

PSY38X2 : Traitement de données en Psychologie - TD N°4

Analyse en Composantes Principales avec Minitab et Modalisa

1) Présentation de Modalisa

1.1 - Qu'est-ce que Modalisa ? Création de la base de données des enquêtes.

Modalisa est un logiciel de traitement d'enquêtes, notamment utilisé par les Sociologues. Il prend en charge la chaîne des travaux depuis la conception du questionnaire, la saisie des réponses des sujets interrogés, et le traitement de ces réponses par des outils statistiques.

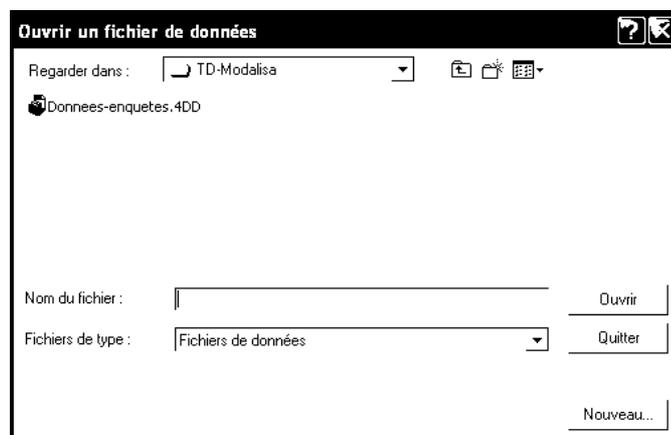
Modalisa est une application construite à partir d'un système de gestion de bases de données, 4ème Dimension. Il en résulte un certain nombre de particularités de fonctionnement qui en font un outil différent d'un tableur ou d'un logiciel de traitement statistique tel que Minitab.

Ainsi, un fichier Modalisa (et ses fichiers annexes) est une base de données rassemblant plusieurs enquêtes. Pour commencer, créons une nouvelle base de données :

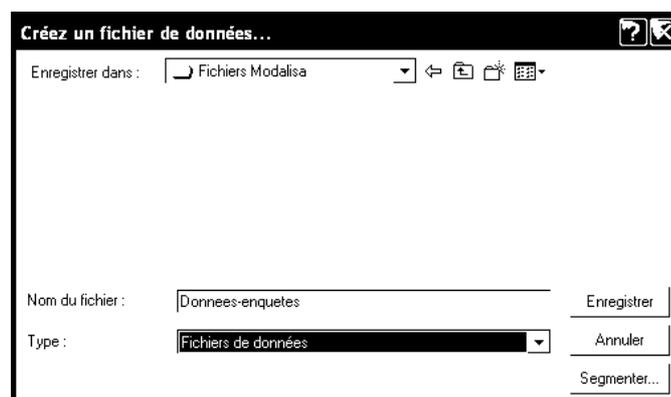
Créez un répertoire "Fichiers Modalisa" sur votre volume du serveur (U:).
Exécutez l'item Modalisa du menu Démarrer.

Le dialogue "Ouvrir un fichier de données" s'affiche.

Cliquez sur le bouton "Nouveau"



Sélectionnez alors le répertoire "Fichiers Modalisa" de votre compte et indiquez "Donnees-enquetes" comme nom de fichier.



Cliquez ensuite sur Enquête-cr ation, ce qui va nous permettre de cr er une enqu te correspondant au cas Mini-ACP vu en cours de statistiques.
Nommez cette enqu te Mini-ACP.

Modalisa a cr e deux fichiers, d'extensions respectives .4DD et .4DR dans le r pertoire indiqu .

1.2 - Traitement de l'exemple Mini-ACP avec Modalisa

1.2.1 - Saisie des donn es

La premi re  tape consiste   d finir les questions composant notre "questionnaire d'enqu te". Pour l'usage que nous en faisons ici, une question est   peu pr s la m me chose qu'une variable statistique. D finissons les quatre questions correspondant au cas Mini-ACP :

Cliquez sur l'onglet "Questions" de la fen tre du pilote, puis sur le bouton "Ajouter".

Pour garder les m mes notations qu'en CM, les questions s'appelleront X1, X2, X3, et X4.

