**36. 判别分析**

**（一）基本原理**

判别分析，是用以判别个体所属类的一种统计方法。其原理是根据已掌握的一批分类明确的样品，建立一个较好的判别函数，使得用该判别函数进行判别时错判事例最少，进而能用此判别函数对给定的一个新样品判别它来自哪个总体。

判别分析方法通常要给出一个判别指标（判别函数），同时还要指定一种判别规则。

**一、距离判别法**

未知总体的样品*x*离哪个总体的距离最近，就判断它属于哪个总体。

1. 对于两个正态总体G1, G2

距离选用马氏（Mahalanobis）距离：

d2(*x*, G1) = (x-μ1)T∑1-1(x-μ1)

d2(*x*, G2) = (x-μ2)T∑2-1(x-μ2)

其中，μ1, μ2, ∑1, ∑2分别为总体G1, G22的均值和协差矩阵。令

W(x) = d2(*x*, G1) - d2(*x*, G2)

称为判别函数，若∑1=∑2时，W(x)是线性函数，此时称为线性判别；若∑1≠∑2，W(x)是二次函数。

2. 多总体情况

设有*m*个总体：G1, …, Gm，其均值、协差阵分别为μi, ∑i. 对给定的样品*x*，按距离最近的准则对*x*进行判别归类：

首先计算样品*x*到*m*个总体的马氏距离di2(*x*), 然后进行比较，把*x*判归距离最小的那个总体，即若dh2(*x*) = min{ di2(*x*) | i = 1,…,m}，则*x*∈Gh.

**二、Fisher线性函数判别法**

为了方便使用，需要寻找尽量简单的判别函数，其中在Fisher准则下的线性判别函数就是只利用总体的一、二阶矩就可求得的判别函数。

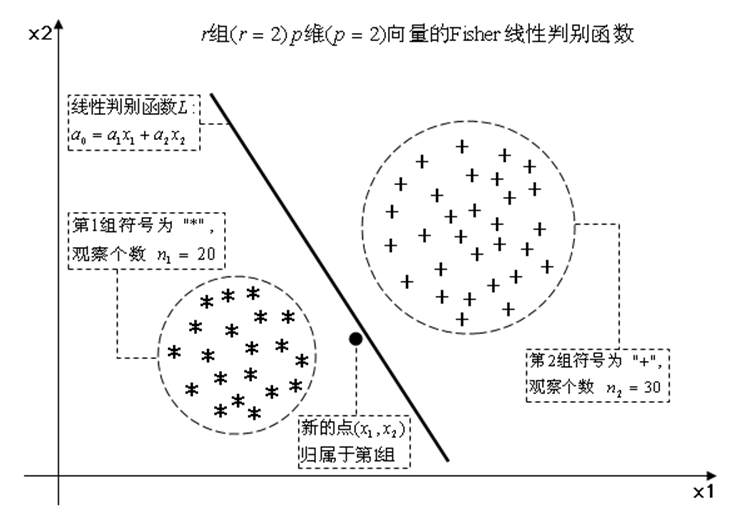


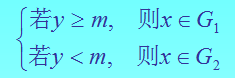
图1 Fisher线性判别分析示意图

下面以两个总体为例来说明Fisher判别的思想。

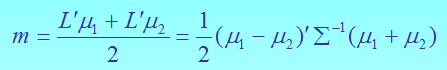
设有两个总体*G*1、*G*2，其均值分别为μ1和μ2，协方差阵分别Σ1和Σ2，并假定Σ1 = Σ2 = Σ，考虑线性组合：*y* = *LTx*。通过寻求合适的*L*向量，使得来自两个总体的数据间的距离较大，而来自同一个总体数据间的差异较小。为此，可以证明，当选*L*=*c*Σ–1(*μ*1–*μ*2)，其中*c* ≠ 0时，所得的投影即满足要求。从而称*c* = 1时的线性函数：

*y* = *LTx* = (*μ*1 – *μ*2)TΣ–1*x*

为Fisher线性判别函数。其判别规则为：



其中，*m*为两个总体均值在投影方向上的中点，即

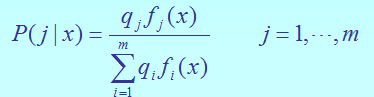


**三、贝叶斯判别法**

设*m*个总体*G*1, …, *Gm*，其分布密度分别为*f*1(*x*), …, *fm*(*x*)，各自的先验概率（可以根据经验事先给出或估出）分别为*q*1,…, *qm*，显然



当抽取了一个未知总体的样品*x*，要判断它属于哪个总体，可用著名的贝叶斯公式计算*x*属于第*j*个总体的后验概率：

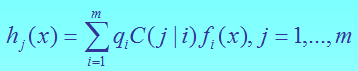


若



则判断*x*属于第*h*个总体。

或者计算按先验概率加权的误判平均损失：



其中，*C*(*j* | *i*)为假定本来属于*Gi*的样品被判为属于*Gj*时造成的损失，当然*C*(*i* | *i*) = 0，*C*(*j* | *i*)≥ 0.

