IBM SPSS Statistics Base 21

Remarque : Avant d'utiliser ces informations et le produit qu'elles concernent, lisez les informations générales sous Remarques sur p. 337.

Cette version s'applique à IBM® SPSS® Statistics 21 et à toutes les publications et modifications ultérieures jusqu'à mention contraire dans les nouvelles versions.

Les captures d'écran des produits Adobe sont reproduites avec l'autorisation de Adobe Systems Incorporated.

Les captures d'écran des produits Microsoft sont reproduites avec l'autorisation de Microsoft Corporation.

Préface

IBM® SPSS® Statistics est un système complet d'analyse de données. Le module complémentaire facultatif Base fournit les techniques d'analyse supplémentaires décrites dans ce manuel. Le module complémentaire Base doit être utilisé avec le système central SPSS Statistics auquel il est entièrement intégré.

A propos de IBM Business Analytics

Le logiciel IBM Business Analytics offre des informations complètes, cohérentes et précises permettant aux preneurs de décision d'améliorer leurs performances professionnelles. Un portefeuille complet de solutions de business intelligence, d'analyses prédictives, de performance financière et de gestion de la stratégie, et d'applications analytiques permet une connaissance claire et immédiate et offre des possibilités d'actions sur les performances actuelles et la capacité de prédire les résultats futurs. En combinant des solutions du secteur, des pratiques prouvées et des services professionnels, les entreprises de toute taille peuvent générer la plus grande productivité, automatiser les décisions en toute confiance et apporter de meilleurs résultats.

Dans le cadre de ce portefeuille, le logiciel IBM SPSS Predictive Analytics aide les entreprises à prédire des événements futurs et à agir de manière proactive en fonction de ces prédictions pour apporter de meilleurs résultats. Des clients dans les domaines commerciaux, gouvernementaux et académiques se servent de la technologie IBM SPSS comme d'un avantage concurrentiel pour attirer ou retenir des clients, tout en réduisant les risques liés à l'incertitude et à la fraude. En intégrant le logiciel IBM SPSS à leurs opérations quotidiennes, les entreprises peuvent effectuer des prévisions, et sont capables de diriger et d'automatiser leurs décisions afin d'atteindre leurs objectifs commerciaux et d'obtenir des avantages concurrentiels mesurables. Pour plus d'informations ou pour contacter un représentant, visitez le site *http://www.ibm.com/spss*.

Support technique

Un support technique est disponible pour les clients du service de maintenance. Les clients peuvent contacter l'assistance technique pour obtenir de l'aide concernant l'utilisation des produits IBM Corp. ou l'installation dans l'un des environnements matériels pris en charge. Pour contacter l'assistance technique, visitez le site IBM Corp. à l'adresse *http://www.ibm.com/support*. Votre nom, celui de votre société, ainsi que votre contrat d'assistance vous seront demandés.

Support technique pour les étudiants

Si vous êtes un étudiant qui utilise la version pour étudiant, personnel de l'éducation ou diplômé d'un produit logiciel IBM SPSS, veuillez consulter les pages Solutions pour l'éducation (*http://www.ibm.com/spss/rd/students/*) consacrées aux étudiants. Si vous êtes un étudiant utilisant une copie du logiciel IBM SPSS fournie par votre université, veuillez contacter le coordinateur des produits IBM SPSS de votre université.

Service clients

Si vous avez des questions concernant votre livraison ou votre compte, contactez votre bureau local. Veuillez préparer et conserver votre numéro de série à portée de main pour l'identification.

Séminaires de formation

IBM Corp. propose des séminaires de formation, publics et sur site. Tous les séminaires font appel à des ateliers de travaux pratiques. Ces séminaires seront proposés régulièrement dans les grandes villes. Pour plus d'informations sur ces séminaires, accédez au site *http://www.ibm.com/software/analytics/spss/training*.

Contenu

1 Livre de codes 1

Onglet Résultats du livre des codes . . . 3
Onglet Statistiques du livre des codes . . . 5

2 Effectifs 8

Statistiques des fréquences . . . 9
Diagrammes des fréquences . . . 11
Format des fréquences . . . 12

3 Descriptives 13

Options Descriptives . . . 14
Fonctionnalités supplémentaires de la commande DESCRIPTIVES . . . 16

4 Explorer 17

Statistiques d'Explorer . . . 18
Diagrammes d'Explorer . . . 19
    Transformations de l'exposant d'Explorer . . . 20
Options d'Explorer . . . 20
Fonctionnalités supplémentaires de la commande EXAMINE . . . 21

5 Tableaux croisés 22

Strates de tableaux croisés . . . 23
Diagrammes en bâtons juxtaposés de tableaux croisés . . . 24
Tableaux croisés affichant les variables de strate dans des strates de tableau . . . 24
Statistiques de tableaux croisés . . . 25
Affichage de cellules (cases) de tableaux croisés . . . 28
Format de tableau croisé . . . 30

6 Récapituler 31

Options de Récapituler 33
Récapituler les statistiques 33

7 Moyennes 36

Moyennes : Options 38

8 Cubes OLAP 41

Cubes OLAP : Statistiques 43
Cubes OLAP : Différences 45
Cubes OLAP : Titre 46

9 Tests T 47

Test T pour échantillons indépendants 47
Définir Groupes Test T pour Echantillons Indépendants 49
Options Test T pour Echantillons Indépendants 49
Test T pour échantillons appariés 50
Options test T pour échantillons appariés 51
Test T pour échantillon unique 52
Options Test T pour échantillon unique 53
Fonctionnalités supplémentaires de la commande T-TEST 53

10 ANOVA à 1 facteur 54

Contrastes ANOVA à 1 facteur 55
Tests Post Hoc ANOVA à 1 facteur 56
Options ANOVA à 1 facteur 58
Fonctionnalités supplémentaires de la commande ONEWAY 59

11 Analyse GLM – Univarié 60

Modèle GLM. . . . 62
Termes construits . . . 62
Somme des carrés. . . . 63
Contrastes GLM . . . 64
Types de contraste . . . 64
Diagrammes de profils GLM . . . 65
Comparaisons post hoc GLM . . . 66
Enregistrement GLM. . . . 68
Options GLM. . . . 70
Fonctionnalités supplémentaires de la commande UNIANOVA . . . 71

12 Corrélations bivariées 73

Options de corrélations bivariées . . . 75
Propriétés supplémentaires des commandes CORRELATIONS et NONPAR CORR . . . 75

13 Corrélations partielles 76

Options Corrélations partielles . . . 77
Fonctionnalités supplémentaires de la commande PARTIAL CORR . . . 78

14 Distances 79

Distances : Mesures de dissimilarité . . . 80
Indices : Mesures de similarité. . . . 81
Fonctionnalités supplémentaires de la commande PROXIMITIES . . . 82

15 Modèles linéaires 83

Pour obtenir un modèle linéaire . . . 84
Objectifs . . . 85
Bases . . . 86
Choix du modèle . . . 87

Ensembles . . . 89
Avancé . . . 90
Options de modèle . . . 90
Récapitulatif de modèle . . . 91
Préparation automatique des données . . . 92
Importance des variables prédites . . . 93
Valeurs prévues en fonction des valeurs observées . . . 94
Résidus . . . 95
Valeurs éloignées . . . 96
Effets . . . 97
Coefficients . . . 98
Moyennes estimées . . . 100
Récapitulatif de création de modèle . . . 101

16 Régression linéaire *102*

Méthodes de sélection des variables de régression linéaire . . . 104
Régression linéaire : Définir la règle . . . 105
Diagrammes de régression linéaire. . . . 105
Régression linéaire : Enregistrer de nouvelles variables . . . 107
Statistiques de régression linéaire . . . 109
Régression linéaire : Options . . . 111
Fonctionnalités supplémentaires de la commande REGRESSION . . . 112

17 Régression ordinale *113*

Régression ordinale : Options . . . 114
Régression ordinale : Résultat . . . 115
Régression ordinale : Emplacement . . . 117
    Termes construits . . . 118
Régression ordinale : Echelle . . . 118
    Termes construits . . . 118
Fonctionnalités supplémentaires de la commande PLUM . . . 119

18 Ajustement de fonctions 120

Modèles d'ajustement de fonctions 122
Enregistrement de l'ajustement de fonctions. 122

19 Régression des moindres carrés partiels 124

Modèle 126
Options 127

20 Analyse du voisin le plus proche 129

Voisins 133
Descriptives 135
Partitions 136
Enregistrer 138
Résultats 139
Options 140
Vue du modèle 141
Espace des descriptives 142
Importance des variables 146
Pairs 147
Distances du voisin le plus proche 147
Carte des quadrants 148
Journal d'erreur de sélection des descriptives 149
Journal d'erreur de la sélection de k 150
Journal d'erreur de sélection de k et des descriptives 151
Le tableau de classification 152
Récapitulatif d'erreur 152

21 Analyse discriminante 153

Définition d'intervalles pour l'analyse discriminante 155
Sélection des observations pour l'analyse discriminante 155
Statistiques de l'analyse discriminante 156
Méthode pas à pas de l'analyse discriminante 157
Analyse discriminante : Classement 158

Enregistrement de l'analyse discriminante 160
Fonctionnalités supplémentaires de la commande DISCRIMINANT. 160

22 Analyse factorielle 161

Sélection des observations pour l'analyse factorielle 162
Descriptives d'analyse factorielle 163
Extraction d'analyse factorielle. 164
Rotation d'analyse factorielle 166
Scores d'analyse factorielle 167
Options d'analyse factorielle. 168
Fonctionnalités supplémentaires de la commande FACTOR. 168

23 Choix d'une procédure de classification 169

24 Analyse TwoStep Cluster 170

Options de la procédure d'analyse TwoStep Cluster 173
Résultats de l'analyse TwoStep Cluster. 175
Le viewer de classes 176
Viewer de classes 177
Navigation dans le viewer de classes 186
Filtrage des enregistrements 187

25 Classification hiérarchique 189

Méthode de classification hiérarchique 190
Statistiques de la classification hiérarchique 191
Diagrammes (graphiques) de classification hiérarchique 192
Sauvegarde des nouvelles variables de classification hiérarchique 193
Fonctionnalités supplémentaires de la syntaxe de commande CLUSTER 193

26 Nuées dynamiques 194

Efficacité de la classification en nuées dynamiques . . . 196
Itération de la classification en nuées dynamiques . . . 196
Enregistrement des analyses de classes de nuées dynamiques . . . 197
Options d'analyses des classes de nuées dynamiques . . . 197
Fonctionnalités supplémentaires de la commande QUICK CLUSTER . . . 198

27 Tests non paramétriques 199

Tests non paramétriques à un échantillon . . . 199
Obtenir des tests non paramétriques à un échantillon . . . 200
Onglet Champs . . . 200
Onglet Paramètres . . . 201
Tests non paramétriques pour échantillons indépendants . . . 206
Obtenir des tests non paramétriques pour échantillons indépendants . . . 207
Onglet Champs . . . 208
Onglet Paramètres . . . 208
Tests non paramétriques pour échantillons liés . . . 211
Obtenir des tests non paramétriques pour échantillons liés . . . 212
Onglet Champs . . . 213
Onglet Paramètres . . . 213
Vue du modèle . . . 218
Récapitulatif d'hypothèses . . . 219
Récapitulatif de l'intervalle de confiance . . . 220
Test à un échantillon . . . 221
Test pour échantillons liés . . . 225
Test pour échantillons indépendants . . . 232
Informations sur les champs qualitatifs . . . 240
Informations sur les champs continus . . . 241
Comparaisons par paire . . . 242
Sous-ensembles homogènes . . . 243
Fonctions supplémentaires de la commande NPTESTS . . . 244
Boîtes de dialogue ancienne version . . . 244
Test du Khi-deux . . . 244
Test binomial . . . 262
Suites en séquences . . . 264
Test Kolmogorov-Smirnov pour un échantillon . . . 266
Tests pour deux échantillons indépendants . . . 268
Tests pour deux échantillons liés . . . 271

Tests pour plusieurs échantillons indépendants . . . 273
Tests pour plusieurs échantillons liés. . . . 276
Test binomial . . . 262
Suites en séquences . . . 264
Test Kolmogorov-Smirnov pour un échantillon . . . 266
Tests pour deux échantillons indépendants . . . 268
Tests pour deux échantillons liés . . . 271
Tests pour plusieurs échantillons indépendants . . . 273
Tests pour plusieurs échantillons liés. . . . 276

28 Analyse des réponses multiples 279

Définition de vecteurs multiréponses . . . 280
Tableaux de fréquences des réponses multiples . . . 281
Tableaux croisés des réponses multiples . . . 283
Définir Intervalles Tableaux croisés De réponses multiples . . . 284
Options Tableaux croisés de réponses multiples . . . 285
Fonctionnalités supplémentaires de la commande MULT RESPONSE . . . 286

29 Tableaux de Résultats 287

Tableaux de bord en lignes . . . 287
Pour obtenir un rapport récapitulatif : Récapitulatifs en lignes . . . 288
Format des Colonnes de données/Ventilations des Tableaux de bord . . . 289
Fonctions récapitulatives des Tableaux pour/Fonctions récapitulatives Finales . . . 289
Options de Ventilation de Tableau de Bord . . . 290
Options du Tableau de bord . . . 291
Présentation du Tableau de bord . . . 291
Titres du Tableau de bord. . . . 293
Tableaux de bord en colonnes . . . 293
Pour obtenir un rapport récapitulatif : Récapitulatifs en colonnes . . . 294
Fonction récapitulative des Colonnes de données . . . 295
Fonction élémentaire des Colonnes de Données pour colonne de total . . . 296
Format des Colonnes du Tableau de bord. . . . 297
Tableaux de bord en Colonnes : Options de Ventilation. . . . 297
Options des Tableaux de bord en Colonnes . . . 298
Présentation du Tableau de bord en Colonnes . . . 298
Fonctionnalités supplémentaires de la commande REPORT. . . . 298

30 Analyse de fiabilité 300

Statistiques de l'analyse de fiabilité . . . 301
Fonctionnalités supplémentaires de la commande RELIABILITY . . . 303

31 Positionnement multidimensionnel 304

Forme des données du positionnement multidimensionnel . . . 306
Positionnement multidimensionnel : créer une mesure . . . 306
Modèle de positionnement multidimensionnel . . . 307
Positionnement multidimensionnel : Options . . . 308
Fonctionnalités supplémentaires de la commande ALSCAL . . . 309

32 Statistiques de ratio 310

Statistiques de ratio . . . 311

33 Courbes ROC 313

Courbe ROC : Options . . . 315

34 Simulation 316

Pour concevoir une simulation basée sur un fichier de modèle . . . 317
Pour concevoir une simulation basée sur des équations personnalisées . . . 317
Pour exécuter une simulation à partir d'un plan de simulation . . . 318
Générateur de simulation . . . 319
    Onglet Modèle . . . 319
    Onglet Simulation . . . 322
Boîte de dialogue Exécuter la simulation . . . 331
    Onglet Simulation . . . 332
    Onglet Résultat . . . 333
Utilisation du résultat graphique créé par la simulation . . . 335
    Options de diagramme . . . 335

Annexe

A Remarques 337

Index 340

Chapitre
1

Livre de codes

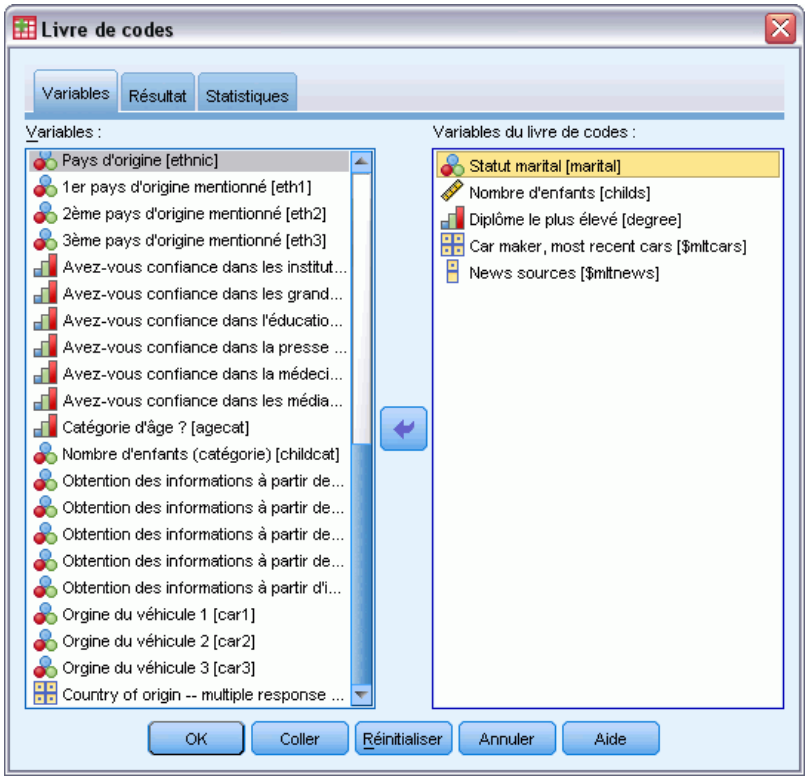
Le livre des codes indique les informations du dictionnaire, telles que les noms de variables, les étiquettes de variables, les étiquettes de valeurs, les valeurs manquantes, ainsi que les statistiques récapitulatives de toutes les variables (ou celles spécifiées) et les vecteurs multiréponses dans l'ensemble de données actif. Pour les variables ordinales et nominales ainsi que pour les vecteurs multiréponses, les statistiques récapitulatives comprennent les effectifs et les pourcentages. Pour les variables d'échelle, les statistiques récapitulatives comprennent la moyenne, l'écart-type et les quartiles.

Remarque : Le livre des codes ignore l'état de fichier scindé. Ceci comprend les groupes de fichiers scindés crées pour l'imputation multiple de valeurs manquantes (disponible dans l'option complémentaire Valeurs manquantes).

Pour obtenir un livre des codes

- A partir des menus, sélectionnez :
  Analyse > Rapports > Livre de codes
- Cliquez sur l'onglet Variables.

Figure 1-1
Boîte de dialogue Livre des codes, onglet Variables



▶ Sélectionnez une ou plusieurs variables et/ou des vecteurs multiréponses.

Sinon, vous pouvez :

- Contrôlez les informations de variables affichées.
- Contrôlez les statistiques affichées (ou excluez toutes les statistiques récapitulatives).
- Contrôlez l'ordre d'affichage des variables et des vecteurs multiréponses.
- Modifiez le niveau de mesure de toute variable dans la liste source afin de modifier les statistiques récapitulatives affichées. Pour plus d'informations, reportez-vous à la section Onglet Statistiques du livre des codes sur p. 5.

Modification des niveaux de mesure

Vous pouvez modifier temporairement le niveau de mesure des variables. (Vous ne pouvez pas modifier celui des vecteurs multiréponses. Ils sont toujours traitées comme nominaux.)

▶ Cliquez avec le bouton droit sur une variable dans la liste source.

▶ Dans le menu contextuel, sélectionnez un niveau de mesure.

Ceci permet de modifier temporairement le niveau de mesure. Concrètement, cela n'est utile que pour les variables numériques. Le niveau de mesure des variables de chaîne est limité aux variables nominales ou ordinales qui sont traitées de la même façon par la procédure du livre des codes.

Onglet Résultats du livre des codes

L'onglet Résultats contrôle les informations de variables disponibles pour chaque variable et vecteurs multiréponses, leur ordre d'affichage et le contenu de la table d'informations des fichiers en option.

Figure 1-2
Boîte de dialogue Livre des codes, onglet Résultats

Informations sur les variables

Ceci permet de contrôler les informations du dictionnaire affichées pour chaque variable.

Position. Un nombre entier qui représente la position de la variable dans l'ordre des fichiers. Non disponible pour les vecteurs multiréponses.

Etiquette. L'étiquette descriptive associée à la variable ou au vecteur multiréponses.

Type. Type de données fondamental. Est *Numérique*, *Chaîne*, ou *Vecteur multiréponses*.

Format. Le format d'affichage de la variable, tel que *A4*, *F8.2* ou *DATE11*. Non disponible pour les vecteurs multiréponses.

Niveau de mesure. Les valeurs possibles sont *Nominale*, *Ordinale*, *Echelle* et *Inconnue*. La valeur affichée est le niveau de mesure stocké dans le dictionnaire et elle n'est pas affectée par tout remplacement de niveau de mesure temporaire spécifié en changeant le niveau de mesure dans la liste de variable source de l'onglet Variables. Non disponible pour les vecteurs multiréponses.

Remarque : Le niveau de mesure des variables numériques peut être « inconnu » avant le premier passage de données lorsque le niveau de mesure n'a pas été explicitement défini, par exemple pour la lecture de données à partir d'une source externe ou des variables récemment créées.

Rôle. Certaines boîtes de dialogue prennent en charge la fonction de présélection de variables pour une analyse basée sur des rôles définis.

Etiquettes de valeurs. Etiquettes descriptives associées à des valeurs de données spécifiques.

- Si l'option Effectif ou Pourcentage est sélectionnée dans l'onglet Statistiques, les étiquettes de valeurs définies sont comprises dans les résultats même si vous ne sélectionnez pas Etiquettes de valeur ici.
- Pour les vecteurs de dichotomies multiples, les « étiquettes de valeur » sont les étiquettes des variables élémentaires du vecteur soit les étiquettes des valeurs comptées, selon la définition du vecteur.

Valeurs manquantes. Valeurs manquantes définies par l'utilisateur. Si l'option Effectif ou Pourcentage est sélectionnée dans l'onglet Statistiques, les étiquettes de valeurs définies sont comprises dans les résultats même si vous ne sélectionnez pas Valeurs manquantes ici. Non disponible pour les vecteurs multiréponses.

Attributs Personnalisés. Attributs de variable personnalisés. Les résultats comprennent à la fois les noms et les valeurs pour tout attribut de variables personnalisé associé à chaque variable. Non disponible pour les vecteurs multiréponses.

Attributs réservés. Attributs de variables système réservés. Vous pouvez afficher les attributs système, mais vous ne devez pas les modifier. Les noms des attributs système commencent par un signe dollar ($) . Les attributs hors affichage, avec les noms qui commencent par « @ » ou « $@ », ne sont pas inclus. Les résultats comprennent à la fois les noms et les valeurs pour tout attribut système associé à chaque variable. Non disponible pour les vecteurs multiréponses.

Informations sur les fichiers

La table d'informations de fichiers en option peut comprendre l'un des attributs de fichiers suivants :

Nom de fichier. Nom du fichier de données IBM® SPSS® Statistics. Si l'ensemble de données n'a jamais été enregistré au format SPSS Statistics, aucun nom de fichier de données n'est disponible. (Si aucun nom de fichier n'est affiché dans la barre de titre de la fenêtre Editeur de données, l'ensemble de données actif ne comporte pas de nom de fichier).

Emplacement. Emplacement du répertoire (dossier) du fichier de données SPSS Statistics. Si l'ensemble de données n'a jamais été enregistré au format SPSS Statistics, aucun n'emplacement n'est disponible.

Nombre d'observations. Nombre d'observations dans l'ensemble de données actif. Ceci est le nombre total d'observations, y compris celles qui peuvent être exclues des statistiques récapitulatives en raison des conditions de filtrage.

Etiquette. Ceci est le fichier d'étiquette (si disponible) défini par la commande `FILE LABEL`.

Documents. Texte de document de fichier de données.

Etat de la pondération : Si la pondération est activée, le nom de la variable de pondération est affiché.

Attributs Personnalisés. Attributs de fichiers de données personnalisés définis par l'utilisateur. Attributs de fichiers de données définis avec la commande `DATAFILE ATTRIBUTE`.

Attributs réservés. Attributs de fichiers de données système réservés. Vous pouvez afficher les attributs système, mais vous ne devez pas les modifier. Les noms des attributs système commencent par un signe dollar ($) . Les attributs hors affichage, avec les noms qui commencent par « @ » ou « $@ », ne sont pas inclus. Les résultats incluent les noms et les valeurs pour tout attribut de fichiers de données système.

Ordre d'affichage des variables

Les alternatives suivantes sont disponibles pour contrôler l'ordre d'affichage des variables et des vecteurs multiréponses :

Alphabétique. Ordre alphabétique par nom de variable.

Fichier. L'ordre d'affichage des variables dans l'ensemble de données (leur ordre d'affichage dans l'Editeur de données). Dans l'ordre croissant, les vecteurs multiréponses sont affichés en dernier, après toutes les variables sélectionnées.

Niveau de mesure. Trier par niveau de mesure. Ceci crée quatre groupes de tri : nominal, ordinal, échelle et inconnu. Les vecteurs multiréponses sont traités comme nominaux

Remarque : Le niveau de mesure des variables numériques peut être « inconnu » avant le premier passage de données lorsque le niveau de mesure n'a pas été explicitement défini, par exemple pour la lecture de données à partir d'une source externe ou des variables récemment créées.

Liste des variables. L'ordre d'affichage des variables et des vecteurs multiréponses dans la liste des variables sélectionnées de l'onglet Variables.

Nom d'attribut personnalisé. La liste des options d'ordre de tri comprend aussi le nom des attributs de variables personnalisés définis par l'utilisateur. Dans l'ordre croissant, les variables dont le tri des attributs ne figure pas en haut, puis celles dont la valeur n'est pas définie pour l'attribut, puis celles avec des valeurs définies pour l'attribut dans l'ordre alphabétique des valeurs.

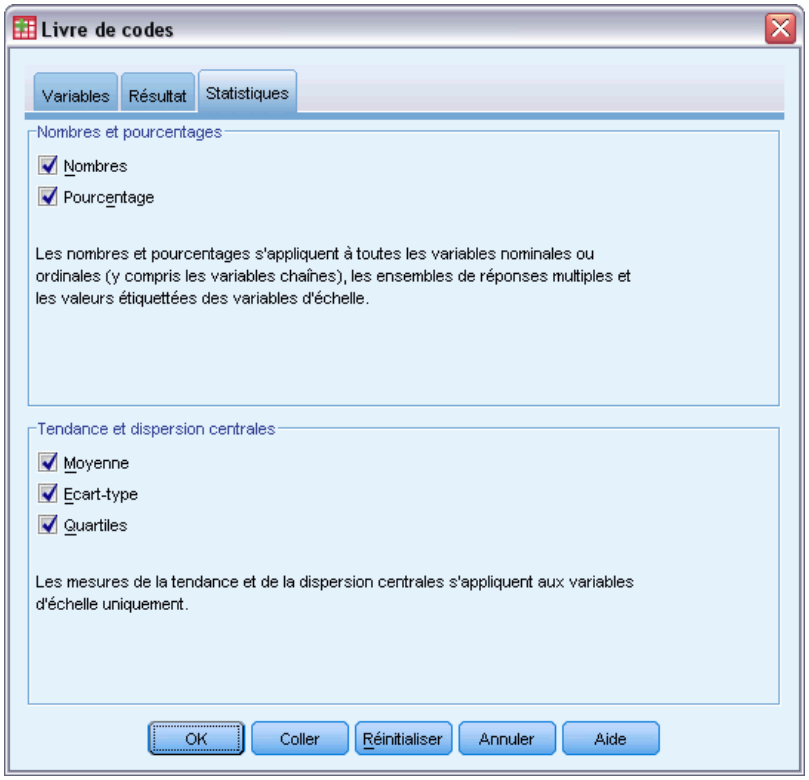
Nombre maximal de catégories

Si les résultats comprennent les étiquettes de valeurs, les effectifs, ou les pourcentages pour chaque valeur unique, vous pouvez supprimer ces informations de la table si le nombre de valeurs dépasse la valeur indiquée. Par défaut, ces informations sont supprimées si le nombre de valeurs uniques de la variable dépasse 200.

Onglet Statistiques du livre des codes

L'onglet Statistiques permet de contrôler les statistiques récapitulatives comprises dans les résultats, ou de supprimer entièrement l'affichage des statistiques récapitulatives.

Figure 1-3
Boîte de dialogue Livre des codes, onglet Statistiques



Nombres et pourcentages

Pour les variables nominales et ordinales, les vecteurs multiréponses et les valeurs d'étiquette des variables d'échelle, les statistiques disponibles sont :

Effectif. Effectif ou nombre d'observations possédant chaque valeur (ou plage de valeurs) d'une variable.

Pourcentage. Pourcentage d'observations ayant une valeur particulière.

Tendance et dispersion centrales

Pour les variables d'échelle, les statistiques disponibles sont :

Moyenne. Mesure de la tendance centrale. Moyenne arithmétique ; somme divisée par le nombre d'observations.

Ecart•type. Mesure de dispersion par rapport à la moyenne. Dans le cas d'une distribution normale, 68 % des observations se situent à l'intérieur d'un écart-type de la moyenne et 95 % se situent à l'intérieur de deux écarts-types. Par exemple, si la moyenne d'âge est de 45 avec un écart-type égal à 10, une distribution normale verra 95 % des observations se situer entre 25 et 65.

Quartiles. Valeurs correspondant aux 25ème, 50ème et 75ème centiles.

Remarque : vous pouvez modifier temporairement le niveau de mesure associé à une variable (et par conséquent modifier les statistiques récapitulatives affichées pour cette variable) dans la liste de variables source de l'onglet Variables.

Chapitre
2

Effectifs

La procédure Fréquences permet d'obtenir des affichages statistiques et graphiques qui servent à décrire de nombreux types de variables. La procédure Fréquences peut jouer un rôle lorsque vous prenez connaissance de vos données.

Pour obtenir un rapport des fréquences et un diagramme en bâtons, vous pouvez trier les différentes valeurs par ordre croissant ou décroissant, ou bien classer les modalités en fonction de leurs fréquences. Le rapport de fréquences peut être supprimé lorsqu'une variable a plusieurs valeurs distinctes. Vous pouvez étiqueter les diagrammes avec des fréquences (par défaut) ou des pourcentages.

Exemple : Quelle est la répartition de la clientèle d'une société selon le type d'industrie dont elle fait partie ? Le résultat pourrait vous apprendre que votre clientèle est composée à 37,5 % d'organismes d'état, à 24,9 % de sociétés commerciales, à 28,1 % d'établissements universitaires et à 9,4 % du secteur de la santé. Pour des données continues et quantitatives, comme par exemple les revenus des ventes, vous pourriez constater que la moyenne de vente par produit est de 3 576 € avec un écart-type de 1 078 €.

Diagrammes et statistiques : Effectifs de fréquence, pourcentages, pourcentages cumulés, moyenne, médiane, mode, somme, écart-type, variance, intervalle, valeurs minimale et maximale, erreur standard de la moyenne, asymétrie et aplatissement (avec leurs erreurs standard), quartiles, centiles choisis par l'utilisateur, diagrammes en bâtons, diagrammes en secteurs et histogrammes.

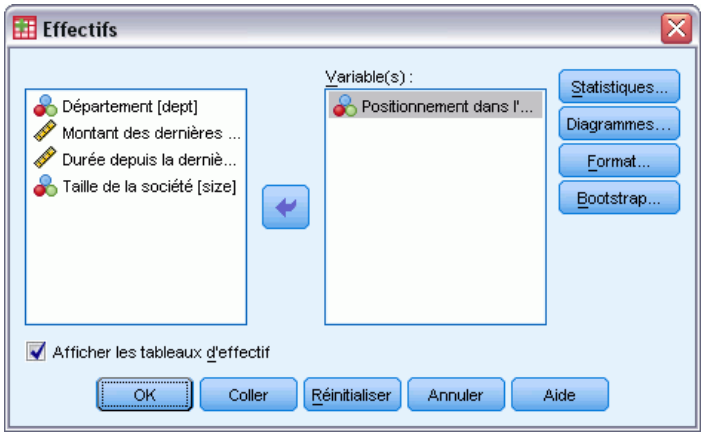
Données. Utilisez des codes numériques ou alphanumériques pour coder les variables qualitatives (mesures de niveau nominal ou ordinal).

Hypothèses : Les tabulations et les pourcentages fournissent une description utile sur les données de n'importe quelle distribution, particulièrement pour les variables disposant de modalités triées ou non. Certaines des statistiques récapitulatives facultatives, telles que la moyenne et l'écart-type, sont fondées sur la théorie de normalité et sont appropriées pour des variables quantitatives avec une distribution symétrique. Les statistiques de base, telles que la médiane, les quartiles et les centiles, sont appropriées pour les variables quantitatives, qu'elles répondent ou non au critère de normalité.

Pour obtenir des tableaux de fréquences

- A partir des menus, sélectionnez :
  Analyse > Statistiques descriptives > Effectifs

Figure 2-1
Boîte de dialogue Fréquences complexes



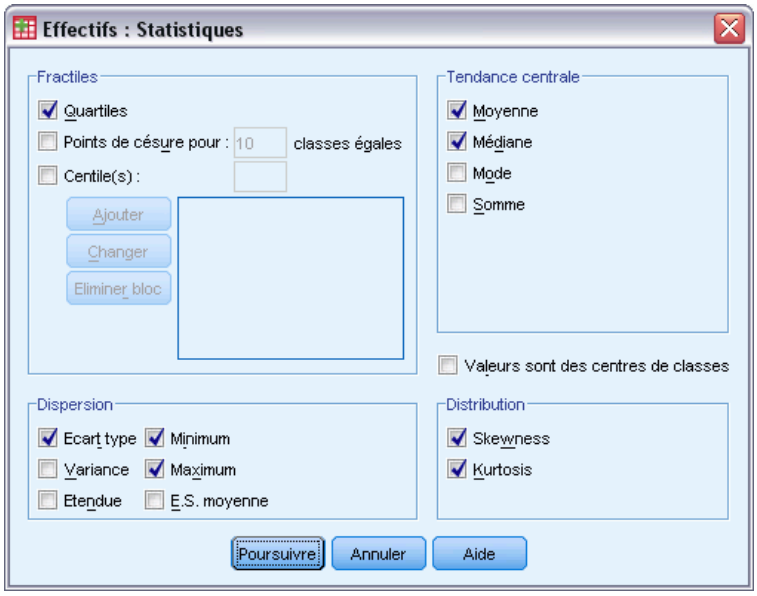
- Sélectionnez une ou plusieurs variables qualitatives ou quantitatives.

Sinon, vous pouvez :

- Cliquer sur Statistiques pour obtenir des statistiques descriptives pour des variables quantitatives.
- Cliquer sur Diagrammes pour obtenir des diagrammes en bâtons, des diagrammes en secteurs ou des histogrammes.
- Cliquer sur Format pour définir l'ordre de présentation des résultats.

Statistiques des fréquences

Figure 2-2
Boîte de dialogue Fréquences : Statistiques



Fractiles : Valeurs d'une variable quantitative qui divisent les données triées en classes par centième. Les quartiles (25ième, 50ième et 75ième centiles) divisent les observations en quatre classes de taille égale. Si vous souhaitez un nombre égal de classes différent de quatre, sélectionnez Partition en n classes égales. Vous pouvez également spécifier des centiles particuliers (par exemple, le 95ième centile, valeur au-dessus de 95 % des observations).

Tendance centrale : Les statistiques décrivant la position de la distribution comprennent la Moyenne, la Médiane, le Mode et la Somme de toutes les valeurs.

- **Moyenne.** Mesure de la tendance centrale. Moyenne arithmétique ; somme divisée par le nombre d'observations.
- **Médiane.** Valeur au-dessus ou au-dessous de laquelle se trouvent la moitié des observations ; 50e centile. Si le nombre d'observations est pair, la médiane correspond à la moyenne des deux observations du milieu lorsqu'elles sont triées dans l'ordre croissant ou décroissant. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées.
- **Mode.** Valeur qui revient le plus fréquemment. Si plusieurs valeurs partagent la plus grande fréquence d'occurrence, chacune d'elles constitue un mode. La procédure Effectifs ne rend compte que du plus petit mode.
- **Somme.** Somme ou total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.

Dispersion : Les statistiques mesurant la variance ou la dispersion dans les données, comprennent l'écart-type, la variance, l'intervalle, le minimum, le maximum et l'erreur standard (ES) de la moyenne.

- **Ecart type.** Mesure de dispersion par rapport à la moyenne. Dans le cas d'une distribution normale, 68 % des observations se situent à l'intérieur d'un écart-type de la moyenne et 95 % se situent à l'intérieur de deux écarts-types. Par exemple, si la moyenne d'âge est de 45 avec un écart-type égal à 10, une distribution normale verra 95 % des observations se situer entre 25 et 65.
- **Variance.** Mesure de dispersion autour de la moyenne, égale à la somme des carrés des écarts par rapport à la moyenne, divisée par le nombre d'observations moins un. La variance se mesure en unités, qui sont égales au carré des unités de la variable.
- **Intervalle.** Différence entre la valeur maximale et la valeur minimale d'une variable numérique (maximum–minimum).
- **Minimum.** Valeur la plus petite d'une variable numérique.
- **Maximum.** Plus grande valeur d'une variable numérique.
- **ES Moyenne.** Mesure du degré de variation de la moyenne d'un échantillon à l'autre au sein d'une même distribution. Cette mesure permet de comparer approximativement la moyenne observée avec une valeur hypothétique (autrement dit, vous pouvez conclure que ces deux valeurs sont différentes si le rapport de la différence avec l'erreur standard est inférieur à -2 ou supérieur à +2).

Distribution : L'Asymétrie et l'Aplatissement sont des statistiques qui décrivent la forme et la symétrie de la distribution. Ces statistiques sont présentées avec leurs erreurs standard.

- **Asymétrie.** Mesure de l'asymétrie d'une distribution. La distribution normale est symétrique et possède une valeur d'asymétrie égale à 0. Une distribution dont la valeur d'asymétrie est positive présente une extrémité droite allongée. Une distribution caractérisée par une importante asymétrie négative présente une extrémité gauche plus allongée. Pour simplifier, une valeur d'asymétrie deux fois supérieure à l'erreur standard correspond à une absence de symétrie.
- **Aplatissement.** Mesure de l'étendue du regroupement des observations autour d'un point central. Dans le cas d'une distribution normale, la valeur de la statistique d'aplatissement est égale à zéro. Un aplatissement positif indique que par rapport a une distribution normale, les observations sont plus regroupées au centre et présentent des extrémités plus fines atteignant les valeurs extrêmes de la distribution. La distribution leptokurtique présente des extrémités plus épaisses que dans le cas d'une distribution normale. Un aplatissement négatif indique que les observations sont moins regroupées au centre et présentent des extrémités plus épaisses atteignant les valeurs extrêmes de la distribution. La distribution platykurtique présentent des extrémités plus fines que dans le cas d'une distribution normale.

Valeurs sont des centres de classes : Si les valeurs dans vos données représentent des centres de classes (par exemple, les âges des individus trentenaires sont représentés par le code 35), sélectionnez cette option pour estimer la médiane et les centiles des données originales, non regroupées.

Diagrammes des fréquences

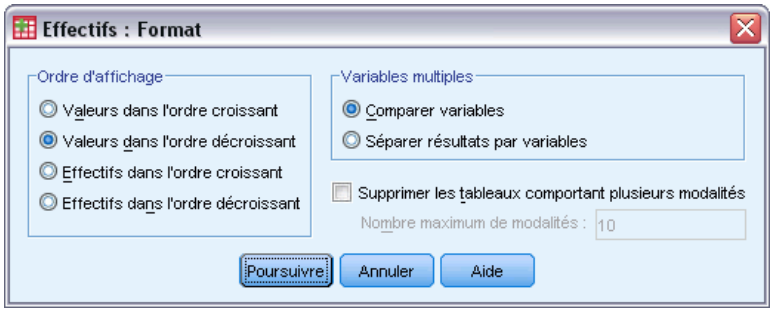
Figure 2-3
Boîte de dialogue Fréquences : Diagrammes

Type de diagramme : Un diagramme en secteurs montre la participation de chaque partie à l'ensemble. Chaque secteur du diagramme correspond à un groupe défini par une simple variable de regroupement. Un diagramme en bâtons montre l'effectif de chaque valeur ou de chaque modalité sous la forme d'un bâton distinct, ce qui vous permet de comparer les modalités visuellement. Un histogramme est également constitué de bâtons mais ils sont répartis à intervalles égaux. La hauteur de chaque bâton représente l'effectif des valeurs d'une variable quantitative appartenant à l'intervalle. Un histogramme montre la forme, le centre et la dispersion de la distribution. Si vous superposez une courbe normale sur l'histogramme, vous pouvez déterminer si les données sont distribuées normalement.

Valeurs du diagramme : Dans les diagrammes en bâtons, l'axe peut être étiqueté par fréquences ou pourcentages de fréquence.

Format des fréquences

Figure 2-4
Boîte de dialogue Fréquences: Format



Ordre d'affichage : Le tableau de fréquences peut être affiché en fonction des valeurs réelles des données ou de l'effectif (fréquence d'occurrence) de ces valeurs et organisé par valeurs croissantes ou décroissantes. Cependant, si vous demandez un histogramme ou des centiles, Effectifs part du principe que la variable est quantitative et affiche ses valeurs par ordre croissant.