Attention   choisir des variables de type num rique et   indiquer correctement les bornes et le nombre de d cimales. En particulier, si on laisse "D cimales"   0 pour la premi re question, il sera impossible de saisir correctement l'observation relative au sujet s5, et  galement impossible de modifier alors la structure de l'enqu te.

Question n 1 Termin  Enregistrer

Nom de variable (60 car.) X1

Libell  de la question (1000 car.)

Abr g  (4 car.) X1 Code formulaire (6 car.)

Type

1   r ponse unique

2   r ponses multiples

3   r ponses ordonn es

4   r ponse num rique

5 texte libre

6 date

Borne inf rieure -100

Borne sup rieure 100

D cimales (0 - 10) 2

Copie d'une question Copie d'un groupe

Apr s avoir d fini les quatre questions, refermez la fen tre   l'aide du bouton Termin , et cliquez sur l'onglet Donn es, puis sur le bouton Ajouter.

N.B. L'ensemble des observations relatives   un sujet donn  est appel  "Questionnaire" par Modalisa.

Saisissez s1 comme identificateur du premier questionnaire, puis saisissez les valeurs observ es des 4 variables sur le sujet N 1.

Questionnaire: 1

Identifiant: s1

Question	Saisie
1	-11
2	-60
3	110
4	

Question n  4 X4

Bornes: 0 / 1000

D cimales: 0

Pour une 'non r ponse', ne saisissez rien.

Libell s

Questions Modalit s

Saisie rapide

Termin  Enregistrer

Passez d'une variable à la suivante à l'aide des flèches, et terminez la saisie par le bouton Enregistrer. Après enregistrement du 6^e questionnaire, cliquez sur le bouton Terminé.
Rappel des données :

Enquête: Mini-ACP

N°	1. X1	2. X2	3. X3	4. X4
1	-11	-60	110	40
2	-12	-62	93	25
3	-15	-80	113	39
4	-14	-75	94	25
5	-14,5	-82	100	30
6	-13	-72	102	32

1.2.2 - ACP proprement dite

Analyse	Plans	Fenêtre	Aide
Aperçu à l'écran			
Aperçu : Série			
Tri à plat			
Série de tris à plat...			
Tri croisé, Khi-deux...			
Série de tris croisés...			
Tableaux juxtaposés			
Profil de modalités : 1 var...			
Profil de variables...			
AFC : Correspondances			
Arbre			
Classes			
Stats 1 variable...			
Stats n variables...			
Stats n sous-populations...			
Stats n Var \ n Sous-pop...			
Test de Kolmogorov - Normalité			
Droite de Henry			
Test t - Comparaison de moy...			
Analyse de la variance...			
Corrélation et régression simple			
Régression multiple			
Matrice de corrélations			
ACP : Composantes principales...			
Tests non paramétriques...			
Typologies...			

Nos données ont été enregistrées dans le fichier `Donnees-enquetes.4DD`. Mais, avant tout traitement, nos données doivent être "compilées", c'est-à-dire indexées et stockées sous une forme convenable.

Exécutez le menu Saisie - Tout recompiler et choisissez l'option Compilation sans sauvegarde.

Déroulez ensuite le menu Analyse : on constate que Modalisa permet de réaliser la plupart des traitements que nous avons étudiés en Statistiques.

Calcul de la matrice des corrélations

Pour ce traitement, comme pour de nombreux autres, Modalisa n'affiche pas les résultats à l'écran, mais, soit les imprime, soit les écrit dans un fichier texte.

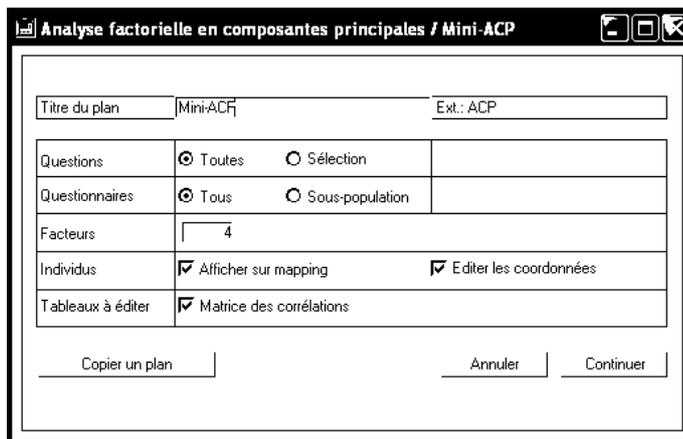
Exécutez le menu Analyse - Matrice de corrélations. Choisissez "fichier" comme destination, puis sélectionnez les 4 variables. Choisissez ensuite "Nouveau Fichier" et indiquez comme nom de fichier Corrélations-Mini-ACP. en veillant à ce que l'enregistrement se fasse dans votre répertoire.