再比较这*m*个误判平均损失的*h*1(*x*), …, *hm*(*x*)的大小，选取其中最小的，就可以判定样品*x*来自该总体。

在实际问题中，错判的损失可以给出定性分析，但很难用数值来表示，但应用贝叶斯判别法时，要求定量给出*C*(*j* | *i*)，*C*(*j* | *i*)的赋值。通常：根据经验人为赋值；假定各种错判的损失都相等。

**错判概率**

当样品*x*∈*Gi*，用判别法判别时，把*x*判给*Gj* (*i*≠*j*)，出现错判。用*P*(*j*|*i*)）表示实属*Gi*的样品错判为*Gj*的概率。

**广义平方距离判别法**

在正态总体的假定下，按贝叶斯判别的思想，在错判造成的损失认为相等的情况下得到的判别函数，其实就是马氏距离判别法在考虑先验概率及协方差阵是否相等情况下的推广，故在SAS的DISCRIM过程中称为广义平方距离判别法。

**四、逐步判别法**

所有变量中，有的变量对区分*k*个总体的判别能力可能很强，有的可能很弱。如果不加区别地用全部变量来建立判别函数，则必增加大量的计算，还可能因为变量间的相关性引起计算上的困难(病态或退化等)及计算精度的降低。另一方面由于一些对区分*k*个总体的判别能力很小的变量的引入，产生干扰，致使建立的判别函数不稳定，反而影响判别效果，因此自然提出一个变量的选择问题。即如何从*m*个变量中挑选出对区分*k*个总体有显著判别能力的变量，来建立判别函数，用以判别归类。

**1. 各变量判别能力的检验**

筛选判别变量和做逐步判别，都需要检验各变量的判别能力。若第*i*个分量间没有显著差异时，说明相应的变量*Xi*对判别分类不起作用，应该剔除。

变量判别能力的度量通常采用删去该变量后考察判别能力的变化，即考察该变量对区分*k*个类是否能提供更多的附加信息，然后由附加信息构造F统计量进行检验。

利用F统计量对假设（H0：第*i*个变量在*k*个总体中的均值相等)作统计检验。若否定H0，表示变量*Xi*对区分*k*个总体的判别能力是显著的(在显著水平α下)。否则，变量*Xi*对区分*k*个总体的判别能力不能提供附加信息，这个变量应剔除。

**2.** **基本思想**

逐个引入变量，每次把一个判别能力最强的变量引入判别式，每引入一个新变量，对判别式中的老变量逐个进行检验，如其判别能力因新变量的引入而变得不显著，应把它从判别式中剔除。这种通过逐步筛选变量使得建立的判别函数中仅保留判别能力显著的变量的方法，就是逐步判别法。

**3. 基本步骤**

（1）逐步筛选变量：根据各变量对区分k个总体的判别能力的大小来筛选变量。SAS中的STEPDISC过程专用于筛选变量子集。该过程利用向前选入、向后剔除或逐步筛选的方法来选择区分k个总体的最佳变量子集。

（2）判别归类：对已选出变量子集，使用前文的某种判别方法对样品进行判别归类。

**（二）判别分析的SAS实现**

**一、PROC STEPDISC过程步（逐步判别分析）**

针对具有一个分类变量和若干数值型变量的数据集，STEPDISC过程执行逐步判别分析：从指定的指标变量（VAR变量）中筛选出一组变量，以用于随后的判别分析。逐步判别分析要求指标变量在各组内服从多元正态分布，并且具有相同的协方差矩阵。

基本语法：

[PROC STEPDISC](http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/statug_stepdisc_sect004.htm) data = 数据集 <可选项> ;

[CLASS](http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/statug_stepdisc_sect006.htm) 变量;

[BY](http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/statug_stepdisc_sect005.htm) 变量;

[VAR](http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/statug_stepdisc_sect008.htm) 变量列表;

[FREQ](http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/statug_stepdisc_sect007.htm) 变量;

[WEIGHT](http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/statug_stepdisc_sect009.htm) 变量;

可选项：

MAXSTEP=n——指定最大步数；

METHOD=BACKWARD/FORWARD/STEPWISE——指定筛选方法：向前选入、向后剔除、逐步筛选；

SHORT——只输出汇总结果（压缩每一步的输出）；

SLE/SLENTRY=p——指定向“向前选入法”的显著水平（默认p=0.15）；

SLS/SLSTAY =p——指定向“后退剔除法”的显著水平（默认p=0.15）；

CLASS语句和BY语句——必不可少，指定分组变量（区别是BY语句需要事先对BY变量排序）；

**二、PROC DISCRIM过程步（判别分析）**

DISCRIM过程根据一个分类变量和若干数值变量的数据计算出各种判别函数(判别准则)，根据这个判别函数，再将该批数据或其他数据中的观测分别归入已知类别中去。

DISCRIM过程用以获得判别准则的数据称之为训练数据集。

若假设每个组内样本的分布为多元正态分布，基于多元正态分布理论的参数法将导出一个线性或二次的判别函数；否则，将采用不基于任何分布的假定的非参数方法（核密度法和K最邻近法），来估计类别的密度从而实现分类的功能。

基本语法：

PROC DISCRIM data=训练数据集 testdata=数据集 <可选项> ;

CLASS变量;

BY 变量;

FREQ 变量;

ID 变量;

PRIORS 概率表;

TESTCLASS 变量;

TESTFREQ 变量;

TESTID 变量;

VAR 变量列表;

WEIGHT变量;

可选项：

 outstat=数据集名——生成数据集，包含各种统计量（均值，标准偏差和相关）；

out/outcross=数据集名——生成(/交叉)数据集，包括来自训练数据集的所有数据，后验概率和每个观测通过重替换被分入的类（当canonical选项指定时，该数据集还包含典型变量得分的新变量）；

outd=数据集名——生成一个包含来自训练数据集的所有数据和每一观测的组密度估计的数据集；

 注：前面命令若加test前缀，则输出testdata数据集生成的相关数据集。

 method＝normal/npar——指定生成分类准则的方法，默认为normal表示基于类内服从多元正态分布，并生成线性或二次判别函数； npar表示采用非参数方法；

 pool= no/ test /yes——确定平方距离的度量，是以合并协方差阵还是组内协方差阵为基础，默认pool= yes采用合并协方差阵得出线性判别函数；当pool＝no时，采用单个组内协方差阵得出二次判别函数；当method= normal时，pool=test要求对组内协方差阵的齐性的似然比检验进行Bartlett修正，当不加可选项short时，线性判别函数会直接给出，而二次型判别函数需通过建立输出数据集方式获得；