Variables multiples : Si vous créez des tableaux statistiques pour des variables multiples, vous pouvez afficher toutes les variables dans un tableau unique (Comparer variables) ou bien afficher un tableau statistique séparé pour chaque variable (Séparer résultats par variables).

Supprimer les tableaux avec plus de n modalités : Cette option évite l'affichage des tableaux ayant plus que le nombre spécifié de valeurs.

Chapitre
3

Descriptives

La procédure Descriptive affiche les résumés de statistiques univariées pour plusieurs variables en un seul tableau et calcule les valeurs standardisées (scores *z*). Les variables peuvent être ordonnées en fonction de la taille de leurs moyennes (en ordre ascendant ou descendant), alphabétiquement ou selon l'ordre dans lequel vous avez sélectionné les variables (par défaut).

Lorsque les scores *z* sont enregistrés, ils sont ajoutés aux données dans l'éditeur de données et sont disponibles pour les diagrammes, les listes de données et les analyses. Lorsque les variables sont enregistrées avec des unités différentes (par exemple, produit domestique brut par personne et pourcentage de la population sachant lire et écrire), une transformation en score *z* place les variables sur une échelle commune pour que la comparaison soit plus facile.

Exemple : Si chaque observation dans vos données contient les totaux des ventes quotidiennes pour chacun des membres du personnel commercial (par exemple, une entrée pour Bob, une pour Kim et une pour Brian) rapportés chaque jour pendant plusieurs mois, la procédure Descriptives peut calculer les ventes quotidiennes moyennes pour chacun des membres du personnel et ordonner les résultats de la moyenne des ventes la plus élevée à la plus basse.

Statistiques : Taille de l'échantillon, moyenne, minimum, maximum, écart-type, variance, intervalle, somme, erreur standard de la moyenne, et aplatissement et asymétrie avec leurs erreurs standards (ES).

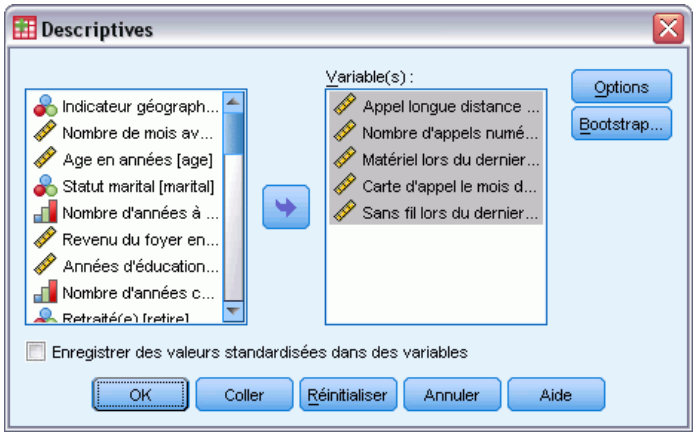
Données. Utilisez des variables numériques après les avoir visualisées graphiquement en cherchant des erreurs d'enregistrement, les valeurs éloignées et les anomalies de distribution. La procédure Descriptives est très efficace pour les gros fichiers (milliers d'observations).

Hypothèses : La plupart des statistiques disponibles (y compris les écarts *z*) sont basées sur une théorie normale et conviennent pour des variables continues (mesures de niveau d'intervalle ou de rapport) avec distribution symétrique. Evitez les variables avec des modalités désordonnées ou des répartitions asymétriques. La distribution des écarts *z* a la même forme que celle des données d'origine. Ainsi, le calcul des écarts *z* n'est pas une solution aux données posant des problèmes.

Pour obtenir des statistiques descriptives

- A partir des menus, sélectionnez :
  Analyse > Statistiques descriptives > Descriptives

Figure 3-1
Boîte de dialogue Descriptives



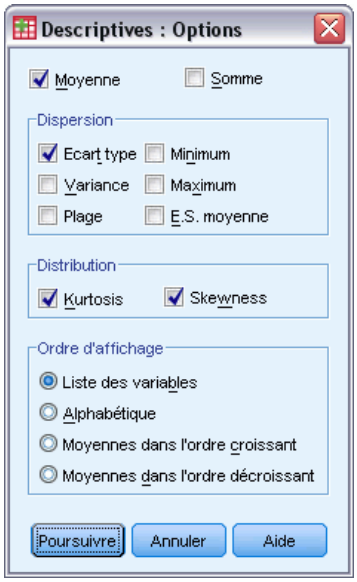
▶ Sélectionnez une ou plusieurs variables.

Sinon, vous pouvez :

- Cliquez sur Enregistrer des valeurs standardisées dans des variables pour enregistrer les écarts *z* comme nouvelles variables.
- Cliquer sur Options pour les statistiques optionnelles et l'ordre d'affichage.

Options Descriptives

Figure 3-2
Boîte de dialogue Descriptives : Options



Moyenne et somme : La moyenne ou moyenne arithmétique s'affiche par défaut.

Dispersion : Les statistiques qui mesurent l'étendue ou les variations dans les données comprennent l'écart-type, la variance, l'intervalle, le minimum, le maximum, et l'erreur standard (ES) de la moyenne.

- **Ecart type.** Mesure de dispersion par rapport à la moyenne. Dans le cas d'une distribution normale, 68 % des observations se situent à l'intérieur d'un écart-type de la moyenne et 95 % se situent à l'intérieur de deux écarts-types. Par exemple, si la moyenne d'âge est de 45 avec un écart-type égal à 10, une distribution normale verra 95 % des observations se situer entre 25 et 65.
- **Variance.** Mesure de dispersion autour de la moyenne, égale à la somme des carrés des écarts par rapport à la moyenne, divisée par le nombre d'observations moins un. La variance se mesure en unités, qui sont égales au carré des unités de la variable.
- **Intervalle.** Différence entre la valeur maximale et la valeur minimale d'une variable numérique (maximum–minimum).
- **Minimum.** Valeur la plus petite d'une variable numérique.
- **Maximum.** Plus grande valeur d'une variable numérique.
- **E.S. moyenne.** Mesure du degré de variation de la moyenne d'un échantillon à l'autre au sein d'une même distribution. Cette mesure permet de comparer approximativement la moyenne observée avec une valeur hypothétique (autrement dit, vous pouvez conclure que ces deux valeurs sont différentes si le rapport de la différence avec l'erreur standard est inférieur à -2 ou supérieur à +2).

Distribution : L'aplatissement et l'asymétrie sont des statistiques qui caractérisent la forme et la symétrie de la distribution. Ces statistiques sont présentées avec leurs erreurs standard.

- **Aplatissement.** Mesure de l'étendue du regroupement des observations autour d'un point central. Dans le cas d'une distribution normale, la valeur de la statistique d'aplatissement est égale à zéro. Un aplatissement positif indique que par rapport a une distribution normale, les observations sont plus regroupées au centre et présentent des extrémités plus fines atteignant les valeurs extrêmes de la distribution. La distribution leptokurtique présente des extrémités plus épaisses que dans le cas d'une distribution normale. Un aplatissement négatif indique que les observations sont moins regroupées au centre et présentent des extrémités plus épaisses atteignant les valeurs extrêmes de la distribution. La distribution platykurtique présentent des extrémités plus fines que dans le cas d'une distribution normale.
- **Asymétrie.** Mesure de l'asymétrie d'une distribution. La distribution normale est symétrique et possède une valeur d'asymétrie égale à 0. Une distribution dont la valeur d'asymétrie est positive présente une extrémité droite allongée. Une distribution caractérisée par une importante asymétrie négative présente une extrémité gauche plus allongée. Pour simplifier, une valeur d'asymétrie deux fois supérieure à l'erreur standard correspond à une absence de symétrie.

Ordre d'affichage : Par défaut, les variables s'affichent dans l'ordre dans lequel vous les avez sélectionnées. En option, vous pouvez afficher les variables alphabétiquement, par moyennes croissantes ou par moyennes décroissantes.

Fonctionnalités supplémentaires de la commande DESCRIPTIVES

Le langage de syntaxe de commande vous permet aussi de :

- Enregistrer les coordonnées standardisées (écarts *z*) pour certaines variables uniquement (à l'aide de la sous-commande `VARIABLES`).
- Spécifier le nom des nouvelles variables contenant des coordonnées standardisées (à l'aide de la sous-commande `VARIABLES`).
- Exclure de l'analyse les observations ayant des valeurs manquantes pour n'importe quelle variable (à l'aide de la sous-commande `MISSING`).
- Trier les variables affichées en utilisant la valeur d'une statistique, et pas uniquement la moyenne (à l'aide de la sous-commande `SORT`).

Reportez-vous à la *Référence de syntaxe de commande* pour une information complète concernant la syntaxe.

Chapitre
4

Explorer

La procédure Explorer produit des résumés statistiques et des affichages graphiques pour toutes vos observations ou séparément pour des groupes d'observations. Il existe plusieurs raisons pour utiliser la procédure Explorer : le filtrage de données, l'identification des valeurs éloignées, la description, la vérification d'hypothèses et la caractérisation des différences parmi les sous populations (groupes d'observations). Le filtrage de données peut vous indiquer les valeurs inhabituelles, les valeurs extrêmes, les trous dans les données ou d'autres particularités. L'exploration des données peut vous aider à déterminer si les techniques statistiques que vous envisagez d'utiliser pour l'analyse de vos données sont appropriées. L'exploration peut indiquer que vous avez besoin de transformer les données si la technique nécessite une répartition gaussienne. Vous pouvez également choisir d'utiliser des tests non paramétriques.

Exemple : Examiner la distribution des temps d'apprentissage pour les souris dans un labyrinthe avec quatre programmes de renforcement. Pour chacun des quatre groupes, vous pouvez voir si la répartition des temps est approximativement gaussienne et si les quatre variances sont égales. Vous pouvez aussi identifier les observations avec les cinq plus grands et les cinq plus petits temps. Les boîtes à moustaches et les diagrammes tige et feuille résument graphiquement la répartition des temps d'apprentissage pour chacun des groupes.

Diagrammes et statistiques : Moyenne, médiane, moyenne tronquée à 5 %, erreur standard, variance, écart-type, minimum, maximum, intervalle, intervalle interquartile, asymétrie et aplatissement avec leurs erreurs standard, intervalle de confiance pour la moyenne (et niveaux de confiance spécifiés), centiles, M-estimateur de Huber, Andrews, Hampel, Tukey, les cinq plus grandes et cinq plus petites valeurs, le Kolmogorov-Smirnov avec un seuil de signification Lilliefors pour tester la normalité, et la statistique Shapiro-Wilk. Boîtes à moustaches, diagrammes tige et feuille, histogrammes, diagrammes de répartition gaussienne, et dispersion/niveaux avec le test de Levene et les transformations.

Données. La procédure d'Explorer peut être utilisée pour les variables quantitatives (Mesures de niveaux d'intervalle ou de rapport). Une variable active (utilisée pour répartir les données en groupes d'observations) doit avoir un nombre raisonnable de valeurs distinctes (modalités). Ces valeurs peuvent être des chaînes de caractères courtes ou numériques. La variable d'étiquette par observation, utilisée pour étiqueter les valeurs extrêmes dans les boîtes à moustache, peut être de courtes chaînes de caractères, de longues chaînes de caractères (15 premiers octets) ou numériques.

Hypothèses : La distribution de vos données ne doit pas obligatoirement être symétrique ou gaussienne.

Pour explorer vos données

- A partir des menus, sélectionnez :
  Analyse > Statistiques descriptives > Explorer

Figure 4-1
Boîte de dialogue Explorer

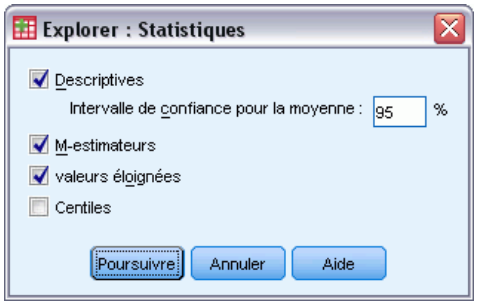
▶ Sélectionnez au moins une variable dépendante.

Sinon, vous pouvez :

- Sélectionner une ou plusieurs variables actives, dont les valeurs définiront les groupes d'observations.
- Sélectionner une variable d'identification pour étiqueter les observations.
- Cliquer sur Statistiques pour les M-estimateurs, les Valeurs éloignées, les Centiles et les tableaux de fréquences.
- Cliquez sur Diagrammes pour les histogrammes, les diagrammes de répartition gaussiens avec tests et la dispersion/niveau avec test de Levene.
- Cliquez sur Options pour le traitement des valeurs manquantes.

Statistiques d'Explorer

Figure 4-2
Boîte de dialogue Statistiques Explorer



Descriptives. Ces mesures de tendance centrale et de dispersion s'affichent par défaut. Les mesures de tendance centrale indiquent la position de la répartition. On y trouve la moyenne, la médiane et la moyenne tronquée à 5 %. Les mesures de dispersion montrent la dissimilarité des valeurs ; on y trouve l'erreur standard, la variance, l'écart-type, le minimum, le maximum,

l'intervalle, et l'intervalle interquartile. Les statistiques descriptives comprennent aussi les mesures de la forme des répartitions. L'asymétrie et l'aplatissement s'affichent avec leurs erreurs standard. L'intervalle du niveau de confiance à 95 % pour la moyenne s'affiche aussi. Vous pouvez spécifier un niveau de confiance différent.

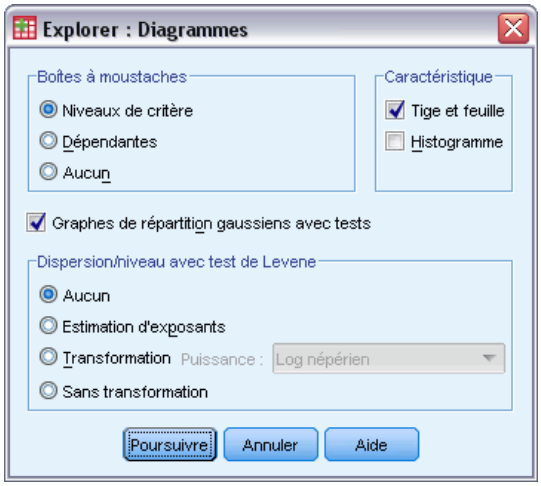
Moyennes pondérées : Estimations de la moyenne et de la médiane de l'échantillon pour estimer la localisation. Les estimateurs calculés diffèrent selon les pondérations qu'ils appliquent aux observations. M-estimateur de Huber, Andrew, Hampel, et Tukey apparaissent.

Valeurs éloignées : Affiche les cinq plus grandes et cinq plus petites valeurs avec les étiquettes d'observations.

Centiles : Affiche les valeurs pour le 5ième, 10ième, 25ième, 50ième, 75ième, 90ième, et 95ième centiles.

Diagrammes d'Explorer

Figure 4-3
Boîte de dialogue Explorer : Diagrammes



Boîtes à moustaches : Ces alternatives contrôlent l'affichage de boîtes à moustaches quand vous avez plus d'une variable dépendante. Niveaux de critère génère un affichage séparé pour chaque variable dépendante. Dans un affichage, les boîtes à moustache sont données pour chacun des groupes définis par une variable active. Dépendantes génère un affichage séparé pour chaque groupe défini par une variable active. Dans un affichage, les boîtes à moustaches s'affichent côte à côte pour chaque variable dépendante. Cet affichage est particulièrement utile lorsque les différentes variables représentent une seule caractéristique mesurée à des moments différents.

Caractéristique : Le groupe caractéristiques vous permet de choisir les diagrammes tige et feuille et les histogrammes.

Graphes de répartition gaussiens avec tests : Affiche les diagrammes de répartition gaussiens et les résidus. La statistique de Kolmogorov-Smirnov avec un seuil de signification Lilliefors pour le test de normalité s'affiche. Si des pondérations non entières sont spécifiées, la statistique Shapiro-Wilk est calculée lorsque la taille d'échantillon pondérée est comprise entre 3 et 50. En

cas de pondérations entières ou en l'absence de pondération, le calcul est effectué lorsque la taille d'échantillon pondérée est comprise entre 3 et 5 000.

Dispersion/niveau avec test de Levene : Contrôle les transformations de données pour les diagrammes de dispersion par niveau. Pour tous les diagrammes de dispersion par niveau, la pente de la ligne de régression et les tests de Levene portant sur l'homogénéité de la variance s'affichent. Si vous sélectionnez une transformation, les tests de Levene sont basés sur les données transformées. Si aucune variable active n'est sélectionnée, les diagrammes de dispersion par niveau ne sont pas produits. Estimation d'exposants produit un diagramme des logs naturels des intervalles interquartile opposés au logs naturels des médianes pour toutes les cellules, en même temps qu'une estimation de la transformation de l'exposant pour arriver à des variances égales dans les cellules. Un diagramme de dispersion par niveau aide à déterminer l'exposant pour qu'une transformation stabilise (rende plus égales) les variances entre groupes. Transformation Exposant vous permet de sélectionner une des alternatives de l'exposant, en suivant éventuellement les recommandations de l'estimation de l'exposant et de produire les diagrammes des données transformées. L'intervalle interquartile et la médiane des données transformées sont dessinés. Sans transformation produit des diagrammes de données brutes. Ceci est équivalent à une transformation avec une puissance de 1.

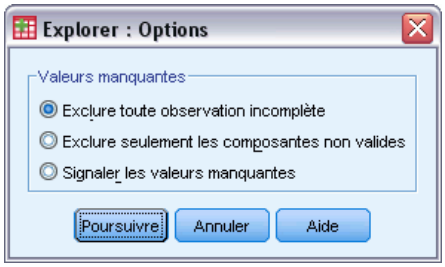
Transformations de l'exposant d'Explorer

Voici les transformations de l'exposant pour les diagrammes de dispersion par niveau. Pour transformer les données, vous devez sélectionner un exposant pour la transformation. Vous avez le choix entre les options suivantes :

- **Log népérien :** Transformation par log naturel. Il s'agit de la valeur par défaut.
- **1/racine carrée :** La réciproque de la racine carrée est calculée pour chaque valeur des données.
- **Réciproque :** La réciproque de chaque valeur des données est calculée.
- **Racine carrée :** La racine carrée de chaque valeur des données est calculée.
- **Carré :** Chaque valeur des données est élevée au carré.
- **Cube :** Chaque valeur des données est élevée au cube.

Options d'Explorer

Figure 4-4
Boîte de dialogue Explorer : Options



Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour l'une ou l'autre des variables dépendantes ou actives sont exclues de toutes les analyses. Il s'agit de la valeur par défaut.
- **Exclure seulement les composantes non valides :** Les observations sans valeur manquante pour une variable dans un groupe (cellule) sont inclues dans l'analyse de ce groupe. L'observation peut avoir des valeurs manquantes pour les variables utilisées dans d'autres groupes.
- **Signaler les valeurs manquantes :** Les valeurs manquantes pour les variables actives sont traitées comme une modalité séparée. Tout résultat est produit pour cette modalité supplémentaire. Les tableaux d'effectifs contiennent les modalités pour les valeurs manquantes. Les valeurs manquantes pour une variable active sont inclues, mais étiquetées comme manquantes.

Fonctionnalités supplémentaires de la commande EXAMINE

La procédure Explorer utilise la syntaxe de la commande `EXAMINE`. Le langage de syntaxe de commande vous permet aussi de :

- demander le total des résultats et des diagrammes en complément des résultats et diagrammes relatifs aux groupes définis par les variables actives (au moyen de la sous-commande `TOTAL`) ;
- spécifier une échelle commune pour un groupe de boîtes à moustaches (au moyen de la sous-commande `SCALE`) ;
- préciser les interactions des variables actives (au moyen de la sous-commande `VARIABLES`) ;
- spécifier les centiles autres que ceux par défaut (au moyen de la sous-commande `PERCENTILES`) ;
- calculer les centiles à l'aide de l'une des cinq méthodes possibles (au moyen de la sous-commande `PERCENTILES`) ;
- spécifier toute transformation de l'exposant pour les diagrammes de dispersion par niveau (au moyen de la sous-commande `PLOT`) ;
- préciser le nombre de valeurs extrêmes à afficher (au moyen de la sous-commande `STATISTICS`) ;
- indiquer les paramètres pour les M-estimateurs d'un emplacement (au moyen de la sous-commande `MESTIMATORS`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
5

Tableaux croisés

La procédure de tableaux croisés établit des tableaux à deux entrées ou à entrées multiples et propose une variété de tests et de mesures d'associations pour les tableaux à deux entrées. La structure du tableau et l'ordre des modalités déterminent quels test ou mesures effectuer.

Les statistiques et les mesures d'association de tableaux croisés ne sont calculées que pour les tableaux à deux entrées. Si vous spécifiez une ligne, une colonne et une strate de facteur (variable de contrôle), la procédure de Crosstabs (tableaux croisés) forme un tableau de statistiques et de mesures pour chaque valeur de la strate de facteur (ou une combinaison de valeurs pour deux variables de contrôle ou plus). Par exemple, si le *sexe* est un facteur de strate pour un tableau *marié* (oui, non) face à la *vie* (est excitante, routinière ou ennuyeuse), les résultats d'un tableau à deux entrées pour les femmes sont calculés séparément de ceux des hommes et affichés sous forme de tableaux consécutifs.

Exemple. Les clients de PME ont-ils plus de probabilités d'être rentables en ventes de services (par exemple, formation et conseil) que ceux de grandes sociétés ? A partir d'une tabulation croisée, vous pourriez apprendre que la majorité des PME (moins de 500 salariés) génèrent des bénéfices de services élevés, alors que la majorité des grandes sociétés (plus de 2 500 salariés) rapportent des bénéfices de services bas.

Statistiques et mesures d'association : Khi-deux de Pearson, Khi-deux du rapport de vraisemblance, test d'association linéaire par linéaire, test exact de Fisher, Khi-deux corrigé de Yates, *r* de Pearson, rho de Spearman, coefficient de contingence, phi, *V* de Cramer, lambdas symétriques et asymétriques, tau de Goodman et Kruskal, coefficient d'incertitude, gamma, *d* de Somer, tau-*b* de Kendall, tau-*c* de Kendall, coefficient êta, Kappa de Cohen, estimation de risque relatif, odds ratio, test de McNemar, statistiques de Cochran et Mantel-Haenszel, ainsi que statistiques des proportions de colonne.

Données. Pour définir les modalités de chaque variable du tableau, utilisez des variables numériques ou des variables sous forme de chaînes (huit caractères ou moins). Par exemple, pour *sexe*, vous pouvez codifier les données avec 1 et 2, ou avec *homme* et *femme*.

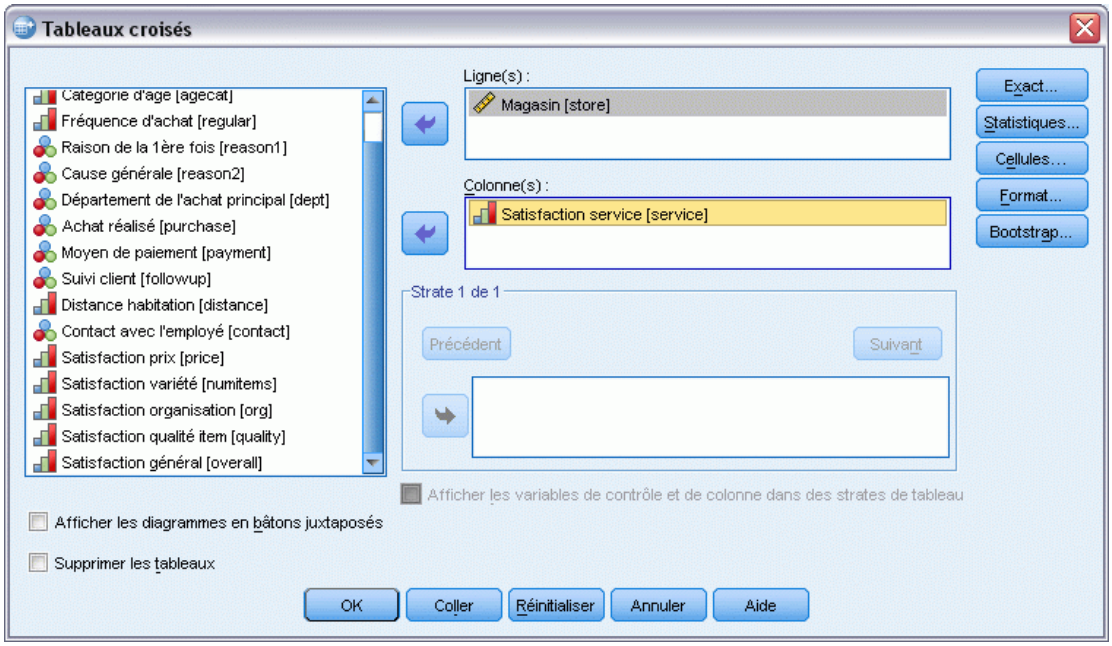
Hypothèses : Des statistiques et des mesures partent du principe de modalités ordonnées (données ordinales) ou de valeurs quantitatives (données d'intervalle ou données de type ratio), tel que décrit dans la section sur les statistiques. D'autres sont valides lorsque les variables du tableau ont des modalités désordonnées (données nominales). Pour les statistiques basées sur le test Khi-deux (phi, *V* de Cramer, coefficient de contingence), les données doivent provenir d'un échantillon aléatoire avec une répartition multinomiale.

Remarque : Les variables ordinales peuvent être des codes numériques représentant des modalités (par exemple, 1 = *faible*, 2 = *moyen*, 3 = *élevé*) ou des valeurs de chaîne. Toutefois, l'ordre alphabétique des valeurs de chaîne est supposé refléter l'ordre réel des modalités. Par exemple, pour une variable chaîne comportant des valeurs *Faible*, *Moyen*, *Elevé*, l'ordre des modalités est interprété comme *Elevé*, *Faible* ou *Moyen*, ce qui ne correspond pas à l'ordre correct. En règle générale, il est recommandé d'utiliser les codes numériques pour représenter les données ordinales.

Pour obtenir des tableaux croisés

▶ A partir des menus, sélectionnez :
Analyse > Statistiques descriptives > Tableaux croisés

Figure 5-1
Boîte de dialogue Tableaux croisés



▶ Sélectionnez des lignes de variables et des colonnes de variables.

Sinon, vous pouvez :

- Sélectionner des variables de contrôle.
- Cliquer sur Statistiques pour les tests et les mesures d'association pour les tableaux à deux entrées ou les sous-tableaux.
- Cliquez sur Cellules pour les valeurs observées et théoriques, les pourcentages et les résidus.
- Cliquez sur Format pour contrôler l'ordre des modalités.

Strates de tableaux croisés

Si vous sélectionnez des variables de strate, un tableau croisé séparé est produit pour chacune des modalités de variable de strate (variable de contrôle). Par exemple, si vous avez une variable de ligne, une variable de colonne, et une variable de strate avec deux modalités, vous obtenez un tableau à deux entrées pour chacune des modalités de la variable de strate. Pour créer une autre strate de variables de contrôle, cliquez sur Suivant. Les sous-tableaux sont produits pour chaque combinaison de catégories pour chaque variable de premier niveau avec chaque variable de second niveau, etc. Si les statistiques et les mesures d'association sont requises, elles ne s'appliquent qu'aux sous-tableaux à deux entrées.

Diagrammes en bâtons juxtaposés de tableaux croisés

Affichage de diagrammes en bâtons juxtaposés : Un diagramme en bâtons juxtaposés vous permet de résumer vos données pour des groupes d'observations. Il y a un regroupement de bâtons pour chaque valeur de la variable que vous avez spécifiée dans Ligne(s). La variable qui définit les bâtons dans chaque regroupement est la variable que vous avez spécifiée dans Colonne(s). Il y a un ensemble de bâtons de couleurs ou de motifs différents pour chaque valeur de cette variable. Si vous spécifiez plus d'une variable dans Colonnes ou Lignes, un diagramme en bâtons juxtaposés est produit pour chaque combinaison de deux variables.

Tableaux croisés affichant les variables de strate dans des strates de tableau

Afficher les variables de strate dans des strates de tableau. Vous pouvez choisir d'afficher les variables de strate (variables de contrôle) sous forme de strates de tableau dans le tableau croisé. Cela vous permet de créer des vues qui montrent les statistiques globales des variables de ligne et de colonne, et de faire défiler les modalités des variables de strate.

Un exemple utilisant le fichier de données *demo.sav* () est montré ci-dessous et a été obtenu comme suit :

- Sélectionnez *Modalité de revenu en milliers (rev_dis)* comme variable de ligne, *Possède un agenda électronique (PDA)* comme variable de colonne et *Niveau d'éducation (educ)* comme variable de strate.
- Sélectionnez l'option Afficher les variables de strate dans des strates de tableau.
- Sélectionnez Colonne dans la sous-boîte de dialogue Contenu des cases.
- Exécutez la procédure Tableaux croisés, double-cliquez sur le tableau croisé et sélectionnez Diplôme universitaire dans la liste déroulante Niveau d'éducation.

Figure 5-2
Tableaux croisés affichant les variables de strate dans des strates de tableau

Tableau croisé Catégories de revenu en milliers ($) * Possesseur d'un agenda électronique * Nombre d'années d'éducation

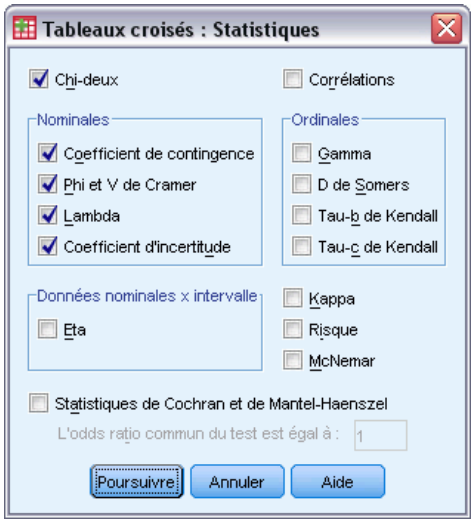
Nombre d'années d'éducation: Bac +3/4

			Possesseur d'un agenda électronique		
		Statistiques	Non	Oui	Total
Catégories de revenu en milliers ($)	Inf à $25	Effectif	146	50	196
		% compris dans Possesseur d'un agenda électronique	15.8%	11.6%	14.5%
	$25 - $49	Effectif	335	155	490
		% compris dans Possesseur d'un agenda électronique	36.3%	35.9%	36.2%
	$50 - $74	Effectif	187	72	259
		% compris dans Possesseur d'un agenda électronique	20.3%	16.7%	19.1%
	$75 - $124	Effectif	255	155	410
		% compris dans Possesseur d'un agenda électronique	27.6%	35.9%	30.3%
Total		Effectif	923	432	1355
		% compris dans Possesseur d'un agenda électronique	100.0%	100.0%	100.0%

La vue sélectionnée du tableau croisé montre les statistiques relatives aux répondants qui possèdent un diplôme universitaire.

Statistiques de tableaux croisés

Figure 5-3
Boîte de dialogue Tableaux croisé : Statistiques



Khi-deux : Pour les tableaux avec deux lignes et deux colonnes, sélectionnez Khi-deux pour calculer le Khi-deux de Pearson, le Khi-deux du rapport de vraisemblance, le test exact de Fisher et le test du Khi-deux de Yates corrigé (correction de continuité). Pour les tableaux 2 × 2, le test exact de Fisher est calculé lorsqu'un tableau qui ne provient pas de lignes ou de colonnes manquantes dans un tableau plus grand présente une cellule avec une fréquence attendue inférieure à 5. Le Khi-deux corrigé de Yates est calculé pour tous les autres tableaux 2 × 2. Pour les tableaux avec n'importe quel nombre de lignes ou de colonnes, sélectionnez Khi-deux pour calculer le Khi-deux de Pearson et le rapport de vraisemblance du Khi-deux. Lorsque les deux variables du tableau sont quantitatives, le Khi-deux donne le test d'association linéaire par linéaire.

Corrélations : Pour les tableaux dans lesquels les lignes et les colonnes contiennent des valeurs ordonnées, les corrélations donnent le coefficient de corrélation de Spearman, rho (données numériques seulement). Le Spearman rho est une mesure d'association entre les ordres de rang. Lorsque les deux variables (facteurs) du tableau sont quantitatives, les corrélations donnent le coefficient de corrélation de Pearson, *r*, une mesure de l'association linéaire entre les variables.

Nominal. Pour les données nominales (sans ordre intrinsèque, comme Catholique, Protestant, Juif), vous pouvez sélectionner le coefficient de contingence, le coefficient Phi et V de Cramér's V, Lambda (lambdas symétriques et asymétriques, et tau de Goodman et Kruskal) et le coefficient d'incertitude.

- **Coefficient de contingence.** Mesure d'association basée sur le Khi-deux. Les valeurs sont toujours comprises entre 0 et 1, 0 indiquant l'absence d'association entre les variables de ligne et de colonne, et les valeurs proches de 1 indiquant un degré d'association élevé entre les variables. La valeur maximale possible dépend du nombre de lignes et de colonnes dans le tableau.
- **Phi et V de Cramer.** Phi est une mesure d'association calculée à partir du Khi-deux. Elle est obtenue en divisant la statistique du Khi-deux par la taille de l'échantillon, puis en prenant la racine carrée du résultat. Le V de Cramer est également une mesure d'association basée sur le Khi-deux.
- **Lambda.** Mesure d'association reflétant la réduction proportionnelle de l'erreur lorsque les valeurs de la variable indépendante sont utilisées pour prévoir la variable dépendante. La valeur 1 signifie que la variable indépendante prévoit parfaitement la variable dépendante. La valeur 0 signifie que la variable indépendante ne prévoit pas du tout la variable dépendante.
- **Coefficient d'incertitude.** Mesure d'association qui indique la réduction proportionnelle de l'erreur lorsque les valeurs d'une variable sont utilisées pour prévoir celles d'une autre. Par exemple, la valeur 0,83 indique que la connaissance d'une variable réduit de 83 % l'erreur dans les prévisions de l'autre variable. Le programme calcule à la fois des versions symétriques et asymétriques de ce coefficient.

Ordinal. Pour les tableaux dont les lignes et les colonnes contiennent des valeurs ordonnées, sélectionnez Gamma (ordre zéro pour les tableaux à 2 entrées et conditionnel pour les tableaux de 3 à 10 entrées), le tau-b de Kendall et le tau-c de Kendall. Pour prévoir les modalités de colonnes à partir des modalités de lignes, sélectionnez le d de Somers.

- **Gamma.** Mesure d'association symétrique entre deux variables ordinales. Cette mesure est située entre -1 et 1. Les valeurs proches d'une valeur absolue de 1 indiquent une relation forte entre les deux variables. Les valeurs proches de 0 indiquent une relation faible ou inexistante. Pour les tableaux d'ordre 2, les gammas d'ordre 0 (zéro) apparaissent. Pour les tableaux d'ordre 3 et les tableaux d'ordre n, les gammas conditionnels apparaissent.

- **d de Somer.** Mesure d'intensité de la relation entre deux variables ordinales, qui s'étend de -1 à 1. Les valeurs proches de 1 indiquent une forte relation entre les deux variables, et celles proches de zéro indiquent une relation faible ou inexistante entre les variables. Le d de Somer est une extension asymétrique du gamma, qui ne diffère de celui-ci que par l'inclusion du nombre de paires non liées à la variable indépendante. Le programme calcule également une version symétrique de cette statistique.
- **Tau•b de Kendall.** Mesure de corrélation non paramétrique pour variables ordinales ou classées qui prend en considération les ex aequo. Le signe du coefficient indique la direction de la relation et sa valeur absolue indique sa force, les valeurs absolues les plus grandes indiquant les relations les plus fortes. Les valeurs peuvent varier de -1 à +1 mais une valeur de -1 ou de +1 ne peut toutefois être obtenue que dans des tableaux carrés.
- **Tau•c de Kendall.** Mesure d'association non paramétrique pour variables ordinales qui ne prend pas en considération les ex aequo. Le signe du coefficient indique la direction de la relation et sa valeur absolue indique sa force, les valeurs absolues les plus grandes indiquant les relations les plus fortes. Les valeurs peuvent varier de -1 à +1 mais une valeur de -1 ou de +1 ne peut toutefois être obtenue que dans des tableaux carrés.

Données nominales x intervalle : Lorsqu'une variable est qualitative et l'autre quantitative, sélectionnez Eta. La variable qualitative doit être codée numériquement.

- **Eta.** Mesure d'association dont les valeurs sont comprises entre 0 et 1, 0 indiquant l'absence d'association entre les variables de ligne et de colonne, et les valeurs proches de 1 indiquant un degré d'association élevé. Eta convient à une variable dépendante continue mesurée sur une échelle d'intervalle (par exemple, le revenu) et une variable indépendante ayant un nombre limité de modalités (par exemple, le sexe). Deux valeurs eta sont calculées : L'une traite la variable de ligne comme variable d'intervalle et l'autre traite la variable de colonne comme variable d'intervalle.

Kappa. Le Kappa de Cohen mesure la concordance entre les évaluations de deux évaluateurs utilisés pour évaluer un même objet. La valeur 1 indique une concordance parfaite. La valeur 0 indique que la concordance ne dépasse pas celle due au hasard. Le Kappa est basé sur un tableau carré dans lequel les valeurs des lignes et des colonnes représentent la même échelle. Chaque cellule qui comprend des valeurs observées pour une variable mais pas une autre se voit affectée un nombre de 0. Le Kappa n'est pas calculé si le type de stockage de données (chaîne ou numérique) n'est pas le même pour les deux variables. Les deux variables chaîne d'une paire doivent être de même longueur.

Risque. Pour les tableaux 2 x 2, mesure de la force de l'association entre la présence d'un facteur et la réalisation d'un événement. Si l'intervalle de confiance de la statistique inclut une valeur de 1, il n'existe aucune association entre le facteur et l'événement. L'odds ratio peut être utilisé comme estimation du risque relatif dans le cas où la réalisation du facteur est rare.

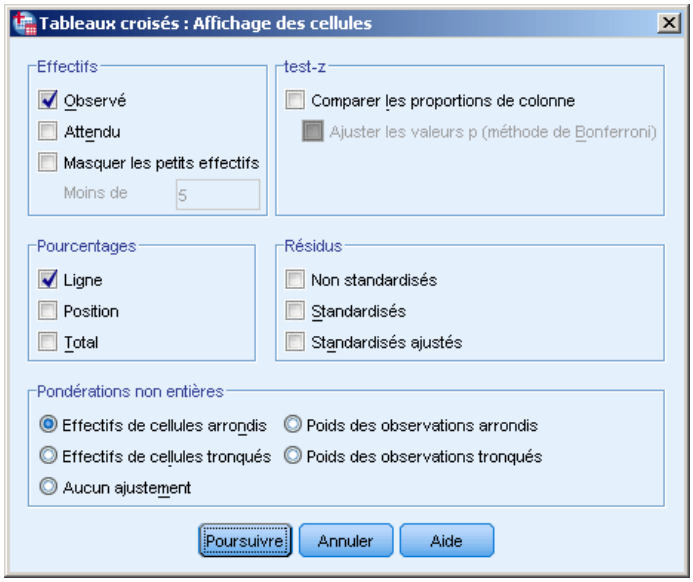
McNemar. Test non paramétrique pour deux variables dichotomiques liées. Il recherche les changements de réponse en utilisant la répartition Khi-deux. Ce test est utile pour détecter les changements avant-après dans les réponses dus à une intervention expérimentale dans les plans. Pour les tableaux carrés plus volumineux, le test McNemar-Bowker de symétrie est reporté.

Statistiques de Cochran et de Mantel•Haenszel. Les statistiques de Cochran et de Mantel-Haenszel servent à tester l'indépendance entre une variable facteur dichotomique et une variable réponse dichotomique, selon les paramètres de covariable définis par des variables (contrôles) de strate.

Remarque : alors que les autres statistiques sont calculées strate par strate, les statistiques de Cochran et de Mantel-Haenszel sont calculées une seule fois pour toutes les strates.

Affichage de cellules (cases) de tableaux croisés

Figure 5-4
Boîte de dialogue Tableaux croisés : Contenu des cases (cellules)



Pour vous aider à découvrir des types dans les données qui contribuent à un test du Khi-deux significatif, la procédure de Tableaux croisés affiche les fréquences attendues et trois types de résidus (déviations) qui mesurent la différence entre les fréquences observées et les fréquences attendues. Chaque cellule du tableau peut contenir toute combinaison d'effectifs, de pourcentages et de résidus sélectionnés.

Effectifs. Nombre d'observations effectivement observées et nombre d'observations attendues si les variables de ligne et de colonne sont indépendantes l'une de l'autre. Vous pouvez choisir de masquer les effectifs inférieurs à un entier spécifique. Les valeurs masquées seront affichées en tant que <N, où N est l'entier spécifié. L'entier spécifié doit être supérieur ou égal à 2, bien que la valeur 0 soit permise et indique qu'aucun effectif n'est masqué.