Modalisa crée un fichier Corrélations-Mini-ACP.txt dans notre répertoire. Il faut ouvrir ce fichier (avec le bloc-notes ou avec WordPad) pour consulter les résultats.

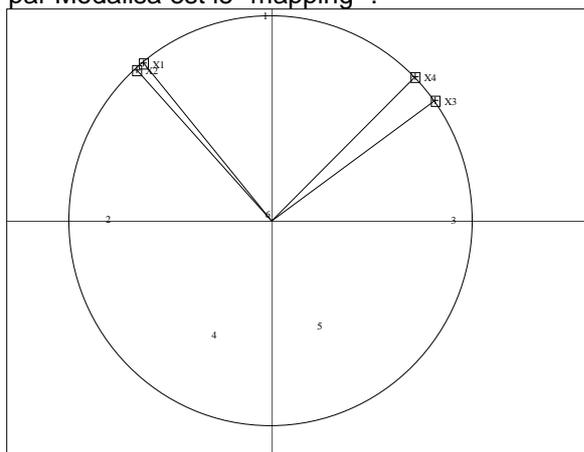
Calcul des résultats de l'analyse en composantes principales

Exécutez le menu Analyse - ACP : Composantes principales... Choisissez "Nouvelle analyse à partir de variables numériques de l'enquête".

Modalisa affiche alors une fenêtre de dialogue que nous pouvons compléter comme suit :



Le principal résultat fourni par Modalisa est le "mapping" :



Remarquez que les variables principales définies par Modalisa sont toutes deux les opposées de celles qui avaient été données en cours.

On peut également afficher les scores des individus à l'aide du menu Individus - Coordonnées individus et (après avoir refermé la fenêtre des coordonnées d'individus) les saturations des variables à l'aide du menu : Variables - Coordonnées.

Individus concernés		
Facteurs	Coordonnées	
	1	2
% Information	50,02%	49,17%
1	-0,070	2,511
2	-2,022	0,029
3	2,246	0,015
4	-0,706	-1,378
5	0,603	-1,271
6	-0,050	0,092

Variables		
Facteurs	Coordonnées	
	1	2
% Information	50,02%	49,17%
x1	-0,628	0,768
x2	-0,665	0,736
x3	0,809	0,585
x4	0,712	0,700

Remarquez que les coordonnées des individus correspondent à une réduction des variables utilisant les écarts types corrigés.

Sauvegarder notre traitement

Modalisa a en fait défini un "Plan d'analyse". Pour sauvegarder un tel plan, on peut exécuter le menu Plans - Lancer les plans d'analyse en complétant les fenêtres de dialogue comme suit :

- destination des données : fichiers
- dossier de destination des fichiers : U:\Fichiers Modalisa.

Modalisa crée alors les fichiers Mini-ACP_ACPi.TXT, Mini-ACP_ACPr.TXT et Mini-ACP_COR.TXT.

Le premier contient les scores des individus et les saturations des variables (en format texte), et le dernier les corrélations entre les variables.

On peut ré-afficher les résultats de l'ACP à l'aide du menu Plans - Lire Résultats. On sélectionne l'item "Analyse factorielle en composantes principales (extension _ACPr.TXT)" puis le fichier Mini-ACP_ACPr.TXT.

D'autres traitements possibles

Modalisa permet d'autres traitements sur notre ACP, tels que :

- Afficher les individus par un symbole, ou par leur numéro (boutons : N°)

- Explorer les axes factoriels au delà des axes 1 et 2 (menus Carte - Facteur X N°... et Carte - Facteur Y N°...)

1.3 - Traitement du cas Budget-temps avec Modalisa

1.3.1 - Importation de données au format texte

Tel qu'est conçu Modalisa, il n'est guère envisageable de créer une base de données pour chaque enquête, et donc de recopier depuis le serveur des bases de données préparées. En revanche, il est assez simple d'ajouter une enquête à notre base de données en important des données au format "texte". C'est ce que nous allons faire pour acquérir les données Budget-temps (cf. exercice 43 des fiches de TD de statistiques).