 slpool=p——指定齐性检验的显著水平（默认p=0.1，只用于pool=test时）；

 threshold=p——指定分类中可以接受的最小后验概率p值，默认p=0，若某观察样品归于某组的最大后验概率值小于该p值，则将该样品归入其它组；

 anova和manova选项——分别要求输出对各类的单个变量与多个变量的均数、均值向量之间进行一元或多元方差分析的结果，其作用就是检验判别函数的判别效果；

 listerr和crosslisterr选项——listerr表示输出由后验概率产生错误分类的样品点的有关信息，crosslisterr表示要求以交叉表的形式输出实际类别与分类结果之间一致和不一致的有关信息；

 var语句——指定用于进行判别分析的变量子集，建立起关于此变量子集的判别函数表达式；

 priors语句——指定先验概率：priors equal（默认），表示各类先验概率相等；priors proportional, 表示各类先验概率等于各类样本频率；priors a=p1 b=p2 c=p3；其中a、b和c是分类标志，p1、p2和p3是先验概率，p1+p2+p3=1.

**三、PROC CANDISC过程步（典型判别分析）**

该过程步计算平方马氏距离并做单变量与多变量的单向方差分析并且计算类均值间基于合并类内协方差阵的平方距离。该过程产生包括典型系数和典型变量得分的输出数据集。

典型判别分析类似于主成份分析和典型相关有关的降维方法。给定若干组带有几个定量变量的观察，典型判别分析得出与组有最大可能多重相关的变量的线性组合。

最大的多重相关叫做第一典型相关。线性组合的系数称为典型系数或典型权重。线性组合定义的变量称为第一典型变量或典型成份。第二典型相关由与第一典型变量无关的线性组合得到。以此类推。

CANDISC过程得出的典型变量，如同主成分概括全变差一样来概括类间变差。基本步骤：

 （1）变换变量使合并的类内协方差阵为单位阵；

 （2）计算变换后的变量的类均值；

 （3）对均值做主成份分析，以每一类中的观察个数为权重；

 （4）特征值等于每个主成份方向上的类间偏差与类内偏差之比；

 （5）把主成份变量反变换到原始变量的空间，获得典型变量。

典型变量间不相关，但典型系数并不正交。因此，典型变量并不代表原始变量空间中的正交方向。对每一个典型相关，CANDISC检验总体中该相关及更小的典型相关为零的假设。采用F近似值比一般的近似值能给出更好的小样本结果。每一类内变量应该具有近似的多元正态分布，为了概率水平有效，方差矩阵应该是共同的。第一典型相关大于等于组与任何一个原始变量间的多重相关。

该过程产生一个包含每一典型变量得分的输出数据集。可以利用print过程列出这些值，还可行以用plot过程作出典型变量对的散点图以助于直观地解释组的不同。另一个输出数据集包含由factor过程利用旋转算法得到的典型系数。

基本语法：

PROC CANDISC data=数据集 <可选项>；

CLASS 变量；

BY变量；

FREQ变量；

VAR变量列表；

WEIGHT 变量；

可选项：

out=数据集名——生成包含原始数据和典型变量得分的数据集；

 outstat=数据集名——生成一个包含各种统计量的数据集；

ncan＝*n*——指定将被计算的典型变量的个数（≤变量个数）；

 prefix＝前缀名——为命名典型变量指定前缀；

singular =p——指定判别全样本相关阵和合并类内协方差阵奇异的标准（０<p <１, 默认p =10-8）；

bcorr/bcov/bsscp——输出类间相关矩阵、协方差矩阵、SSCP矩阵；

 pcorr/pcov/psscp——合并类内相关矩阵（基于合并类内协方差的偏相关）、协方差矩阵、修正SSCP矩阵；

 tcorr/tcov/tsscp——全样本相关矩阵、协方差矩阵、修正SSCP矩阵；

 wcorr/wcov/wsscp——每一类水平的类内相关矩阵、协方差矩阵、SSCP矩阵；

anova——检验总体中每一变量类均值相等的假设的单变量统计量；

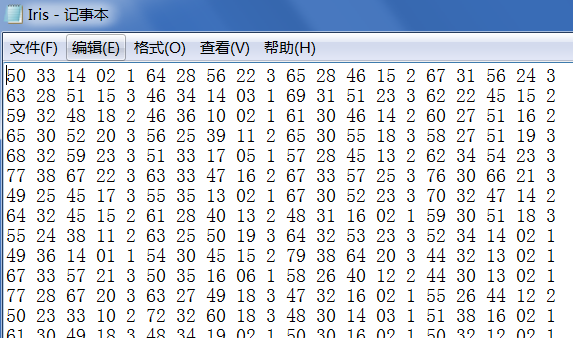
 distance——类均值间的平方马氏（Mahalanobis）距离；

 simple——全样本和类内的简单描述性统计量；

 stdmean——全样本和合并的类内标准化类均值；

short——只打印典型相关表和多元检验统计数字。

**例1** Fisher于1936年发表的鸢尾花（Iris）数据被广泛地作为判别分析的例子。数据（C:\MyRawData\Iris.txt）是对3种鸢尾花：刚毛鸢尾花（setos）、变色鸢尾花（versicolor）和佛吉尼亚鸢尾花（virginica）各抽取一个容量为50的样本，测量其花萼长（sepallen）x1、花萼宽（sepalwid）x2 、花瓣长（petallen）x3、花瓣宽（petalwid）x4，单位为mm，分组标记为S. 部分数据如下：

****

以这个容量为150的样本作为训练数据集，建立判别函数。

**代码：**

**proc** **format**;

value specname **1** = 'Setosa'