Comparer les proportions de colonne. Cette option calcule les comparaisons par paire des proportions de colonne et indique quelles paires de colonnes (pour une ligne donnée) sont significativement différentes. Les différences significatives sont indiquées dans le tableau croisé dans un format de style APA à l'aide d'indices et sont calculées au niveau de signification 0,05. *Remarque* : Si cette option est spécifiée sans sélectionner les effectifs observés ou les pourcentages de colonne, alors les effectifs observés sont inclus dans le tableau croisé, les indices de style APA indiquant les résultats des tests de proportion de colonne.

- **Ajustement des valeurs p (méthode Bonferroni).** Les comparaisons par paire des proportions de colonne utilisent la correction Bonferroni, qui ajuste le taux de signification observé pour les comparaisons multiples.

Pourcentages : Les pourcentages peuvent s'additionner par ligne ou par colonne. Les pourcentages du nombre total d'observations représentées dans le tableau (une strate) sont également disponibles. *Remarque* : Si l'option Masquer les petits effectifs est sélectionnée dans le groupe Effectifs, les pourcentages associés aux effectifs masqués sont aussi masqués.

Résidus : Les résidus non standardisés donnent la différence entre les valeurs observées et les valeurs théoriques. Les résidus standardisés et standardisés ajustés sont également disponibles.

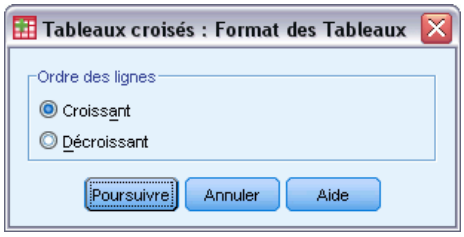
- **Non standardisés.** Différence entre la valeur observée et la valeur théorique. La valeur théorique correspond au nombre d'observations attendues dans la cellule quand il n'existe pas de relation entre les deux variables. Un résidu positif indique que la cellule contient plus d'observations que si les variables de ligne et de colonne étaient indépendantes.
- **Standardisés.** Résidu, divisé par une estimation de son erreur standard. Egalement appelés résidus de Pearson, les résidus standardisés ont une moyenne de 0 et un écart-type de 1.
- **Standardisés ajustés.** Résidu d'une cellule (valeur observée moins valeur théorique) divisé par une estimation de son erreur standard. Le résidu standardisé qui en résulte est exprimé en écarts par rapport à la moyenne.

Pondérations non entières : En général, les effectifs de cellules sont des valeurs entières, car ils représentent le nombre d'observations figurant dans chaque cellule. Toutefois, si le fichier de données est pondéré par une variable de pondération avec des fractions (par exemple, 1,25), les effectifs de cellules peuvent également être des fractions. Vous pouvez tronquer ou arrondir les valeurs avant ou après le calcul des effectifs de cellules, ou utiliser des effectifs de cellules non entiers pour l'affichage des tableaux et les calculs statistiques.

- **Arrondi des effectifs des cellules.** Les poids des observations sont utilisés tels quels, mais les poids accumulés dans les cellules sont arrondis avant le calcul des statistiques.
- **Troncature des effectifs des cellules.** Les poids des observations sont utilisés tels quels, mais les poids accumulés dans les cellules sont arrondis avant le calcul des statistiques.
- **Arrondi des poids des observations.** Les poids des observations sont arrondis avant leur utilisation.
- **Troncature des poids des observations.** Les poids des observations sont tronqués avant leur utilisation.
- **Aucun ajustement.** Les pondérations des observations sont utilisées telles quelles et les comptes des cellules fractionnées sont utilisées. Toutefois, lorsque des statistiques exactes (disponibles uniquement avec l'option Tests exacts) sont demandées, les pondérations cumulées dans les cellules sont tronquées ou arrondies avant le calcul des statistiques du test exact.

Format de tableau croisé

Figure 5-5
Boîte de dialogue Tableaux croisés : Format



Vous pouvez arranger les lignes par ordre croissant ou décroissant de valeur de la variable de ligne.

Chapitre
6

Récapituler

La procédure Récapituler calcule les statistiques de sous-groupes pour les variables à l'intérieur des modalités de variables de regroupement. Tous les niveaux de variables de regroupement sont à tabulation croisée. Vous pouvez choisir l'ordre dans lequel les statistiques sont affichées. Les statistiques récapitulatives sont affichées pour chaque variable à travers toutes les modalités. Les valeurs des données dans chaque modalité peuvent être listées ou supprimées. Avec d'importants fichiers de données, vous pouvez choisir de lister seulement les premières observations *n*.

Exemple : Quel est le montant moyen de ventes de produits par région et par secteur de clientèle ? Vous pouvez découvrir que le montant moyen des ventes est légèrement plus élevé dans la région Ouest que dans les autres régions, avec des sociétés commerciales dans la région Ouest apportant le montant moyen de ventes le plus élevé.

Statistiques : Somme, nombre d'observations, moyenne, médiane, médiane groupée, erreur standard pour la moyenne, minimum, maximum, plage, valeur de la variable pour la première modalité de la variable de regroupement, valeur de la variable pour la dernière modalité de la variable de regroupement, écart-type, variance, aplatissement, erreur standard d'aplatissement, asymétrie, erreur standard d'asymétrie, pourcentage de la somme totale, pourcentage de *N* total, pourcentage de la somme dans, pourcentage de *N* dans, moyennes géométrique et harmonique.

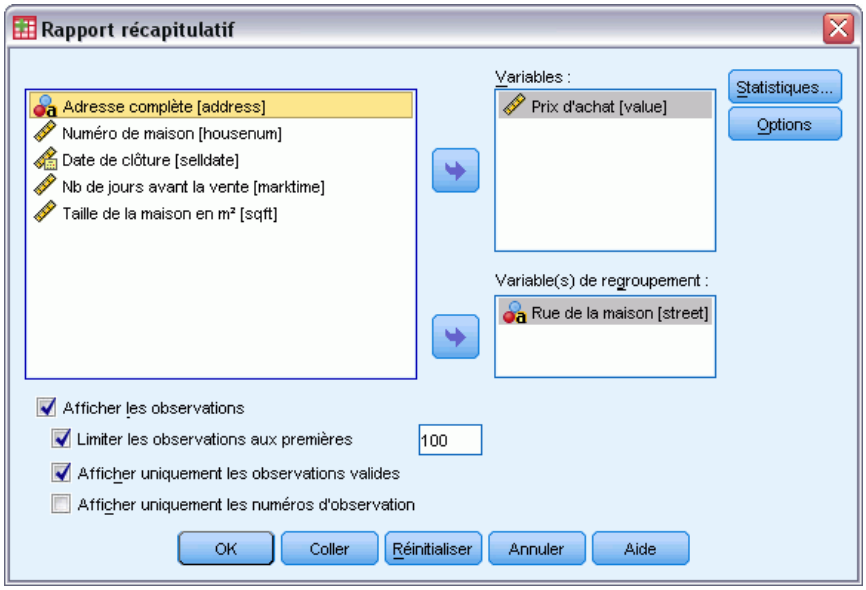
Données. Les variables de regroupement sont des variables qualitatives dont les valeurs peuvent être numériques ou chaîne. Le nombre de modalités doit être raisonnablement limité. Les autres variables doivent pouvoir être classées.

Hypothèses : Certains des sous-groupes statistiques optionnels, tels que la moyenne et l'écart-type sont basés sur la théorie normale et conviennent aux variables quantitatives ayant une distribution symétrique. Les statistiques robustes telles que la médiane et l'intervalle, conviennent aux variables quantitatives qui confirment ou infirment l'hypothèse de normalité.

Obtenir des récapitulatifs des observations

- A partir des menus, sélectionnez :
  Analyse > Rapports > Récapitulatif des observations

Figure 6-1
Boîte de dialogue Rapport récapitulatif



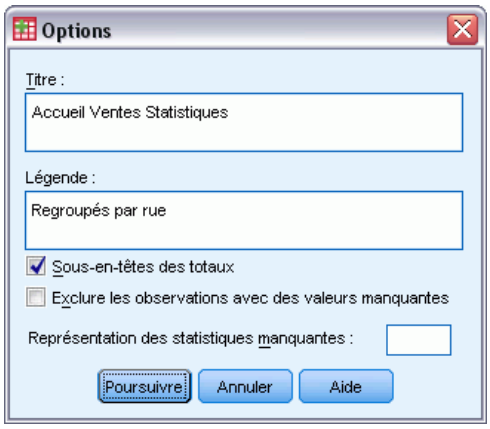
▶ Sélectionnez une ou plusieurs variables.

Sinon, vous pouvez :

- Sélectionner au moins une variable de regroupement afin de diviser vos données en sous-groupes.
- Cliquer sur Options afin de modifier le titre du résultat, ajouter une légende au-dessous du résultat, ou exclure les observations ayant des valeurs manquantes.
- Cliquer sur Statistiques pour obtenir des statistiques facultatives.
- Sélectionner Afficher les observations afin de répertorier les observations dans chaque sous-groupe. Par défaut, le système ne liste que les 100 premières observations de votre fichier. Vous pouvez augmenter ou diminuer la valeur de l'option Limiter les observations aux *n* premières ou désélectionner cet élément pour répertorier toutes les observations.

Options de Récapituler

Figure 6-2
Options



Récapituler vous permet de modifier le titre de votre résultat ou d'ajouter une légende qui apparaîtra en dessous du tableau de sortie. Vous pouvez contrôler les sauts de ligne dans les titres et légendes en tapant \n à tous les endroits où vous voulez insérer un saut de ligne dans le texte.

Vous pouvez également choisir d'afficher ou de supprimer les sous-en-têtes des totaux et d'inclure ou d'exclure les observations ayant des valeurs manquantes pour toute variable prise en compte dans toute analyse. Il est souvent souhaitable de marquer les observations manquantes dans le résultat par un point ou un astérisque. Saisir un caractère, une phrase, ou un code que vous souhaitez voir apparaître lorsqu'une valeur manque, sinon, aucun traitement spécial ne s'applique aux observations manquantes dans le résultat.

Récapituler les statistiques

Figure 6-3
Boîte de dialogue Statistiques de rapport récapitulatives

Vous pouvez choisir l'une des statistiques de sous-groupe suivantes pour les variables à l'intérieur de chaque modalité de chacune des variables de regroupement : Somme, nombre d'observations, moyenne, médiane, médiane groupée, erreur standard pour la moyenne, minimum, maximum, intervalle, valeur de la variable pour la première modalité de la variable de regroupement, valeur de la variable pour la dernière modalité de la variable de regroupement, écart-type, variance, aplatissement, erreur standard d'aplatissement, asymétrie, erreur standard d'asymétrie, pourcentage de somme totale, pourcentage de *N* total, pourcentage de la somme dans, pourcentage de *N* dans, moyenne géométrique, moyenne harmonique. L'ordre dans lequel les statistiques apparaissent dans la liste Variables correspond à celui dans lequel elles seront affichées dans le résultat. Les statistiques récapitulatives sont aussi affichées pour chaque variable à travers toutes les modalités.

Première. Affiche la première valeur rencontrée dans le fichier de données.

Moyenne géométrique. Racine nième du produit des valeurs de données, n représentant le nombre d'observations.

Médiane de groupes. Médiane calculée pour les données codées dans des groupes. Par exemple, pour les données d'âge, si chaque valeur de la trentaine est codée 35, chaque valeur de la quarantaine est codée 45, etc., la médiane de groupes est la médiane calculée à partir des données codées.

Moyenne harmonique. Fonction utilisée pour estimer la taille moyenne d'un groupe lorsque la taille des échantillons diffère d'un groupe à l'autre. La moyenne harmonique correspond au nombre total d'échantillons divisé par la somme des réciproques des tailles de l'échantillon.

Aplatissement. Mesure de l'étendue du regroupement des observations autour d'un point central. Dans le cas d'une distribution normale, la valeur de la statistique d'aplatissement est égale à zéro. Un aplatissement positif indique que par rapport a une distribution normale, les observations sont plus regroupées au centre et présentent des extrémités plus fines atteignant les valeurs extrêmes de la distribution. La distribution leptokurtique présente des extrémités plus épaisses que dans le cas d'une distribution normale. Un aplatissement négatif indique que les observations sont moins regroupées au centre et présentent des extrémités plus épaisses atteignant les valeurs extrêmes de la distribution. La distribution platykurtique présentent des extrémités plus fines que dans le cas d'une distribution normale.

Dernière. Affiche la dernière valeur rencontrée dans le fichier de données.

Maximum. Plus grande valeur d'une variable numérique.

Moyenne. Mesure de la tendance centrale. Moyenne arithmétique ; somme divisée par le nombre d'observations.

Médiane. Valeur au-dessus ou au-dessous de laquelle se trouvent la moitié des observations ; 50e centile. Si le nombre d'observations est pair, la médiane correspond à la moyenne des deux observations du milieu lorsqu'elles sont triées dans l'ordre croissant ou décroissant. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées.

Minimum. Valeur la plus petite d'une variable numérique.

Nombre d'observations. Nombre d'observations (ou d'enregistrements).

Pourcentage de N total. Pourcentage du nombre total d'observations dans chaque modalité.

Pourcentage de la somme totale. Pourcentage de la somme totale dans chaque modalité.

Intervalle. Différence entre la valeur maximale et la valeur minimale d'une variable numérique (maximum–minimum).

Asymétrie. Mesure de l'asymétrie d'une distribution. La distribution normale est symétrique et possède une valeur d'asymétrie égale à 0. Une distribution dont la valeur d'asymétrie est positive présente une extrémité droite allongée. Une distribution caractérisée par une importante asymétrie négative présente une extrémité gauche plus allongée. Pour simplifier, une valeur d'asymétrie deux fois supérieure à l'erreur standard correspond à une absence de symétrie.

Erreur standard du Kurtosis. Le rapport de l'aplatissement à son erreur standard peut servir de test de normalité (il y a anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur d'aplatissement positive importante indique que les extrémités de la distribution sont plus allongées que celles d'une distribution normale ; une valeur d'aplatissement négative présente des extrémités plus courtes (semblables à celles d'une distribution uniforme sous forme de boîtes).

Erreur standard du Skewness. Rapport de l'asymétrie avec son erreur standard, qui peut servir de test de normalité (il y a anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur d'asymétrie positive importante indique une extrémité allongée vers la droite ; une valeur négative extrême produit une extrémité allongée vers la gauche.

Somme. Somme ou total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.

Variance. Mesure de dispersion autour de la moyenne, égale à la somme des carrés des écarts par rapport à la moyenne, divisée par le nombre d'observations moins un. La variance se mesure en unités, qui sont égales au carré des unités de la variable.

Chapitre
7

Moyennes

La procédure des moyennes calcule les moyennes de sous-groupes et les statistiques univariées correspondantes pour des variables dépendantes au sein des modalités d'une ou de plusieurs variables indépendantes. Vous pouvez également obtenir une analyse à un facteur de la variance, un coefficient êta et des tests de linéarité.

Exemple : Mesurez la quantité moyenne de lipides absorbée par trois différents types d'huile alimentaire et effectuez une analyse à un facteur de la variance pour voir si les moyennes divergent.

Statistiques : Somme, nombre d'observations, moyenne, médiane, médiane groupée, erreur standard pour la moyenne, minimum, maximum, plage, valeur de la variable pour la première modalité de la variable de regroupement, valeur de la variable pour la dernière modalité de la variable de regroupement, écart-type, variance, aplatissement, erreur standard d'aplatissement, asymétrie, erreur standard d'asymétrie, pourcentage de la somme totale, pourcentage de *N* total, pourcentage de la somme dans, pourcentage de *N* dans, moyennes géométrique et harmonique. Les options comportent une analyse de la variance, un coefficient êta, un coefficient êta-carré, ainsi que des tests de linéarité *R* et R^2.

Données. Les variables dépendantes sont quantitatives et les variables indépendantes sont qualitatives. Les valeurs des variables qualitatives peuvent être soit numériques, soit des chaînes.

Hypothèses : Certains des sous-groupes statistiques optionnels, tels que la moyenne et l'écart-type sont basés sur la théorie normale et conviennent aux variables quantitatives ayant une distribution symétrique. Les statistiques robustes, telles que la médiane, conviennent aux variables continues qui confirment ou infirment l'hypothèse de normalité. L'analyse de la variance résiste aux écarts par rapport à la normalité, à condition que les données de chaque cellule soient symétriques. L'analyse de la variance part également du principe que les groupes sont issus de populations ayant la même variance. Pour vérifier cette hypothèse, utilisez le test d'homogénéité de la variance de Levene, disponible dans la procédure ANOVA à un facteur.

Pour obtenir des moyennes de sous-groupes

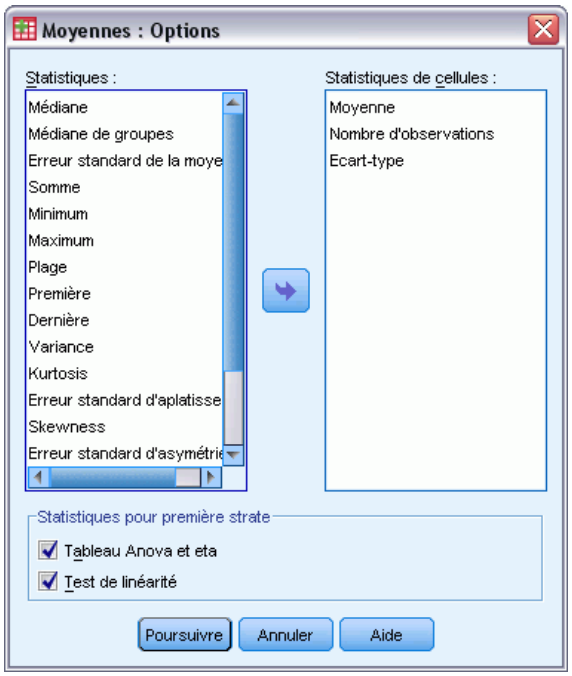
- A partir des menus, sélectionnez :
  Analyse > Comparer les moyennes > Moyennes

Figure 7-1
Boîte de dialogue Moyennes

- Sélectionnez au moins une variable dépendante.
- Utilisez l'une des méthodes suivantes pour sélectionner des variables indépendantes qualitatives :
  - Sélectionnez une ou plusieurs variables indépendantes. Des résultats distincts sont présentés pour chaque variable indépendante.
  - Sélectionnez une ou plusieurs strates des variables indépendantes. Chaque strate divise une nouvelle fois l'échantillon. Si vous avez une variable indépendante dans la strate 1 et une dans la strate 2, les résultats sont présentés dans un tableau croisé, par opposition aux tableaux séparés pour chaque variable indépendante.
- Cliquez sur Options pour obtenir des statistiques facultatives, telles que la table d'analyse de la variance, eta, eta-carré, R et R^2.

Moyennes : Options

Figure 7-2
Boîte de dialogue Moyennes : Options



Vous pouvez choisir l'une des statistiques de sous-groupe suivantes pour les variables à l'intérieur de chaque modalité de chacune des variables de regroupement : Somme, nombre d'observations, moyenne, médiane, médiane groupée, erreur standard pour la moyenne, minimum, maximum, intervalle, valeur de la variable pour la première modalité de la variable de regroupement, valeur de la variable pour la dernière modalité de la variable de regroupement, écart-type, variance, aplatissement, erreur standard d'aplatissement, asymétrie, erreur standard d'asymétrie, pourcentage de somme totale, pourcentage de *N* total, pourcentage de la somme dans, pourcentage de *N* dans, moyenne géométrique, moyenne harmonique. Vous pouvez changer l'ordre de présentation des statistiques des sous-groupes. L'ordre dans lequel les statistiques apparaissent dans la liste Cellule Statistiques correspond à celui dans lequel elles seront affichées dans le résultat. Les statistiques récapitulatives sont aussi affichées pour chaque variable à travers toutes les modalités.

Première. Affiche la première valeur rencontrée dans le fichier de données.

Moyenne géométrique. Racine nième du produit des valeurs de données, n représentant le nombre d'observations.

Médiane de groupes. Médiane calculée pour les données codées dans des groupes. Par exemple, pour les données d'âge, si chaque valeur de la trentaine est codée 35, chaque valeur de la quarantaine est codée 45, etc., la médiane de groupes est la médiane calculée à partir des données codées.

Moyenne harmonique. Fonction utilisée pour estimer la taille moyenne d'un groupe lorsque la taille des échantillons diffère d'un groupe à l'autre. La moyenne harmonique correspond au nombre total d'échantillons divisé par la somme des réciproques des tailles de l'échantillon.

Aplatissement. Mesure de l'étendue du regroupement des observations autour d'un point central. Dans le cas d'une distribution normale, la valeur de la statistique d'aplatissement est égale à zéro. Un aplatissement positif indique que par rapport a une distribution normale, les observations sont plus regroupées au centre et présentent des extrémités plus fines atteignant les valeurs extrêmes de la distribution. La distribution leptokurtique présente des extrémités plus épaisses que dans le cas d'une distribution normale. Un aplatissement négatif indique que les observations sont moins regroupées au centre et présentent des extrémités plus épaisses atteignant les valeurs extrêmes de la distribution. La distribution platykurtique présentent des extrémités plus fines que dans le cas d'une distribution normale.

Dernière. Affiche la dernière valeur rencontrée dans le fichier de données.

Maximum. Plus grande valeur d'une variable numérique.

Moyenne. Mesure de la tendance centrale. Moyenne arithmétique ; somme divisée par le nombre d'observations.

Médiane. Valeur au-dessus ou au-dessous de laquelle se trouvent la moitié des observations ; 50e centile. Si le nombre d'observations est pair, la médiane correspond à la moyenne des deux observations du milieu lorsqu'elles sont triées dans l'ordre croissant ou décroissant. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées.

Minimum. Valeur la plus petite d'une variable numérique.

Nombre d'observations. Nombre d'observations (ou d'enregistrements).

Pourcentage de N total. Pourcentage du nombre total d'observations dans chaque modalité.

Pourcentage de N total. Pourcentage de la somme totale dans chaque modalité.

Intervalle. Différence entre la valeur maximale et la valeur minimale d'une variable numérique (maximum–minimum).

Asymétrie. Mesure de l'asymétrie d'une distribution. La distribution normale est symétrique et possède une valeur d'asymétrie égale à 0. Une distribution dont la valeur d'asymétrie est positive présente une extrémité droite allongée. Une distribution caractérisée par une importante asymétrie négative présente une extrémité gauche plus allongée. Pour simplifier, une valeur d'asymétrie deux fois supérieure à l'erreur standard correspond à une absence de symétrie.

Erreur standard du Kurtosis. Le rapport de l'aplatissement à son erreur standard peut servir de test de normalité (il y a anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur d'aplatissement positive importante indique que les extrémités de la distribution sont plus allongées que celles d'une distribution normale ; une valeur d'aplatissement négative présente des extrémités plus courtes (semblables à celles d'une distribution uniforme sous forme de boîtes).

Erreur standard du Skewness. Rapport de l'asymétrie avec son erreur standard, qui peut servir de test de normalité (il y a anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur d'asymétrie positive importante indique une extrémité allongée vers la droite ; une valeur négative extrême produit une extrémité allongée vers la gauche.

Somme. Somme ou total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.

Variance. Mesure de dispersion autour de la moyenne, égale à la somme des carrés des écarts par rapport à la moyenne, divisée par le nombre d'observations moins un. La variance se mesure en unités, qui sont égales au carré des unités de la variable.

Statistiques pour première strate

Tableau Anova et eta. Affiche un tableau d'analyse unifactorielle de la variance et calcule êta et êta carré (mesures de l'association) pour chaque variable indépendante de la première couche.

Test de linéarité. Calcule la somme des carrés, les degrés de liberté et le carré moyen associés aux composants linéaires et non linéaires, ainsi que le rapport F, le R et le R-deux. La linéarité n'est pas calculée si la variable indépendante est une chaîne courte.

Chapitre
8

Cubes OLAP

La procédure de Cubes OLAP (Online Analytical Processing) calcule les totaux, les moyennes et autres statistiques univariées pour des variables récapitulatives continues à l'intérieur de modalités d'une ou plusieurs variables de regroupement qualitatives. Une strate séparée dans le tableau est créée pour chaque modalité de chaque variable de regroupement.

Exemple : Ventes totales et moyennes pour différentes régions et lignes de produits à l'intérieur de chaque région.

Statistiques. Somme, nombre d'observations, moyenne, médiane, médiane groupée, erreur standard pour la moyenne, minimum, maximum, intervalle, valeur de la variable pour la première modalité de la variable de regroupement, valeur de la variable pour la dernière modalité de la variable de regroupement, écart-type, variance, aplatissement, erreur standard d'aplatissement, asymétrie, erreur standard d'asymétrie, pourcentage des observations totales, pourcentage de somme totale, pourcentage des observations totales dans les variables de regroupement, pourcentage de la somme totale dans les variables de regroupement, moyenne géométrique et moyenne harmonique.

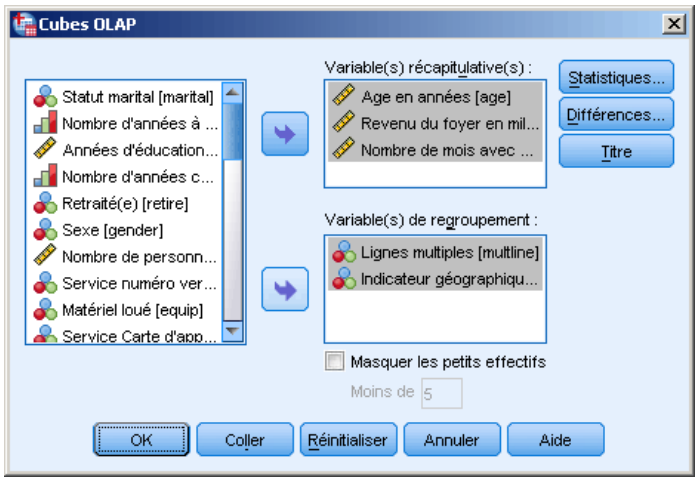
Données. Les variables récapitulatives sont quantitatives (variables continues mesurées sur une échelle d'intervalle ou de rapport) et les variables de regroupement sont qualitatives. Les valeurs des variables qualitatives peuvent être soit numériques, soit des chaînes.

Hypothèses : Certains des sous-groupes statistiques optionnels, tels que la moyenne et l'écart-type sont basés sur la théorie normale et conviennent aux variables quantitatives ayant une distribution symétrique. Les statistiques robustes telles que la médiane et l'intervalle, conviennent aux variables quantitatives qui confirment ou infirment l'hypothèse de normalité.

Pour obtenir des cubes OLAP

- A partir des menus, sélectionnez :
  Analyse > Rapports > Cubes OLAP

Figure 8-1
Boîte de dialogue Cubes OLAP



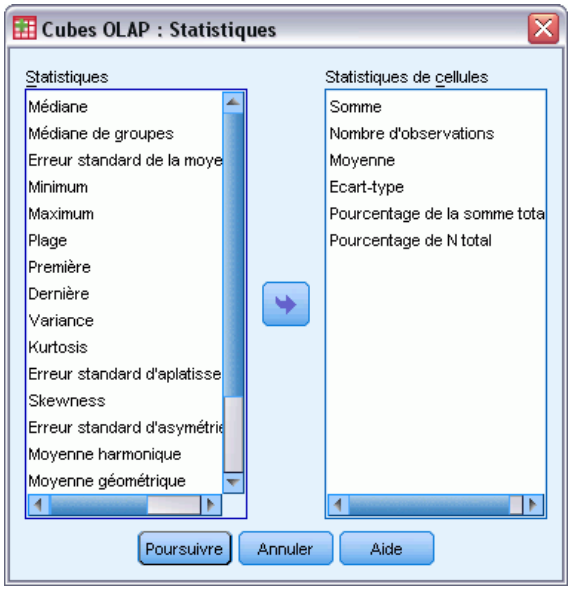
- Sélectionnez une ou plusieurs variables récapitulatives continues.
- Sélectionnez une ou plusieurs variables qualitatives.

Eventuellement :

- Sélectionner d'autres statistiques récapitulatives (cliquez sur Statistiques...). Vous devez sélectionner un ou plusieurs critères de regroupement pour pouvoir sélectionner les statistiques récapitulatives.
- Calculer les différences entre des paires de variables et des paires de groupes définies par un critère de regroupement (cliquez sur Différences).
- Créer des titres de tableaux personnalisés (cliquez sur Titre).
- Masquer les effectifs inférieurs à un entier spécifique. Les valeurs masquées seront affichées en tant que <N, où N est l'entier spécifié. L'entier spécifié doit être supérieur ou égal à 2.

Cubes OLAP : Statistiques

Figure 8-2
Boîte de dialogue Cubes OLAP : Statistiques



Vous pouvez choisir l'une des statistiques de sous-groupe suivantes pour les variables récapitulatives à l'intérieur de chaque modalité de chacune des variables de regroupement : Somme, Nombre d'observations, Moyenne, Médiane, Médiane de groupes, Erreur std de la moyenne, Minimum, Maximum, Intervalle, Premier (valeur de la variable pour la première modalité de la variable de regroupement), Dernier (valeur de la variable pour la dernière modalité de la variable de regroupement), Ecart type, Variance, Aplatissement, Erreur standard de l'aplatissement, Asymétrie, Erreur std d'asymétrie, Pourcentage de N (observations) totales, Pourcentage de somme tot., Pourcentage des observations totales dans les variables de regroupement, Pourcentage de la somme totale dans les variables de regroupement, Moyenne géométrique et Moyenne harmonique.

Vous pouvez changer l'ordre de présentation des statistiques des sous-groupes. L'ordre dans lequel les statistiques apparaissent dans la liste Cellule Statistiques correspond à celui dans lequel elles seront affichées dans le résultat. Les statistiques récapitulatives sont aussi affichées pour chaque variable à travers toutes les modalités.

Première. Affiche la première valeur rencontrée dans le fichier de données.

Moyenne géométrique. Racine nième du produit des valeurs de données, n représentant le nombre d'observations.

Médiane de groupes. Médiane calculée pour les données codées dans des groupes. Par exemple, pour les données d'âge, si chaque valeur de la trentaine est codée 35, chaque valeur de la quarantaine est codée 45, etc., la médiane de groupes est la médiane calculée à partir des données codées.

Moyenne harmonique. Fonction utilisée pour estimer la taille moyenne d'un groupe lorsque la taille des échantillons diffère d'un groupe à l'autre. La moyenne harmonique correspond au nombre total d'échantillons divisé par la somme des réciproques des tailles de l'échantillon.

Aplatissement. Mesure de l'étendue du regroupement des observations autour d'un point central. Dans le cas d'une distribution normale, la valeur de la statistique d'aplatissement est égale à zéro. Un aplatissement positif indique que par rapport a une distribution normale, les observations sont plus regroupées au centre et présentent des extrémités plus fines atteignant les valeurs extrêmes de la distribution. La distribution leptokurtique présente des extrémités plus épaisses que dans le cas d'une distribution normale. Un aplatissement négatif indique que les observations sont moins regroupées au centre et présentent des extrémités plus épaisses atteignant les valeurs extrêmes de la distribution. La distribution platykurtique présentent des extrémités plus fines que dans le cas d'une distribution normale.

Dernière. Affiche la dernière valeur rencontrée dans le fichier de données.

Maximum. Plus grande valeur d'une variable numérique.

Moyenne. Mesure de la tendance centrale. Moyenne arithmétique ; somme divisée par le nombre d'observations.

Médiane. Valeur au-dessus ou au-dessous de laquelle se trouvent la moitié des observations ; 50e centile. Si le nombre d'observations est pair, la médiane correspond à la moyenne des deux observations du milieu lorsqu'elles sont triées dans l'ordre croissant ou décroissant. La médiane est une mesure de tendance centrale et elle n'est pas, à l'inverse de la moyenne, sensible aux valeurs éloignées.

Minimum. Valeur la plus petite d'une variable numérique.

Nombre d'observations. Nombre d'observations (ou d'enregistrements).

Pourcentage de N dans. Pourcentage du nombre d'observations pour le critère de regroupement spécifié dans les modalités des autres critères de regroupement. Si vous n'avez qu'un seul critère de regroupement, cette valeur est identique au pourcentage du nombre total d'observations.

Pourcentage de la somme dans. Pourcentage de la somme de la variable de regroupement définie dans les modalités des autres variables de regroupement. Si vous n'avez qu'un seul critère de regroupement, cette valeur est identique au pourcentage de la somme totale.

Pourcentage de N total. Pourcentage du nombre total d'observations dans chaque modalité.

Pourcentage de la somme totale. Pourcentage de la somme totale dans chaque modalité.

Intervalle. Différence entre la valeur maximale et la valeur minimale d'une variable numérique (maximum–minimum).

Asymétrie. Mesure de l'asymétrie d'une distribution. La distribution normale est symétrique et possède une valeur d'asymétrie égale à 0. Une distribution dont la valeur d'asymétrie est positive présente une extrémité droite allongée. Une distribution caractérisée par une importante asymétrie négative présente une extrémité gauche plus allongée. Pour simplifier, une valeur d'asymétrie deux fois supérieure à l'erreur standard correspond à une absence de symétrie.

Erreur standard du Kurtosis. Le rapport de l'aplatissement à son erreur standard peut servir de test de normalité (il y a anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur d'aplatissement positive importante indique que les extrémités de la distribution sont plus allongées que celles d'une distribution normale ; une valeur d'aplatissement négative présente des extrémités plus courtes (semblables à celles d'une distribution uniforme sous forme de boîtes).

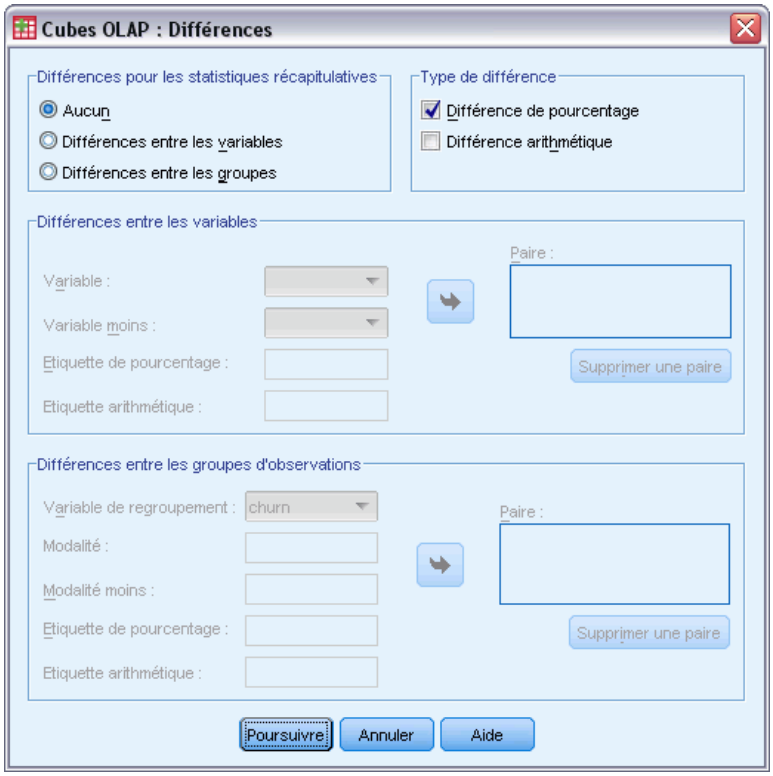
Erreur standard du Skewness. Rapport de l'asymétrie avec son erreur standard, qui peut servir de test de normalité (il y a anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur d'asymétrie positive importante indique une extrémité allongée vers la droite ; une valeur négative extrême produit une extrémité allongée vers la gauche.

Somme. Somme ou total des valeurs, pour toutes les observations n'ayant pas de valeur manquante.

Variance. Mesure de dispersion autour de la moyenne, égale à la somme des carrés des écarts par rapport à la moyenne, divisée par le nombre d'observations moins un. La variance se mesure en unités, qui sont égales au carré des unités de la variable.

Cubes OLAP : Différences

Figure 8-3
Boîte de dialogue Cubes OLAP : Différences



Cette boîte de dialogue vous permet de calculer les différences arithmétiques et de pourcentage qui existent entre des variables récapitulatives ou entre des groupes définis par un critère de regroupement. Les différences sont calculées pour toutes les mesures sélectionnées dans la boîte de dialogue Cubes OLAP : Statistiques.

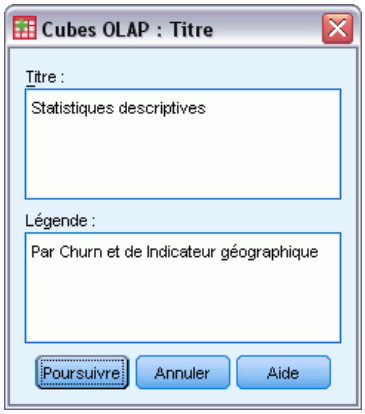
Différences entre les variables. Calcule les différences existant entre des paires de variables. Les valeurs des statistiques récapitulatives de la seconde variable (variable moins) de chaque paire sont soustraites des valeurs des statistiques récapitulatives de la première variable de la paire. Pour les différences de pourcentage, la valeur de la caractéristique de la variable moins est utilisée en

tant que dénominateur. Vous devez sélectionner plusieurs variables récapitulatives dans la boîte de dialogue principale avant d'indiquer les différences entre les variables.

Différences entre les groupes d'observations. Calcule les différences existant entre les paires de groupes définies par une variable de regroupement. Les valeurs des statistiques récapitulatives de la seconde modalité (modalité moins) dans chaque paire sont soustraites des valeurs des statistiques récapitulatives de la première modalité de la paire. Les différences de pourcentage utilisent la valeur de la statistique récapitulative pour la modalité moins en tant que dénominateur. Vous devez sélectionner au moins une variable de regroupement dans la boîte de dialogue principale avant d'indiquer les différences entre groupes.

Cubes OLAP : Titre

Figure 8-4
Boîte de dialogue Cubes OLAP : Titre



Il vous est possible de modifier le titre de votre sortie ou d'ajouter une légende qui apparaîtra au dessous du tableau de sortie. Vous pouvez également contrôler la répartition des titres et légendes sur plusieurs lignes en tapant \n partout où vous souhaitez insérer un saut de ligne dans le texte.

Chapitre
9

Tests T

Il existe trois types de test *t* :

Test T pour échantillons indépendants (test t pour deux échantillons) : Permet de comparer la moyenne d'une variable de deux groupes d'observations. Les statistiques descriptives pour chaque groupe et le test de Levene permettant d'obtenir l'égalité des variances sont disponibles ainsi que les valeurs *t* de variance égale et inégale, et qu'un intervalle de confiance de 95 % pour la différence des moyennes.

Test T pour échantillons appariés (test t dépendant) : Permet de comparer la moyenne de deux variables pour un seul groupe. Ce test sert aussi pour les plans d'études appariées ou de contrôle d'observation. Les résultat incluent les statistiques descriptives pour les variables de test, leurs corrélations, les statistiques descriptives pour les différences appariées, le test *t* et un intervalle de confiance de 95 %.

Test T pour échantillon unique : Permet de comparer la moyenne d'une variable avec une valeur connue ou supposée. Les statistiques descriptives des variables tests sont affichées avec le test *t*. Un intervalle de confiance de 95 % pour la différence entre la moyenne de la variable test et la valeur test supposée fait partie du résultat par défaut.

Test T pour échantillons indépendants

La procédure du Test T pour échantillons indépendants permet de comparer la moyenne de deux groupes d'observations. Idéalement, pour ce test, les sujets doivent être attribués de manière aléatoire à deux groupes, de manière à ce que toute différence dans la réponse soit due au traitement (ou à un manque de traitement) et non pas à d'autres facteurs. Ceci n'est pas le cas si l'on compare un revenu moyen pour les hommes et les femmes. Une personne n'est pas aléatoirement désignée comme devant être un homme ou une femme. Dans de telles situations, il faut s'assurer que les différences dans les autres facteurs ne cachent pas ou n'augmentent de différence significative dans les moyennes. Les différences de revenu moyen peuvent être influencées par des facteurs tels que l'éducation et non par le sexe seul.

Exemple : Les patients souffrant d'hypertension se voient assignés de façon aléatoire à un groupe placebo et à un groupe auquel on donne un traitement. Les sujets du groupe placebo reçoivent une pilule inactive et les sujets du groupe auquel on donne un traitement reçoivent un nouveau médicament supposé réduire l'hypertension. Après que les sujets ont suivi le traitement pendant deux mois, le test *t* pour deux échantillons est utilisé pour comparer la tension artérielle moyenne du groupe placebo à celle du groupe qui suit le traitement. Chaque patient est examiné une fois et appartient à un groupe.

Statistiques : Pour chaque variable, on a les éléments suivants : taille de l'échantillon, moyenne, écart-type, et erreur standard de la moyenne. Pour la différence de la moyenne: moyenne, erreur standard, et intervalle de confiance (vous pouvez spécifier le niveau de confiance). Tests : Test de Levene sur l'égalité des variances et tests *t* des variances combinées et séparées pour l'égalité des moyennes.