Le fichier Budget-temps.txt

Visualisez le fichier W:\PSY3\TD-Modalisa\Budget-temps.txt à l'aide d'un éditeur de texte. Observez la manière dont les données sont structurées (première ligne contenant les noms des variables, tabulations séparant les colonnes, etc).

Import des données Budget-temps.txt

- Affichez la fenêtre du pilote de Modalisa
- Exécutez le menu Fichier - Import ASCII
- Choisissez l'item "Importer des questionnaires dans une nouvelle enquête..."
- Sélectionnez le fichier W:\PSY3\TD-Modalisa\Budget-temps.txt

Modalisa affiche alors un certain nombre de fenêtres indiquant de quelle manière les données doivent être importées.

- Indiquez à Modalisa d'ignorer la première colonne (identifiants de individus)
- Laissez les options par défaut pour les autres colonnes de données.
- Donnez le nom "Budget-temps" à la nouvelle enquête

Remarquez que Modalisa ne permet pas d'importer les identifiants des individus (on peut toutefois les stocker dans l'une des variables de l'enquête). On peut les rétablir de la façon suivante :

- A partir du pilote, cliquez sur l'onglet Données.
- Exécutez le menu Saisie - Modifier Codes questionnaires
- Saisir, dans l'ordre, les codes des individus :
HAU, FAU, FNU, HMU, FMU, HCU, FCU, HAW, FAW, FNW, HMW, FMW, HCW, FCW, HAY, FAY, FNY, HMY, FMY, HCY, FCY, HAE, FAE, FNE, HME, FME, HCE, FCE

Exécutez ensuite une ACP sur les données de cette enquête, en spécifiant 3 ou 4 facteurs. Comparez les résultats obtenus avec ceux figurant dans la fiche de TD de statistiques.

2) Analyse en Composantes Principales avec Minitab

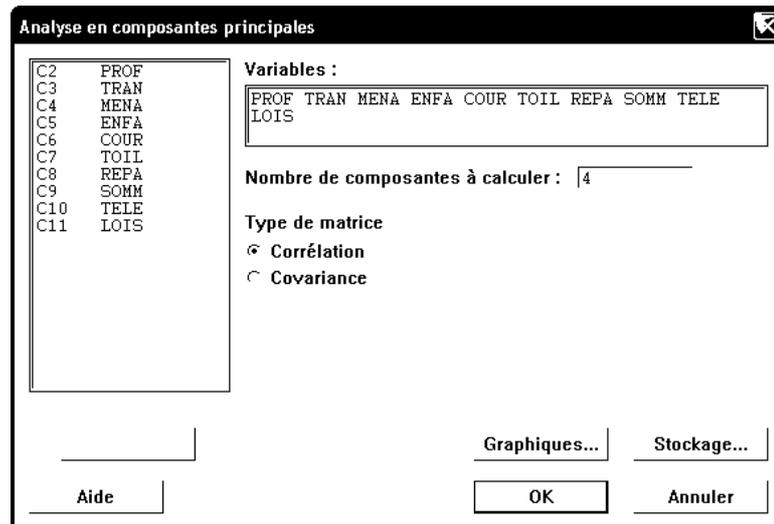
2.1 - Traitement du cas Budget-temps avec Minitab

2.1.1 - Afficher le nuage des individus et le diagramme des valeurs propres

Chargez Minitab et ouvrez le projet W:\PSY3\TD-Minitab\Budget-temps.mpj

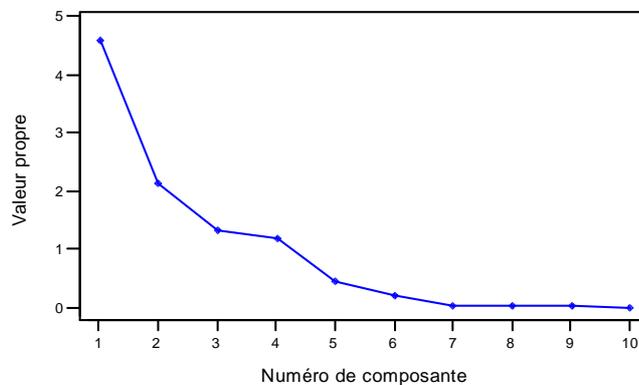
Exécutez le menu Stat - Multivarié - Composantes Principales...