**2** = 'Versicolor'

**3** = 'Virginica';

**data** iris;

infile 'C:\MyRawData\Iris.txt';

input sepallen sepalwid petallen petalwid species @@;

format species specname.;

label sepallen = 'Sepal Length in mm.'

sepalwid = 'Sepal Width in mm.'

petallen = 'Petal Length in mm.'

petalwid = 'Petal Width in mm.';

**run**;

**proc** **print** data = iris;

title 'Discriminant Analysis of Fisher (1936) Iris Data';

**run**;

**proc** **stepdisc** data = iris short sle=**0.3** sls=**0.05**;

class species;

var sepallen sepalwid petallen petalwid;

**run**;

**proc** **discrim** data = iris method = normal pool = test anova short crosslisterr;

class species;

var petallen;

**run**;

**proc** **discrim** data = iris outstat = plotiris method = normal pool = test manova listerr crosslisterr;

class species;

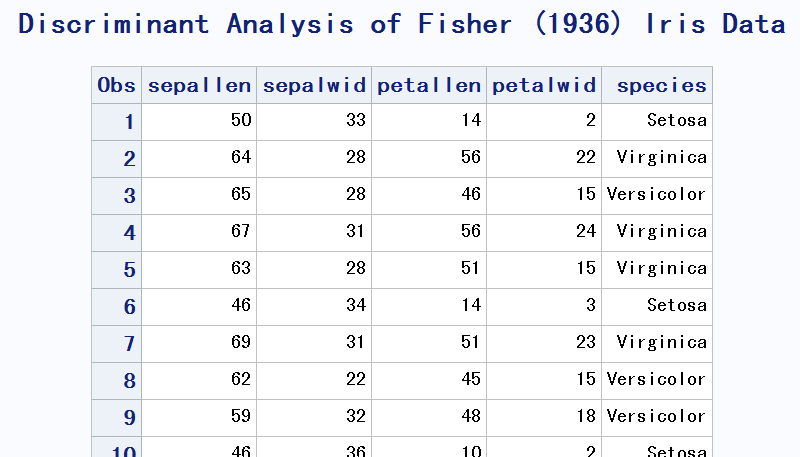
var petallen petalwid sepalwid;

**run**;

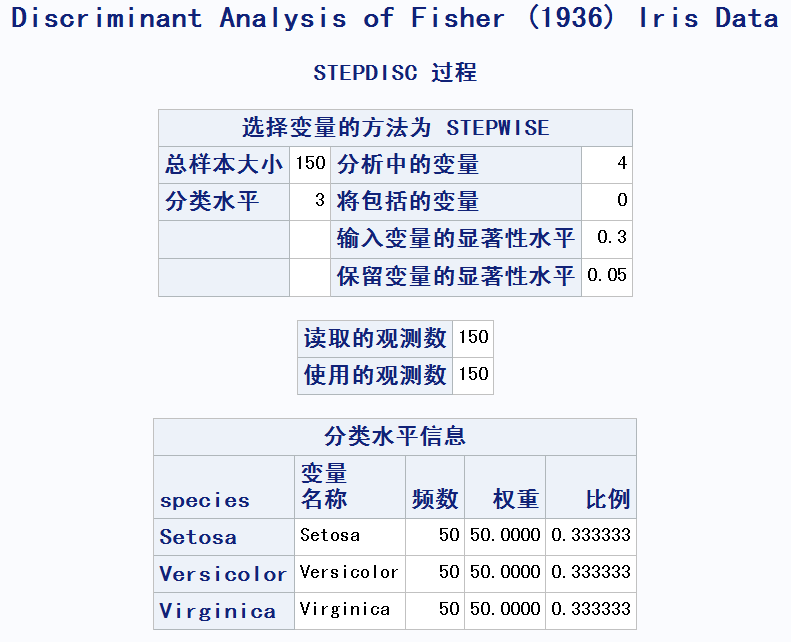
**proc** **print** data = plotiris;

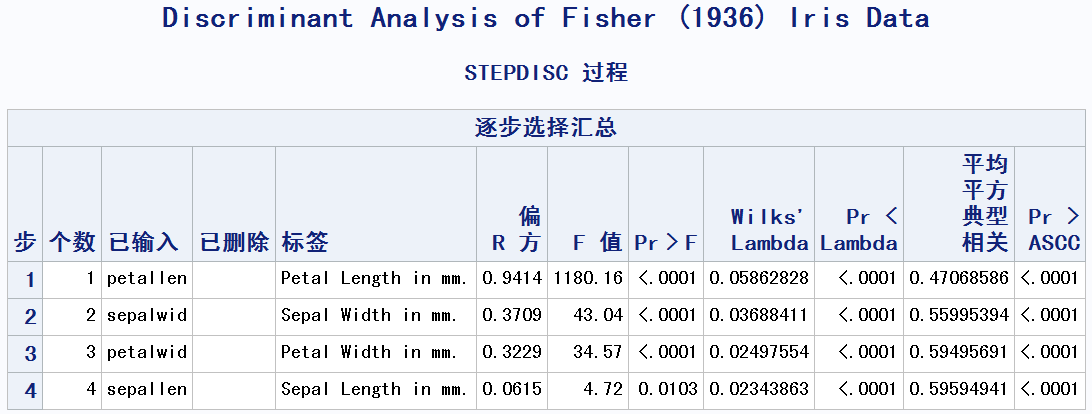
**run**;

**运行结果及说明：**



这是部分数据，species为分类变量，共3个分类。





先用STEPDISC逐步判别过程做变量筛选，short选项用来压缩逐步判别的输出（只输出汇总结果），sle=0.3 sls=0.05用来设置“向前选入法”和“后退剔除法”的显著水平。