Données. Les valeurs de la variable quantitative qui vous intéresse se trouvent dans une seule colonne du fichier de données. La procédure utilise une variable de regroupement à deux valeurs pour séparer les observations en deux groupes. Le critère de regroupement peut être numérique (on peut avoir des valeurs telles que 1 et 2, ou 6,25 et 12,5) ou alphanumérique (telles que *oui* et *non*). Vous pouvez utiliser également une variable quantitative, telle que l'*âge*, pour séparer les observations en deux groupes en précisant une césure (la césure 21 provoque un groupe dont l'*âge* est inférieur à 21 ans et un groupe dont l'âge est supérieur à 21 ans).

Hypothèses : Pour le test *t* de variance égale, les observations doivent être indépendantes, et les échantillons aléatoires de distribution normale doivent avoir la même variance de population. Pour le test *t* de variance égale, les observations doivent être indépendantes, et les échantillons aléatoires doivent avoir une distribution normale. Le test *t* pour deux échantillons est assez robuste pour se départir de la normalité. Lors de la vérification graphique des distributions, vérifiez qu'elles sont symétriques et n'ont pas de valeurs éloignées.

Obtenir un test t pour échantillons indépendants

- A partir des menus, sélectionnez :
  Analyse > Comparer les moyennes > Test T pour échantillons indépendants

Figure 9-1
Boîte de Dialogue Test T pour échantillons indépendants

- Sélectionnez au moins une variable test quantitative. Un test *t* distinct est alors calculé pour chaque variable.
- Sélectionnez un seul critère de regroupement et cliquez sur Définir groupes pour spécifier deux codes pour les groupes à comparer.
- Vous pouvez également cliquer sur Options pour contrôler le traitement des données manquantes et le niveau de l'intervalle de confiance.

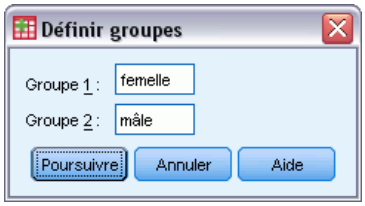
Définir Groupes Test T pour Echantillons Indépendants

Figure 9-2
Boîte de dialogue Définir groupes pour variables numériques

Pour les critères de regroupement numérique, définissez les deux groupes du test *t* en spécifiant deux valeurs ou un point de séparation :

- **Utiliser les valeurs spécifiées :** Saisissez une valeur pour le Groupe 1 et une autre pour le Groupe 2. Les observations qui ont une autre valeur sont exclues de l'analyse. Il n'est pas nécessaire que les nombres soient des entiers (par exemple, 6,25 et 12,5 sont valides).
- **Césure.** Vous avez également la possibilité de saisir un nombre qui sépare les valeurs de la variable de regroupement en deux groupes. Toutes les observations ayant des valeurs inférieures à la césure constituent un groupe et les observations ayant des valeurs supérieures ou égales à la césure constituent l'autre groupe.

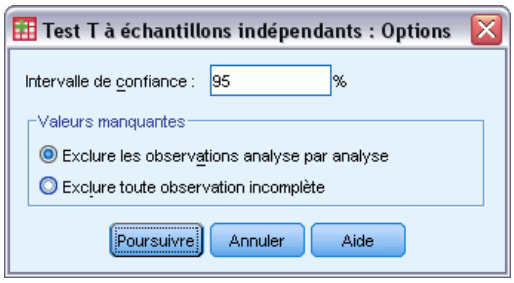
Figure 9-3
Définir la boîte de dialogue Groupes pour les variables caractères



Pour les critères de regroupement alphanumériques, entrez une chaîne pour le Groupe 1 et une autre pour le Groupe 2, par exemple *oui* et *non*. Les observations avec d'autres chaînes sont exclues de l'analyse.

Options Test T pour Echantillons Indépendants

Figure 9-4
Boîte de dialogue Test T pour échantillons indépendants : Options



Intervalle de confiance : Par défaut, un intervalle de confiance de 95 % pour la différence dans les moyennes est affiché. Saisir une valeur comprise entre 1 et 99 pour demander un niveau de confiance différent.

Valeurs manquantes : Quand vous testez plusieurs variables et que des données sont manquantes pour au moins une variable, vous pouvez indiquer à la procédure les observations à inclure (ou exclure).

- **Exclure les observations analyse par analyse :** Chaque test *t* utilise toutes les observations qui ont des données valides pour les variables testées. La taille des échantillons peut varier d'un test à l'autre.
- **Exclure toute observation incomplète :** Chaque test *t* utilise seulement les observations qui ont des données valides pour toutes les variables utilisées dans les tests *t* requis. La taille des échantillons est constante durant les tests.

Test T pour échantillons appariés

La procédure du Test T pour Echantillons Appariés compare la moyenne de deux variables pour un seul groupe. Elle permet de calculer les différences entre les valeurs des deux variables pour chaque observation et de tester si la moyenne diffère de 0.

Exemple : Dans le cadre d'une étude sur l'hypertension, des mesures sont prises sur tous les patients au début de l'étude, un traitement est administré, puis on procède à une nouvelle mesure. Par conséquent, chaque sujet est l'objet de deux mesures, souvent nommées mesures *avant* et *après*. Il existe une alternative à ce test, il s'agit d'une étude appariée ou de contrôle d'observation dans laquelle chaque déclaration dans le fichier de données contient la réponse du patient ainsi que celle de son sujet de contrôle apparié. Dans le cadre d'une étude sur la tension artérielle, les patients et les contrôles peuvent être appariés selon l'âge (un patient âgé de 75 ans avec un membre du groupe de contrôle âgé de 75 ans).

Statistiques. Pour chaque variable, on a les éléments suivants : moyenne, taille d'échantillon, écart-type, et erreur standard de la moyenne. Pour chaque paire de variables, on a les éléments suivants : Corrélation, différence moyenne de moyennes, test *t* et intervalle de confiance pour la différence moyenne (vous pouvez préciser le niveau de confiance). Ecart-type et erreur standard de la différence moyenne.

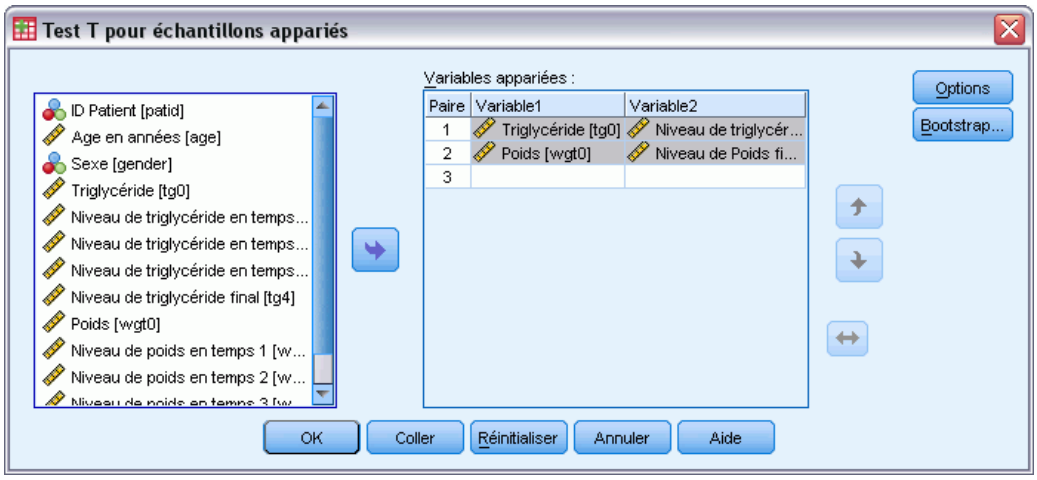
Données. Pour chaque test apparié, précisez deux variables continues (niveau d'intervalle de mesure ou niveau de ratio de mesure). Dans le cadre d'une étude appariée ou de contrôle d'observation, la réponse pour chaque sujet test et son sujet de contrôle apparié doit être dans la même observation du fichier de données.

Hypothèses : Les observations pour chaque paire devraient être réalisées dans les mêmes conditions. Les différences moyennes devraient suivre une distribution normale. Les variances de chaque variable peuvent être égales ou inégales.

Obtenir un test t pour échantillons appariés

▶ A partir des menus, sélectionnez :
Analyse > Comparer les moyennes > Test T pour échantillons appariés

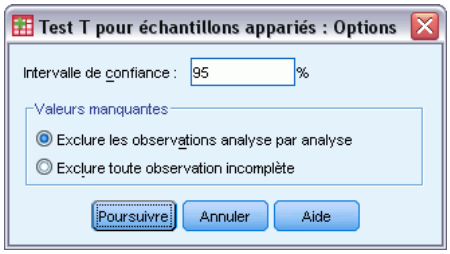
Figure 9-5
Boîte de dialogue Test T pour échantillons appariés



- Sélectionnez une ou plusieurs paires de variables
- Vous pouvez également cliquer sur Options pour contrôler le traitement des données manquantes et le niveau de l'intervalle de confiance.

Options test T pour échantillons appariés

Figure 9-6
Boîte de dialogue Test T pour échantillons appariés



Intervalle de confiance : Par défaut, un intervalle de confiance de 95 % pour la différence dans les moyennes est affiché. Saisir une valeur comprise entre 1 et 99 pour demander un niveau de confiance différent.

Valeurs manquantes : Quand vous testez plusieurs variables et que des données sont manquantes pour au moins une variable, vous pouvez indiquer à la procédure les observations à inclure (ou exclure) :

- **Exclure les observations analyse par analyse :** Chaque test *t* utilise toutes les observations qui ont des données valides pour la paire de variables testées. La taille des échantillons peut varier d'un test à l'autre.
- **Exclure toute observation incomplète :**Chaque test *t* utilise seulement les observations qui ont des données valides pour toutes les paires de variables testées. La taille des échantillons est constante durant les tests.

Test T pour échantillon unique

La procédure du Test T pour échantillon unique permet de tester si la moyenne d'une seule variable diffère d'une constante spécifiée.

Exemples : Un chercheur souhaite tester si le QI moyen d'un groupe d'étudiants diffère de 100. Un fabricant céréalier prélève un échantillon de boîtes à partir d'une chaîne de production et vérifie si le poids moyen des échantillons diffère de 1,3 livres à l'intervalle de confiance 95 %.

Statistiques : Pour chaque variable test : moyenne, écart-type, et erreur standard de la moyenne. Différence moyenne entre chaque valeur de donnée et la valeur test supposée, le test *t* vérifie que cette différence est égale à 0 et vérifie également l'intervalle de confiance pour cette différence (vous pouvez préciser le niveau de confiance).

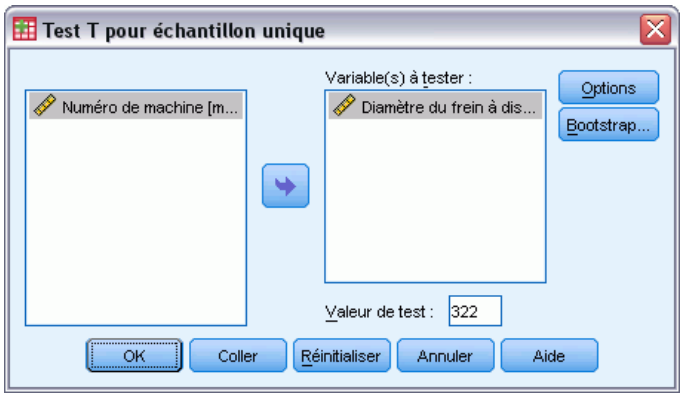
Données. Afin de tester les valeurs d'une variable quantitative par rapport à une valeur test supposée, choisissez une variable quantitative et saisissez une valeur test supposée.

Hypothèses : Ce test suppose que les données sont distribuées normalement ; cependant, ce test résiste convenablement à la normalité.

Obtenir un test t pour échantillon unique

- A partir des menus, sélectionnez :
  Analyse > Comparer les moyennes > Test T pour échantillon unique

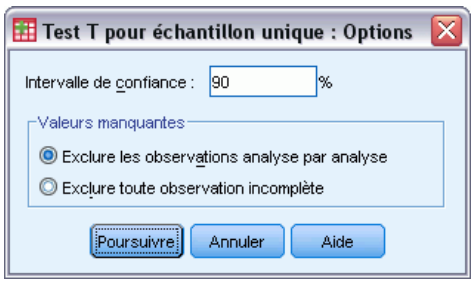
Figure 9-7
Boîte de dialogue Test T pour échantillon unique



- Sélectionnez au moins une variable à tester par rapport à la même valeur supposée.
- Entrez une valeur test numérique à laquelle vous souhaitez comparer chaque moyenne d'échantillon.
- Vous pouvez également cliquer sur Options pour contrôler le traitement des données manquantes et le niveau de l'intervalle de confiance.

Options Test T pour échantillon unique

Figure 9-8
Boîte de Dialogue Options Test T pour échantillon unique



Intervalle de confiance : Par défaut, un intervalle de confiance de 95 % pour la différence entre la moyenne et la valeur de test supposée est affiché. Saisir une valeur comprise entre 1 et 99 pour demander un niveau de confiance différent.

Valeurs manquantes : Quand vous testez plusieurs variables et que des données sont manquantes pour au moins une variable, vous pouvez indiquer à la procédure les observations à inclure (ou exclure).

- **Exclure les observations analyse par analyse :** Chaque test *t* utilise toutes les observations qui ont des données valides pour les variables testées. La taille des échantillons peut varier d'un test à l'autre.
- **Exclure toute observation incomplète :** Chaque test *t* utilise seulement les observations qui ont des données valides pour toutes les variables utilisées dans n'importe lequel des tests *t* requis. La taille des échantillons est constante durant les tests.

Fonctionnalités supplémentaires de la commande T-TEST

Le langage de syntaxe de commande vous permet aussi de :

- Produire à la fois des tests t pour un échantillon et pour des écahntillons indépendants en exécutant une commande unique.
- Tester une variable avec chacune des variables d'une liste dans un test t apparié (avec la sous-commande `PAIRS`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
10

ANOVA à 1 facteur

La procédure de l'analyse de variance ANOVA à 1 facteur permet d'effectuer une analyse de variance univariée sur une variable quantitative dépendante par une variable critère simple (indépendant). L'analyse de variance sert à tester l'hypothèse d'égalité des moyennes. Cette technique est une extension du test *t* pour deux échantillons.

Déterminer que des différences existent parmi les moyennes ne vous suffit peut-être pas. Vous voulez éventuellement savoir quelles sont les moyennes qui diffèrent. Il existe deux types de tests pour comparer les moyennes : les contrastes a priori et les tests post hoc. Les contrastes sont des tests définis *avant* l'expérience, et les tests post hoc sont effectués *après* l'expérience. Vous pouvez aussi tester les tendances à travers les modalités.

Exemple : Les beignets absorbent la graisse dans des proportions variées lorsqu'ils sont cuisinés. Une expérience est conduite à partir de l'utilisation de trois types de graisse : huile d'arachide, huile de maïs, et saindoux. L'huile d'arachide et l'huile de maïs sont des graisses non saturées, et le saindoux une graisse saturée. En plus du fait de déterminer si la quantité de graisse absorbée dépend du type de graisse utilisée, vous pouvez créer un contraste a priori afin de déterminer si le degré d'absorption de graisse diffère pour les graisses saturées et non saturées.

Statistiques. Pour chaque groupe : nombre d'observations, moyenne, écart type, erreur standard pour la moyenne, minimum, maximum et intervalle de confiance à 95 % pour la moyenne. Test de Levene pour l'homogénéité de la variance, tableau d'analyse de la variance et tests d'égalité des moyennes pour chaque variable dépendante, contrastes a priori spécifiés par l'utilisateur et tests d'intervalle et comparaisons multiples post hoc : Bonferroni, Sidak, test de Tukey, GT2 de Hochberg, Gabriel, Dunnett, test *F* de Ryan-Einot-Gabriel-Welsch (R-E-G-W *F*), test d'intervalle de Ryan-Einot-Gabriel-Welsch (R-E-G-W *Q*), T2 de Tamhane, T3 de Dunnett, Games-Howell, test *C* de Dunnett, test de Duncan, Student-Newman-Keuls (S-N-K), *B* de Tukey, Waller-Duncan, Scheffé et différence la moins significative.

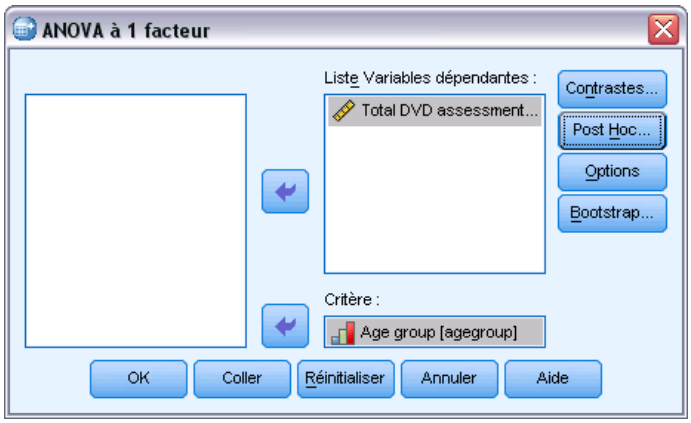
Données. Les valeurs de la variable active devraient être des nombres entiers, et la variable dépendante devrait être quantitative (niveau d'intervalle de mesures).

Hypothèses : Chaque groupe est un échantillon aléatoire indépendant extrait d'une population normale. L'analyse de la variance supporte les écarts à la normalité, bien que les données doivent être symétriques. Les groupes devraient être composés de populations à variance égale. Pour tester cette hypothèse, utiliser le test d'homogénéité de variance de Levene.

Obtenir une analyse de variance à un facteur

- A partir des menus, sélectionnez :
  Analyse > Comparer les moyennes > ANOVA à 1 facteur

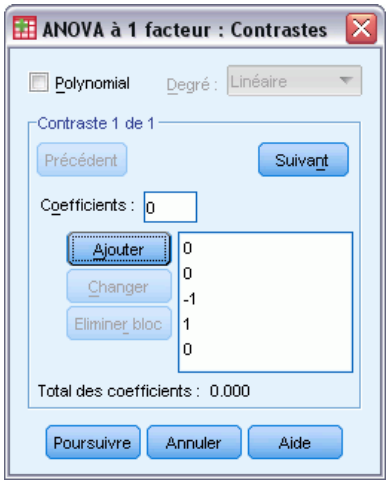
Figure 10-1
Boîte de dialogue ANOVA à 1 facteur



- Sélectionnez au moins une variable dépendante.
- Sélectionnez une variable active indépendante simple.

Contrastes ANOVA à 1 facteur

Figure 10-2
Boîte de dialogue ANOVA à 1 facteur : Contrastes



Vous pouvez diviser les sommes des carrés inter-groupes en tendances composants ou spécifier les contrastes a priori.

Modèle polynomial : Diviser les sommes des carrés inter-groupes en tendances composants. Vous pouvez tester la tendance d'une variable dépendante à travers les niveaux ordonnés de la variable active. Par exemple, vous pourriez tester la tendance linéaire (croissante ou décroissante) des salaires perçus les plus élevés à travers les niveaux ordonnés.

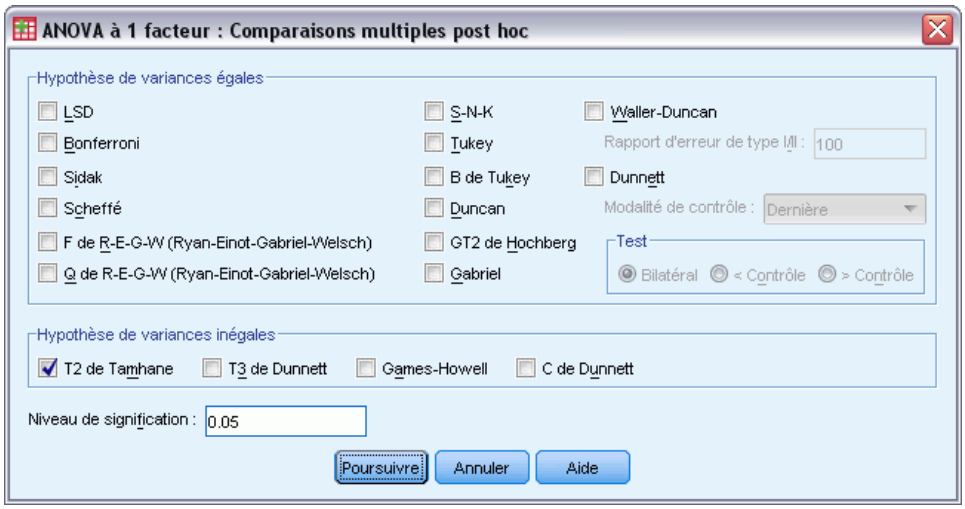
- **Degré :** Vous pouvez choisir un polynôme de premier, deuxième, troisième, quatrième ou cinquième degré.

Coefficients. Contrastes a priori spécifiés à tester par la statistique *t*. Saisissez un coefficient pour chaque groupe (modalité) de la variable active et cliquez sur Ajouter après chaque saisie. Chaque nouvelle valeur s'ajoute au bas de la liste des coefficients. Pour spécifier des groupes de contrastes supplémentaires, cliquez sur Suivant. Utilisez Suivant et Précédent pour vous déplacer entre les groupes de contrastes.

L'ordre des coefficients est important car il correspond à l'ordre croissant des valeurs de modalité de la variable active. Le premier coefficient de la liste correspond à la valeur la plus petite de la variable active, et le dernier coefficient correspond à la valeur la plus élevée. Par exemple, s'il y a six modalités de variables actives, les coefficients –1, 0, 0, 0, 0,5 et 0,5 mettent en contraste le premier groupe avec les cinquième et sixième groupes. Pour la plupart des applications, les coefficients devraient s'élever à 0. Les groupes qui n'atteignent pas 0 peuvent aussi être utilisés, mais un message d'avertissement s'affiche.

Tests Post Hoc ANOVA à 1 facteur

Figure 10-3
ANOVA à 1 facteur : Comparaisons multiples a posteriori



Lorsque vous avez déterminé qu'il existe des différences parmi les moyennes, les tests d'intervalles post hoc et de comparaisons multiples par paire peuvent déterminer les moyennes qui diffèrent. Les tests d'intervalle identifient les sous-groupes homogènes de moyennes qui ne diffèrent pas les uns des autres. Les comparaisons multiples appariées testent la différence entre les moyennes appariées et engendrent une matrice pour laquelle les astérisques indiquent les moyennes de groupes significativement différentes au niveau alpha 0.05.

Hypothèse de variances égales

Le test de Tukey, le GT2 de Hochberg, le test de Gabriel et le test de Scheffé sont des tests de comparaisons multiples et d'intervalle. Il existe d'autres tests d'intervalle, tels que le test *B* de Tukey, le S-N-K (Student-Newman-Keuls), le Duncan, le R-E-G-W *F* (*F* de Ryan-Einot-Gabriel-Welsch), le R-E-G-W *Q* (test d'intervalle de Ryan-Einot-Gabriel-Welsch) et le Waller-Duncan. Les tests de comparaison multiple disponibles sont les suivants : Bonferroni,

test de différence significative de Tukey, Sidak, Gabriel, Hochberg, Dunnett, Scheffé et LSD (différence la moins significative).

- **LSD.** Utilisation de tests t pour effectuer toutes les comparaisons par paire entre des moyennes de groupe. Le taux d'erreur n'est pas corrigé dans le cas de comparaisons multiples.
- **Bonferroni.** Utilise des tests t pour effectuer des comparaisons par paire entre les moyennes de groupes, mais contrôle le taux d'erreur global en spécifiant comme taux d'erreur pour chaque test le taux d'erreur empirique divisé par le nombre total de tests. Le seuil de signification observé est ainsi ajusté en raison des comparaisons multiples réalisées.
- **Sidak.** Test de comparaisons multiples par paire reposant sur la statistique t. Le test Sidak ajuste le seuil de signification en fonction des comparaisons multiples et fournit des bornes plus étroites que le test Bonferroni.
- **Scheffé.** Exécute des comparaisons par paire simultanées pour toutes les paires de moyennes possibles. Utilise la distribution d'échantillonnage F. Peut servir à examiner toutes les combinaisons linéaires possibles de moyennes de groupe, et pas seulement des comparaisons par paire.
- **F de R-E-G-W (Ryan-Einot-Gabriel-Welsch).** Procédure multiple descendante de Ryan-Einot-Gabriel-Welsch basée sur un test F.
- **Q de R-E-G-W (Ryan-Einot-Gabriel-Welsch).** Procédure multiple descendante de Ryan-Einot-Gabriel-Welsch basée sur un intervalle de Student.
- **S-N-K.** Ce test effectue toutes les comparaisons de moyennes par paire, à l'aide de la distribution des intervalles de Student. Lorsque la taille des échantillons est égale, il compare aussi les moyennes par paire dans les sous-ensembles homogènes, en utilisant une procédure pas à pas. Les moyennes sont triées dans l'ordre décroissant et les différences extrêmes sont testées en premier.
- **Tukey.** Utilise la statistique de plages de Student pour effectuer toutes les comparaisons de groupes par paire. Fixe le taux d'erreur expérimental au niveau du taux d'erreur de l'ensemble pour toutes des comparaisons par paire.
- **B de Tukey :.** Utilise la distribution de plages de Student pour effectuer des comparaisons de groupes par paire. La valeur critique est la moyenne de la valeur correspondante du test de Tukey et du test de Student-Newman-Keuls.
- **Duncan.** Réalise des comparaisons par paires en suivant un ordre pas à pas identique à celui utilisé dans le test de Student-Newman-Keuls, mais établit un niveau de protection du taux d'erreur pour l'ensemble des tests, plutôt que pour chaque test en particulier. Utilise la statistique d'intervalle de Student.
- **GT2 de Hochberg.** Test de multiples comparaisons et intervalles appariés utilisant le modulus maximum de Student. Similaire au test de Tukey.
- **Gabriel.** Test de comparaison par paire qui utilise le modulus maximum de Student. Il est plus efficace que le GT2 de Hochberg lorsque les tailles des cellules sont inégales. Le test de Gabriel offre plus de souplesse lorsque les tailles des cellules divergent beaucoup.
- **Waller-Duncan.** Test de comparaisons multiples reposant sur une statistique t et utilisant une approche bayésienne.
- **Dunnett.** Test-t de comparaisons multiples par paires comparant un ensemble de traitements à une moyenne de contrôle unique. La dernière modalité est la modalité de contrôle par défaut. Vous pouvez également choisir la première modalité. L'option Bilatéral teste que la moyenne

à un certain niveau (hormis la modalité de contrôle) du facteur n'est pas égale à celle de la modalité de contrôle. L'option <Contrôle permet de tester si la moyenne est inférieure, à un certain niveau du facteur, à celle de la modalité de contrôle. L'option > Contrôle permet de tester si la moyenne est supérieure, à un certain niveau du facteur, à celle de la modalité de contrôle.

Hypothèse de variances inégales

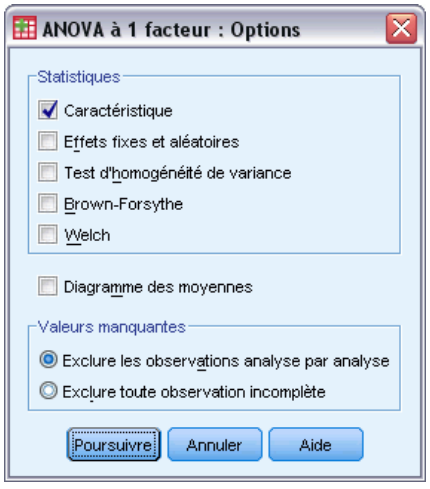
Les tests de comparaison multiple qui ne supposent pas de variances égales sont le T2 de Tamhane, le T3 de Dunnett, Games-Howell et le *C* de Dunnett.

- **T2 de Tamhane.** Test de comparaison par paire conservatif reposant sur le test T. Ce test est opportun lorsque les variances sont inégales.
- **T3 de Dunnett.** Test des comparaisons par paires basé sur le module maximal de Student. Ce test est opportun lorsque les variances sont inégales.
- **Games-Howell.** Test des comparaisons appariées qui peut parfois être souple. Ce test est opportun lorsque les variances sont inégales.
- **C de Dunnett.** Test des comparaisons par paires basé sur l'intervalle de Student. Ce test est opportun lorsque les variances sont inégales.

Remarque : Il peut vous paraître plus facile d'interpréter le résultat à partir de tests post hoc si vous désactivez l'option Masquer les lignes et les colonnes vides dans la boîte de dialogue Propriétés du tableau (dans le tableau pivotant activé, choisissez Propriétés du tableau dans le menu Format).

Options ANOVA à 1 facteur

Figure 10-4
Boîte de dialogue ANOVA à 1 facteur : Options



Statistiques. Choisissez une ou plusieurs des options suivantes :

- **Caractéristique :** La procédure calcule le nombre d'observations, la moyenne, l'écart type, l'erreur standard de la moyenne, le minimum, le maximum, et les intervalles de confiance à 95 % pour chaque variable dépendante de chaque groupe.

- **Effets fixes et aléatoires :** La procédure affiche l'écart type, l'erreur standard et l'intervalle de confiance à 95 % pour le modèle à effets fixes, ainsi que l'erreur standard, l'intervalle de confiance à 95 % et l'estimation de la variance inter-composants pour le modèle à effets aléatoires.
- **Test d'homogénéité de variance :** La procédure calcule la statistique de Levene pour tester l'égalité des variances de groupe. Ce test ne dépend pas de l'hypothèse de normalité.
- **Brown-Forsythe :** La procédure calcule la statistique de Brown-Forsythe pour tester l'égalité des moyennes de groupe. Il est préférable d'utiliser cette statistique (au lieu de la statistique *F*) lorsque l'hypothèse d'égalité des variances n'est pas satisfaite.
- **Welch :** Calcule la statistique de Welch pour tester l'égalité des moyennes de groupe. Il est préférable d'utiliser cette statistique (au lieu de la statistique *F*) lorsque l'hypothèse d'égalité des variances n'est pas satisfaite.

Diagramme des moyennes : Affiche un diagramme qui représente les moyennes de sous-groupes (les moyennes de chaque groupe définies par les valeurs de la variable active).

Valeurs manquantes. Contrôle le traitement des valeurs manquantes.

- **Exclure les observations analyse par analyse :** Aucune observation avec valeur manquante n'est utilisée, que ce soit pour la variable dépendante ou pour la variable active d'une analyse donnée. De même, on n'utilise pas d'observation en dehors de l'intervalle spécifié pour la variable active.
- **Exclure toute observation incomplète.** Les observations ayant des valeurs manquantes pour la variable active ou pour toute variable dépendante contenue dans la liste dépendante de la boîte de dialogue principale sont exclues de toutes les analyses. Si vous n'avez pas spécifié de variables multiples dépendantes, cela est sans effet.

Fonctionnalités supplémentaires de la commande ONEWAY

Le langage de syntaxe de commande vous permet aussi de :

- Obtenir des statistiques à effets fixes et aléatoires Écart type, erreur standard de la moyenne et intervalles de confiance de 95 % pour le modèle à effets fixes. Erreur standard, intervalles de confiance de 95 % et estimation de la variance inter-composants pour le modèle à effets aléatoires (en utilisant `STATISTICS=EFFECTS`).
- Spécifier les niveaux alpha pour la différence de moindre signification, tests de comparaison multiple Bonferroni, Duncan et Scheff (avec la sous-commande `RANGES`).
- Ecrire une matrice des moyennes, des écarts-types et des fréquences ou lire une matrice des moyennes, des fréquences, des variances combinées et des degrés de liberté des variances combinées. Ces matrices peuvent être utilisées à la place des données brutes pour obtenir une analyse à un facteur de la variance (avec la sous-commande `MATRIX`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
11

Analyse GLM – Univarié

GLM – Univarié fournit un modèle de régression et une analyse de la variance pour plusieurs variables dépendantes par un ou plusieurs facteurs ou variables. Les variables actives divisent la population en groupes. Cette procédure de régression linéaire généralisée vous permet de tester les hypothèses nulles à propos des effets des autres variables sur la moyenne de différents regroupements de la variable dépendante. Vous pouvez rechercher les interactions entre les facteurs ainsi que les effets des différents facteurs, certains d'entre eux étant aléatoires. En outre, les effets et les interactions des covariables avec les facteurs peuvent être inclus. Pour l'analyse de la régression, les variables indépendantes (explicatives) sont spécifiées comme covariables.

Vous pouvez tester les modèles équilibrés comme déséquilibrés. Un modèle est équilibré si chaque cellule de ce modèle contient le même nombre d'observations. L'analyse GLM – Univarié teste non seulement les hypothèses mais elle produit également des estimations.

Vous disposez de contrastes a priori communs pour effectuer les tests d'hypothèse. En outre, lorsqu'un test *F* global se révèle significatif, vous pouvez utiliser les tests post hoc pour évaluer les différences entre les moyennes spécifiques. Les moyennes marginales estimées fournissent des estimations des valeurs moyennes estimées pour les cellules dans le modèle et les diagrammes des profils (diagrammes d'interaction) de ces moyennes vous permettent de visualiser plus facilement certaines des relations.

Les résidus, les prévisions, la distance de Cook et les valeurs influentes peuvent être enregistrées sous forme de nouvelles variables dans votre fichier de données pour vérifier les hypothèses.

Poids WLS. Vous permet de spécifier une variable utilisée pour pondérer les observations pour une analyse pondérée (WLS) des moindres carrés, peut-être pour compenser les différents niveaux de précision des mesures.

Exemple : Des données sont collectées sur les différents participants au Marathon de Paris sur plusieurs années. Le temps effectué par chaque participant est la variable dépendante. Les autres facteurs comprennent le temps (froid, modéré, chaud), le nombre de mois d'entraînement, le nombre de marathons précédemment effectués et le sexe. L'âge est considéré comme co-variable. Vous devez trouver que le sexe a un effet significatif et que l'interaction du sexe avec le temps est significatif.

Méthodes. Les sommes des carrés de type I, II, III et IV peuvent servir à évaluer les différentes hypothèses. Le type III est la valeur par défaut.

Statistiques : Tests d'intervalle post hoc et comparaisons multiples : La différence la moins significative, Bonferroni, Sidak, Scheffé, *F* multiple de Ryan-Einot-Gabriel-Welsch, l'intervalle multiple de Ryan-Einot-Gabriel-Welsch, Student-Newman-Keuls, le test de Tukey, *b* de Tukey, Duncan, GT2 de Hochberg, Gabriel, le test *t* de Waller Duncan, Dunnett (unilatéral, bilatéral), T2 de Tamhane, T3 de Dunnett, Games-Howell et *C* de Dunnett. Statistiques descriptives : moyenne observée, écart-type et effectifs pour toutes les variables dépendantes dans toutes les cellules. Le test de Levene pour l'homogénéité de la variance.

Diagrammes. Dispersion par niveau, résiduels et profils (interaction).

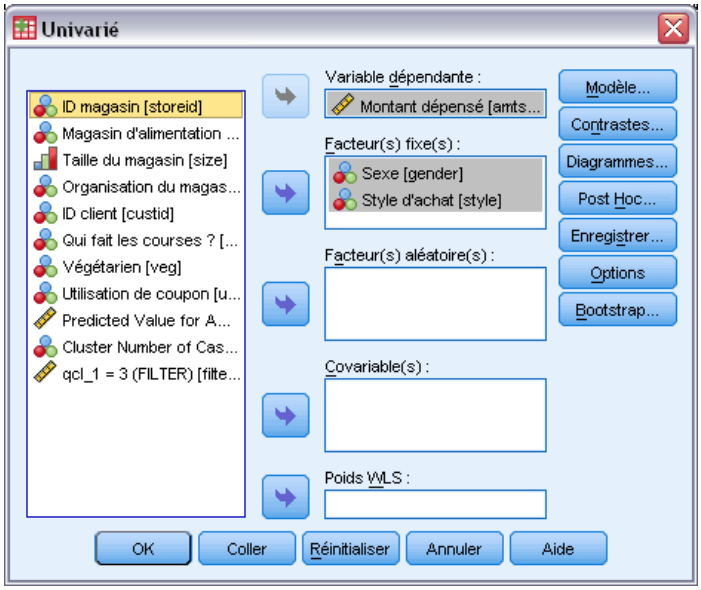
Données. La variable dépendante est quantitative. Les facteurs sont qualitatifs. Il peut s'agir de valeurs numériques ou alphanumériques de 8 caractères au maximum. Les covariables sont des variables quantitatives liées à la variable dépendante.

Hypothèses : Les données forment un échantillon aléatoire d'une population normale ou gaussienne. Dans cette population, toutes les variances de cellule sont égales. L'analyse de la variance supporte les écarts à la normalité, bien que les données doivent être symétriques. Pour vérifier les hypothèses, vous pouvez utiliser les tests d'homogénéité de la variance et les diagrammes de dispersion par niveau. Vous pouvez également étudier les résidus et les diagrammes de résidus.

Pour obtenir des tables GLM - Univarié

- A partir des menus, sélectionnez :
  Analyse > Modèle linéaire général > Univarié

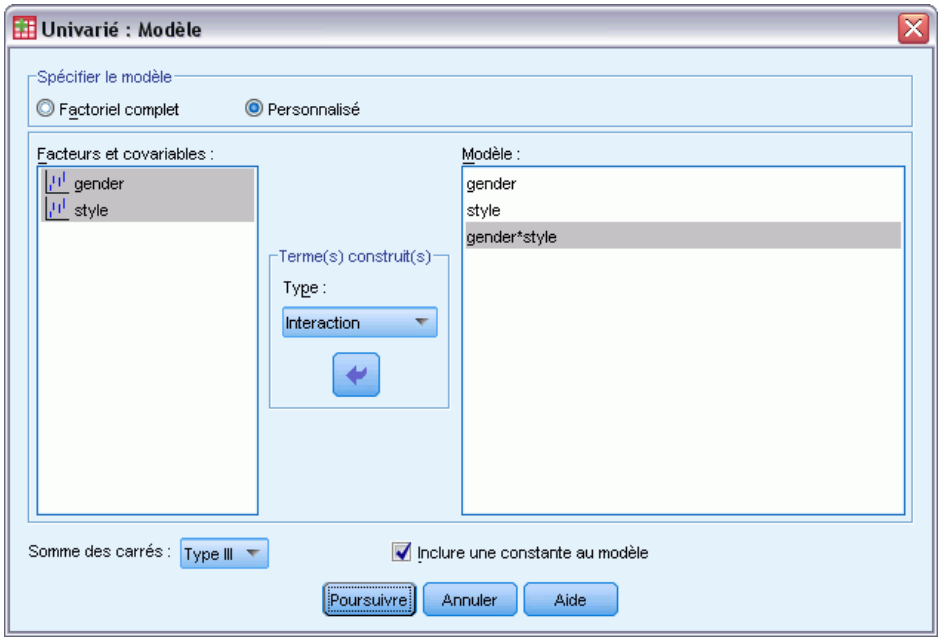
Figure 11-1
Boîte de dialogue GLM Univarié



- Sélectionnez une variable dépendante.
- Sélectionnez des variables pour Facteurs fixés, Facteurs aléatoires et Covariables en fonction de vos données.
- En option, vous pouvez utiliser WLS Weight pour préciser une variable de pondération pour l'analyse des moindres carrés pondérés. Si la valeur de la variable de pondération est nulle, négative ou manquante, l'observation est exclue de l'analyse. Une variable déjà utilisée dans le modèle ne peut pas servir de variable de pondération.

Modèle GLM

Figure 11-2
Boîte de dialogue Modèle univarié



Spécifier le modèle : Un modèle factoriel général contient tous les effets principaux des facteurs, des covariables et toutes les interactions facteur/facteur. Il ne contient pas de d'interactions de covariable. Sélectionnez Autre pour indiquer un sous-ensemble d'interactions ou des interactions variable active/covariable. Vous devez indiquer tous les termes à inclure dans le modèle.

Critères et covariables :SFM Les facteurs et les covariables sont répertoriés.

Modèle : Le modèle dépend de la nature de vos données. Après avoir sélectionné Autre, vous pouvez choisir les effets principaux et les interactions qui présentent un intérêt pour votre analyse.

Somme des carrés Méthode de calcul des sommes des carrés. Pour les modèles équilibrés ou non, auxquels aucune cellule ne manque, le type III est la méthode le plus fréquemment utilisée.

Inclure une constante au modèle : L'ordonnée est généralement incluse dans le modèle. Si vous partez du principe que les données passent par l'origine, vous pouvez exclure la constante.

Termes construits

Pour les facteurs et covariables sélectionnés :

Interaction : Crée le terme d'interaction du plus haut niveau de toutes les variables sélectionnées. Il s'agit de la valeur par défaut.

Effets principaux : Crée un terme d'effet principal pour chaque variable sélectionnée.

Toutes d'ordre 2 : Crée toutes les interactions d'ordre 2 possibles des variables sélectionnées.

Toutes d'ordre 3 : Crée toutes les interactions d'ordre 3 possibles des variables sélectionnées.

Toutes d'ordre 4 : Crée toutes les interactions d'ordre 4 possibles des variables sélectionnées.

Toutes d'ordre 5 : Crée toutes les interactions d'ordre 5 possibles des variables sélectionnées.

