的轮廓比较相象，这样就构成了一组特征，因而与主成分分析中的因素相当。最后，也是最重要的，对应分析给行与列分类提供维数相当的表达。缺点或局限性：首先，这一方法是描述性的因而不适于假设检验。若要定量描述分类间的关系，应当使用对数线性模型等方法。对应分析最适于探索性分析。其次，同其它降维方法一样，并不存在确定一个合适维数的方法，研究者应在可以解释与数据表达的简明之间进行权衡。最后，对应分析对于异常值仍比较敏感。conjoint 分析特别用于分析应答者如何对某种想法、产品或服务的趋向性。其基本假设是消费者是在综合了这些特征之后进行评价，在这一意义上它与析因设计相当。在了解了应答者的趋向性之后，我们可以对这一倾向性进行剖析，估价单个特征的贡献多大。因此，它有别于判别分析和回归分析，因为后者只是根据多个特征与总的倾向性进行关联分析。这一区别可以称作“组合——分解(decomposional /compositional)”的不同。

§2.3.10 生存分析

1. 在医学和可靠性分析中，某个对象的观察常常与时间有关，如观察某种肿瘤由诊断到死亡的病程，某种产品从使用到失效的时间。这些数据常用生存分析来处理。不同于以往前的统计分析，生存分析研究取值大于零，并且有截尾的随机变量。

   现用T 表示失效时间，记其分布函数为$F(t) = P(T \leq t)$，表示了到时刻t 失效的概率；生存函数为$S(t) = P(T > t) = 1 - F(t) = exp(-H(t))$，表示t 时刻仍然存活的概率，其中H(t)=$\int_0^t h(u)du$称为累积风险函数。如果变量$T$的密度函数$f(t)$存在，那么风险函数$h(t) = f(t)/S(t) = f(t)/[1 - F(t)]$,它表示在时刻$t$活着的条件下，时刻$t$后失效的概率。

2. Kaplan-Meier 模型是一种非参方法，设有k个观察时间$t_1 < t_2 < ... < t_k$, 每个观察时间t_i有n_i个观察对象，其中有d_i个失效，$i = 1, 2, ..., k$, 生存函数的估计为

   $$\hat{S(t)} = \Pi_{t_j<t}\frac{n_j - d_j}{n_j}, \quad S.E.[\hat{S(t)}] = \hat{S(t)}\sqrt{\Sigma_{t_j<t}\frac{d_j}{n_j(n_j - d_j)}}$$

   其误差估计公式又称做Greenwood 公式。

3. log-rank检验

   现要对两组处理进行比较，两组分别有M 和N 个观察，看其生存情况是否相同，假设在$M + N$ 个对象的失效时间是不重复时，可依据每个时间区间把两组的情况列于表2.14：

表 2.14 生存率比较计算表

处理组	t_j 时的失效	t_{j-0}时处于风险的数目
第一组	m_j	M_j
第二组	n_j	N_j

在边缘值给定下n_j 的条件分布是超几何分布，即在有限总体$M_j + N_j$ 中有$m_j + n_j$个有某种特征，然后从N_j 的样本中观察到n_j 个具有特征。因此n_j的条件期望与方差是：

$$E_j = \frac{N_j(m_j + n_j)}{M_j + N_j}$$

$$V_j = \frac{M_J N_j(m_j + n_j)(M_j + N_j - m_j - n_j)}{(M_j + N_j - 1)(M_j + N_j)/2}$$

记$w_1, w_2, ..., w_J$ 是一些已知的常数，假设$n_1, n_2, ..., n_J$ 相互独立且具有正态分布，其矩由上式给出，则$\Sigma w_j(n_j - E_j)$ 也具有正态分布。进一步有$Q(w) = (\Sigma_j w_j(n_j - E_j))^2/\Sigma_j V_j w_j^2$服从自由度为1 的卡方分布。

当$w_j = 1$ 时，就是比例风验检验：$Q_{PH} = (O - E)^2/V, O = \Sigma n_j$ 是第二处理组有观察失效数目，E 是相应的期望值，这一检验拥有众多的名字：log- rank、Mantel-Heanszel、Generalized Savage 以及exponential order scores test。

当$w_j = M_j + N_j$时，就成为generalized Wilcoxon 检验。这一方法可以推广到多组。另外，对于风险函数的检验也是很重要的。第6章介绍了对Gehan关于白血病数据分析的结果。