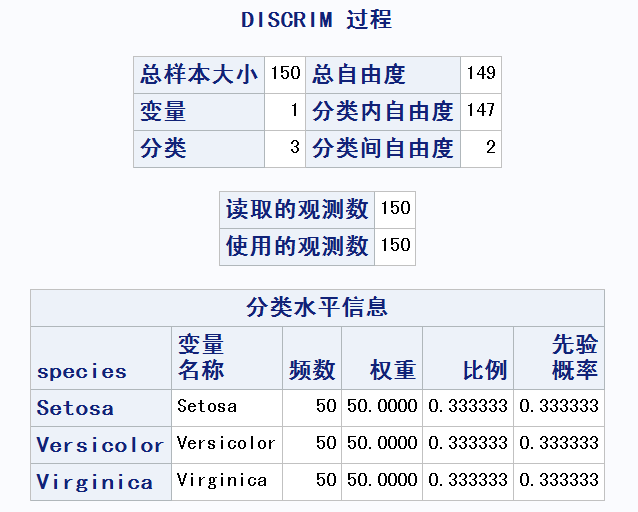
四个变量逐步被选入判别分析过程的次序为*petallen*、*sepalwid*、*petalwid*、*sepallen*，并且没有变量被剔除出来。说明最有效的判别函数是由这四个变量组成的。

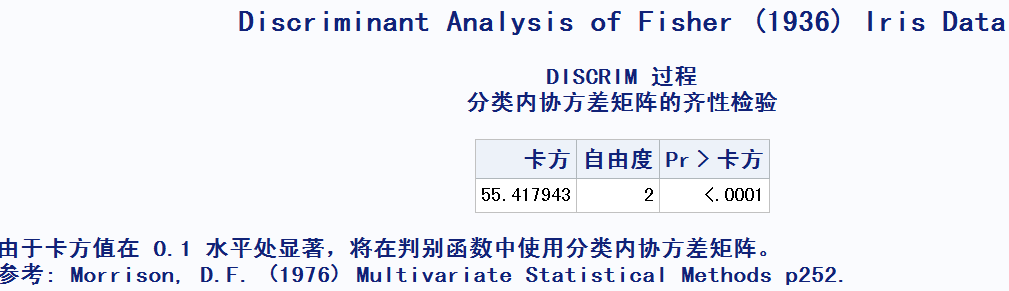
但是，第一个进入的*petallen*变量却已经能很好区分类别信息。它的偏R2达到0.9414，F值很大达到1180.161，平均平方典型相关值为0.47068586. 将它与四个变量全部用于判别分析时的平均平方典型相关值0.59594941比较，其实相差并不是非常大。

特别注意，用前三个变量时的平均平方典型相关值0.59495691与用全部四变量时的0.59594941相比较，几乎没有差异。用偏R2分析也能得到相同的结论。

因此，从能得出的判别函数既简洁又高效的角度看，用*petallen*、*sepalwid*和*petalwid*三个变量作为指标建立判别函数式是最优的。另外，如果用一个变量作为指标建立判别函数式，我们应该首先使用变量*petallen*。

另外，从逐步判别分析的第一步还可看到（只要去掉stepdisc过程中的short选项），使用*petalwid* 变量也能较好地建立判别函数式，因为它的R2也高达0.9289，与0.9414相差并不多。













以上输出了用单个变量*petallen*作为指标建立判别函数式的结果。选择*pool*=test时，表示要先经过原假设H0：3组的方差齐性，检验若满足齐性则合并，反之则不合并。

因卡方值=55.417943，自由度=2，*P*值=0.0001，在*α*=0.1水平上拒绝接受 H0. 所以将用各类内部的协方差矩阵计算类间的平方距离，并建立判别函数。单变量*petallen*所产生的总的、合并的和组间的标准差为：17.6530、4.3033、20.9070，R2= 0.9414，F=1180.161，P值=0.0001检验结果表明*petallen*具有显著区别3组总体的能力（与前面的逐步判别分析的结果应该是一致的）。

假设各类先验概率均为1/3时（DISCRIM过程的默认值），用全部数据建立起来的二次判别函数，再用来判别每一个样品的理论归属，最后得到与实际归属比较后的误判率=0.0467，150×0.0467=7例误判，符合率=1—0.0467=95.33%。同时还给出了按实际分类与理论分类吻合与否的交叉表，主对角线上的50、46、47为两种分类一致的记录数，及占各类实际总数的百分比为100%、92%、94%，也就是说Versicolor类错判4例，Virginica类错判3例。

利用二次判别函数还导出互相证实的结果。所谓互相证实，就是在共有*N*个样品中，每次留下一个样品作为新样品，由*N*—1个样品建立判别函数，然后将留下的这个样品代入判别函数，判别其归属。对每一个样品都留下来一次作为新样品来判别其归属。这样有利于减小用全部数据建立的判别函数再对全部数据进行回代判别所产生的偏差。误判率=0.0467，正好等于用全部数据建立起来的二次判别函数的误判率，但要注意这两者的误判率常常是不相等的，通常互相证实的误判率要高些。

7例误判的样品号分别为：12、25、63、83、118、131、148。分析118号样品：其原来实际归属为Versicolor类，理论误判为Virginica类，是由于分别按二次判别函数计算出来的这个样品归属3类的事后概率得到的，归属于Setosa的事后概率为0.0000，归属于Versicolor的事后概率为0.4710，归属于Virginica的事后概率为0.5290，取3个概率值中最大者所对应的样品类为理论归类结果。

所以实际Versicolor类中的50例，有4例（12、118、131、148）误判到Virginica类，46例实际分类与理论分类吻合，还有3例（25、63、83）由Virginica类误判到Versicolor类，最终结果Versicolor类=46+3=49例。