Somme des carrés

Pour ce modèle, vous pouvez choisir un type de sommes des carrés. Le type III est le plus courant et c'est la valeur par défaut.

Type I : Cette méthode est également appelée décomposition hiérarchique de la somme des carrés. Chaque terme est ajusté uniquement pour le terme qui le précède dans le modèle. La somme des carrés de type I est généralement utilisée pour :

- Une analyse de la variance équilibrée dans laquelle tout effet principal est spécifié avant les effets d'interaction de premier ordre, et chaque effet de premier ordre spécifié avant ceux de second ordre, et ainsi de suite.
- Un modèle de régression polynomial dans lequel les termes d'ordre inférieur sont spécifiés avant ceux d'ordre supérieur.
- Un modèle par imbrication pur dans lequel le premier effet spécifié est imbriqué dans le second et le second spécifié dans le troisième, etc. (Cette forme d'imbrication peut être spécifiée par la syntaxe uniquement.)

Type II : Cette méthode calcule les sommes des carrés d'un effet dans le modèle ajusté pour tous les autres effets « appropriés ». Un effet approprié est un effet qui correspond à tous les effets qui ne contiennent pas l'effet à étudier. La méthode des sommes des carrés de type II sert généralement pour :

- Une analyse de la variance équilibrée.
- Tout modèle qui contient un effet principal uniquement.
- Tout modèle de régression.
- Un modèle par emboîtement pur. (Cette forme d'emboîtement peut être spécifiée par la syntaxe.)

Type III : Valeur par défaut. Cette méthode calcule les sommes des carrés d'un effet dans le modèle comme les sommes des carrés ajustée pour tout autre effet qui ne le contient pas et orthogonal à chaque effet qui le contient. Les sommes de carrés de type III présentent l'avantage essentiel qu'elles ne varient pas avec les fréquences de cellule tant que la forme générale d'estimabilité reste constante. Ce type de somme des carrés est donc souvent considéré comme utile pour les modèles déséquilibrés auxquels aucune cellule ne manque. Dans le modèle factoriel sans cellule manquante, cette méthode est équivalente à la technique de Yates des carrés moyens pondérés. La méthode des sommes des carrés de type III sert généralement pour :

- Tous les modèles énumérés dans les types I et II.
- Tous les modèles équilibrés ou non qui ne contiennent pas de cellules vides.

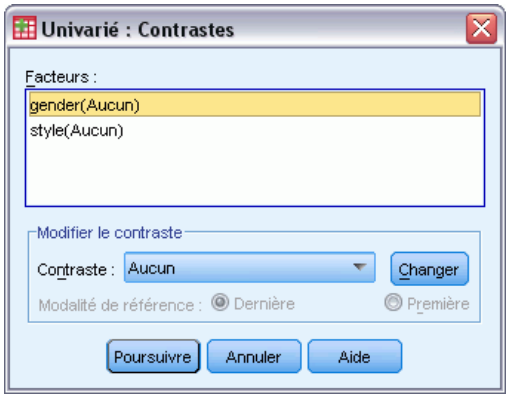
Type IV : Cette méthode est conçue pour une situation dans laquelle il manque des cellules. Pour chaque effet *F* dans le modèle, si *F* n'est inclus dans aucun autre effet, Type IV = Type III = Type II. Si *F* est inclus dans d'autres effets, le Type IV distribue les contrastes à effectuer parmi les

paramètres dans *F* sur tous les effets de niveau supérieur de façon équitable. La méthode des sommes des carrés de type IV sert généralement pour :

- Tous les modèles énumérés dans les types I et II.
- Tous les modèles équilibrés ou non qui contiennent des cellules vides.

Contrastes GLM

Figure 11-3
Boîte de dialogue GLM Univarié : Contrastes



Les contrastes servent à tester les différences entre les niveaux d'un facteur. Vous pouvez spécifier un contraste pour chaque facteur du modèle (dans un modèle de mesures répétées, pour chaque facteur inter-sujets). Les contrastes représentent des combinaisons linéaires des paramètres.

Le test des hypothèses est fondé sur l'hypothèse nulle **LB** =0, **L** étant la matrice des coefficients de contraste et **B** le vecteur de paramètre. Lorsqu'un contraste est spécifié, une matrice **L** est créée. Les colonnes de la matrice **L** correspondantes au facteur concordent avec le contraste. Les colonnes restantes sont ajustées de telle sorte que la matrice **L** puisse être estimée.

Le résultat reprend une statistique *F* pour chaque ensemble de contrastes. Pour les différences de contraste, le système affiche également les intervalles de confiance simultanés de type Bonferroni fondés sur la distribution *t* de Student.

Contrastes possibles

Les contrastes fournis sont écart, simple, différence, Helmert, répétée et modèle polynomial. Pour les contrastes d'écart et simple, vous pouvez choisir si la modalité de référence est la première ou la dernière.

Types de contraste

Ecart : Compare la moyenne de chaque niveau (hormis une modalité de référence) à la moyenne de tous les niveaux (grande moyenne). Les niveaux du facteur peuvent être de n'importe quel ordre.

Simple : Compare la moyenne de chaque niveau à celle d'un niveau donné. Ce type de contraste est utile lorsqu'il y a un groupe de contrôle. Vous pouvez prendre la première ou la dernière modalité en référence.

Différence : Compare la moyenne de chaque niveau (hormis le premier) à la moyenne des niveaux précédents. (Parfois appelé contrastes d'Helmert inversé.)

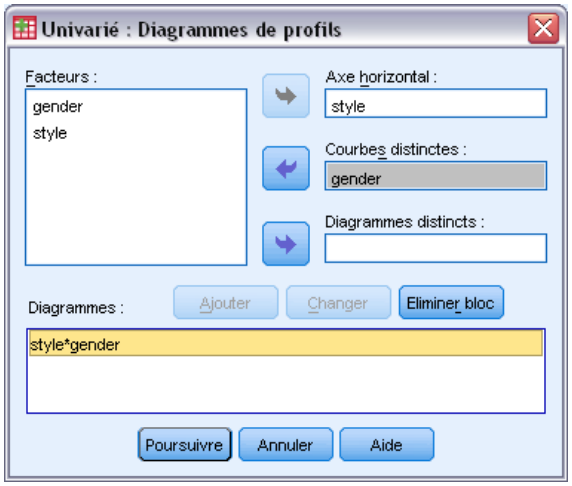
Helmert : Compare la moyenne de chaque niveau de facteur (hormis le dernier) à la moyenne des niveaux suivants.

Répété : Compare la moyenne de chaque niveau (hormis le premier) à la moyenne du niveau suivant.

Modèle polynomial : Compare l'effet linéaire, l'effet quadratique, l'effet cubique etc. Le premier degré de liberté contient l'effet linéaire sur toutes les modalités, le second degré l'effet quadratique, etc. Ces contrastes servent souvent à estimer les tendances polynomiales.

Diagrammes de profils GLM

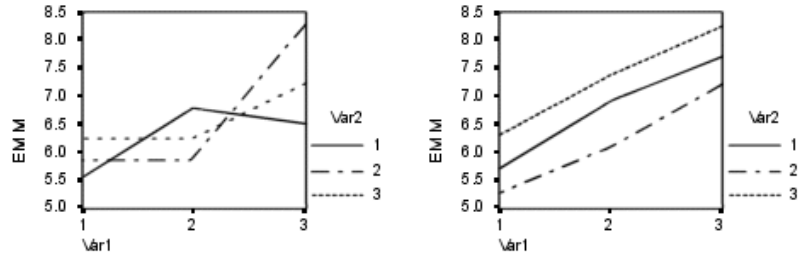
Figure 11-4
Boîte de dialogue GLM – Univarié : Diagrammes des protocoles



Les diagrammes des profils (diagrammes d'interaction) sont utiles pour comparer les moyennes marginales dans votre modèle. Un diagramme des profils est une courbe dont chaque point indique la moyenne marginale estimée d'une variable dépendante (ajustée pour les covariables) à un niveau du facteur. Les niveaux d'un second facteur peuvent servir à dessiner des courbes distinctes. Chaque niveau dans un troisième facteur peut servir à créer un diagramme distinct. Tous les facteurs fixés et aléatoire sont disponibles pour les diagrammes. Pour les analyses multivariées, les diagrammes des profils sont créés pour chaque variable dépendante. Dans une analyse à mesures répétées, à la fois les facteurs inter-sujets et intra-sujets peuvent être utilisés dans les diagrammes des profils. GLM - Multivarié et GLM - Mesures Répétées ne sont disponibles que si vous avez installé l'option Statistiques avancées.

Un diagramme des profils pour un facteur montre si la moyenne marginale estimée est croissante ou décroissante sur les niveaux. Pour au moins deux facteurs, des courbes parallèles indiquent qu'il n'y a pas d'interaction entre les facteurs, ce qui signifie que vous recherchez les niveaux d'un seul facteur. Les courbes non parallèles indiquent une interaction.

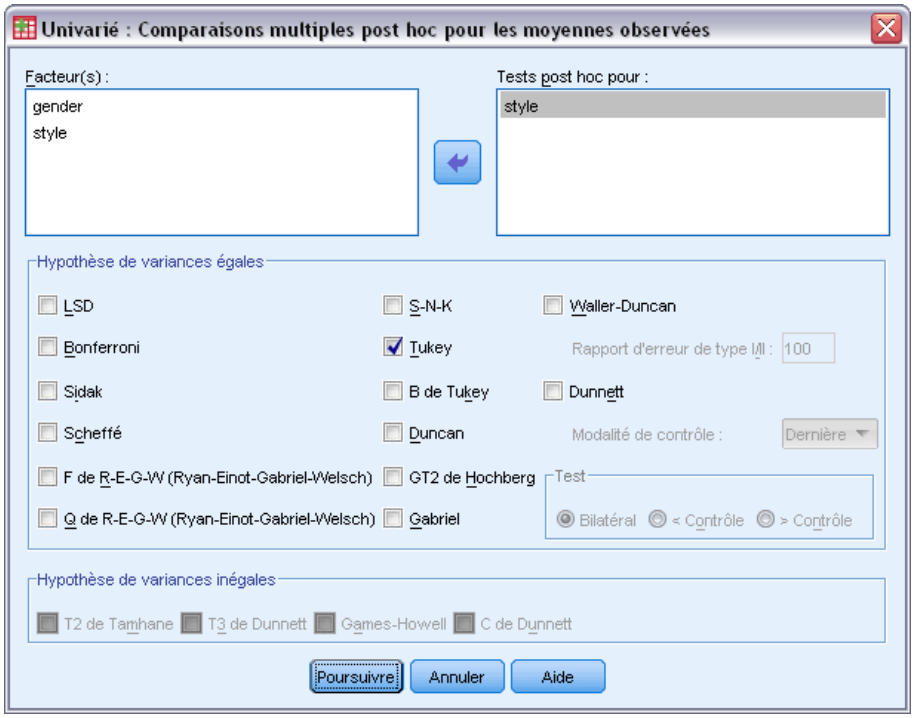
Figure 11-5
Diagramme non parallèle (gauche) et diagramme parallèle (droite)



Après avoir sélectionné des facteurs pour l'axe horizontal afin de spécifier un diagramme et, éventuellement, des facteurs pour des courbes ou des diagrammes distincts, vous devez ajouter le diagramme à la liste Diagrammes.

Comparaisons post hoc GLM

Figure 11-6
Boîte de dialogue Post Hoc



Test de comparaison multiple post hoc : Lorsque vous avez déterminé qu'il existe des différences parmi les moyennes, les tests d'intervalles post hoc et de comparaisons multiples par paire peuvent déterminer les moyennes qui diffèrent. Les comparaisons sont effectuées sur des valeurs non-ajustées. Ces tests servent aux facteurs inter-sujets fixés seulement. Dans GLM - Mesures répétées, ces tests ne sont pas disponibles s'il n'y a pas de facteurs inter-sujets. Les tests de comparaisons multiples post hoc sont effectués pour la moyenne de tous les niveaux des facteurs intra-sujets. Pour GLM - Multivarié, les tests post hoc sont effectués séparément pour chaque

variable dépendante. GLM - Multivarié et GLM - Mesures Répétées ne sont disponibles que si vous avez installé l'option Statistiques avancées.

Les tests de différence significative de Bonferroni et Tukey servent généralement comme tests de comparaison multiples. Le **test de Bonferroni**, fondé sur la statistique *t* de Student, ajuste le niveau de signification observé en fonction du nombre de comparaisons multiples qui sont effectuées. Le **test t de Sidak** ajuste également le niveau de signification et fournit des limites plus strictes que le test de Bonferroni. Le **test de Tukey** utilise la statistique d'intervalle selon Student pour effectuer des comparaisons par paire entre les groupes et fixe le taux d'erreur empirique au taux d'erreur du regroupement de toutes les comparaisons par paire. Lorsque vous testez un grand nombre de paires de moyennes, le test de Tukey est plus efficace que celui de Bonferroni. Lorsqu'il y a peu de paires, Bonferroni est plus efficace.

Le **GT2 de Hochberg** est similaire au test de Tukey mais il utilise un modulus maximum selon Student. Le test de Tukey est généralement plus efficace. Le **test de comparaison par paire de Gabriel** utilise également le modulus maximum selon Student. Il est plus efficace que le GT2 de Hochberg lorsque les tailles des cellules sont inégales. Le test de Gabriel offre plus de souplesse lorsque les tailles des cellules divergent beaucoup.

Le **test de comparaison multiple de Dunnett** compare un ensemble de traitements à une simple moyenne de contrôle. La dernière modalité est la modalité de contrôle par défaut. Vous pouvez également choisir la première modalité. Vous pouvez également choisir un test unilatéral ou bilatéral. Pour tester que la moyenne à un certain niveau (hormis la modalité de contrôle) du facteur n'est pas égale à celle de la modalité de contrôle, utilisez le test double-face. Pour tester si la moyenne est inférieure, à un certain niveau du facteur, à celle de la modalité de contrôle, sélectionnez < Contrôle. Pour tester si la moyenne est supérieure, à un certain niveau du facteur, à celle de la catégorie de contrôle, sélectionnez > Contrôle.

Ryan, Einot, Gabriel et Welsch (R-E-G-W) ont développé deux tests d'intervalles multiples descendants. Les procédures multiples descendantes testent d'abord que toutes les moyennes sont égales. Si toutes les moyennes ne sont pas égales, l'égalité est testée sur des sous-ensembles de moyennes. Le **F de R-E-G-W** est fondé sur le test *F* et le **Q de R-E-G-W** est fondé sur l'intervalle selon Student. Ces tests sont plus efficaces que le test d'intervalles multiples de Duncan et Student-Newman-Keuls (procédures multiples descendantes), mais ils sont conseillés lorsque les cellules sont de taille inégale.

Lorsque les variances sont inégales, utilisez le **T2 de Tamhane** (test de comparaisons par paire conservatif fondé sur un test *t*), le **T3 de Dunnett** (comparaison par paire fondée sur le modulus maximal selon Student), le **test de comparaison par paire** de **Games-Howell** (parfois flexible) ou le **C de Dunnett** (test de comparaison par paire fondé sur l'intervalle selon Student). Notez que s'il y a plusieurs facteurs dans le modèle, ces tests ne sont pas valides et ne seront pas produits.

Le **test d'intervalles multiples de Duncan**, Student-Newman-Keuls (**S-N-K**) et le **b de Tukey** sont des tests d'intervalle qui classifient les moyennes de groupe et calculent une valeur d'intervalle. Ces tests ne sont pas utilisés aussi souvent que les tests évoqués précédemment.

Le **test t de Waller-Duncan** utilise une approche de Bayes. Ce test d'intervalle utilise la moyenne harmonique de la taille de l'échantillon lorsque les échantillons sont de tailles différentes.

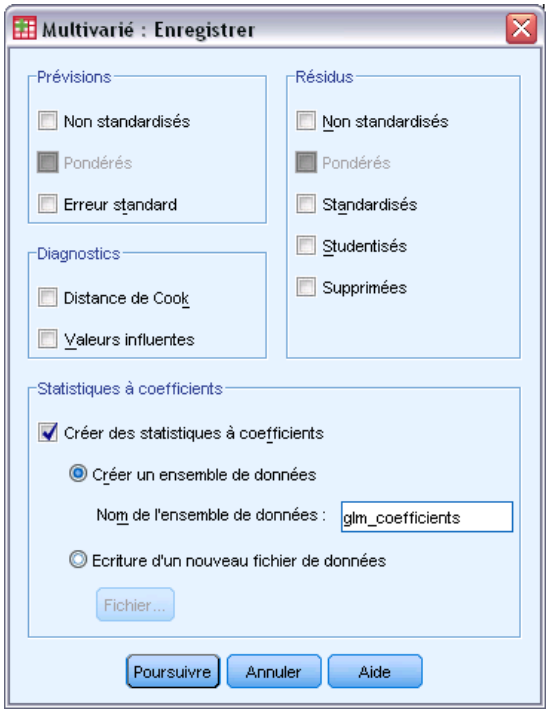
Le niveau de signification du **test de Scheffé** est conçu pour permettre toutes les combinaisons linéaires possibles des moyennes de groupe à tester, pas seulement par paire, disponibles dans cette fonction. Il en résulte que le test de Scheffé est souvent plus strict que les autres, ce qui signifie qu'une plus grande différence de moyenne est nécessaire pour être significative.

Le test de comparaison multiple par paire de différence la moins significative (**LSD**) est équivalent aux divers tests *t* individuels entre toutes les paires des groupes. L'inconvénient de ce test est qu'il n'essaie pas d'ajuster le niveau d'importance observée pour les comparaisons multiples.

Tests affichés : Les comparaisons par paire sont proposées pour LSD, Sidak, Bonferroni, Games et Howell, T2 et T3 de Tamhane, *C* et T3 de Dunnett. Des sous-ensembles homogènes pour les tests d'intervalle sont proposés pour S-N-K, *b* de Tukey, Duncan, *F* et *Q* de R-E-G-W et Waller. Le test de Tukey, le GT2 de Hochberg, le test de Gabriel et le test de Scheffé sont à la fois des tests de comparaison multiple et des tests d'intervalle.

Enregistrement GLM

Figure 11-7
Enregistrer



Vous pouvez enregistrer les prévisions par le modèle, les résidus et les mesures associées sous forme de nouvelles variables dans l'éditeur de données. La plupart de ces variables peuvent servir à étudier les hypothèses relatives aux données. Pour enregistrer les valeurs afin de les utiliser dans une autre session IBM® SPSS® Statistics, vous devez enregistrer le fichier de données en cours.

Prévisions : Valeurs que le modèle estime pour chaque observation.

- **Non standardisés :.** Valeur prévue par le modèle pour la variable dépendante.
- **Pondéré.** Valeurs estimées non standardisées pondérées. Disponibles uniquement lorsqu'une variable WLS a été préalablement sélectionnée.
- **Erreur standard.** Estimation de l'écart-type de la valeur moyenne de la variable dépendante, pour des observations ayant la même valeur pour les variables indépendantes.

Diagnostics : Mesures permettant d'identifier les observations avec des combinaisons inhabituelles de valeurs pour les variables indépendantes et les observations qui peuvent avoir un impact important sur le modèle.

- **Distance de Cook.** Mesure du degré dont les résidus de toutes les observations sont modifiés si une observation donnée est exclue des calculs des coefficients de régression. Si la distance de Cook est élevée, l'exclusion d'une observation changerait substantiellement la valeur des coefficients.
- **Valeurs influentes.** Valeurs influentes non centrées. Mesure de l'influence d'un point sur l'ajustement de la régression.

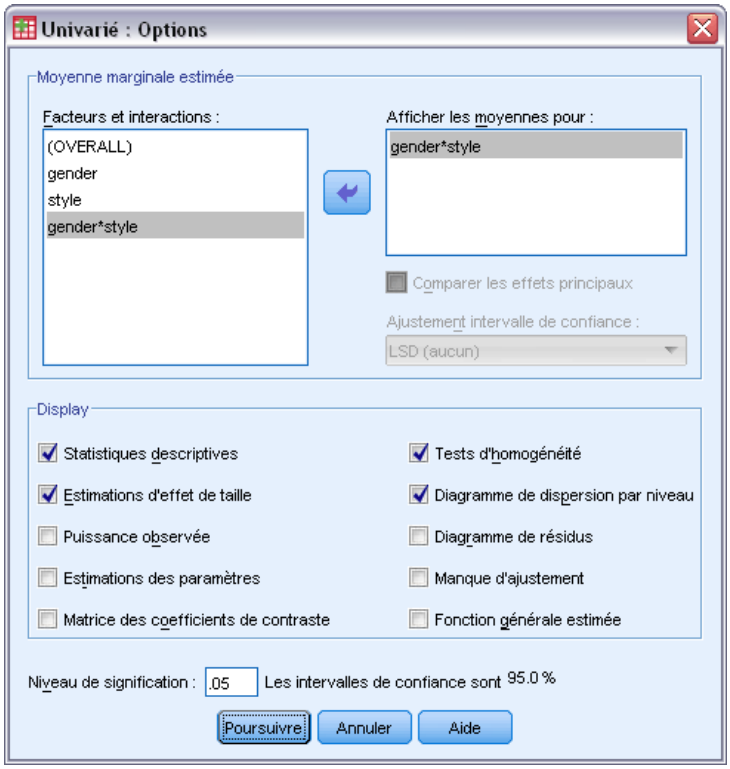
Résidus : Un résidu non standardisé correspond à la valeur réelle de la variable dépendante moins la valeur estimée par le modèle. Les résidus standardisés, selon Student et supprimés sont également disponibles. Si vous avez choisi une variable de pondération, les résidus standardisés pondérés sont disponibles.

- **Non standardisés :.** Différence entre la valeur observée et la valeur prévue par le modèle.
- **Pondéré.** Résolus non normalisés pondérés. Disponibles uniquement lorsqu'une variable WLS a été préalablement sélectionnée.
- **Standardisé.** Résidu, divisé par une estimation de son erreur standard. Egalement appelés résidus de Pearson, les résidus standardisés ont une moyenne de 0 et un écart-type de 1.
- **Studentisés.** Résidu, divisé par une estimation de son écart-type, qui varie d'une observation à l'autre, selon la distance entre les valeurs et la moyenne des variables indépendantes pour chaque observation.
- **Supprimées.** Résidu d'une observation lorsque celle-ci est exclue du calcul des coefficients de régression. Il s'agit de la différence entre la valeur de la variable dépendante et la prévision ajustée.

Statistiques à coefficients. Ecrit une matrice variance-covariance des estimations des paramètres du modèle dans un nouvel ensemble de données de la session en cours ou dans un fichier de données externe au format SPSS Statistics. D'autre part, pour chaque variable dépendante, il y aura une ligne d'estimations, une ligne de valeurs de signification pour les statistiques *t* correspondant aux estimations et une ligne de degrés de liberté résiduels. Pour un modèle multivarié, il y a les mêmes lignes pour chaque variable dépendante. Vous pouvez utiliser ces fichiers de matrice dans les autres procédures qui lisent des fichiers de matrice.

Options GLM

Figure 11-8
Options



Des statistiques facultatives sont disponibles à partir de cette boîte de dialogue. Ces statistiques sont calculées à l'aide de modèle à effets fixes.

Moyenne marginale estimée : Sélectionnez les facteurs et les interactions pour lesquels vous souhaitez obtenir des estimations de la moyenne marginale de la population dans les cellules. Ces moyennes sont ajustées pour les covariables, si elles existent.

- **Comparer les effets principaux :** Propose des comparaisons par paire non corrigées des moyennes marginales estimées pour tout effet principal dans le modèle, à la fois pour les facteurs inter-sujets et intra-sujets. Ceci n'est valable que si les effets principaux sont sélectionnés dans la liste Afficher les moyennes.
- **Ajustement intervalle de confiance :** Sélectionnez l'ajustement aux intervalles et à la significativité des intervalles en adoptant l'une des méthodes suivantes : la différence de moindre signification (LSD), l'ajustement Bonferroni ou l'ajustement de Sidak. Cet élément est disponible uniquement si Comparer les effets principaux est sélectionné.

Afficher :Sélectionnez Statistiques descriptives pour produire des moyennes, des écarts-types et des effectifs pour toutes les variables dépendantes de toutes les cellules. L'option Estimation d'effet de taille fournit une valeur partielle de Eta carré pour chaque effet et chaque estimation. La statistique d'Eta carré décrit la proportion de la variabilité totale imputable au facteur. Sélectionnez Puissance observée pour obtenir la puissance du test lorsque l'autre hypothèse est définie sur la base de la valeur observée. Sélectionnez Estimation des paramètres pour produire des estimations de

paramètres, des erreurs standard, des tests *t*, des intervalles de confiance et la puissance observée de chaque test. Sélectionnez Matrice des coefficients de contraste pour obtenir la matrice **L**.

L'option des tests d'homogénéité produit le test de Levene d'homogénéité de la variance pour chaque variable dépendante sur toutes les combinaisons de niveaux des facteurs inter-sujets, uniquement pour les facteurs inter-sujets. Les options des diagrammes de dispersion par niveau et de résidus sont utiles pour vérifier les hypothèses sur les données. Ceci n'est pas valable s'il n'y a pas de facteurs. Sélectionnez Diagrammes résiduels pour produire un diagramme résiduel observé/estimé/standardisé pour chaque variable dépendante. Ces diagrammes sont utiles pour vérifier l'hypothèse de variance égale. Sélectionnez Manque d'ajustement pour vérifier si la relation entre la variable dépendante et les variables indépendantes peut être convenablement décrite par le modèle. Fonction générale estimée vous permet de construire des tests d'hypothèses personnalisés basés sur la fonction générale estimée. Les lignes de n'importe quelle matrice des coefficients de contraste sont des combinaisons linéaires de la fonction générale estimée.

Niveau de signification : Vous souhaitez peut-être ajuster le niveau de signification utilisé dans les tests post hoc et le niveau de confiance utilisé pour construire des intervalles de confiance. La valeur spécifiée est également utilisée pour calculer l'intensité observée pour le test. Lorsque vous spécifiez un niveau de signification, le niveau associé des intervalles de confiance est affiché dans la boîte de dialogue.

Fonctionnalités supplémentaires de la commande UNIANOVA

Le langage de syntaxe de commande vous permet aussi de :

- Spécifier les effets en cascade dans un modèle (à l'aide de la sous-commande `DESIGN`).
- Spécifier les tests d'effets par rapport à une combinaison linéaire d'effets ou une valeur (à l'aide de la sous-commande `TEST`).
- Spécifier de multiples contrastes (à l'aide de la sous-commande `CONTRAST`).
- Inclure les valeurs manquantes pour l'utilisateur (à l'aide de la sous-commande `MISSING`).
- Spécifier les critères EPS (à l'aide de la sous-commande `CRITERIA`).
- Construisez une matrice **L** personnalisée, une matrice **M** ou une matrice **K** (à l'aide des sous-commandes `LMATRIX`, `MMATRIX` et `KMATRIX`).
- Pour les contrastes simples ou d'écart, spécifier une modalité de référence intermédiaire (à l'aide de la sous-commande `CONTRAST`).
- Spécifier les mesures pour les contrastes polynomiaux (à l'aide de la sous-commande `CONTRAST`).
- Spécifier des termes d'erreur pour les comparaisons post hoc (à l'aide de la sous-commande `POSTHOC`).
- Calculer les moyennes marginales estimées pour chaque facteur ou interaction de facteurs parmi les facteurs de la liste (à l'aide de la sous-commande `EMMEANS`).
- Attribuer des noms aux variables temporaires (à l'aide de la sous-commande `SAVE`).
- Construire un fichier de matrice de corrélation (à l'aide de la sous-commande `OUTFILE`).

- Construire un fichier de type matrice de données qui contient les statistiques provenant de la table ANOVA inter-sujets (à l'aide de la sous-commande `OUTFILE`).
- Enregistrer la matrice du plan dans un nouveau fichier de données (à l'aide de la sous-commande `OUTFILE`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
12

Corrélations bivariées

La procédure de corrélations bivariées calcule le coefficient de corrélation de Pearson, le rho de Spearman et le tau-*b* de Kendall avec leurs seuils de signification. Les corrélations mesurent comment les variables ou les ordres de rang sont liés. Avant de calculer un coefficient de corrélation, parcourez vos données pour rechercher les valeurs éloignées (qui peuvent provoquer des résultats erronés) et les traces d'une relation linéaire. Le coefficient de corrélation de Pearson est une mesure d'association linéaire. Deux variables peuvent être parfaitement liées, mais si la relation n'est pas linéaire, le coefficient de corrélation de Pearson n'est pas une statistique appropriée pour mesurer leur association.

Exemple : Le nombre de matchs de basket-ball remportés par une équipe est-il lié au nombre moyen de points marqués par match ? Un diagramme de dispersion indique qu'il existe une relation linéaire. L'analyse des données de la saison NBA 1994–1995 démontre que le coefficient de corrélation de Pearson (0,581) est significatif au niveau 0,01. On peut penser que plus on a gagné de matchs dans une saison, moins l'adversaire a marqué de points. Ces variables sont liées négativement (–0,401) et la corrélation est significative au niveau 0,05.

Statistiques : Pour chaque variable, on a les éléments suivants : nombre d'observations avec des valeurs non manquantes, moyenne, et écart-type. Pour chaque paire de variables, on a les éléments suivants : coefficient de corrélation de Pearson, rho de Spearman, tau-*b* de Kendall, produits des écarts et covariance.

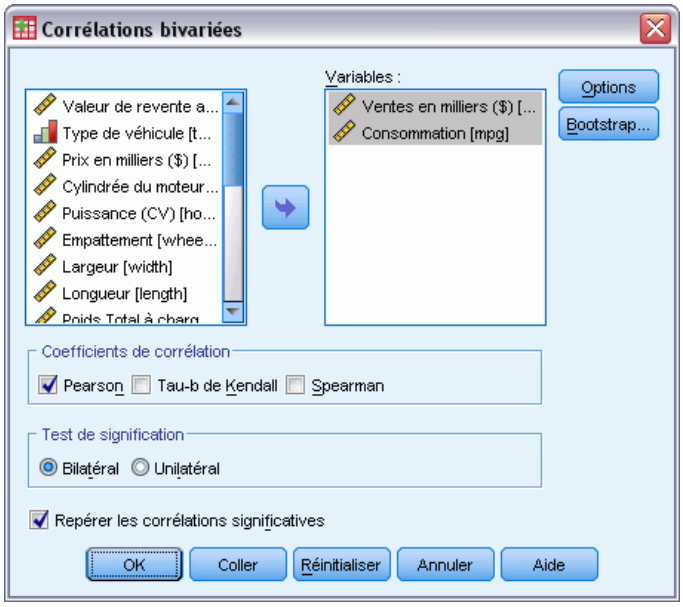
Données. Utilisez des variables quantitatives symétriques pour le coefficient de corrélation de Pearson, et des variables quantitatives ou des variables avec des modalités ordonnées pour le rho de Spearman et le tau-*b* de Kendall.

Hypothèses : Le coefficient de corrélation de Pearson part du principe que chaque paire de variables est gaussienne bivariée.

Pour obtenir des corrélations bivariées

A partir des menus, sélectionnez :
Analyse > Corrélation > Bivariée

Figure 12-1
Boîte de dialogue Corrélations bivariées



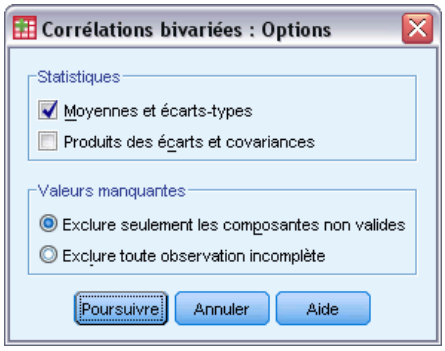
▶ Sélectionnez plusieurs variables numériques.

Les options suivantes sont également disponibles :

- **Coefficients de corrélation :** Pour des variables quantitatives, normalement distribuées, choisissez le coefficient de corrélation de Pearson. Si vos données ne sont pas distribuées normalement ou si elles comportent des modalités ordonnées, choisissez le Tau-b de Kendall ou la corrélation de Spearman, qui mesure l'association entre les ordres de rangs. Les coefficients de corrélation vont de la valeur –1 (relation négative parfaite) à +1 (relation positive parfaite). La valeur 0 indique l'absence de relation linéaire. Lors de l'interprétation de vos résultats, vous ne pouvez pas, à partir de l'existence d'une corrélation significative, conclure en l'existence d'une relation de cause à effet.
- **Test de signification :** Vous pouvez choisir des probabilités bilatérales ou unilatérales. Si la direction de l'association est connue à l'avance, choisissez Unilatéral. Sinon, sélectionnez Bilatéral.
- **Repérer les corrélations significatives :** Les coefficients de corrélation significatifs au niveau 0,05 sont identifiés par un seul astérisque et ceux qui sont significatifs au niveau 0,01 sont identifiés par deux astérisques.

Options de corrélations bivariées

Figure 12-2
Boîte de dialogue Corrélations bivariées : Options



Statistiques : Pour les corrélations de Pearson, vous pouvez choisir l'une des options suivantes (ou les deux) :

- **Moyennes et écarts-types :**Affichés pour chaque variable. Le nombre d'observations avec valeurs non manquantes est également affiché. Les valeurs manquantes sont examinées variable par variable quel que soit votre réglage des valeurs manquantes.
- **Produits des écarts et covariances :** Indiqués pour chaque paire de variables. Le produit des écarts est égal à la somme des produits des variables moyennes corrigées. Ceci est le numérateur du coefficient de corrélation de Pearson. La covariance est une mesure non standardisée de la relation entre deux variables, égale au produit des écarts divisé par *N*–1.

Valeurs manquantes : Vous pouvez choisir l'un des éléments suivants :

- **Exclure seulement les composantes non valides :** Les observations avec des valeurs manquantes pour l'une ou les deux variables d'une paire pour un coefficient de corrélation sont exclues de l'analyse. Etant donné que chaque coefficient est basé sur toutes les observations ayant des codes valides pour cette paire particulière de variables, la quantité maximale d'informations disponibles est utilisée dans chaque calcul. Ceci peut aboutir à un jeu de coefficients basé sur un nombre variable d'observations.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour une variable sont exclues de toutes les analyses.

Propriétés supplémentaires des commandes CORRELATIONS et NONPAR CORR

Le langage de syntaxe de commande vous permet aussi de :

- Ecrire une f pour les corrélations de Pearson qui peut être utilisée à la place de données brutes pour obtenir d'autres analyses comme une analyse factorielle (avec la sous-commande `MATRIX`).
- Obtenir des corrélations de chaque variable dans une liste avec chaque variable d'une seconde liste (en utilisant le mot clé `WITH` avec la sous-commande `VARIABLES`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
13

Corrélations partielles

La procédure des corrélations partielles calcule les coefficients de corrélation partielle qui décrivent le rapport linéaire entre deux variables tout en contrôlant les effets d'une ou plusieurs autres variables. Les corrélations sont des mesures d'association linéaire. Deux variables peuvent être parfaitement liées mais, si leur rapport n'est pas linéaire, un coefficient de corrélation n'est pas une statistique adaptée pour mesurer leur association.

Exemple : Existe-t-il une relation entre le financement associé aux soins de santé et les taux d'attaque ? Contre toute attente, une étude fait état d'une corrélation *positive* : Lorsque le financement associé aux soins de santé augmente, les taux d'attaque augmentent. Cependant, le contrôle du taux de visite aux fournisseurs de soins de santé supprime presque la corrélation positive observée. Le financement lié aux soins de santé et les taux d'attaque sont associés de manière positive car le nombre de personnes ayant accès aux soins de santé augmente en même temps que le financement. De ce fait, le nombre de maladies déclarées par les docteurs et les hôpitaux augmente également

Statistiques : Pour chaque variable, on a les éléments suivants : nombre d'observations avec des valeurs non manquantes, moyenne, et écart-type. Matrices de corrélation partielle et simple, avec degrés de liberté et seuils de signification.

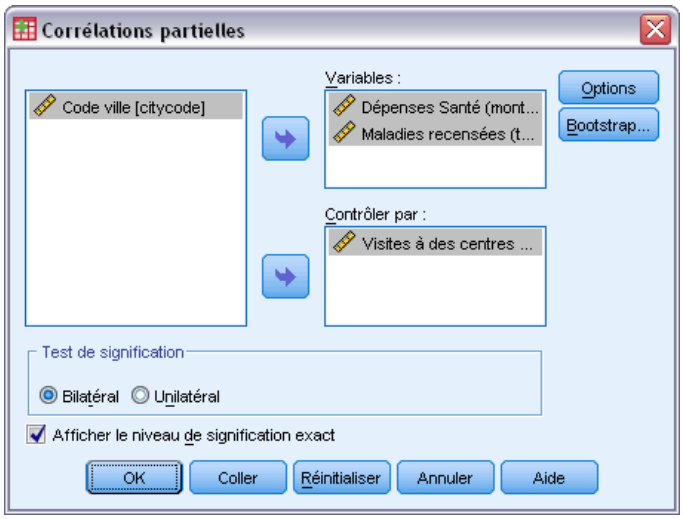
Données. Utiliser des variables quantitatives et symétriques.

Hypothèses : La procédure des Corrélations Partielles suppose que chaque paire de variables présente une corrélation normale.

Obtenir des corrélations partielles

- A partir des menus, sélectionnez :
  Analyse > Corrélation > Partielle

Figure 13-1
Boîte de dialogue Corrélations partielles



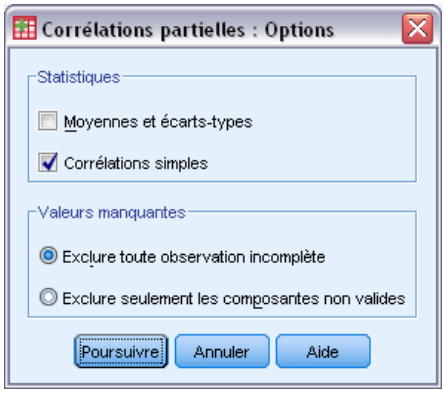
- Sélectionnez au moins deux variables numériques pour lesquelles vous voulez calculer des corrélations partielles.
- Sélectionnez une ou plusieurs variables numériques de contrôle.

Les options suivantes sont également disponibles :

- **Test de signification :** Vous pouvez choisir des probabilités bilatérales ou unilatérales. Si la direction de l'association est connue à l'avance, choisissez Unilatéral. Sinon, sélectionnez Bilatéral.
- **Afficher le seuil exact de signification :** La probabilité et les degrés de liberté sont affichés par défaut pour chaque coefficient de corrélation. Si vous désélectionnez cette option, les coefficients significatifs au seuil 0,05 sont identifiés par une astérisque, les coefficients significatifs au seuil de 0,01 par deux astérisques, et les degrés de liberté sont supprimés. Cette configuration affecte aussi bien les matrices de corrélation partielle que simple.

Options Corrélations partielles

Figure 13-2
Boîte de dialogue Corrélations partielles : Options



Statistiques : Vous avez le choix entre les deux options suivantes :

- **Moyennes et écarts-types :**Affichés pour chaque variable. Le nombre d'observations avec valeurs non manquantes est également affiché.
- **Corrélations simples :** Une matrice de corrélations simples entre toutes les variables, y compris les variables de contrôle, s'affiche.

Valeurs manquantes : Vous avez le choix entre les options suivantes :

- **Exclure toute observation incomplète :** Les observations ayant des valeurs manquantes pour une variable quelconque, y compris une variable de contrôle, sont exclues de tous les calculs.
- **Exclure seulement les composantes non valides :** Pour le calcul des corrélations simples sur lesquelles se basent les corrélations partielles, une observation ayant des valeurs manquantes pour une composante ou les deux composantes d'une paire de variables ne sera pas utilisée. La suppression des composantes non valides seulement utilise autant de données que possible. Le nombre d'observations peut toutefois différer selon les coefficients. Lorsque la suppression des composantes non valides seulement est sélectionnée, les degrés de liberté d'un coefficient partiel donné sont basés sur le plus petit nombre d'observations utilisées dans le calcul de l'une quelconque des corrélations d'ordre zéro.

Fonctionnalités supplémentaires de la commande PARTIAL CORR

Le langage de syntaxe de commande vous permet aussi de :

- Lire une matrice de corrélation d'ordre zéro ou écrire une matrice de corrélation d'ordre zéro (avec la sous-commande `MATRIX`).
- Obtenir des corrélations partielles entre deux listes de variables (en utilisant le mot-clé `WITH` dans la sous-commande `VARIABLES`).
- Obtenir des analyses multiples (avec les sous-commandes `VARIABLES`).
- Spécifier les valeurs des ordres à demander (par exemple, à la fois les corrélations partielles de premier et de second ordre) lorsque vous avez deux variables de contrôle (avec la sous-commande `VARIABLES`).
- Supprimer les coefficients redondants (avec la sous-commande `FORMAT`).
- Afficher une matrice de corrélations simples lorsque certains coefficients ne peuvent pas être calculés (avec la sous-commande `STATISTICS`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
14

Distances

Cette procédure permet de calculer de très nombreuses statistiques mesurant les similitudes ou les différences (distances) entre des paires de variables ou d'observations. Vous pourrez ensuite utiliser ces mesures de similarité ou de dissimilarité avec d'autres procédures, comme l'analyse factorielle, la classification ou le positionnement multidimensionnel, afin de simplifier l'analyse des fichiers de données complexes.