Complétez la fenêtre de dialogue comme suit :



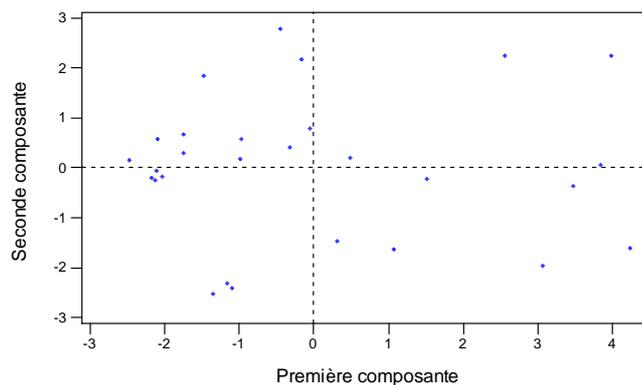
Sélectionnez le sous-dialogue Graphiques... et demandez les deux graphiques proposés.

Diagramme en cône de PROF-LOIS



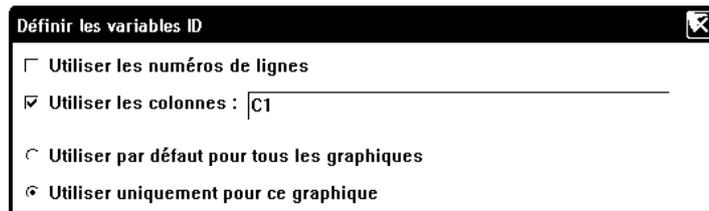
La décroissance brutale entre la 4^e et la 5^e valeur propre justifie que nous ne demandions que 4 composantes.

Diagramme des scores du PROF-LOIS



Minitab nous fournit également la projection du nuage des individus sur le plan défini par les deux premières composantes. Il aurait été intéressant de pouvoir afficher les étiquettes des individus. Ce n'est pas possible, mais Minitab nous offre la solution de remplacement suivante :

- Cliquez avec le bouton droit de la souris sur le graphique et sélectionnez l'item Balayer.
- Cliquez de nouveau avec le bouton droit de la souris et sélectionnez l'item Définir les variables ID...
- Complétez, par exemple, la fenêtre de dialogue comme suit :

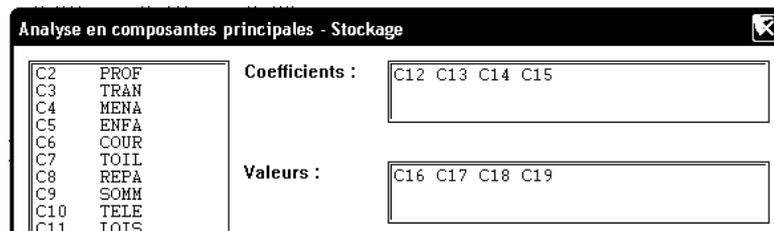


Lorsqu'on clique sur un point de graphique, ou lorsqu'on sélectionne un groupe de points en faisant glisser la souris, une petite fenêtre indique le ou les identifiants des sujets correspondants.

2.1.2 - Obtenir des données numériques relatives à l'ACP

La fenêtre session affiche les valeurs propres et les "loadings" ou coefficients des variables. On peut obtenir les scores des individus en utilisant le bouton "Stockage" de la fenêtre de dialogue de l'ACP.

- Ré-exécutez comme précédemment le menu Stat - Multivarié - Composantes Principales...
- Cliquez sur le bouton Stockage et indiquez, par exemple :



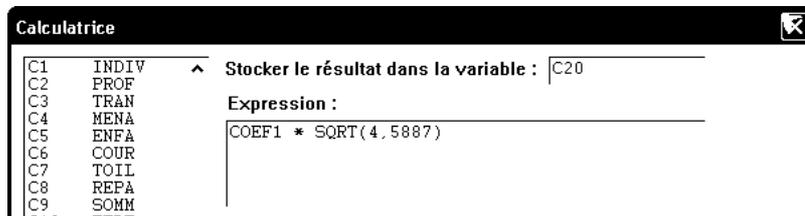
N.B. Dans les deux zones d'édition, il faut indiquer autant de colonnes que de composantes principales demandées. Les colonnes pourront ensuite être renommées COEF1 à COEF4 et SCORE1 à SCORE4.