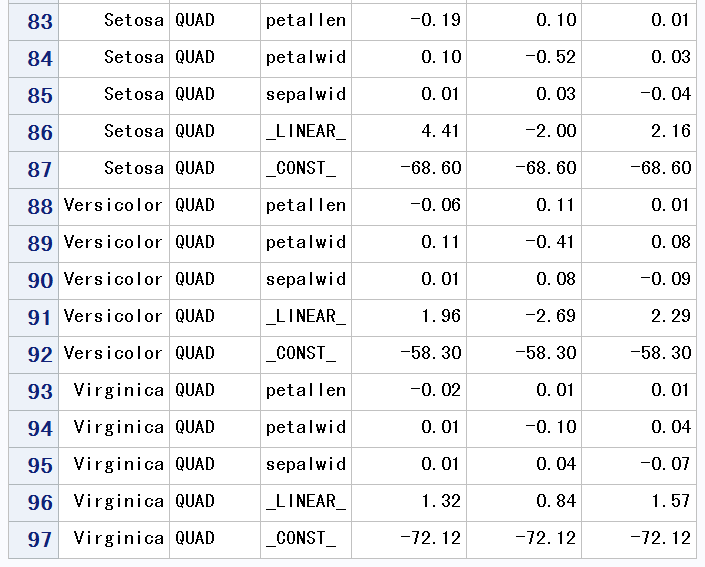












以上是用3 个变量petallen、sepalwid和petalwid做判别分析的结果。误判率为0.0333，误判4例：5、8、9、12.

上面最后一个表输出了二次判别函数的系数。二次型判别函数的系数存放在由“*outstat*=数据集名”选项规定的输出数据集中，其中\_type\_为quad的输出行中存放所有二次型判别函数的系数。例如，本例中计算setosa类的二次型的判别函数为：

setosa = -68.60+4.41*petallen*-2.00*petalwid*+2.16*sepalwid*

-0.19*petallen*2+0.10*petallen*×*petalwid+*0.01*petallen*×*sepalwid*

*-0.52petalwid*2+0.03*petalwid*×*sepalwid*-0.04*sepalwid*2

将需要判别归属分类的数据四个变量值代入上式，求得setosa值，用同样的方法求得versicolor和 virginica值，比较三者谁最大，就归属于谁。

**例2** 用**例1**中的训练数据集和判别函数，对新数据集（原数据去掉speices列）做判别分析。

**代码：**

**data** newiris (drop=species);

set iris;

**run**;

**proc** **print** data=plotdata;

**run**;

**proc** **discrim** data=iris testdata=newiris testout=plotp testoutd=plotd method=normal pool=yes short noclassify crosslisterr;

class species;

var petallen ;

title2 'Using Normal Density Estimates with Equal Variance';

**run**;

**proc** **discrim** data=iris testdata=newiris testout=plotp testoutd=plotd method=normal pool=no short noclassify crosslisterr;

class species;

var petallen ;

title2 'Using Normal Density Estimates with Unequal Variance';

**run**;

**proc** **print** data=plotp;

**run**;

**proc** **print** data=plotd;

**run**;

**proc** **discrim** data=iris testdata=newiris testout=plotp testoutd=plotd method=npar kernel=normal r=**.4** pool=yes short noclassify crosslisterr;

class species;

var petallen ;

title2 'Using Kernel Density Estimates with Equal Bandwidth';

**run**;

**proc** **discrim** data=iris testdata=newiris testout=plotp testoutd=plotd method=npar kernel=normal r=**.4** pool=no short noclassify crosslisterr;

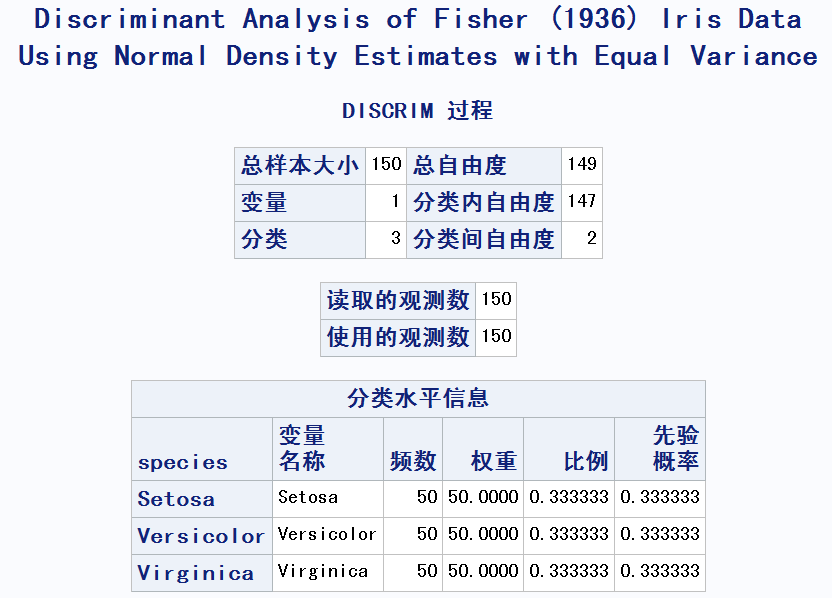
class species;