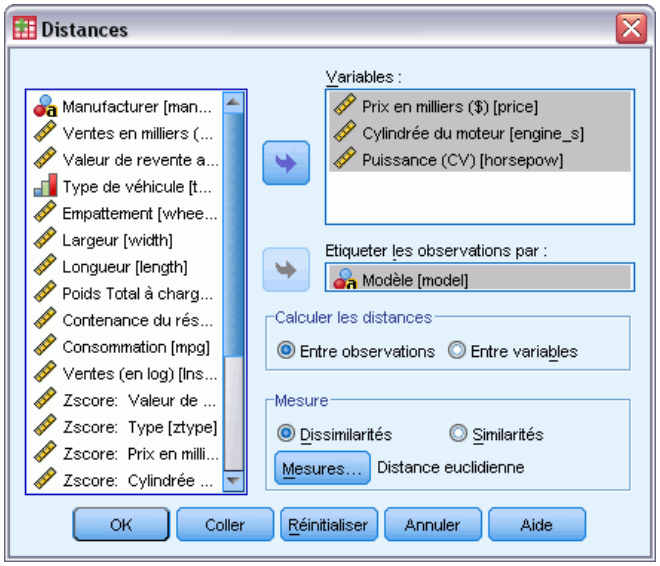
Exemple : Est-il possible de mesurer les similarités entre des paires de voitures en fonction de certaines caractéristiques, comme le nombre de cylindres, la consommation et la puissance ? En calculant les similarités existant entre des voitures, vous pouvez déterminer les voitures qui sont semblables et celles qui sont différentes. Dans l'optique d'une analyse plus formelle, vous pouvez appliquer une classification hiérarchique ou un positionnement multidimensionnel aux similarités afin d'examiner la structure sous-jacente.

Statistiques : Pour les données d'intervalle, les mesures de dissimilarité sont la distance Euclidienne, le carré de la distance Euclidienne, la distance de Tchebycheff, la distance de Manhattan (bloc), la distance de Minkowski ou une mesure personnalisée. Pour les données d'effectif, les mesures sont khi-deux et phi-deux. Pour les données binaires, les mesures de dissimilarité sont la distance Euclidienne, le carré de la distance Euclidienne, l'écart de taille, la différence de motif, la variance, la forme, ou la mesure de Lance et Williams. Pour les données d'intervalles, les mesures de similarité sont la corrélation de Pearson ou cosinus. Pour les données binaires, il s'agit des mesures suivantes : Russel et Rao, indice de Sokal et Michener, Jaccard, Dice, Rogers et Tanimoto, Sokal et Sneath 1, Sokal et Sneath 2, Sokal et Sneath 3, Kulczynski 1, Kulczynski 2, Sokal et Sneath 4, Hamann, lambda, *D* d'Anderberg, *Y* de Yule, *Q* de Yule, Ochiai, Sokal et Sneath 5, corrélation phi tétrachorique ou dispersion.

Pour obtenir des matrices de distance

- A partir des menus, sélectionnez :
  Analyse > Corrélation > Indices

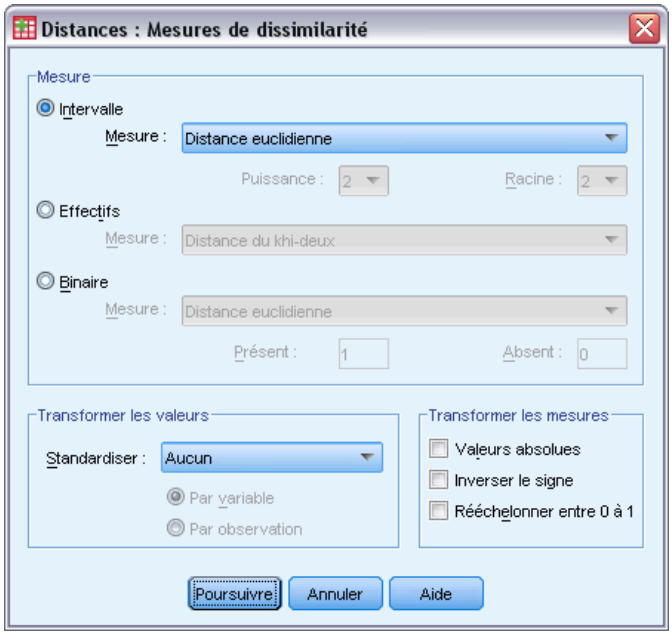
Figure 14-1
Boîte de dialogue Distances



- Sélectionnez au minimum une ou deux variables numériques pour calculer respectivement les distances existant entre des observations ou des variables.
- Sélectionnez une possibilité dans le groupe Calculer les distances pour calculer les proximités existant entre des observations ou des variables.

Distances : Mesures de dissimilarité

Figure 14-2
Boîte de dialogue Indices : Mesures de dissimilarité



Dans le groupe Mesure, sélectionnez la possibilité qui correspond au type de vos données (intervalle, effectif ou binaire). Ensuite, dans la liste déroulante, sélectionnez l'une des mesures correspondant à ce type de données. Les mesures disponibles sont, par type de données :

- **Intervalle :** Distance Euclidienne, Carré de la distance Euclidienne, Distance de Tchebycheff, Distance de Manhattan, Distance de Minkowski ou Autre.
- **Effectifs :** Distance du Khi-deux ou Distance du phi-deux.
- **Binaire :** Distance Euclidienne, Carré de la distance Euclidienne, Ecart de taille, Différence de motif, Variance, Forme, ou Lance et Williams. (Entrez des valeurs dans les champs Présent et Absent pour indiquer les deux valeurs significatives. Aucune autre valeur ne sera prise en compte dans Distances.)

Le groupe Transformer les valeurs vous permet de standardiser les valeurs des données pour les observations ou les variables *avant* le calcul des proximités. Ces transformations ne s'appliquent pas aux données binaires. Les méthodes de standardisation disponibles sont *Centrer-réduire*, Entre –1 et 1, Entre 0 et 1, Maximum = 1, Moyenne = 1 et Ecart type = 1.

Le groupe Transformer les mesures vous permet de transformer les valeurs générées par la mesure de distance. Elles sont appliquées après le calcul de la mesure d'indice. Les options possibles sont Valeurs absolues, Inverser le signe, et Rééchelonner entre 0 et 1.

Indices : Mesures de similarité

Figure 14-3
Boîte de dialogue Indices : Mesures de similarité

Dans le groupe Mesure, sélectionnez la possibilité qui correspond au type de vos données (intervalle ou binaire). Ensuite, dans la liste déroulante, sélectionnez l'une des mesures correspondant à ce type de données. Les mesures disponibles sont, par type de données :

- **Intervalle :** Corrélation de Pearson ou Cosinus.
- **Binaire :** Russel et Rao, Indice de Sokal et Michener, Jaccard, Dice, Rogers et Tanimoto, Sokal et Sneath 1, Sokal et Sneath 2, Sokal et Sneath 3, Kulczynski 1, Kulczynski 2, Sokal et Sneath 4, Hamann, Lambda, *D* d'Anderberg, *Y* de Yule, *Q* de Yule, Ochiai, Sokal et Sneath 5, Corrélation phi tétrachorique ou Dispersion. (Entrez des valeurs dans les champs Présent et Absent pour indiquer les deux valeurs significatives. Aucune autre valeur ne sera prise en compte dans Distances.)

Le groupe Transformer les valeurs vous permet de standardiser les valeurs des données pour les observations ou les variables avant le calcul des proximités. Ces transformations ne s'appliquent pas aux données binaires. Les méthodes de standardisation disponibles sont *Centrer-réduire*, Entre –1 et 1, Entre 0 et 1, Maximum = 1, Moyenne = 1 ou Ecart type = 1.

Le groupe Transformer les mesures vous permet de transformer les valeurs générées par la mesure de distance. Elles sont appliquées après le calcul de la mesure d'indice. Les options possibles sont Valeurs absolues, Inverser le signe, et Rééchelonner entre 0 et 1.

Fonctionnalités supplémentaires de la commande PROXIMITIES

La procédure Distances utilise la syntaxe de la commande `PROXIMITIES`. Le langage de syntaxe de commande vous permet aussi de :

- Indiquez un nombre entier comme la puissance pour la mesure de distance de Minkowski.
- Indiquez des nombres entiers comme la puissance et la racine pour une mesure de distance personnalisée.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
15

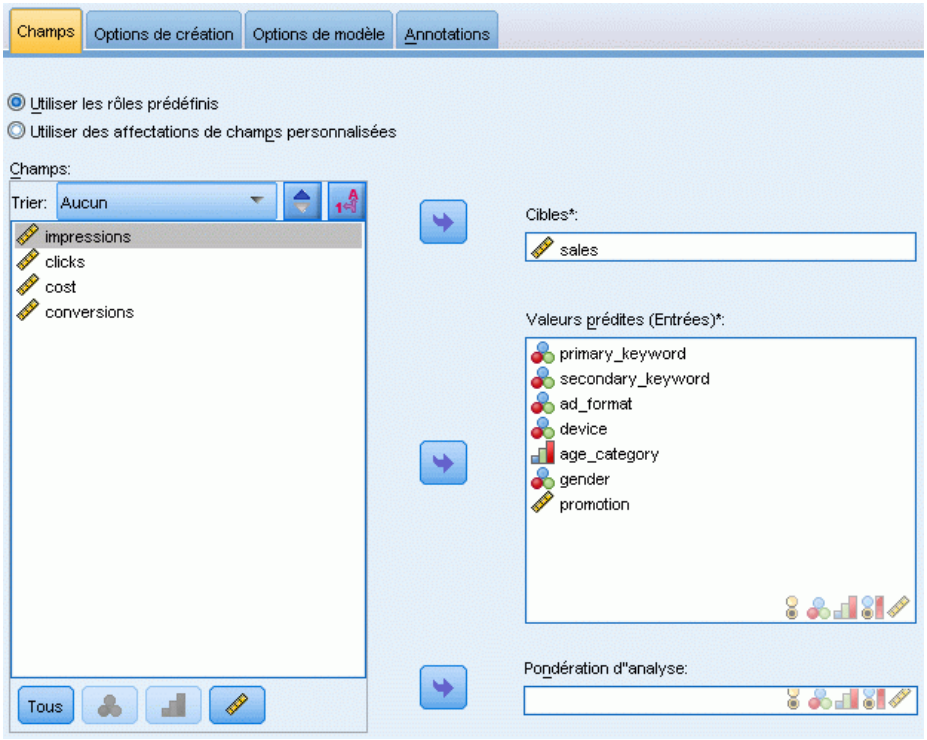
Modèles linéaires

Les modèles linéaires prédisent une cible continue en fonction de relations linéaires entre la cible et une ou plusieurs variables prédites.

Les modèles linéaires sont relativement simples et produisent une formule mathématique pouvant facilement être évaluée. Les propriétés de ces modèles sont bien comprises et peuvent généralement être créées très rapidement par rapport à d'autres types de modèles (tels que les réseaux neuronaux ou les arbres de décision) sur le même ensemble de données.

Exemple : Une compagnie d'assurances disposant de ressources restreintes pour enquêter sur les demandes de remboursement des propriétaires de biens immobiliers, souhaite construire un modèle pour estimer les coûts des demandes d'indemnisation. Le déploiement de ce modèle dans les centres de service permettra aux représentants d'entrer les informations des demandes d'indemnisation alors qu'ils sont au téléphone avec les clients et de recevoir immédiatement le coût "prévu" de la demande d'indemnisation, sur la base de données anciennes.

Figure 15-1
Onglet Champs



Exigences concernant les champs. Il doit y avoir une cible et au moins une entrée. Par défaut, les champs avec les rôles prédéfinis Les deux ou Aucun ne sont pas utilisés. La cible doit être continue (échelle). Il n'y a pas de restriction sur les niveaux de mesure des variables prédites

(variables d'entrée) ; les champs qualitatifs (nominaux et ordinaux) sont utilisés comme facteurs dans le modèle et les champs continus sont utilisés comme covariables.

Remarque : si un champ qualitatif comprend plus de 1000 modalités, la procédure ne s'exécute pas et aucun modèle n'est créé.

Pour obtenir un modèle linéaire

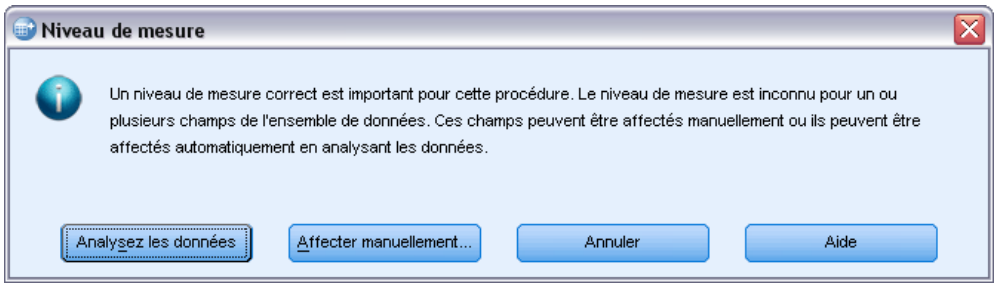
Cette fonction nécessite l'option Statistiques de base.

A partir des menus, sélectionnez :
Analyse > Régression > Modèles linéaires automatiques...

- Vérifiez qu'il existe au moins une cible et une entrée.
- Cliquez sur Options de création pour spécifier les paramètres optionnels de création et de modèle.
- Cliquez sur Options du modèle pour enregistrer les scores dans l'ensemble de données actif et exporter le modèle vers un fichier externe.
- Cliquez sur Exécuter pour exécuter la procédure et créer les objets du modèle.

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Figure 15-2
Alerte du niveau de mesure



- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Objectifs

Quel est votre objectif principal ?

- **Créer un modèle standard.** Cette méthode permet de créer un modèle unique afin de prédire la cible à l'aide de valeurs prédites. En général, les modèles standards sont plus faciles à interpréter et peuvent être plus rapidement évalués que des ensembles de données boostés, de bagging ou de grande taille.
- **Améliorer la précision d'un modèle (boosting).** Cette méthode permet de créer un modèle d'ensemble à l'aide du boosting, qui génère une séquence de modèles afin d'obtenir des prédictions plus précises. Les ensembles peuvent être plus longs à construire et l'obtention de leurs résultats peut être plus longue qu'avec un modèle standard.

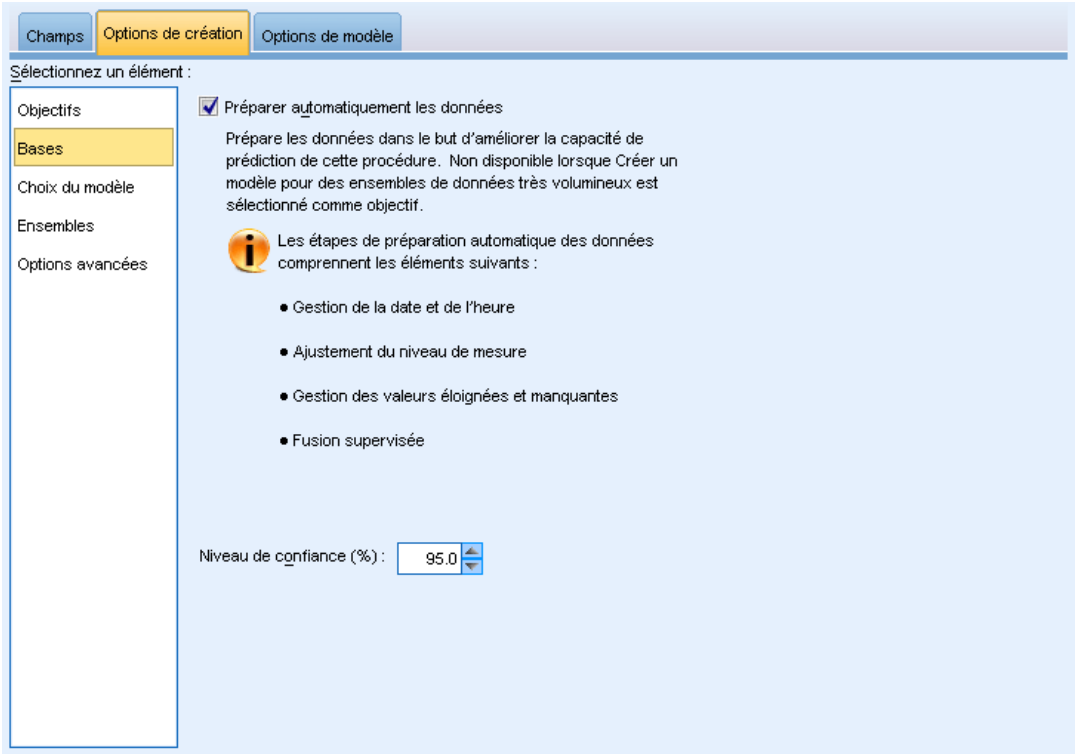
  Le boosting produit une succession de « modèles de composant », chacun étant construit à partir de la totalité de l'ensemble de données. Avant la création successive de chaque modèle de composant, les enregistrements sont pondérés en fonction des résidus des modèles de composant précédents. Les observations présentant de grands résidus se voient attribuées des pondérations d'analyse relativement plus élevées, de sorte que le modèle de composant suivant se concentre aussi sur la prédiction de ces enregistrements. Ensemble, ces modèles de composant forment un modèle d'ensemble. Le modèle d'ensemble évalue les nouveaux enregistrements à l'aide d'une règle de combinaison ; les règles disponibles dépendent du niveau de mesure de la cible.
- **Améliorer la stabilité du modèle (bagging).** Cette méthode permet de créer un modèle d'ensemble à l'aide du bagging (agrégation par bootstrap), qui génère plusieurs modèles afin d'obtenir des prédictions plus fiables. Les ensembles peuvent être plus longs à construire et l'obtention de leurs résultats peut être plus longue qu'avec un modèle standard.

  L'agrégation de bootstrap (bagging) produit des réplicats de l'ensemble de données d'apprentissage en effectuant un échantillonnage avec remplacement à partir de l'ensemble de données d'origine. Ceci crée des échantillons de bootstrap de taille identique à celle de l'ensemble de données d'origine. Ensuite, un modèle de composant est créer à partir de chaque réplicat. Ensemble, ces modèles de composant forment un modèle d'ensemble. Le modèle d'ensemble évalue les nouveaux enregistrements à l'aide d'une règle de combinaison ; les règles disponibles dépendent du niveau de mesure de la cible.
- **Créer un modèle pour des ensembles de données très volumineux (nécessite IBM® SPSS® Statistics Server).** Cette méthode permet de créer un modèle d'ensemble en scindant l'ensemble de données en blocs de données distincts. Choisissez cette option si votre ensemble de données est trop important pour que vous puissiez créer l'un des modèles ci-dessus, ou pour la construction d'un modèle incrémental. La construction de cette option peut être moins longue, mais l'obtention des résultats peut être plus longue qu'avec un modèle standard. Cette option nécessite une connexion à SPSS Statistics Server .

Pour plus d'informations sur les paramètres relatifs au boosting, au bagging et aux ensembles de données volumineux, reportez-vous à Ensembles sur p. 89.

Bases

Figure 15-3
Paramètres de base



Préparer automatiquement les données. Cette option permet à la procédure de transformer la cible et les variables prédites en interne afin de maximiser la puissance de prédiction du modèle ; toutes les transformations sont enregistrées avec le modèle et appliquées aux nouvelles données pour l'évaluation. Les versions originales de champs transformés sont exclues du modèle. Par défaut, les préparations automatiques de données suivantes sont réalisées.

- **Gestion de la date et de l'heure.** Chaque variable prédite de date est transformée en une nouvelle variable prédite continue qui contient la durée écoulée depuis une date de référence (01/01/1970). Chaque variable prédite d'heure est transformée en une nouvelle variable prédite continue qui contient la durée écoulée depuis une heure de référence (00:00:00).
- **Régler le niveau de mesure.** Les variables prédites continues ayant moins de 5 valeurs distinctes sont reconverties en variables prédites ordinales. Les variables prédites ordinales ayant plus de 10 valeurs distinctes sont reconverties en variables prédites continues.
- **Gestion des valeurs éloignées.** Les valeurs de variables prédites continues qui se trouvent au-delà d'une valeur de césure (écart-type de 3 par rapport à la moyenne) sont définies sur la valeur de césure.
- **Gestion des valeurs manquantes.** Les valeurs manquantes de variables prédites nominales sont remplacées par le mode de la partition d'apprentissage. Les valeurs manquantes de variables prédites ordinales sont remplacées par la médiane de la partition d'apprentissage.

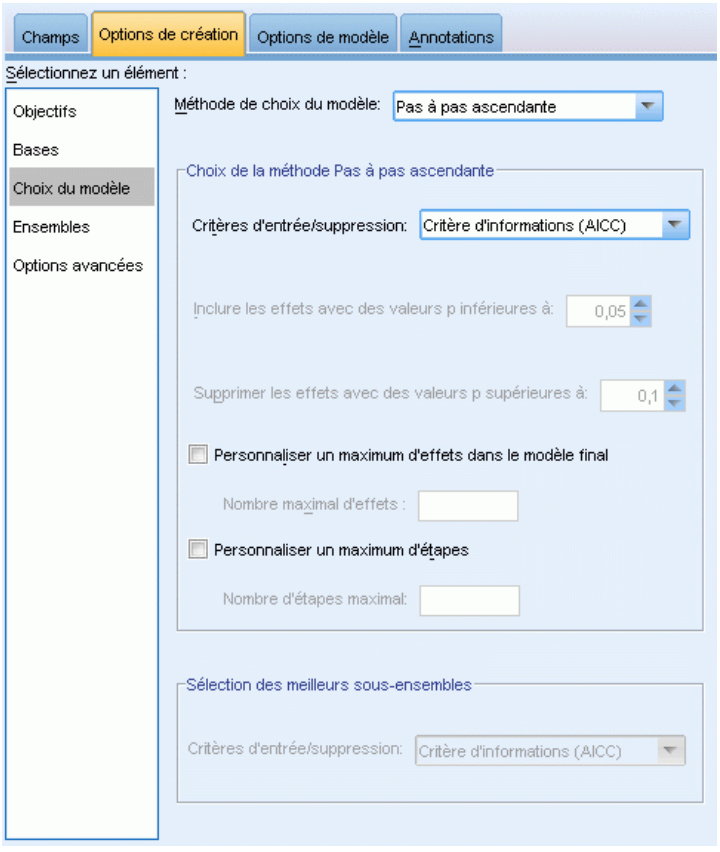
Les valeurs manquantes de variables prédites continues sont remplacées par la moyenne de la partition d'apprentissage.

- **Fusion supervisée.** Ceci crée un modèle plus petit en réduisant le nombre de champs à traiter en association avec la cible. Les modalités similaires sont identifiées en fonction de la relation entre l'entrée et la cible. Les modalités ne différant pas de manière significative (c'est-à-dire ayant une valeur p supérieure à 0,1), sont fusionnées. Si toutes les catégories sont fusionnées en une seule, les versions d'origine et dérivées du champ sont exclues du modèle car elles n'ont pas de valeur de variable prédite.

Niveau de confiance. Il s'agit du niveau de confiance utilisé pour calculer les estimations d'intervalle des coefficients de modèle dans la vue Coefficients. Définissez une valeur supérieure à 0 et inférieure à 100. La valeur par défaut est 95.

Choix du modèle

Figure 15-4
Paramètres du choix du modèle



Méthodes de choix du modèle. Choisissez l'une des méthodes de sélection du modèle (détails ci-dessous) ou Inclure toutes les valeurs prédites, qui entre simplement toutes les variables prédites disponibles en tant que termes du modèle des effets principaux. Le modèle Pas à pas ascendant est utilisé par défaut.

Choix de la méthode Pas à pas ascendante. Elle commence sans effet dans le modèle et ajoute et supprime des effets une étape à la fois jusqu'à ce qu'aucune autre ne puisse être ajoutée ou supprimée en fonction des critères pas à pas.

- **Critères d'entrée/suppression.** Il s'agit des statistiques utilisées pour savoir si un effet doit être ajouté ou supprimé du modèle. Critère d'information (AICC) est basé sur la vraisemblance du modèle fourni à l'ensemble d'apprentissage, et est ajusté afin de pénaliser des modèles trop complexes. Statistiques F est basé sur un test statistique de l'amélioration dans l'erreur d'un modèle. R-deux ajusté est basé sur l'adéquation de l'ensemble d'apprentissage, et est ajusté afin de pénaliser des modèles trop complexes. Le Critère de prévention du surajustement (ASE) est basé sur l'adéquation (carré de l'erreur moyenne, ou ASE) de l'ensemble de prévention de surajustement. L'ensemble de prévention de surajustement est un sous-échantillon aléatoire d'environ 30 % de l'ensemble de données original qui n'est pas utilisé pour former le modèle.

  Si un autre critère que Statistiques F est sélectionné, à chaque étape l'effet qui correspond à l'accroissement positif le plus important dans le critère est ajouté au modèle. Tous les effets du modèle qui correspondent à une diminution du critère sont supprimés.

  Si Statistiques F est sélectionné en tant que critère, à chaque étape l'effet ayant la plus petite valeur *p* inférieure au seuil spécifié, Inclure les effets avec des valeurs p inférieures à, est ajouté au modèle. La valeur par défaut est 0,05. Tous les effets du modèle ayant une valeur *p* supérieure au seuil spécifié, Supprimer les effets ayant des valeurs p supérieures à, sont supprimés. La valeur par défaut est 0.10.
- **Personnaliser le nombre maximum d'effets dans le modèle final.** Par défaut, tous les effets disponibles peuvent être entrés dans le modèle. Si l'algorithme pas à pas se termine à une étape avec le nombre spécifié d'effets, l'algorithme s'arrête à l'ensemble d'effets en cours.
- **Personnaliser le nombre maximal d'étapes.** L'algorithme pas à pas s'arrête après un certain nombre d'étapes. Par défaut, il s'agit de 3 fois le nombre d'effets disponibles. Vous pouvez également spécifier un nombre entier positif maximum d'étapes.

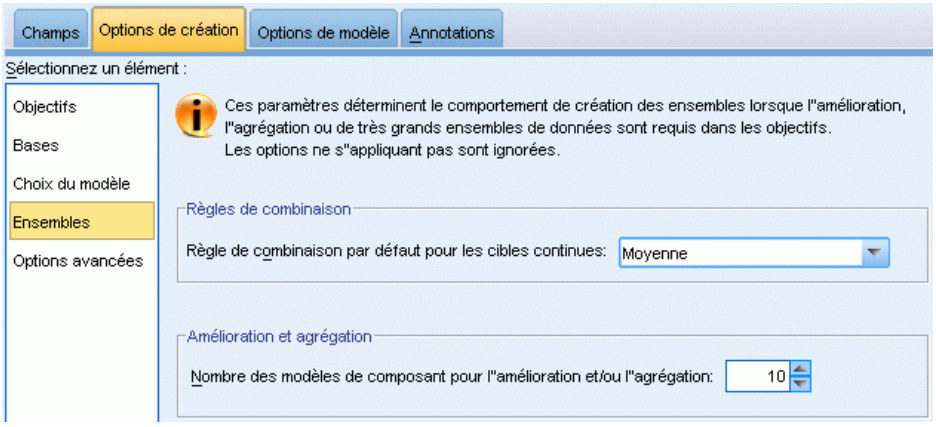
Sélection des meilleurs sous-ensembles. Ceci permet de vérifier "tous les modèles possibles" ou au moins un sous-ensemble plus important des modèles possibles qu'en pas à pas ascendant, pour choisir le meilleur en fonction du critère des meilleurs sous-ensembles. Critère d'information (AICC) est basé sur la vraisemblance du modèle fourni à l'ensemble d'apprentissage, et est ajusté afin de pénaliser des modèles trop complexes. R-deux ajusté est basé sur l'adéquation de l'ensemble d'apprentissage, et est ajusté afin de pénaliser des modèles trop complexes. Le Critère de prévention du surajustement (ASE) est basé sur l'adéquation (carré de l'erreur moyenne, ou ASE) de l'ensemble de prévention de surajustement. L'ensemble de prévention de surajustement est un sous-échantillon aléatoire d'environ 30 % de l'ensemble de données original qui n'est pas utilisé pour former le modèle.

Le modèle ayant la plus grande valeur de critère est sélectionné comme meilleur modèle.

Remarque : La sélection des meilleurs sous-ensembles demande plus de ressources de calcul que la sélection pas à pas ascendante. Lorsque la sélection des meilleurs sous-ensembles est effectuée en conjonction avec le boosting, le bagging ou le traitement d'ensembles très volumineux, elle peut être plus longue que la création d'un modèle standard à l'aide de la sélection pas à pas ascendante.

Ensembles

Figure 15-5
Paramètres des ensembles



Ces paramètres déterminent le comportement d'assemblage qui se produit lors du boosting, du bagging ou lorsque que des ensembles volumineux de données sont requis dans les objectifs. Les options qui ne s'appliquent pas à l'objectif sélectionné sont ignorées.

Bagging et très grands ensembles de données. Lors de l'évaluation d'un ensemble, il s'agit de la règle utilisée pour combiner les valeurs prédites à partir des modèles de base pour calculer la valeur de score d'un ensemble.

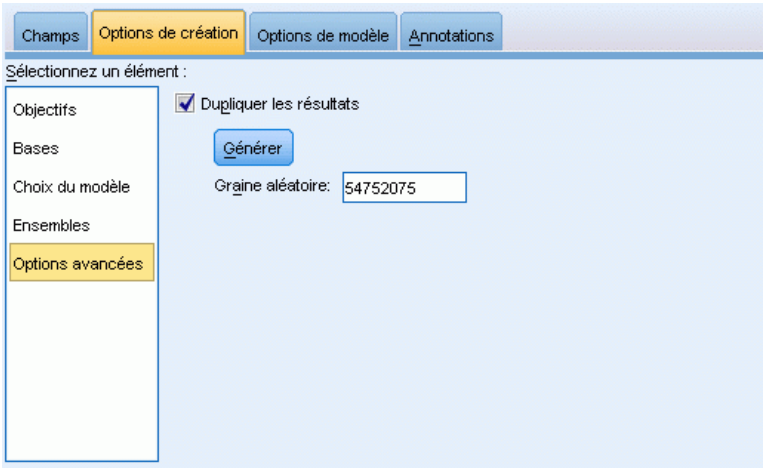
- **Règles de combinaison par défaut pour les cibles continues.** Des valeurs prédites d'ensemble pour des cibles continues peuvent être combinées à l'aide de la moyenne ou de la médiane des valeurs prédites à partir des modèles de base.

Veuillez noter que lorsque l'objectif consiste à améliorer la précision du modèle, les sélections de règles de combinaisons sont ignorées. Le boosting utilise toujours un vote majoritaire pondéré pour évaluer des cibles catégorielles et une médiane pondérée pour évaluer des cibles continues.

Boosting et bagging. Spécifiez le nombre de modèles de base à créer lorsque l'objectif est d'améliorer la précision ou la stabilité du modèle ; pour le bagging, il s'agit du nombre d'échantillons de bootstrap. Il doit s'agir d'un entier positif.

Avancé

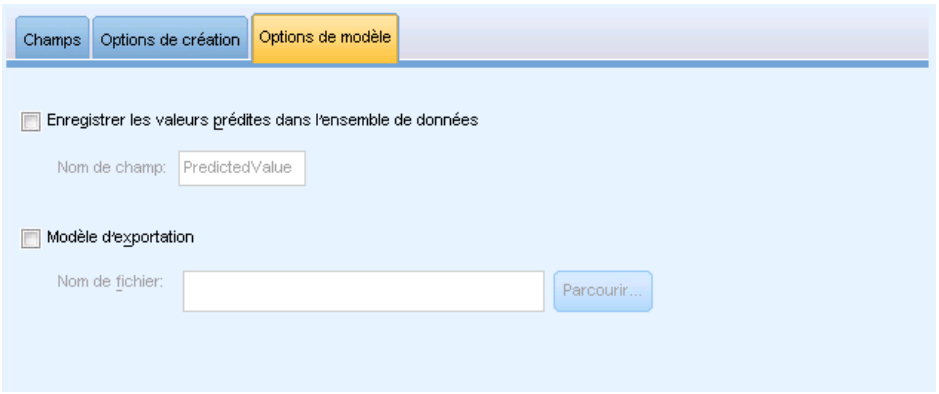
Figure 15-6
Paramètres avancés



Dupliquer les résultats. Définir une valeur de départ aléatoire vous permet de dupliquer des analyses. Le générateur de nombres aléatoires est utilisé pour choisir les enregistrements de l'ensemble de prévention de surajustement. Spécifiez un entier ou cliquez sur Générer, ce qui crée un entier pseudo-aléatoire compris entre 1 et 2147483647, inclus. La valeur par défaut est 54752075.

Options de modèle

Figure 15-7
Onglet Options de modèle

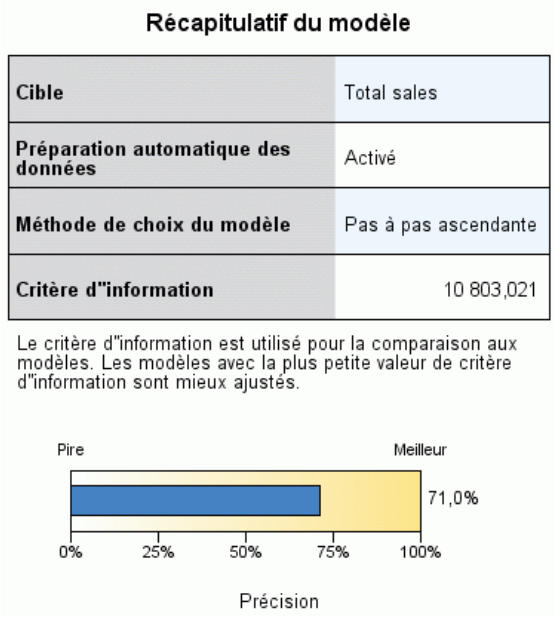


Enregistre les valeurs prédites dans l'ensemble de données. Le nom par défaut de la variable est *PredictedValue*.

Exporter le modèle. Cette option écrit le modèle sur un fichier *.zip* externe. Vous pouvez utiliser ce fichier de modèle pour appliquer les informations du modèle aux autres fichiers de données à des fins d'évaluation. Spécifiez un nom de fichier valide et unique. Si la spécification du fichier pointe vers un fichier existant, le fichier est écrasé.

Récapitulatif de modèle

Figure 15-8
Vue récapitulative du modèle



La vue récapitulative du modèle est un instantané, permettant de consulter en un coup d'œil le modèle et son ajustement.

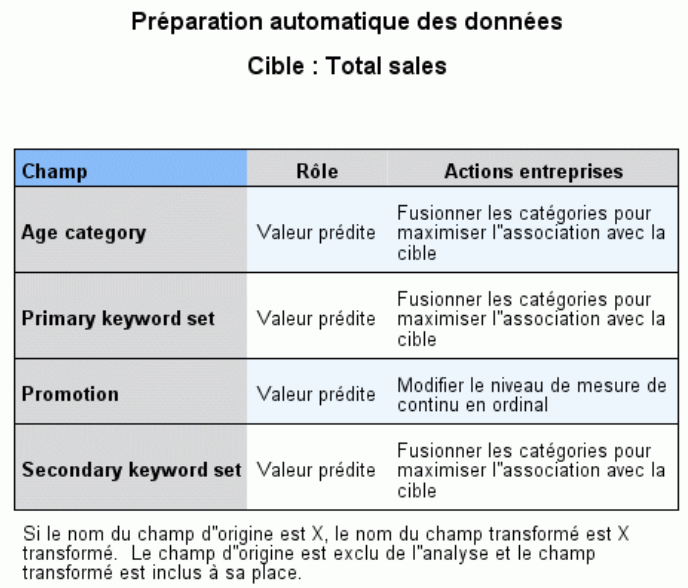
Tableau. Le tableau identifie certains paramètres de modèle de haut niveau, dont :

- Le nom de la cible spécifié sur l'onglet Champs,
- Si la préparation automatique des données a été réalisée comme spécifié dans les paramètres de base,
- La méthode de sélection du modèle et le critère de sélection spécifiés dans les paramètres de sélection du modèle. La valeur du critère de sélection du modèle final est également affichée et elle est présentée en plus petit, disposant d'un meilleur format.

Diagramme. Le diagramme affiche la précision du modèle final, qui est présenté en plus grand, disposant d'un meilleur format. La valeur est de 100 × le R^2 ajusté pour le modèle final.

Préparation automatique des données

Figure 15-9
Vue Préparation automatique des données



Préparation automatique des données

Cible : Total sales

Champ	Rôle	Actions entreprises
Age category	Valeur prédite	Fusionner les catégories pour maximiser l"association avec la cible
Primary keyword set	Valeur prédite	Fusionner les catégories pour maximiser l"association avec la cible
Promotion	Valeur prédite	Modifier le niveau de mesure de continu en ordinal
Secondary keyword set	Valeur prédite	Fusionner les catégories pour maximiser l"association avec la cible

Si le nom du champ d"origine est X, le nom du champ transformé est X transformé. Le champ d"origine est exclu de l"analyse et le champ transformé est inclus à sa place.

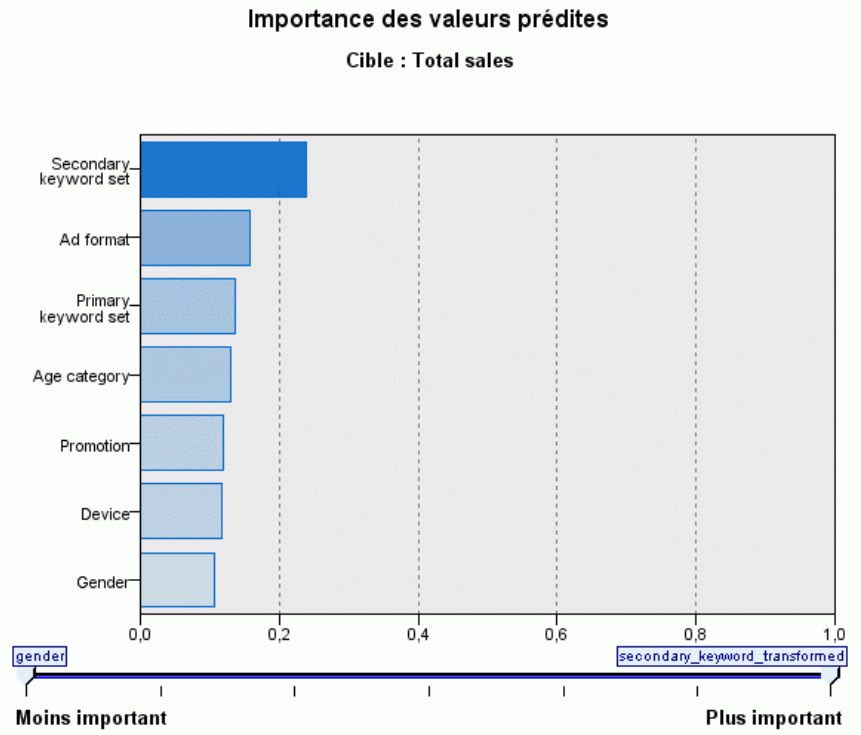
Cette vue affiche des informations concernant les champs qui ont été exclus et la façon dont les champs transformés ont été dérivés dans l'étape de préparation automatique des données (ADP). Pour chaque champ transformé ou exclu, le tableau répertorie le nom du champ, son rôle au sein de l'analyse et l'action entreprise par l'étape ADP. Les champs sont triés selon l'ordre alphabétique croissant des noms de champ. Les actions possibles pour chaque champ comprennent :

- Calculer la durée : en mois calcule le temps écoulé en mois à partir des valeurs d'un champ contenant des dates et de la date système actuelle.
- Calculer la durée : en heures calcule le temps écoulé en heures à partir des valeurs d'un champ contenant des heures et de l'heure système actuelle.
- Modifier le niveau de mesure de continu en ordinal reconvertit les champs continus possédant moins de 5 valeurs uniques en champs ordinaux.
- Modifier le niveau de mesure d'ordinal en continu reconvertit les champs ordinaux possédant plus de 10 valeurs uniques en champs continus.
- Supprimer les valeurs éloignées définit les valeurs des variables prédites continues qui se trouvent au-delà d'une valeur de césure (écart-type de 3 par rapport à la moyenne) sur la valeur de césure.
- Remplacer les valeurs manquantes remplace les valeurs manquantes des champs nominaux par le mode, celles des champs ordinaux par la médiane, et celles des champs continus par la moyenne.

