PROC DISCRIM

ANALYSE DISCRIMINANTE

Dans une optique exploratoire, l'analyse discriminante procède, sur la base de la distance euclidienne, à une étude typologique ayant pour objectif de différencier les individus en fonction des variables figurant dans un tableau de données. Comme l'indique le tableau ci-dessous, la commande `PROC CORRESP` est suivie d'abord de l'option DATA= puis des options `optnum` séparées chacune par un caractère blanc. Les instructions `VAR, CLASS, PRIOR, WEIGHT, FREQ, ID, TESTCLASS, TESTFREQ, TESTID et BY`, séparées chacune par un point-virgule, permettent d'affiner l'analyse.

```
PROC DISCRIM DATA=fic1 optinum;
  VAR var1 var2 var3 var4 var5 var6 …;
  CLASS var1 var2…;
  PRIOR var2 …;
  WEIGHT var3;
  FREQ  var4 …;
  ID var5;
  TESTCLASS var1 …;
  TESTFREQ var4 …;
  TESTID var5;
  BY var6 …;
```

Panorama des options disponibles

Lorsque l'analyse statistique ne porte pas sur la dernière table (Data) Sas mémorisée, la commande PROC DISCRIM doit être suivie de l'option `DATA=nomtab1 où nomtab1` est le nom du tableau d'entrée (Data) Sas contenant les données à étudier. Si l'option DATA=`nomtab1` est absente, l'analyse porte alors sur la dernière table Sas mémorisée. Les options `optnum`, figurant après l'éventuelle option `DATA=nomtab1`, permettent à la fois de préciser les conditions techniques de l'analyse,

Option	Utilité de l'option
`CANONICAL`	réalise une analyse discriminante canonique
`METHOD=NORMAL`	effectue une analyse discriminante basée sur l'hypothèse d'une distribution normale des variables (option par défaut)
`METHOD=NPAR`	effectue une analyse discriminante basée sur une méthode non paramétrique
`ANOVA`	teste l'hypothèse d'égalité des moyennes des classes en mode univarié
`MANOVA`	teste l'hypothèse d'égalité des moyennes des classes en mode multivarié
`ALL`	retient toutes les options possibles

les tableaux Sas (Data) et les fichiers utiles,

Option	Utilité de l'option
OUT=nomtab2	permet de donner le nom du fichier de sortie (ici,*nomtab2*)
OUTCROSS=nomtab3	permet de donner le nom du fichier de sortie (ici, *nomtab3*) contenant les données de base, les probabilités ainsi que les classifications réalisées
OUTSTAT=nomtab4	permet de donner le nom du fichier de sortie (ici, *nomtab4*) contenant les différentes statistiques relatives aux variables sélectionnées
TESTDATA=nomtab5	permet de donner le nom du tableau (ici, *nomtab5*) contenant le classement des variables (utile seulement si on utilise l'instruction TESTCLASS, TESTFREQ OU TESTID)

ainsi que les matrices de corrélation ou les matrices de variances-covariances désirées des classes construites lors de l'analyse.

Option	Utilité
BCORR	crée la matrice des corrélations inter-classes
PCORR	crée la matrice des corrélations intra-classes
WCORR	crée les matrices de corrélations intra-classes pour chaque classe
BCOV	crée la matrice variance-covariance inter-classes
PCOV	crée la matrice variance-covariance intra-classes
WCOV	crée les matrices variance-covariance intra-classes pour chaque classe

Panorama des instructions disponibles

De nombreuses instructions, séparées chacune par un point-virgule, peuvent figurer après la commande PROC DISCRIM comme l'indique le tableau ci-dessous qui précise l'utilité de chacune de ces instructions.

Instruction	Utilité
VAR	fixe la liste des variables retenues (*var1, var2, var3, var4, var5, var6*). En l'absence de cette instruction, toutes les variables sont retenues
CLASS	réalise les calculs par classe de variables (ici, *var1, ...*) données à priori (si les résultats de l'analyse sont identiques à ceux obtenus avec l'instruction BY, ils sont toutefois sont présentés sous une forme légèrement différente)
PRIOR	détermine la probabilité associée à la variable indiquée (ici, *var2,...*)
WEIGHT	crée la variable (ici, *var3*) servant de facteur de pondération aux autres variables
FREQ	estime en pourcentages, simples et cumulés, les variables (ici, *var4*, …)
ID	retient une variable (ici, *var5*) comme identificateur (par défaut, l'identificateur est le numéro de lignes _N_ attribué à chaque observation
TESTCLASS	vérifie que les variables citées (ici, *var1*) sont bien classées
TESTFREQ	vérifie que les fréquences associées aux variables indiquées (ici, *var4, ...*) sont bonnes
TESTID	vérifie que l'identificateur (ici, *var 5*) est correct
BY	réalise les calculs par classe de variables données a posteriori (ici, *var6,...*).

N.B. : L'analyse discriminante est parfois suivie par une classification réalisée à partir des résultats obtenus au terme de l'analyse (à l'aide des procédures PROC CLUSTER et PROC TREE).