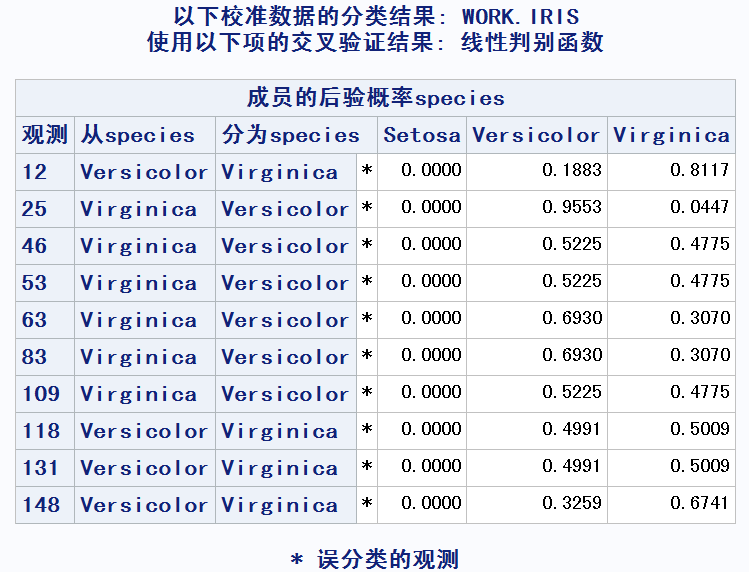
var petallen ;

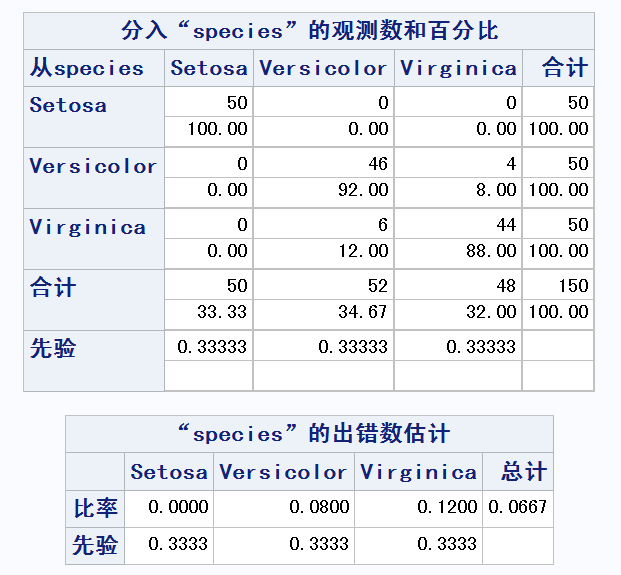
title2 'Using Kernel Density Estimates with Unequal Bandwidth';

**run**;

**运行结果及说明：**







调用了四次判别分析discrim过程。前二个过程假设组内服从多元正态分布，是参数分析方法（method=normal）。第一个过程假设三个类的方差相等（pool=yes），第二个过程假设三个类的方差不相等（pool=no）。后二个过程采用非参数分析方法（method=npar），在正态核估计密度中分别考虑带宽相等（pool=yes）和带宽不等（pool=no）的两种情况。对新数据集中的记录进行分类后，每一条记录的分类标志存放在输出的plotp数据集中。

上面是第一个过程的输出结果，其它过程输出结果也是类似的（略）。误判例数分别为：10,7,7,7.

**例3** 还是**例1**的数据，调用典型判别candisc过程和输出典型变量散布图。

**代码：**

**proc** **candisc** data=iris out=outcan distance anova;

class species;

var sepallen sepalwid petallen petalwid;

title 'Using Canonical Discriminant Analysis';

**run**;

**proc** **print** data=outcan;

title 'outcan';

**run**;

**proc** **format**;

value specfmt **1** = '+'

**2** = 'c'

**3** = '\*';

**run**;

**proc** **plot** data=outcan formchar='|----|---' vpct=**50** hpct=**80**;

plot can2\*can1=species;

format species specfmt.;

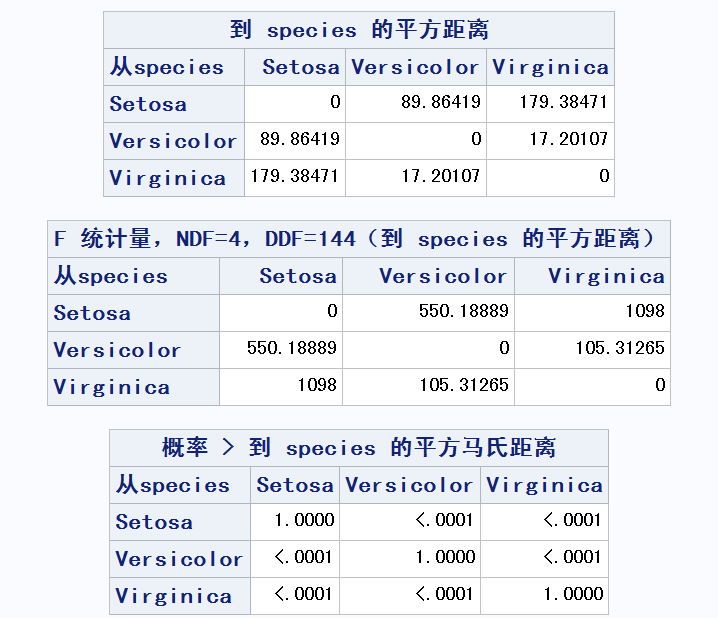
title2 'Plot of Canonical Variables';