- Fusionner les catégories pour augmenter l'association avec la cibleidentifie les catégories de variables prédites similaires en fonction de la relation entre l'entrée et la cible. Les catégories ne différant pas de manière significative (c'est-à-dire ayant une valeur *p* supérieure à 0,05), sont fusionnées.
- Exclure les valeurs prédites constantes / après le traitement des valeurs éloignées / après la fusion des catégories supprime les valeurs prédites qui possèdent une valeur unique, éventuellement une fois les autres actions ADP effectuées.

Importance des variables prédites

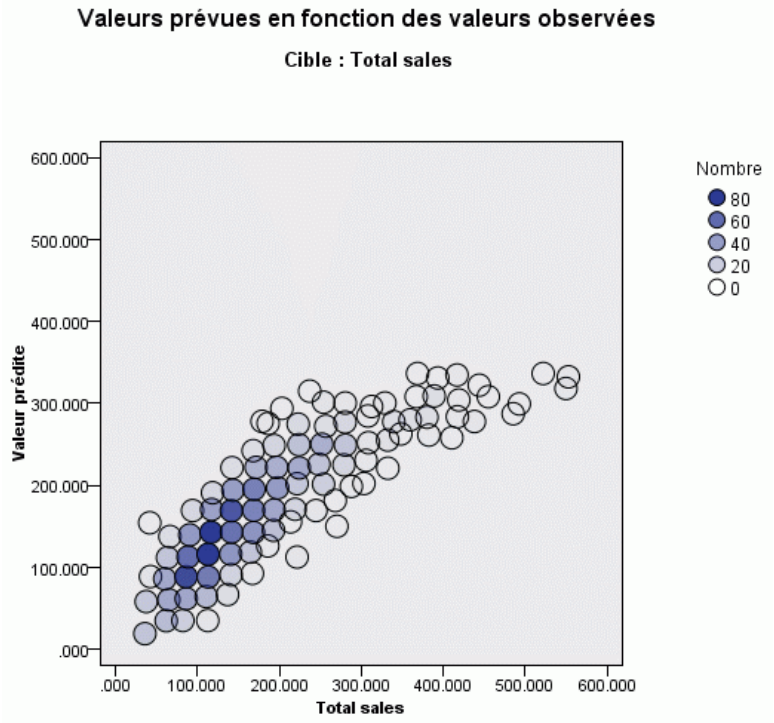
Figure 15-10
Vue Importance des variables prédites



Généralement, vous souhaitez concentrer vos efforts de modélisation sur les champs prédicteurs les plus importants et vous envisagez d'exclure et d'ignorer les moins importants. Le diagramme d'importance des valeurs prédites peut vous y aider en indiquant l'importance relative de chaque valeur prédite en estimant le modèle. Etant donné que les valeurs sont relatives, la somme des valeurs pour l'ensemble des valeurs prédites affichée est 1,0. L'importance des valeurs prédites n'a aucun rapport avec la précision du modèle. Elle est juste liée à l'importance de chaque valeur prédite pour la réalisation d'une prévision, peu importe si cette dernière est précise ou non.

Valeurs prévues en fonction des valeurs observées

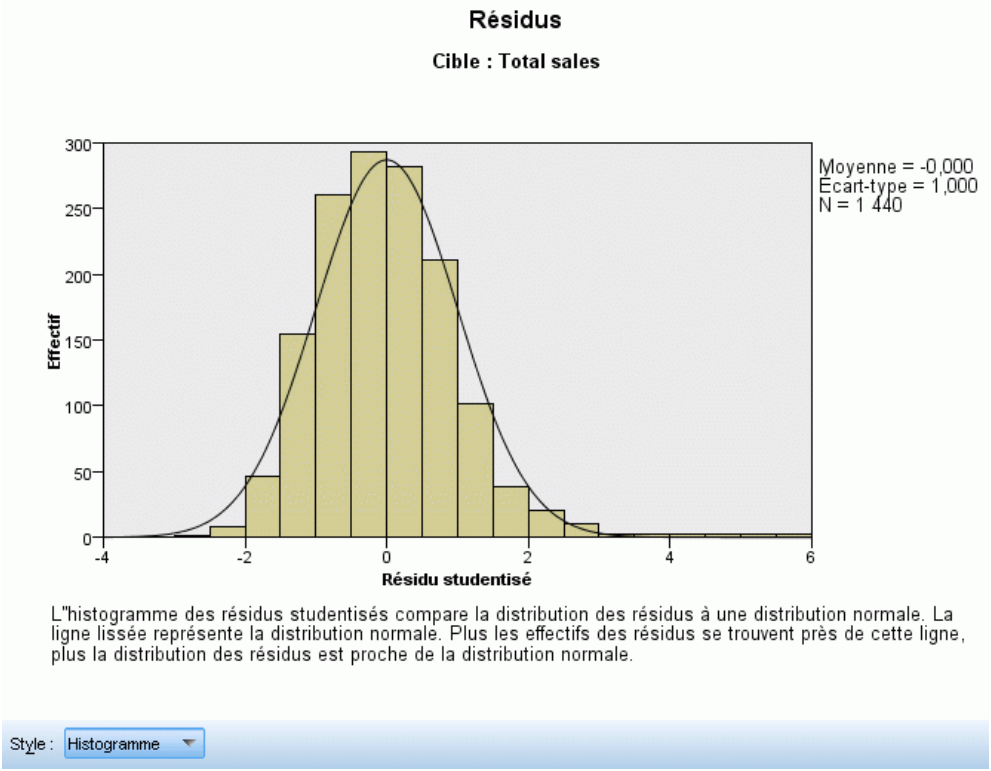
Figure 15-11
Vue Valeurs prédites en fonction des valeurs observées



Ceci affiche un diagramme de dispersion mis en intervalles des valeurs prédites sur l'axe vertical par les valeurs observées sur l'axe horizontal. Idéalement, les points devraient se trouver sur une ligne de 45 degrés ; cette vue peut indiquer si des enregistrements sont particulièrement mal prédits par le modèle.

Résidus

Figure 15-12
Vue Résidus, style d'histogramme



Ceci affiche un diagramme de diagnostique des résidus du modèle.

Styles de diagramme. Il existe différents styles d'affichage accessibles depuis la liste déroulante Style .

- **Histogramme.** Il s'agit d'un histogramme à intervalles des résidus studentisés avec une superposition de la distribution normale. Les modèles linéaires supposent que les résidus ont une distribution normale, de sorte que l'histogramme doit, dans l'idéal, approcher étroitement la ligne de lissage.
- **Diagramme P-P.** Il s'agit d'un diagramme probabilité-probabilité mis en intervalles qui compare des résidus studentisés à une distribution normale. Si la pente des points représentés est moins forte que la ligne normale, les résidus affichent une plus grande variabilité qu'une distribution normale ; si la pente est plus forte, les résidus affichent une moins grande variabilité qu'une distribution normale. Si les points représentés ont une courbe en S, la distribution des résidus est asymétrique.

Valeurs éloignées

Figure 15-13
Vue Valeurs éloignées

Valeurs éloignées

Cible : Total sales

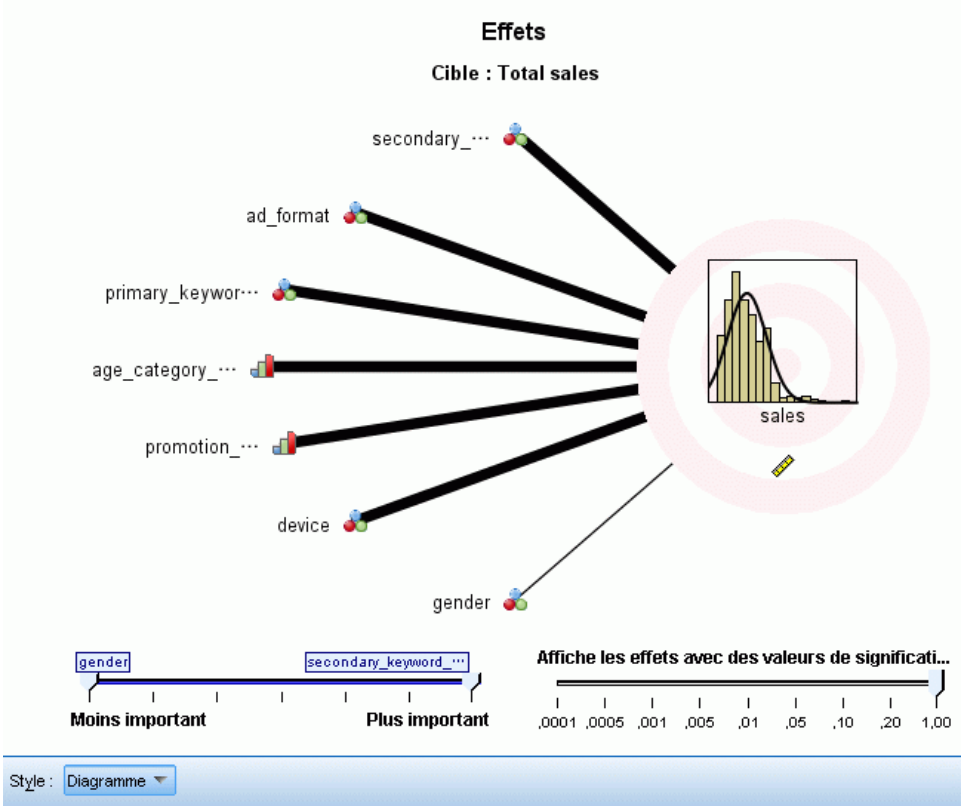
Total sales	Distance de Cook
560.040	0,026
566.440	0,025
548.990	0,018
539.630	0,018
485.430	0,014
543.240	0,014

Ce tableau répertorie les enregistrements qui exercent une influence excessive sur le modèle et affiche l'ID d'enregistrement (s'il est spécifié dans l'onglet Champs), la valeur cible et la distance de Cook. La distance de Cook est une mesure du degré de modification des résidus de tous les enregistrements si un enregistrement donné est exclu des calculs des coefficients de modèle. Une distance de Cook importante signifie que l'exclusion d'un enregistrement modifie de manière importante les coefficients et doit donc être considérée comme ayant une influence.

Les enregistrements ayant une influence doivent être examinés soigneusement afin de déterminer si vous pouvez leur octroyer une pondération inférieure dans l'estimation du modèle, tronquer les valeurs éloignées à un seuil acceptable ou supprimer complètement les enregistrements ayant une influence.

Effets

Figure 15-14
vue Effets, style de diagramme



Cette vue affiche la taille de chaque effet dans le modèle.

Styles. Il existe différents styles d'affichage accessibles depuis la liste déroulante Style .

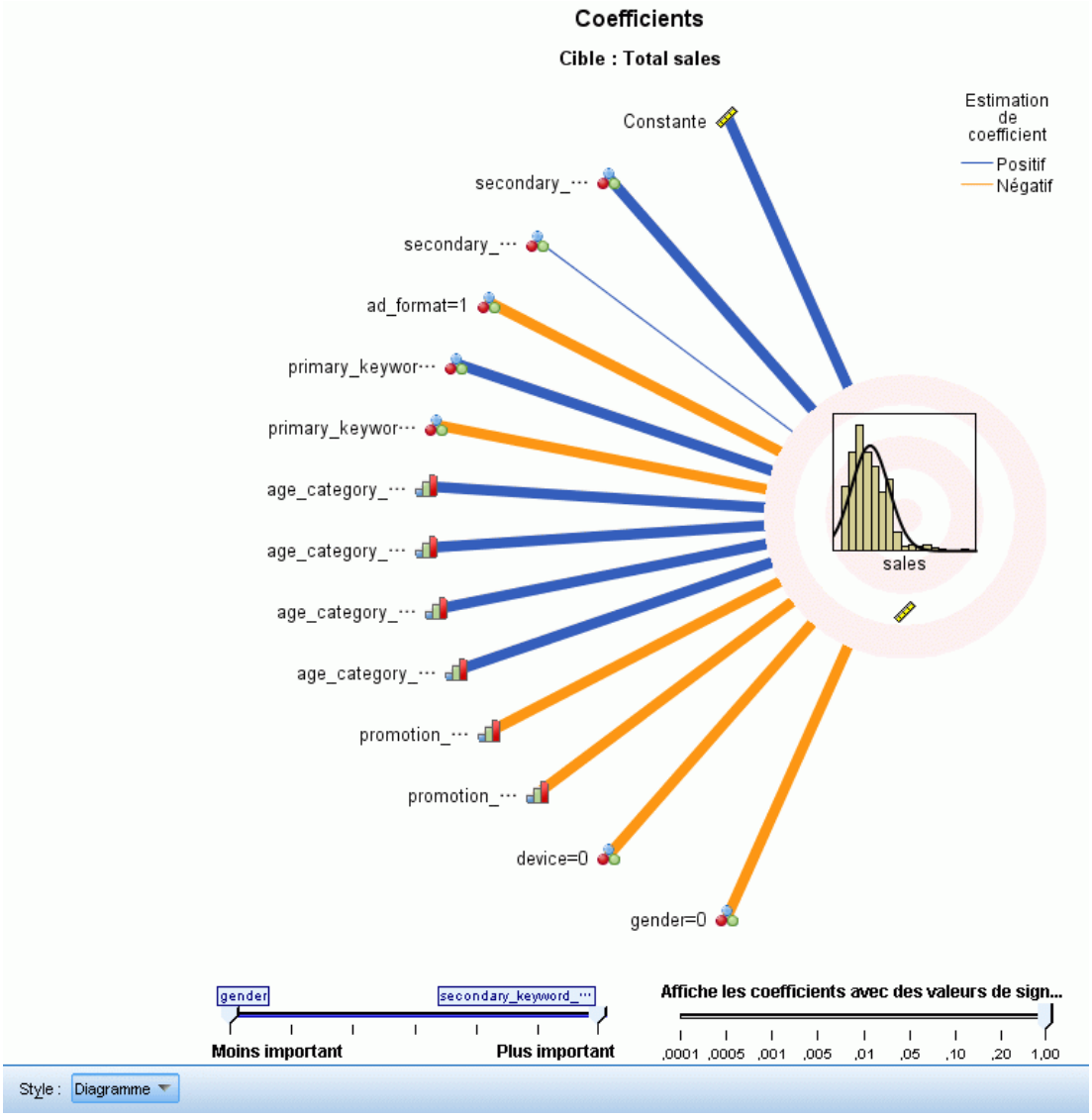
- **Diagramme.** Il s'agit d'un diagramme dans lequel les effets sont triés de haut en bas en diminuant l'importance de la variable prédite. Les lignes de connexion du diagramme sont pondérées en fonction de la signification de l'effet, une largeur de ligne plus importante correspondant à des effets plus importants (valeurs *p* plus petites). Lorsque vous passez la souris sur une ligne de connexion, une info-bulle affiche la valeur-*p* et l'importance de l'effet. Il s'agit de la valeur par défaut.
- **Tableau.** Il s'agit d'un tableau ANOVA pour le modèle général et les effets de modèle individuels. Il s'agit d'effets individuels triés de haut en bas en diminuant l'importance de la variable prédite. Notez, que par défaut, le tableau est réduit et n'affiche que les résultats du modèle global. Pour consulter les résultats des effets du modèle individuel, cliquez sur la cellule Modèle corrigé dans le tableau.

Importance des variables prédites. Il existe un curseur de l'importance des variables prédites qui contrôle celles qui sont affichées dans la vue. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les variables prédites les plus importantes. Par défaut, les 10 premiers effets sont affichés.

Signification. Il existe un curseur de signification qui offre des commandes plus avancées sur les effets affichés dans la vue, en plus de celles affichées en fonction de l'importance des variables prédites Les effets ayant des valeurs de signification plus grandes que la valeur du curseur sont masqués. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les effets les plus importants. Par défaut, la valeur est de 1,00, de sorte qu'aucun effet n'est filtré en fonction de la signification.

Coefficients

Figure 15-15
Vue Coefficients, style de diagramme



Cette vue affiche la valeur de chaque coefficient du modèle. Veuillez noter que les facteurs (variables prédites catégorielles) sont codés par un indicateur dans le modèle, de sorte que les **effets** comportant des facteurs ont généralement plusieurs **coefficients** associés, un pour chaque catégorie exceptée la catégorie correspondant au paramètre redondant (de référence).

Styles. Il existe différents styles d'affichage accessibles depuis la liste déroulante Style .

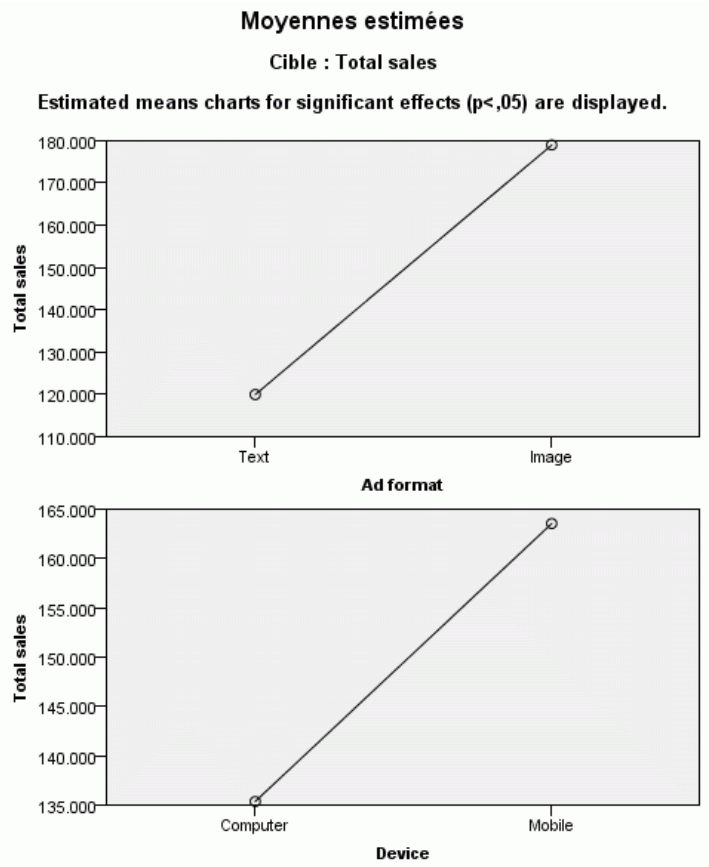
- **Diagramme.** Il s'agit d'un diagramme qui affiche d'abord la constante, puis trie les effets de haut en bas en diminuant l'importance de la variable prédite. Au sein des effets contenant des facteurs, les coefficients sont triés dans l'ordre croissant de la valeur des données. Les lignes de connexion du diagramme sont coloriées en fonction du signe du coefficient (voir le diagramme) et sont pondérées en fonction de la signification du coefficient, une largeur de ligne plus importante correspondant à des coefficient plus significatifs (valeurs-*p* plus petites). Lorsque vous passez la souris sur une ligne de connexion, une info-bulle affiche la valeur du coefficient, sa valeur-*p* et l'importance de l'effet auquel est associé le paramètre. Il s'agit du style par défaut.
- **Tableau.** Affiche les valeurs, les tests de signification et les intervalles de confiance des coefficients de modèles individuels. Après la constante, les effets sont triés de haut en bas en diminuant l'importance de la variable prédite. Au sein des effets contenant des facteurs, les coefficients sont triés dans l'ordre croissant de la valeur des données. Notez, que par défaut, le tableau est réduit et n'affiche que le coefficient, la signification et l'importance de chaque paramètre du modèle. Pour consulter l'erreur standard, la statistique *t* et l'intervalle de confiance, cliquez sur la cellule Coefficient dans le tableau. Lorsque vous passez la souris sur le nom d'un paramètre du modèle dans le tableau, une info-bulle affiche le nom du paramètre, l'effet auquel il est associé, et (pour les valeurs prédites catégorielles) les étiquettes des valeurs associées au paramètre du modèle. Ceci est particulièrement utile pour afficher les nouvelles catégories créées lorsque la préparation automatique des données fusionne les catégories similaires d'une valeur prédite catégorielle.

Importance des variables prédites. Il existe un curseur de l'importance des variables prédites qui contrôle celles qui sont affichées dans la vue. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les variables prédites les plus importantes. Par défaut, les 10 premiers effets sont affichés.

Signification. Il existe un curseur de signification qui offre des commandes plus avancées sur les coefficients affichés dans la vue, en plus de celle affichée en fonction de l'importance des variables prédites. Les coefficients ayant des valeurs de signification plus grandes que la valeur du curseur sont masqués. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les coefficients les plus importants. Par défaut, la valeur est de 1,00, de sorte qu'aucun coefficient n'est filtré en fonction de la signification.

Moyennes estimées

Figure 15-16
Vue Moyennes estimées



Il s'agit de diagrammes affichés pour des variables prédites significatives. Le diagramme affiche la valeur estimée par le modèle de la cible sur l'axe vertical pour chaque valeur de la variable prédite de l'axe horizontal en conservant toutes les autres variables prédites. Il offre une visualisation pratique des effets des coefficients de chaque variable prédite sur la cible.

Remarque : si aucune variable prédite n'est significative, aucune moyenne estimée n'est générée.

Récapitulatif de création de modèle

Figure 15-17
Vue Récapitulatif de création de modèle, algorithme pas à pas ascendant

Récapitulatif de création de modèle

Cible : Total sales

		Etape						
		1	2	3	4	5	6	7
Critère d"information		11 949,413	11 597,758	11 347,000	11 118,878	10 965,287	10 816,338	10 803,021
Effet	secondary_keyword_transformed	✓	✓	✓	✓	✓	✓	✓
	ad_format		✓	✓	✓	✓	✓	✓
	primary_keyword_transformed			✓	✓	✓	✓	✓
	age_category_transformed				✓	✓	✓	✓
	promotion_transformed					✓	✓	✓
	device						✓	✓
	gender							✓

La méthode de création de modèle est la méthode ascendante pas à pas avec le critère d"information.
Une coche signifie que l"effet est présent dans le modèle à cette étape.

Lorsqu'un algorithme de sélection de modèle différent de Aucun est sélectionné dans les paramètres de sélection de modèles, il propose certains détails concernant le processus de création du modèle.

Pas à pas ascendant. Lorsque l'algorithme de sélection est pas à pas ascendant, le tableau affiche les 10 dernières étapes de l'algorithme pas à pas. Pour chaque étape, la valeur du critère de sélection et les effets du modèle à cette étape sont affichés. Ceci vous offre un aperçu de l'ampleur de la contribution de chaque étape au modèle. Chaque colonne vous permet de trier les lignes afin que vous puissiez voir plus facilement les effets qui se trouvent dans le modèle à chaque étape donnée.

Meilleurs sous-ensembles. Lorsque l'algorithme de sélection est Meilleurs sous-ensembles, le tableau affiche les 10 meilleurs modèles. Pour chaque modèle, la valeur du critère de sélection et les effets du modèle sont affichés. Ceci vous donne un aperçu de la stabilité des meilleurs modèles ; s'ils ont tendance à avoir des effets similaires avec quelques différences, vous pouvez alors avoir une confiance raisonnable dans le « meilleur » modèle ; s'ils ont tendance à avoir des effets très différents, certains des effets peuvent être trop similaires et doivent être associés (ou l'un d'entre eux doit être supprimé). Chaque colonne vous permet de trier les lignes afin que vous puissiez voir plus facilement les effets qui se trouvent dans le modèle à chaque étape donnée.

Chapitre
16

Régression linéaire

La régression linéaire estime les coefficients de l'équation linéaire, impliquant une ou plusieurs variables indépendantes, qui estiment le mieux la valeur de la variable dépendante. Par exemple, vous pouvez essayer d'estimer les ventes annuelles globales d'un commercial (la variable dépendante) à partir de variables indépendantes telles que l'âge, l'éducation et les années d'expérience.

Exemple : Le nombre de matches gagnés par une équipe de basket-ball au cours d'une saison est-il lié au nombre moyen de points marqués par l'équipe à chaque match ? Un diagramme de dispersion indique que ces variables ont un lien linéaire. Le nombre de matches gagnés et le nombre moyen de points marqué par l'équipe adverse ont également un lien linéaire. Ces variables ont une relation négative. Lorsque le nombre de matches gagnés augmente, le nombre moyen de points marqués par les adversaires diminue. A l'aide de la régression linéaire, vous pouvez modéliser la relation entre ces variables. Un bon modèle peut être utilisé pour prévoir combien de matches les équipes vont gagner.

Statistiques : Pour chaque variable, on a les éléments suivants : nombre d'observations valides, moyenne et écart-type. Pour chaque modèle : coefficients de régression, matrice de corrélations, mesures et corrélations partielles, R multiple, R^2, R^2 ajusté, variation de R^2, erreur standard de l'estimation, tableau d'analyse de la variance, prévisions et résidus. En plus, intervalles de confiance à 95 % pour chaque coefficient de régression, matrice variances-covariances, facteur d'inflation de la variance, tolérance, test de Durbin-Watson, mesures de distances (Mahalanobis, Cook, et valeurs influentes), DfBêta, différence de prévision, intervalles d'estimation et diagnostics des observations. Diagrammes : dispersion, diagrammes partiels, histogrammes et diagrammes de répartition gaussiens.

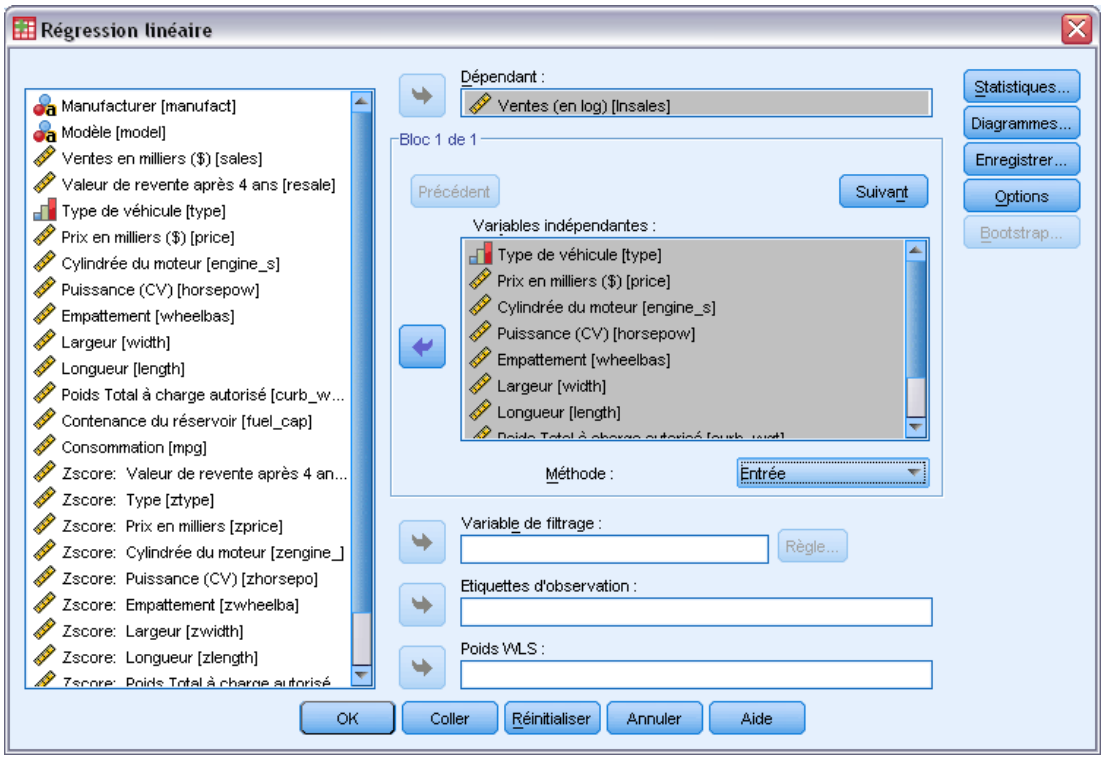
Données : Les variables dépendantes et indépendantes doivent être quantitatives. Les variables qualitatives, comme la religion, la qualification, la zone de résidence, doivent être enregistrées sous forme de variables binaires (muettes) ou sous de tout autre type de variables de contraste.

Hypothèses : Pour chaque valeur de la variable indépendante, la distribution de la variable dépendante doit être normale. La variance de la distribution de la variable dépendante doit être constante pour toutes les valeurs de la variable indépendante. La relation entre la variable dépendante et chaque variable indépendante doit être linéaire et toutes les observations doivent être indépendantes.

Obtenir une analyse de régression linéaire

- A partir des menus, sélectionnez :
  Analyse > Régression > Linéaire

Figure 16-1
Boîte de dialogue Régression linéaire



- ▶ Dans la boîte de dialogue Régression linéaire, sélectionnez une variable numérique dépendante.
- ▶ Sélectionnez une ou plusieurs variables indépendantes.

Sinon, vous pouvez :

- Grouper des variables indépendantes en blocs et spécifier différentes méthodes d'entrée pour différents sous-groupes de variables.
- Choisir une variable de sélection pour limiter l'analyse à un sous-groupe d'observations ayant une ou des valeurs particulières pour cette variable.
- Sélectionner une variable d'identification d'observations pour identifier des points sur les diagrammes.
- Sélectionnez une variable de pondération WLS numérique pour une analyse des moindres carrés pondérés.

WLS. Permet d'obtenir un modèle des moindres carrés pondéré. Les points de données sont pondérés par l'inverse de leur variance. Ainsi, les observations dont la variance est élevée ont moins d'impact sur l'analyse que celles dont la variance est faible. Si la valeur de la variable de pondération est nulle, négative ou manquante, l'observation est exclue de l'analyse.

Méthodes de sélection des variables de régression linéaire

La sélection d'une méthode vous permet de spécifier la manière dont les variables indépendantes sont entrées dans l'analyse. En utilisant différentes méthodes, vous pouvez construire divers modèles de régression à partir du même groupe de variables.

- **Introduire (régression).** Procédure de sélection de variables au cours de laquelle toutes les variables d'un bloc sont introduites en une seule opération.
- **Pas à pas.** A chaque étape, le programme saisit la variable indépendante exclue de l'équation ayant la plus petite probabilité de F, si cette probabilité est suffisamment faible. Les variables déjà comprises dans l'équation de régression sont éliminées si leur probabilité de F devient trop grande. Le processus s'arrête lorsqu'aucune variable ne peut plus être introduite ou éliminée.
- **Eliminer bloc.** Procédure de sélection de variables dans laquelle toutes les variables d'un bloc sont supprimées en une seule étape.
- **Elimination descendante.** Procédure de sélection de variables au cours de laquelle toutes les variables sont entrées dans l'équation, puis éliminées une à une. La variable ayant la plus petite corrélation partielle avec la variable dépendante est la variable dont l'élimination est étudiée en premier. Si elle répond aux critères d'élimination, elle est supprimée. Une fois la première variable éliminée, l'élimination de la variable suivante restant dans l'équation et ayant le plus petit coefficient de corrélation partielle est étudiée. La procédure prend fin quand plus aucune variable de l'équation ne satisfait aux critères d'élimination.
- **Introduction ascendante.** Procédure de sélection pas à pas de variables, dans laquelle les variables sont introduites séquentiellement dans le modèle. La première variable considérée est celle qui a la plus forte corrélation positive ou négative avec la variable dépendante. Cette variable n'est introduite dans l'équation que si elle satisfait le critère d'introduction. Si la première variable est introduite dans l'équation, la variable indépendante externe à l'équation et qui présente la plus forte corrélation partielle est considérée ensuite. La procédure s'interrompt lorsqu'il ne reste plus de variables satisfaisant au critère d'introduction.

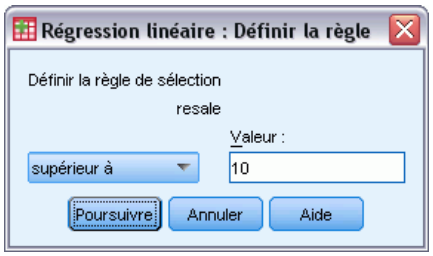
Les valeurs de significativité dans vos résultats sont basées sur l'adéquation à un modèle unique. Par conséquent, les valeurs de significativité ne sont généralement pas valables lorsqu'on utilise une méthode progressive (pas à pas, ascendante ou descendante).

Toutes les variables doivent respecter le critère de tolérance pour être entrées dans l'équation, quelle que soit la méthode d'entrée spécifiée. Le niveau de tolérance par défaut est 0,0001. Une variable n'est pas entrée si elle fait passer la tolérance d'une autre variable déjà entrée dans le modèle en dessous du seuil de tolérance.

Toutes les variables indépendantes sélectionnées sont ajoutées dans un seul modèle de régression. Cependant, vous pouvez spécifier différentes méthodes d'entrée pour les sous-groupes de variables. Par exemple, vous pouvez entrer un bloc de variables dans le modèle de régression en utilisant la sélection pas à pas, et un second bloc en utilisant la sélection ascendante. Pour ajouter un second bloc de variables au modèle de régression, cliquez sur Suivant.

Régression linéaire : Définir la règle

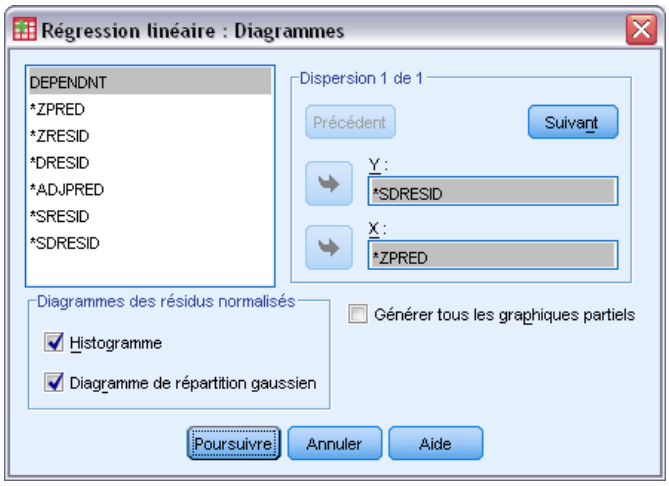
Figure 16-2
Boîte de dialogue Régression linéaire : Définir loi



Les observations définies par la règle de sélection sont incluses dans l'analyse. Par exemple, si vous sélectionnez une variable, choisissez égale et saisissez 5 pour la valeur, alors seules les observations pour lesquelles la variable sélectionnée a une valeur égale à 5 seront incluses dans l'analyse. Une valeur chaîne est également permise.

Diagrammes de régression linéaire

Figure 16-3
Boîte de dialogue Régression linéaire : Graphiques (diagrammes)



Les diagrammes peuvent aider à valider les hypothèses de normalité, linéarité et d'égalité des variances. Les diagrammes sont également utiles pour détecter les valeurs éloignées, les observations éloignées et les observations influentes. Après avoir été enregistrés comme variables nouvelles, les prévisions, résidus et autres diagnostics sont disponibles dans l'éditeur de données pour construire des diagrammes avec les variables indépendantes. Les diagrammes suivants sont disponibles :

Diagrammes de dispersion : Vous pouvez afficher deux des éléments suivants : la variable dépendante, les prévisions standardisées, les résidus standardisés, les résidus supprimés, les prévisions ajustées, les résidus standardisés et les résidus supprimés de Student. Affichez les résidus standardisés par rapport aux prévisions standardisées pour vérifier la linéarité et l'égalité des variances.

Liste des variables sources :. Répertorie la variable dépendante (DEPENDNT), ainsi que les variables prévues et les résidus suivants : prévisions standardisées (*ZPRED), résidus standardisés (*ZRESID), résidus supprimés (*DRESID), prévisions ajustées (*ADJPRED), résidus de Student (*SRESID), résidus supprimés de Student (*SDRESID).

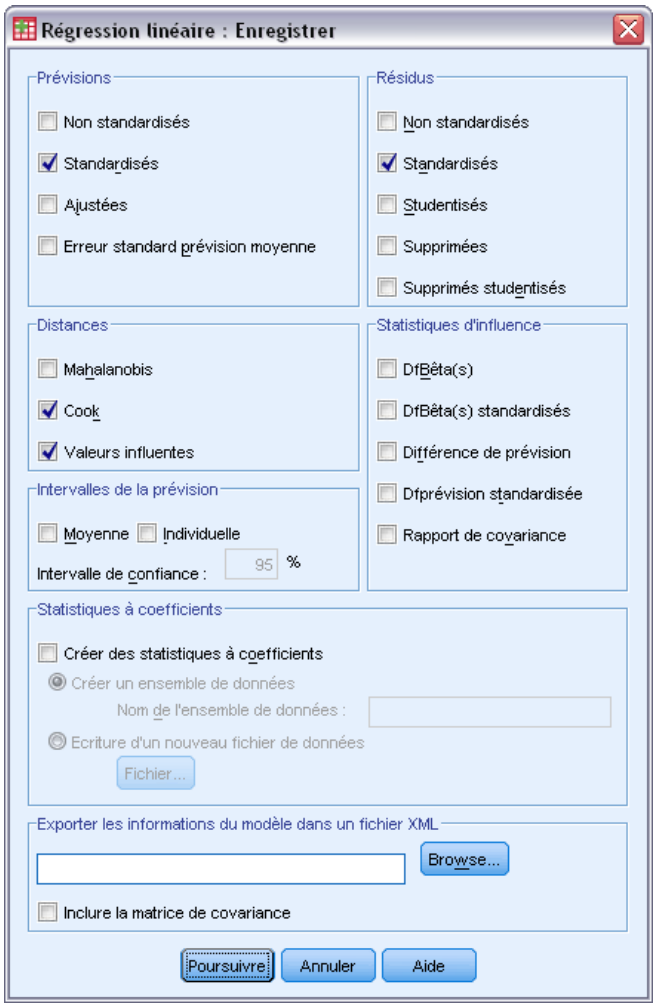
Générer tous les graphiques partiels : Affiche des diagrammes de dispersion des résidus de chaque variable indépendante et les résidus de la variable dépendante lorsque les deux variables sont régressées séparément par rapport au reste des variables indépendantes. Au moins deux variables indépendantes doivent être dans l'équation pour produire un diagramme partiel.

Diagrammes des résidus standardisés : Vous pouvez obtenir des histogrammes des résidus standardisés et des diagrammes de répartition gaussiens en comparant la répartition des résidus standardisés à une répartition gaussienne.

Si vous demandez des diagrammes, des statistiques récapitulatives sont affichées pour les prévisions standardisées et les résidus standardisés (*ZPRED et *ZRESID).

Régression linéaire : Enregistrer de nouvelles variables

Figure 16-4
Boîte de dialogue Régression linéaire : Enregistrer les nouvelles variables



Vous pouvez enregistrer les prévisions, les résidus et autres statistiques utiles pour les diagnostics. Chaque sélection ajoute une ou plusieurs variables à votre fichier de données actif.

Prévisions : Valeurs prévues par le modèle de régression pour chaque observation.

- **Non standardisés.** Valeur prévue par le modèle pour la variable dépendante.
- **Standardisés.** Transformation de chaque prévision en sa forme normalisée. La prévision moyenne est soustraite de la prévision, et la différence est divisée par l'écart-type des prévisions. Les prévisions standardisées ont une moyenne de 0 et un écart-type de 1.
- **Ajustées.** Prévision pour une observation lorsqu'une observation est exclue du calcul des coefficients de régression.
- **Erreur standard prévision moyenne.** Erreurs standard des prévisions. Estimation de l'écart-type de la valeur moyenne de la variable expliquée pour les unités statistiques qui ont les mêmes valeurs pour les valeurs explicatives.

Distances : Mesures permettant d'identifier les observations avec des combinaisons inhabituelles de valeurs pour les variables indépendantes et les observations qui peuvent avoir un impact important sur le modèle.

- **Mahalanobis.** Mesure de la distance entre les valeurs d'une observation et la moyenne de toutes les observations sur les variables indépendantes. Une distance de Mahalanobis importante identifie une observation qui a des valeurs extrêmes pour des variables indépendantes.
- **Cook.** Mesure du degré dont les résidus de toutes les observations sont modifiés si une observation donnée est exclue des calculs des coefficients de régression. Si la distance de Cook est élevée, l'exclusion d'une observation changerait substantiellement la valeur des coefficients.
- **Valeurs influentes.** Mesures de l'influence d'un point sur l'ajustement de la régression. La valeur influente centrée varie de 0 (aucune influence sur la qualité de l'ajustement) à (N-1)/N.

Intervalles de la prévision : Les limites supérieure et inférieure pour les intervalles de la prévision moyenne et individuelle.

- **Moyenne.** Limites inférieure et supérieure (deux variables) de l'intervalle de prévision de la réponse moyenne prévue.
- **Individuelle.** Limites inférieure et supérieure (deux variables) de l'intervalle de prévision de la variable dépendante pour une seule observation.
- **Intervalle de confiance.** Entrez une valeur comprise entre 1 et 99,99 pour spécifier le seuil de confiance pour les deux intervalles de la prévision. Vous devez sélectionner Moyenne ou Individuelle avant d'entrer cette valeur. Les seuils d'intervalle de confiance typiques sont 90, 95 et 99.

Résidus : La valeur réelle de la variable indépendante moins la valeur prévue par l'équation de régression.