表 2.15 比较多组生存情况的计算表

处理组	t_j 失效数	t_{j-0} 的失效数
0	n_{0j}	N_{0j}
1	n_{1j}	N_{1j}
.	. .	
k	n_{kj}	N_{kj}
总计	$n_{.j}$	$N_{.j}$

当组数多于两个，有类似的公式。现把各处理组的失效情况列如表 2.15 ：

$$n_j = (n_{1j}, n_{2j}, ..., n_{kj})', E_j = (E_{1j}, E_{2j}, ..., E_{kj})'$$

$$E_{ij} = N_{ij} n_{.j} / N_{.j}$$

$$V_{il} = [n_{.j} N_{ij} (N_{.j} - n_{.j})(N_{.j}\delta_{il} - N_{lj})]/[(N_{.j} - 1)N_{.j}^2]$$

其中δ_{il} 是Kronecker 记号。第k组的广义比例风险统计量

$$Q_{PH} = (O - E)' V^{-1} (O - E)$$

在H：$h_{0(t)} = h_{1(t)} = ...h_{k(t)}$ 的假设下，Q_{PH}近似服从自由度为k 的χ^2分布。

4. 参数模型最简单的是指数分布，$f(t) = \lambda \exp(-\lambda_t), h(t) = \lambda, S(t) = \exp(-\lambda_t), t > 0$。常用的参数模型列于表2.16：

表 2.16 几种生存分布的风险函数与生存函数

分　　布	h(t)	S(t)
指数(exponential)	λ	$\exp(-\lambda t)$
威布尔(Weibull)	λt^{γ}	$\exp(-\lambda/(\gamma+1)t^{\gamma+1})$
Gompertz	$\lambda \exp(\gamma t)$	$\exp[-(\lambda/\gamma)(\exp(\gamma t) - 1)]$
伽马(Gamma)	单调或定常	
对数正态(log normal)	增至最大，然后下降	

各模型的参数的估计通常采用极大似然方法。设有n个样本在某时刻观察到d个失效，记$t_i, i = 1, ..., n$为观察到的相应的失效或截尾时间，假定$t_1, ..., t_d$为失效时间，则似然函数为：

$$L = \Pi_{i=1}^{d} f(t_i) \Pi_{i=d+1}^{n} S(t_i)$$

标准误由参数的Fisher 信息矩阵而得。

对指数分布，似然函数为

$$L(\lambda) = \lambda^d \exp(\lambda \sum t_i)$$

极大似然估计为$\lambda = d/\sum t_i$，利用正态近似，有相应的检验统计量：

$$(\ln\hat{\lambda} - \ln\lambda)/\sqrt{(1/d)}$$

两个指数分布相等的检验统计量是：

$$(\ln\hat{\lambda}_1 - \ln\hat{\lambda}_2)/\sqrt{1/d_1 + 1/d_2}$$

均为正态性z检验。

生存分析的寿命表法在SAS PHREG、BMDP1L 和SPSS/PC+ 实现。

5. Cox 回归模型

Cox 模型是一种半参数或称比例风险模型(proportional hazard model, PH model)，其风险函数为$h(t;x) = \lambda(t)\exp(\beta' x), \lambda(t)$ 是一个任意的函数，称为基线风险函数。X称为协变量，β是未知参数；又因为$h(t;x_1)/h(t;x_2) = \exp[\beta'(x_1 - x_2)]$ 对所有时刻t 成立，所以是比例风险。设观察到k个失效时间$t_1 < t_2 ... < t_k$, 令x_i为其相应的协变量值，Cox偏似然函数为

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta x_i)}{\sum_{j\in R(t_i)}\exp(\beta x_j)}$$

其中$R(t_i)$是在先于时刻t_i还没有失效的样本集。参数β的统计推断基于Cox偏似然函数。

当协变量是时间的函数，则模型称时间协变量的Cox 模型。当失效出现重复时，问题要复杂，常用的处理方法如Breslow 方法、离散logistic 方法、Efron 方法等，这些方法可由SAS PHREG 和SPSS SURVIVAL 实现。

SAS生存分析的过程为LIFEREG、LIFETEST和PHREG。SPSS 过程为SURVIVAL、KM和COXREG，分别进行寿命表、Kaplan-Meier和Cox生存分析。BMDP2L可以进行固定协变量和时变协变量的Cox回归分析。