**run**;

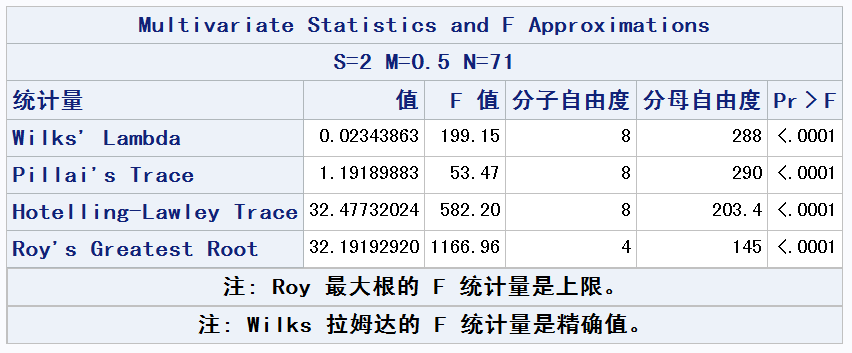
**运行结果及说明：**

distance选项指定类均值之间的平方距离为马氏距离。anova选项要求检验总体中每一个变量类均值相等的假设。









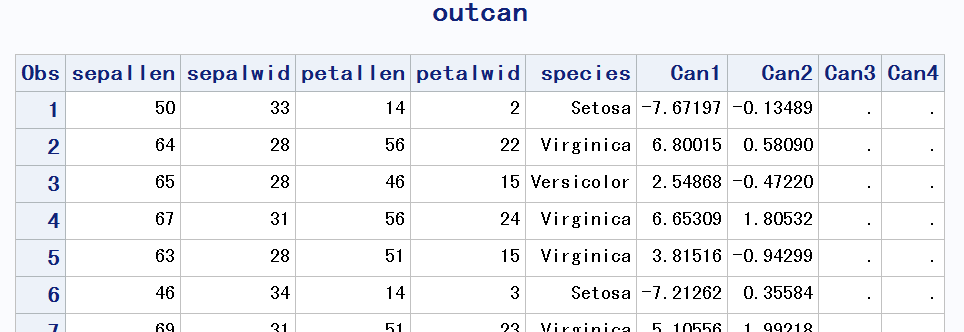












典型变量*Can*1和*Can*2与原始变量的关系可写成：

*Can1* = -0.0829377642*sepallen-*0.1534473068*sepalwid*

+0.2201211656*petallen*+0.281046039*petalwid*+*c*1

*Can*2 = 0.0024102149*sepallen*+0.2164521235*sepalwid*

-0.0931921210 *petallen*+0.2839187853*petalwid*+*c*2

每一条观察的四个变量值转换成两个典型变量的具体数值。将某观察的四个变量值和两个典型变量值代入，即可解出*c*1和*c*2。例如，对于第2观察来说（obs=2），*sepallen*=64，*sepalwid*=28，*petallen*=56，*petalwid*=22，*can*1= 6.80015，*can*2=0.58090，代入上二式即可得到：

*c*1＝－2.1051，*c*2＝－6.66147

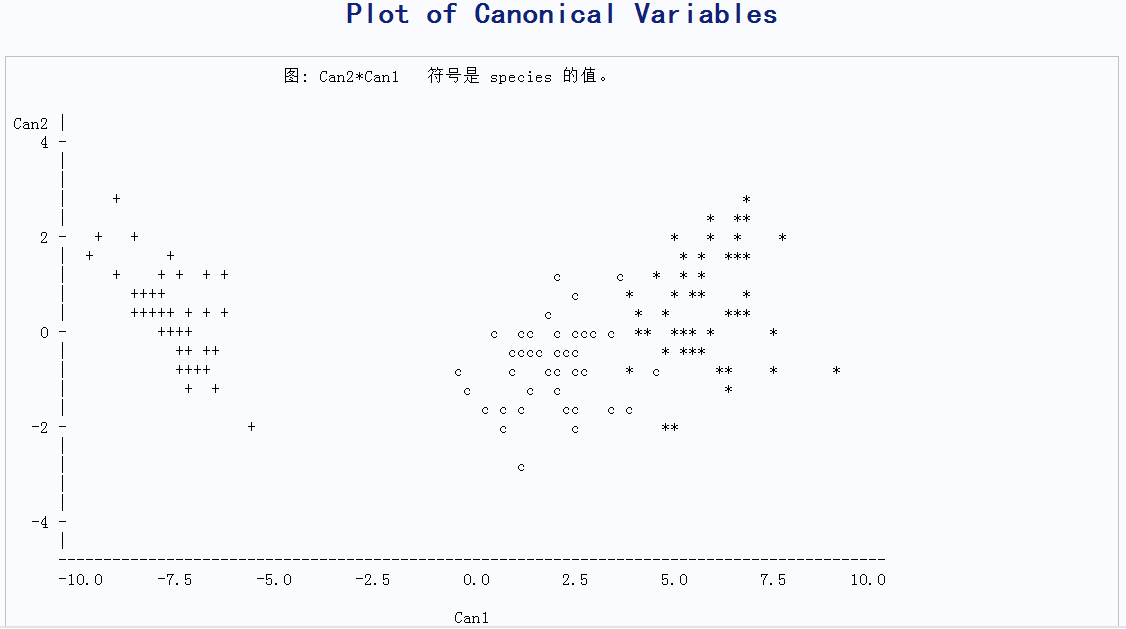
三类的线性判别函数分别为：

Setosa = -7.607599927 *Can*1+0.215133017 *Can*2

Versicolor = 1.825049490 *Can*1-0.727899622 *Can*2

Virginica = 5.782550437 *Can*1+ 0.512766605 *Can*2

例如，第1条记录数据*sepallen*=50，*sepalwid*=33，*petallen*=14，*petalwid*=2，计算出的*Can*1=-7.67197，*Can*2= -0.13489，再分别代入3个线性判别函数计算得出：setosa=58.3362, versicolor=-13.9035，virginica=-44.4327，由于算出的setosa值最大，所以归属于setosa类。



用典型变量*Can*1和*Can*2所画的散点图，可见这两个典型变量能较好地反映这三类总体的区别。

从输出结果可以看出，典型变量*Can*1和*Can*2作为原始变量的线性组合，集中了原始变量的绝大部分信息，并很好地概括了各组别之间的差异。典型判别是一种很好的降维技术，它用少量的典型变量代替原始变量，把样品分组情况直观地表现了出来。