Les colonnes C12 à C15 affichent les coefficients des composantes (loadings) tandis que les colonnes C16 à C19 affichent les valeurs des composantes principales sur les individus, c'est-à-dire les scores des individus.

Les saturations des variables ne sont pas calculées, mais on pourra obtenir, par exemple, la saturation de la variable PROF sur la composante CP1 en calculant la corrélation entre PROF et SCORE1. On obtient :

Corrélation de PROF et SCORE1 = -0,977; Niveau de p = 0,000

On peut aussi calculer les saturations de l'ensemble des variables sur la première composante principale en multipliant les valeurs de la colonne COEF1 par la racine carrée de la première valeur propre. Utilisez la calculatrice pour stocker le résultat dans la colonne C20, que vous renommerez SAT1 :



N.B. Le nombre 4,5887 est la valeur de la première valeur propre, lue dans la fenêtre Session.

Procédez de même pour les saturations selon la seconde composante, stockées en colonne C21, renommée SAT2.

On pourra également reconstruire le nuage des individus dans le plan (CP1,CP2) , avec cette fois-ci des étiquettes représentant les individus, en utilisant le menu Graphiques - Diagramme... de la manière suivante :

Variables du graphique :

Graph.	Y	X
1	SCORE2	SCORE1
2		
3		

Affichage de données :

Elém.	Affichage	Pour chaque	Variables de groupe
1	Symbole	Graph.	
2			
3			

Editer les attributs...

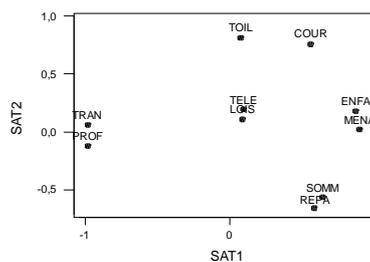
Annotation Cadre Régions

Il faut ensuite dérouler le menu local Annotation, choisir l'item Etiquettes des individus et spécifier la colonne C1 comme source des étiquettes.

L'utilisation de ce menu n'est évidemment pas limitée aux deux premières composantes principales, et on pourra essayer de représenter les individus dans les plans (CP1, CP3), (CP2,CP3), etc

On pourra aussi utiliser ce menu pour construire un diagramme relatif aux variables. Mais, il faut utiliser les saturations des variables pour obtenir un diagramme analogue à celui produit par Modalisa, ou celui vu en cours. Par exemple :

- Saisir en C22 les noms des variables
- Utiliser le menu Graphiques - Diagramme ... pour construire le diagramme suivant :



Après avoir enregistré ce fichier, envoyez-le par mail à votre enseignant.

3) Travail de monitorat à rendre

Le fichier U:\PSY3\TD-Minitab\Basket.mpj est un projet Minitab contenant les données relatives au cas suivant :

On s'intéresse au profil de 18 basketteurs de 14 ans. Ils ont passé un certain nombre de tests relatifs aux qualités physiques requises pour la pratique de cette discipline.

TAI : taille en cm

VIT : vitesse sur 30 m (en secondes)

DET : détente verticale en cm : sauter le plus haut possible, le bras tendu

PAS : passe en mètres : lancer un ballon de basket le plus loin possible

LEG : endurance, en litres/mn/kg : test Le Luc Léger

STA : adresse statique, en nombre de paniers.

La variable VIT est codée systématiquement avec un signe "-" afin que, comme pour les autres variables, une valeur élevée traduise une bonne performance.

Le fichier U:\PSY3\TD-Modalisa\Basket.txt contient les mêmes données en format texte, en vue de l'importation dans une base de données Modalisa.

Utilisez Minitab et Modalisa pour effectuer une ACP sur ces données.

Rédigez ensuite sous Word un rapport commentant les résultats trouvés (liens entre les composantes principales et les variables observées, oppositions entre individus, etc). Vous pourrez illustrer ce rapport en copiant les représentations graphiques que vous jugerez intéressantes parmi celles produites par Minitab ou Modalisa.

Envoyez par mail le fichier Word seul (inutile de joindre le fichier Minitab ou la base de données Modalisa).