- **Non standardisés.** Différence entre la valeur observée et la valeur prévue par le modèle.
- **Standardisés.** Résidu, divisé par une estimation de son erreur standard. Egalement appelés résidus de Pearson, les résidus standardisés ont une moyenne de 0 et un écart-type de 1.
- **Studentisés.** Résidu, divisé par une estimation de son écart-type, qui varie d'une observation à l'autre, selon la distance entre les valeurs et la moyenne des variables indépendantes pour chaque observation.
- **Supprimées.** Résidu d'une observation lorsque celle-ci est exclue du calcul des coefficients de régression. Il s'agit de la différence entre la valeur de la variable dépendante et la prévision ajustée.
- **Supprimés studentisés.** Résidu supprimé d'une observation, divisé par son erreur standard. La différence entre le résidu supprimé de Student et le résidu de Student associé indique l'impact de l'élimination d'une observation sur sa propre prédiction.

Influences individuelles : Modification des coefficients de régression (DfBêta[s]) et des prévisions (différence de prévision) qui résulte de l'exclusion d'une observation particulière. Les valeurs DfBêtas et de différence de prévision standardisées sont également disponibles ainsi que le rapport de covariance.

- **Différence de bêta.** La différence de bêta correspond au changement des coefficients de régression qui résulte du retrait d'une observation particulière. Une valeur est calculée pour chaque terme du modèle, y compris la constante.

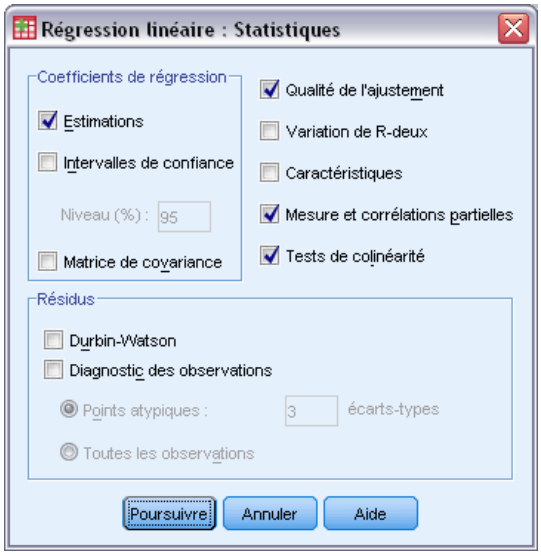
- **DfBêta(s) standardisée.** Différence normalisée de la valeur bêta. Modification du coefficient de régression, résultant de l'exclusion d'une observation donnée. Vous pouvez par exemple examiner les observations ayant des valeurs absolues supérieures à 2, divisées par la racine carrée de N, N représentant le nombre d'observations. Une valeur est calculée pour chaque terme du modèle, y compris la constante.
- **Différence d'ajustement.** La différence d'ajustement est le changement de la prévision résultant de l'exclusion d'une observation donnée.
- **Dfprévision standardisée.** Différence normalisée de la valeur ajustée. Modification de la prévision qui résulte de l'exclusion d'une observation donnée. Vous pouvez par exemple examiner les valeurs standardisées dont la valeur absolue est supérieure à 2 fois la racine carrée de p/N, p correspondant au nombre de paramètres du modèle et N, au nombre d'observations.
- **Rapport de covariance.** Rapport entre le déterminant de la matrice de variance-covariance si une observation donnée a été exclue du calcul des coefficients de régression et le déterminant de la matrice de covariance avec toutes les observations incluses. Si le rapport est proche de 1, l'observation modifie peu la matrice de covariance.

Statistiques à coefficients. Enregistre les coefficients de régression dans un ensemble de données ou dans un fichier de données. Les ensembles de données sont disponibles pour utilisation ultérieure dans la même session mais ne sont pas enregistrés en tant que fichiers sauf si vous le faites explicitement avant la fin de la session. Le nom des ensembles de données doit être conforme aux règles de dénomination de variables.

Exporter les informations du modèle dans un fichier XML : Les estimations de paramètres et leurs covariances (facultatif) sont exportées vers le fichier spécifié au format XML (PMML). Vous pouvez utiliser ce fichier de modèle pour appliquer les informations du modèle aux autres fichiers de données à des fins d'évaluation.

Statistiques de régression linéaire

Figure 16-5
Boîte de dialogue Statistiques



Les statistiques suivantes sont disponibles :

Coefficients de régression : L'option Estimations affiche le coefficient de régression *B*, l'erreur standard de *B*, le coefficient bêta standardisé, la valeur *t* de *B* et le niveau de signification bilatéral de *t*. Les Intervalles de confiance affichent les intervalles de confiance avec le niveau spécifié de confiance pour chaque coefficient de regréssion ou une matrice de covariance. L'option Matrice de covariance affiche la matrice de variance-covariance des coefficients de régression avec les covariances hors de la diagonale et les variances dans la diagonale. Une matrice de corrélation est également affichée.

Qualité de l'ajustement : Les variables entrées et supprimées du modèle sont listées et les statistiques de la qualité de l'ajustement suivantes sont affichées : *R* multiple, R^2 et R^2 ajusté, erreur standard de l'estimation et un tableau d'analyse de variance.

Variation de R-deux : Variation de la statistique du R^2 obtenue en ajoutant ou en enlevant une variable indépendante. Si la variation du R^2 associée à une variable est importante, cela signifie que la variable est une bonne explication de la variable dépendante.

Descriptives. Fournit le nombre d'observations valides, la moyenne et l'écart-type de chaque variable de l'analyse. Une matrice de corrélations avec le seuil de signification unilatéral et le nombre d'observations pour chaque corrélation sont également affichés.

Corrélation partielle. Corrélation résiduelle entre deux variables après l'élimination de la corrélation due à leur association mutuelle avec les autres variables. Il s'agit de la corrélation entre la variable dépendante et une variable indépendante lorsque les effets linéaires des autres variables indépendantes du modèle ont été éliminés des deux variables.

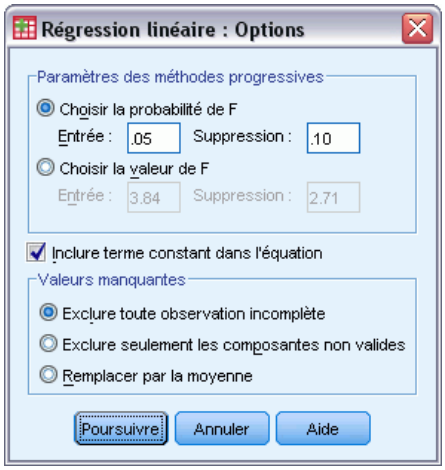
Corrélation partielle. Il s'agit de la corrélation entre la variable dépendante et une variable indépendante lorsque les effets linéaires des autres variables indépendantes du modèle ont été éliminés de la variable indépendante. Elle est liée à la modification du R-deux lorsqu'une variable est ajoutée à une équation. (Parfois appelée corrélation semi-partielle.)

Tests de colinéarité : La colinéarité (ou multicolinéarité) est la situation indésirable où une variable indépendante est une fonction linéaire d'autres variables indépendantes. Les valeurs propres de la matrice des produits croisés dimensionnés et non centrés, les indices de conditionnement et les proportions de décomposition de variance sont affichés ainsi que les facteurs d'inflation de la variance (VIF) et les tolérances pour les variables individuelles.

Résidus : Affiche le test de Durbin-Watson de corrélation sérielle des résidus et le diagnostic des observations correspondant au critère de sélection (valeurs éloignées de *n* écarts-types).

Régression linéaire : Options

Figure 16-6
Boîte de dialogue Régression linéaire : Options



Les options suivantes sont disponibles :

Paramètres des méthodes progressives : Ces options sont valables lorsque la méthode de sélection ascendante, descendante ou progressive a été sélectionnée. Des variables peuvent être entrées ou supprimées du modèle soit en fonction de la signification (probabilité) de la valeur *F*, soit en fonction de la valeur *F* elle-même.

- **Utiliser la probabilité de F.** Une variable est entrée dans le modèle si le seuil de signification de la valeur F est inférieur à la valeur Entrée ; la variable est éliminée si ce seuil est supérieur à la valeur Elimination. La valeur Entrée doit être inférieure à la valeur Elimination et toutes deux doivent être positives. Pour introduire davantage de variables dans le modèle, diminuez la valeur Entrée. Pour éliminer davantage de variables du modèle, réduisez la valeur Elimination.
- **Utiliser la valeur de F.** Une variable est introduite dans un modèle si sa valeur F est supérieure à la valeur Entrée et elle est éliminée si la valeur F est inférieure à la valeur Elimination. La valeur Entrée doit être supérieure à la valeur Elimination et toutes deux doivent être positives. Pour introduire davantage de variables dans le modèle, réduisez la valeur du champ Entrée. Pour éliminer davantage de variables dans le modèle, augmentez la valeur du champ Elimination.

Inclure terme constant dans l'équation : Par défaut, le modèle de régression inclut un terme constant. Désélectionner cette option force la régression jusqu'à l'origine, ce qui est rarement utilisé. Certains résultats de la régression jusqu'à l'origine ne sont pas comparables aux résultats de la régression incluant une constante. Par exemple, R^2 ne peut pas être interprété de la manière habituelle.

Valeurs manquantes : Vous pouvez choisir l'un des éléments suivants :

- **Exclure toute observation incomplète :** Seules les observations dont les valeurs sont valides pour toutes les variables sont incluses dans les analyses.

- **Exclure seulement les composantes non valides :** Les observations pour lesquelles les données sont complètes pour la paire de variables corrélées sont utilisées pour calculer le coefficient de corrélation sur lequel l'analyse de régression est basée. Les degrés de liberté sont basés sur le minimum *N* par paire.
- **Remplacer par la moyenne :** Toutes les observations sont utilisées pour les calculs, en substituant la moyenne de la variable aux observations manquantes.

Fonctionnalités supplémentaires de la commande REGRESSION

Le langage de syntaxe de commande vous permet aussi de :

- Ecrire une matrice de corrélation ou lire une matrice à la place de données brutes afin d'obtenir une analyse de régression (avec la sous-commande `MATRIX`).
- Spécifier des niveaux de tolérance (avec la sous-commande `CRITERIA`).
- Obtenir plusieurs modèles pour des variables dépendantes différentes ou identiques (avec les sous-commandes `METHOD` et `DEPENDENT`).
- Obtenir des statistiques supplémentaires (avec les sous-commandes `DESCRIPTIVES` et `STATISTICS`).

Reportez-vous à la *Référence de syntaxe de commande* pour une information complète concernant la syntaxe.

Chapitre
17

Régression ordinale

La régression ordinale vous permet d'effectuer un modèle dont la variable dépendante est une réponse ordinale sur un groupe de prédicteurs pouvant être soit des facteurs, soit des covariables. Le concept de régression ordinale repose sur la méthodologie de McCullagh (1980, 1998) ; vous trouverez cette procédure sous le nom de `PLUM` dans la syntaxe.

L'analyse de la régression linéaire standard implique la réduction des différences de sommes des carrés entre une variable de réponse (dépendante) et une combinaison pondérée des prédicteurs (variables indépendantes). Les coefficients estimés reflètent le mode d'affectation de la réponse due aux modifications des prédicteurs. La réponse est numérique, dans le sens où les changements de niveau de réponse sont équivalents pour l'ensemble des intervalles de réponse. Par exemple, la différence de taille existant entre une personne de 150 cm et une de 140 cm est de 10 cm, ce qui correspond à la même différence existant entre une personne de 210 cm et une de 200 cm. Ces relations n'existent peut-être pas pour les variables ordinales, pour lesquelles le choix et le nombre de modalités de réponse peut être relativement arbitraire.

Exemple : La régression ordinale peut être utilisée en vue d'étudier la réaction des patients à certaines doses de médicament. §AA Les réactions possibles peuvent être classées en *aucune*, *légère*, *modérée* ou *grave*. La différence entre une réaction légère et une réaction modérée est difficile, voire impossible à quantifier car elle repose sur une perception. De plus, la différence entre une réponse légère et une réponse modérée peut être supérieure ou inférieure à la différence existant entre une réponse modérée et une réponse grave.

Diagrammes et statistiques : Fréquences observées et théoriques, et fréquences cumulées, résidus de Pearson pour les fréquences cumulées, probabilités observées et théoriques, probabilités observées et cumulées théoriques de chaque modalité de réponse par type de covariable, corrélation asymptotique et matrices de covariance des estimations des paramètres, Khi-deux de Pearson et Khi-deux du rapport de vraisemblance, qualité d'ajustement, historique des itérations, test d'hypothèses de lignes parallèles, estimations des paramètres, erreurs standard, intervalles de confiance, statistiques de Cox et Snell, de Nagelkerke et R^2 de McFadden.

Données. La variable dépendante est considérée comme ordinale, mais peut être soit numérique, soit chaîne. L'ordre est déterminé par le tri des valeurs de la variable dépendante par ordre croissant. La plus petite valeur définit la première modalité. Les variables actives sont supposées être qualitatives. Les covariables doivent être numériques. Notez que l'utilisation de plusieurs covariables continues peut engendrer la création d'un tableau volumineux de probabilités par cellule.

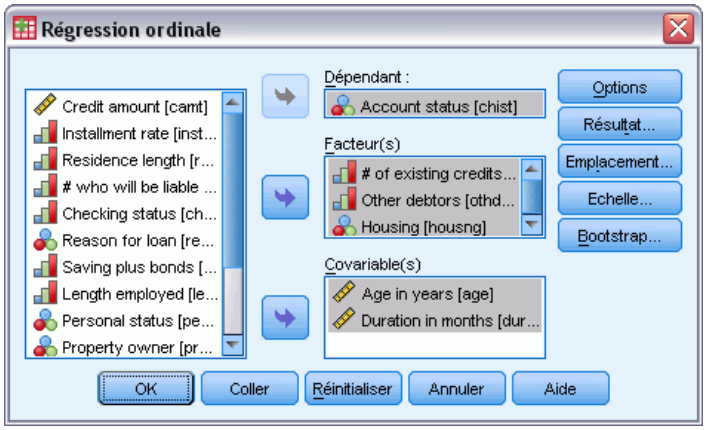
Hypothèses : Seule une variable de réponse est autorisée et doit donc être spécifiée. De plus, pour chaque type de valeurs distinct parmi les variables explicatives, les réponses sont supposées être des variables multinomiales explicatives.

Procédures apparentées : La régression logistique nominale utilise des modèles similaires pour les variables dépendantes nominales.

Obtention d'une régression ordinale

- A partir des menus, sélectionnez :
  Analyse > Régression > Ordinale...

Figure 17-1
Boîte de dialogue Régression ordinale

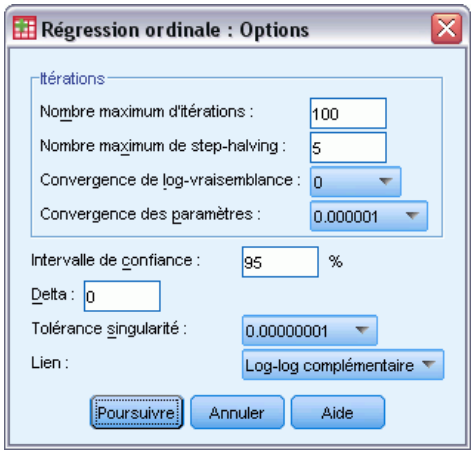


- Sélectionnez une variable dépendante.
- Cliquez sur OK.

Régression ordinale : Options

La boîte de dialogue Options vous permet d'ajuster des paramètres utilisés dans l'algorithme d'estimation itératif, de choisir un niveau de confiance pour vos estimations de paramètres et de sélectionner une fonction de lien.

Figure 17-2
Boîte de dialogue Régression ordinale : Options



Itérations : Vous pouvez personnaliser l'algorithme itératif.

- **Nombre maximum d'itérations :** Spécifiez un nombre entier non négatif. Si vous indiquez 0, la procédure renvoie aux estimations initiales.
- **Nombre maximum de dichotomie :** Spécifiez un nombre entier positif.
- **Convergence de log-vraisemblance :** L'algorithme s'interrompt si la modification absolue ou relative apportée au log-vraisemblance est inférieure à cette valeur. Le critère n'est pas utilisé si 0 est spécifié.
- **Convergence des paramètres :** L'algorithme s'interrompt si la modification absolue ou relative apportée à chaque estimation de paramètres est inférieure à cette valeur. Le critère n'est pas utilisé si 0 est spécifié.

Intervalle de confiance : Spécifiez une valeur supérieure ou égale à 0 et inférieure à 100.

Delta. Valeur ajoutée aux fréquences zéro par cellule. Spécifiez une valeur non négative, inférieure à 1.

Tolérance singularité : Option utilisée pour vérifier les prédicteurs à haute dépendance. Sélectionnez une valeur dans la liste des options.

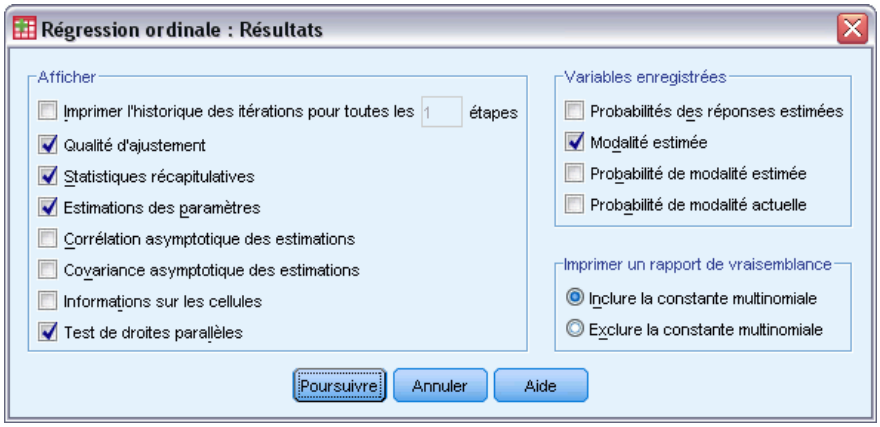
Fonction de lien. La fonction de lien consiste en une transformation des probabilités cumulées permettant d'estimer le modèle. Les cinq fonctions de lien disponibles sont récapitulées dans le tableau suivant.

Fonction	Forme	Application standard
Logit	$\log(\xi / (1-\xi))$	Modalités réparties de façon égale
Log-log complémentaire	$\log(-\log(1-\xi))$	Modalités supérieures les plus probables
Log-log négatif	$-\log(-\log(\xi))$	Modalités inférieures plus probables
Probit	$\Phi^{-1}(\xi)$	Variable de latence normalement répartie
Cauchit (Cauchy inverse)	$\tan(\pi(\xi-0,5))$	Variable de latence avec de nombreux extrema

Régression ordinale : Résultat

La boîte de dialogue Résultat vous permet de générer des tableaux d'affichage dans le Viewer et d'enregistrer des variables dans le fichier de travail.

Figure 17-3
Boîte de dialogue Régression ordinale : Résultat



Afficher :Génère des tableaux pour :

- **Historique des itérations d'impression :** Le log-vraisemblance et les estimations des paramètres sont imprimés en fonction de la fréquence des itérations d'impression spécifiée. La première et la dernière itération sont toujours imprimées.
- **Qualité d'ajustement :** Khi-deux de Pearson et khi-deux du rapport de vraisemblance. Ces éléments sont calculés en fonction du classement indiqué dans la liste des variables.
- **Statistiques récapitulatives :** Statistiques de Cox et Snell, de Nagelkerke et R^2 de McFadden.
- **Estimations des paramètres :** Estimations des paramètres, erreurs standard et intervalles de confiance.
- **Corrélation asymptotique des estimations :** Matrice de corrélations des estimations de paramètres.
- **Covariance asymptotique des estimations :** Matrice de covariances des estimations de paramètres.
- **Informations sur les cellules :** Fréquences observées et théoriques, et fréquences cumulées, résidus de Pearson pour les fréquences cumulées, probabilités observées et théoriques, probabilités observées et cumulées de chaque modalité de réponse par type de covariable. Notez que pour les modèles comportant plusieurs types de covariable (par exemple, les modèles avec covariables continues), cette option peut générer un tableau très volumineux et donc, difficile à gérer.
- **Test de droites parallèles :** Test d'hypothèse selon laquelle les paramètres d'emplacement sont équivalents pour tous les niveaux de la variable dépendante. Cette option est uniquement disponible pour le modèle d'emplacement.

Variables enregistrées : Enregistre les variables suivantes dans le fichier de travail :

- **Probabilités des réponses estimées :** Probabilités estimées sur un modèle de classement d'un type de facteur/covariable dans les modalités de réponse. Il existe autant de probabilités que de nombre de modalités de réponse.
- **Modalité estimée :** Modalité de réponse contenant la probabilité estimée maximale pour un type de facteur/covariable.

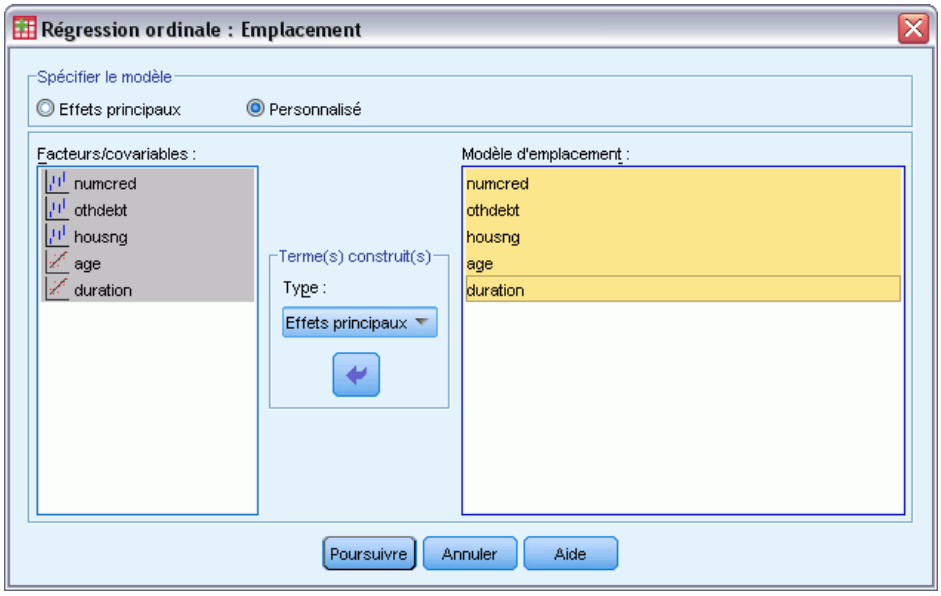
- **Probabilité de modalité estimée :** Probabilité estimée de classement d'un type de facteur/covariable au sein d'une modalité prévue. Cette probabilité représente également le maximum de probabilités estimées du type de facteur/covariable.
- **Probabilité de modalité actuelle :** Probabilité estimée de classement d'un type de facteur/covariable au sein de la modalité actuelle.

Imprimer un rapport de vraisemblance : Commande l'affichage du rapport de log-vraisemblance. Inclure la constante multinomiale vous indique la valeur complète de la vraisemblance. Pour comparer vos résultats parmi les produits n'incluant pas de constante, vous pouvez choisir de l'exclure.

Régression ordinale : Emplacement

La boîte de dialogue Emplacement vous permet de spécifier le modèle d'emplacement de l'analyse.

Figure 17-4
Boîte de dialogue Régression ordinale : Emplacement



Spécifier un modèle : Un modèle comportant des effets principaux contient des effets principaux de covariable et de facteur, mais aucun effet d'interaction. Vous pouvez créer un modèle personnalisé pour définir des sous-groupes d'interactions de facteurs ou de covariables.

Facteurs/covariables : §FM Les facteurs et les covariables sont répertoriés.

Modèle d'emplacement : Le modèle dépend des effets principaux et des effets d'interaction que vous sélectionnez.

Termes construits

Pour les facteurs et covariables sélectionnés :

Interaction : Crée le terme d'interaction du plus haut niveau de toutes les variables sélectionnées. Il s'agit de la valeur par défaut.

Effets principaux : Crée un terme d'effet principal pour chaque variable sélectionnée.

Toutes d'ordre 2 : Crée toutes les interactions d'ordre 2 possibles des variables sélectionnées.

Toutes d'ordre 3 : Crée toutes les interactions d'ordre 3 possibles des variables sélectionnées.

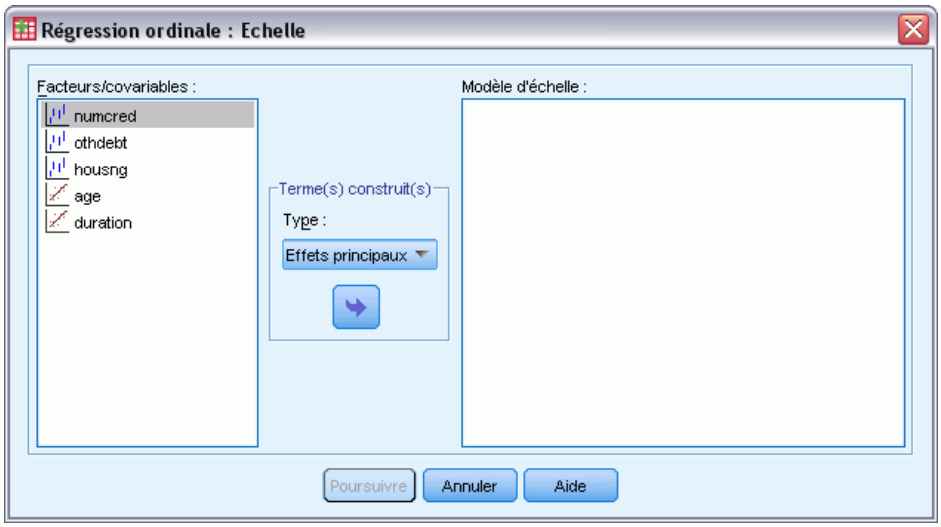
Toutes d'ordre 4 : Crée toutes les interactions d'ordre 4 possibles des variables sélectionnées.

Toutes d'ordre 5 : Crée toutes les interactions d'ordre 5 possibles des variables sélectionnées.

Régression ordinale : Echelle

La boîte de dialogue Echelle vous permet de spécifier le modèle d'échelle de l'analyse.

Figure 17-5
Boîte de dialogue Régression ordinale : Echelle



Facteurs/covariables : §FM Les facteurs et les covariables sont répertoriés.

Modèle d'échelle : Le modèle dépend des effets principaux et des effets d'interaction que vous sélectionnez.

Termes construits

Pour les facteurs et covariables sélectionnés :

Interaction : Crée le terme d'interaction du plus haut niveau de toutes les variables sélectionnées. Il s'agit de la valeur par défaut.

Effets principaux : Crée un terme d'effet principal pour chaque variable sélectionnée.

Toutes d'ordre 2 : Crée toutes les interactions d'ordre 2 possibles des variables sélectionnées.

Toutes d'ordre 3 : Crée toutes les interactions d'ordre 3 possibles des variables sélectionnées.

Toutes d'ordre 4 : Crée toutes les interactions d'ordre 4 possibles des variables sélectionnées.

Toutes d'ordre 5 : Crée toutes les interactions d'ordre 5 possibles des variables sélectionnées.

Fonctionnalités supplémentaires de la commande PLUM

Vous pouvez personnaliser la régression ordinale en collant vos sélections dans une fenêtre de syntaxe et en modifiant la syntaxe de commande `PLUM`. Le langage de syntaxe de commande vous permet aussi de :

- Créer des tests d'hypothèse personnalisés en spécifiant des hypothèses nulles comme combinaisons linéaires de paramètres.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
18

Ajustement de fonctions

La procédure d'ajustement de fonctions produit des statistiques de régression d'ajustement de fonctions et les diagrammes relatifs pour 11 modèles différents de régression d'ajustement de fonctions. Un modèle différent est produit pour chaque variable dépendante. Vous pouvez aussi enregistrer les prévisions, les résidus et les intervalles de la prévision comme nouvelles variables.

Exemple : Un fournisseur de services Internet suit le pourcentage dans le temps du trafic de messages électroniques infectés par un virus sur ses réseaux. Un diagramme de dispersion révèle que la relation n'est pas linéaire. Vous pouvez ajuster un modèle quadratique ou cubique en fonction des données, vérifier la validité des hypothèses et la qualité d'ajustement du modèle.

Statistiques : Pour chaque modèle : coefficients de régression, R multiples, R^2, R^2 ajusté, erreur standard de la prévision, tableau d'analyse de la variance, prévisions, résidus et intervalles de prévision. Modèles : linéaire, logarithmique, inverse, quadratique, cubique, de puissance, composé, en S, logistique, de croissance et exponentiel.

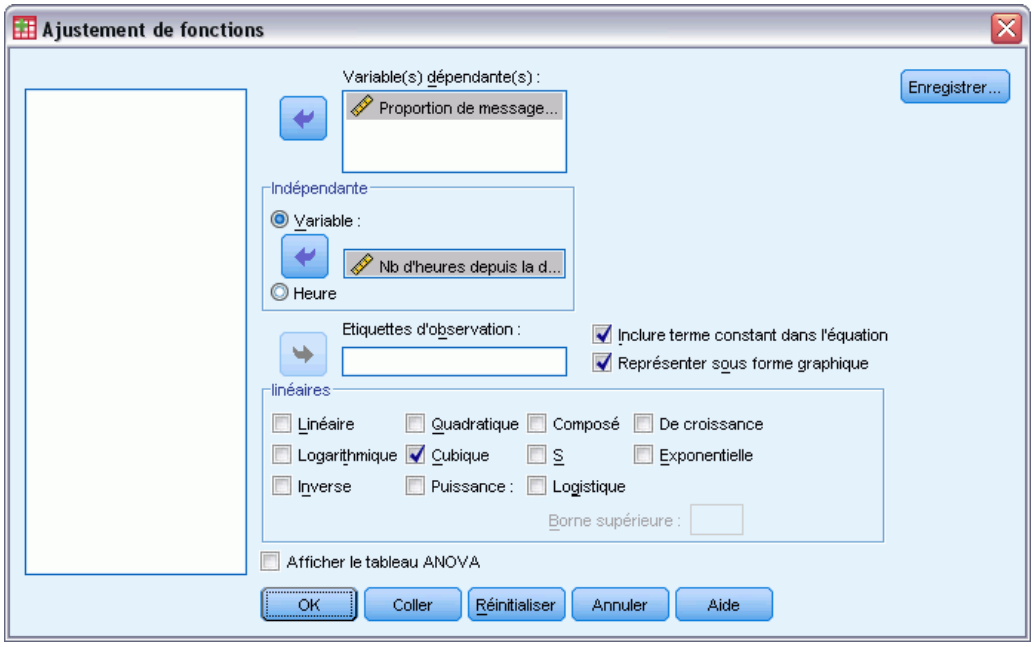
Données : Les variables dépendantes et indépendantes doivent être quantitatives. Si vous sélectionnez Temps à partir de l'ensemble de données actif comme variable indépendante (au lieu de sélectionner une variable), la procédure Ajustement de fonctions génère une variable de temps où la durée entre les observations est uniforme. Si Temps est sélectionné, la variable dépendante doit être une mesure de séries chronologiques. L'analyse des séries chronologiques nécessite une structure de fichier de données dans lequel chaque observation (rangée) représente un ensemble d'observations à des moments différents et où la durée entre les observations est uniforme.

Hypothèses : Vérifiez vos données graphiquement pour déterminer comment sont reliées les variables indépendantes et dépendantes (de manière linéaire ou exponentielle, etc.). Les résidus d'un bon modèle doivent être répartis aléatoirement et doivent être normaux. Si vous utilisez un modèle linéaire, les hypothèses suivantes doivent être vérifiées : pour chaque valeur de la variable indépendante, la distribution de la variable dépendante doit être normale. La variance de la distribution de la variable dépendante doit être constante pour toutes les valeurs de la variable indépendante. La relation entre la variable dépendante et la variable indépendante doit être linéaire, et toutes les observations doivent être indépendantes.

Pour obtenir un ajustement de fonctions

- A partir des menus, sélectionnez :
  Analyse > Régression > Ajustement de fonctions

Figure 18-1
Boîte de dialogue Ajustement de fonctions



- Sélectionnez au moins une variable dépendante. Un modèle différent est produit pour chaque variable dépendante.
- Sélectionnez une variable indépendante (une variable dans le fichier de travail ou dans l'ensemble de données actif ou Temps).
- Eventuellement :
  - Sélectionner une variable pour étiqueter des observations dans les diagrammes de dispersion. Pour chaque point du diagramme de dispersion, utilisez l'outil de sélection de points pour afficher la valeur de la variable avec Etiquette d'observation.
  - Cliquez sur Enregistrer pour enregistrer les prévisions, les résidus et les intervalles de prévision comme nouvelles variables.

Les options suivantes sont également disponibles :

- **Inclure terme constant dans l'équation :** Evalue un terme constant dans l'équation de régression. La constante est incluse par défaut.
- **Représenter sous forme graphique :** Représente graphiquement les valeurs de la variable dépendante et chaque modèle sélectionné face à la variable indépendante. Un diagramme séparé est produit pour chaque variable dépendante.
- **Afficher le tableau ANOVA :** Affiche un tableau récapitulatif de l'analyse de la variance pour chaque modèle sélectionné.

Modèles d'ajustement de fonctions

Vous pouvez choisir des modèles de régression d'ajustement de fonctions. Pour déterminer quel modèle utiliser, représentez vos données sous forme graphique. Si vos variables semblent être liées linéairement, utilisez un modèle de régression linéaire simple. Lorsque vos variables ne sont pas liées linéairement, essayez de transformer vos données. Lorsque la transformation n'améliore pas les choses, vous devrez peut-être utiliser un modèle plus élaboré. Observez un diagramme de dispersion de vos données. Si le diagramme ressemble à une fonction mathématique que vous reconnaissez, ajustez vos données en fonction de ce type de modèle. Par exemple, si vos données ressemblent à une fonction exponentielle, utilisez un modèle exponentiel.

Linéaire. Modèle dont l'équation est Y = b0 + (b1 * t). Les valeurs de la série sont modélisées comme fonction linéaire du temps.

Logarithmique. Modèle dont l'équation est Y = b0 + (b1 * ln(t)).

Inverse. Modèle dont l'équation est Y = b0 + (b1 / t).

Quadratique. Modèle dont l'équation est Y = b0 + (b1 * t) + (b2 * t**2). Le modèle quadratique peut être utilisé pour modéliser une série qui « décolle » ou qui s'amortit.

Cubique. Modèle défini par l'équation Y = b0 + (b1 * t) + (b2 * t**2) + (b3 * t**3).

Exposant. Modèle dont l'équation est Y = b0 * (t**b1) ou ln(Y) = ln(b0) + (b1 * ln(t)).

Composé. Modèle dont l'équation est la suivante : Y = b0 * (b1**t) ou ln(Y) = ln(b0) + (ln(b1) * t).

En S. Modèle dont l'équation est Y = e**(b0 + (b1/t)) ou ln(Y) = b0 + (b1/t).

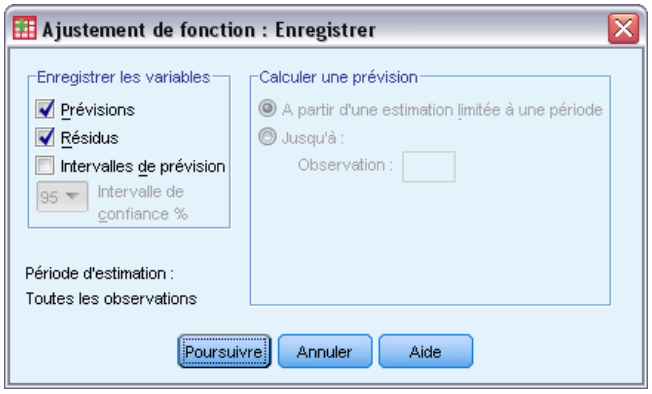
Logistique. Modèle dont l'équation est Y = 1 / (1/u + (b0 * (b1**t))) ou ln(1/y-1/u)= ln (b0) + (ln(b1)*t), u étant la valeur de la borne supérieure. Après avoir sélectionné la logistique, précisez la valeur de la borne supérieure à utiliser dans l'équation de régression. La valeur doit être un nombre positif supérieur à la plus grande valeur de la variable dépendante.

Croissance. Modèle dont l'équation est Y = e**(b0 + (b1 * t)) ou ln(Y) = b0 + (b1 * t).

Exponentielle. Modèle dont l'équation est Y = b0 * (e**(b1 * t)) ou ln(Y) = ln(b0) + (b1 * t).

Enregistrement de l'ajustement de fonctions

Figure 18-2
Boîte de dialogue Ajustement de fonctions : Enregistrer



Enregistrer les variables : Pour chaque modèle sélectionné, vous pouvez enregistrer les prévisions, les résidus (valeur observée de la variable dépendante moins la prévision du modèle) et les intervalles de prévision (limites supérieure et inférieure). Les nouveaux noms de variable et les étiquettes descriptives s'affichent dans un tableau dans la fenêtre de résultats.

Calculer une prévision : Si, dans l'ensemble de données actif, vous sélectionnez Temps à la place d'une variable comme variable indépendante, vous pouvez spécifier une période de prévision au-delà de la fin de la série chronologique. Vous avez le choix entre les options suivantes :

- **A partir d'une estimation limitée à une période :** Prévoit les valeurs pour toutes les observations du fichier, à partir des observations de la période d'estimation. La période d'estimation qui s'affiche en bas de la boîte de dialogue est définie avec la boîte de sous dialogue Intervalle de l'option Sélectionner des observations du menu Données. Si aucune période d'estimation n'a été définie, toutes les observations sont utilisées pour prévoir les valeurs.
- **Jusqu'à :** Prévoit les valeurs jusqu'à la date, l'heure ou le numéro de l'observation spécifié, à partir des observations de la période d'estimation. Cette fonctionnalité peut être utilisée pour prévoir les valeurs au-delà de la dernière observation de la série chronologique. Les variables courantes de date définies déterminent les zones de texte disponibles pour la spécification de la fin de la période de prévision. Si aucune variable de date n'est définie, vous pouvez spécifier le numéro de l'observation finale.

Utilisez l'option Définir des dates... dans le menu Données pour créer des variables de date.

Chapitre
19

Régression des moindres carrés partiels

La procédure de régression des moindres carrés partiels estime les modèles de régression des moindres carrés partiels (également connus sous le nom de « projection to latent structure », PLS). La technique de prévision PLS constitue une solution de remplacement par rapport à la régression par les moindres carrés classiques, à la corrélation canonique ou à la modélisation d'équation structurelle, particulièrement utile lorsque les variables prédites présentent une forte corrélation ou lorsque le nombre de variables prédites dépasse le nombre d'observations.

PLS combine des fonctionnalités d'analyse des composants principaux et la régression multiple. Un ensemble de facteurs latents expliquant autant que possible la covariance entre les variables indépendantes et dépendantes est extrait. Ensuite, une étape de régression prévoit les valeurs des variables dépendantes à l'aide de la décomposition des variables indépendantes.

Disponibilité. PLS est une commande d'extension qui requiert l'installation du IBM® SPSS® Statistics - Integration Plug-In for Python sur le système où vous prévoyez d'exécuter PLS. Le module d'extension PLS doit être installé séparément, et peut être téléchargé depuis *http://www.ibm.com/developerworks/spssdevcentral*.

Tableaux. Proportion de variance expliquée (par facteur latent), pondérations de facteurs latents, corrélations de facteurs latents, importance de la variable indépendante dans la projection (VIP), et estimations des paramètres de régression (par variable dépendante) sont tous générés par défaut.

Diagrammes : Importance de la variable dans la projection (VIP), facteurs, pondération des trois premiers facteurs latents et distance au modèle sont tous générés depuis l'onglet Options.

Niveau de mesure. Les variables dépendantes et indépendantes (prédites) peuvent être échelle, nominal ou ordinal. La procédure considère que le niveau de mesure approprié a été assigné à toutes les variables, bien que vous puissiez changer provisoirement le niveau de mesure d'une variable en cliquant avec le bouton droit de la souris sur la variable dans la liste des variables source, puis en sélectionnant un niveau de mesure dans le menu contextuel. Les variables qualitatives (nominales ou ordinales) sont traitées de manière équivalente par la procédure.

Codage des variables indicatrices. La procédure recode provisoirement les variables dépendantes qualitatives via le codage un-de-*c* pendant la durée de la procédure. S'il existe des modalités *c* d'une variable, cette dernière est stockée comme vecteurs *c*, la première modalité étant identifiée par (1,0,...,0), la suivante par (0,1,0,...,0), ... et la dernière par (0,0,...,0,1). Les variables dépendantes qualitatives sont représentées à l'aide du codage de façon fictive, c'est-à-dire, elles omettent simplement l'indicateur correspondant à la modalité de référence.

Pondérations d'effectif. Les valeurs de pondération sont arrondies au nombre entier le plus près avant utilisation. Les observations avec des pondérations manquantes ou des pondérations inférieures à 0,5, ne sont pas utilisées dans les analyses.

Valeurs manquantes : Les valeurs manquantes spécifiées par l'utilisateur et par le système sont traitées comme non valides.

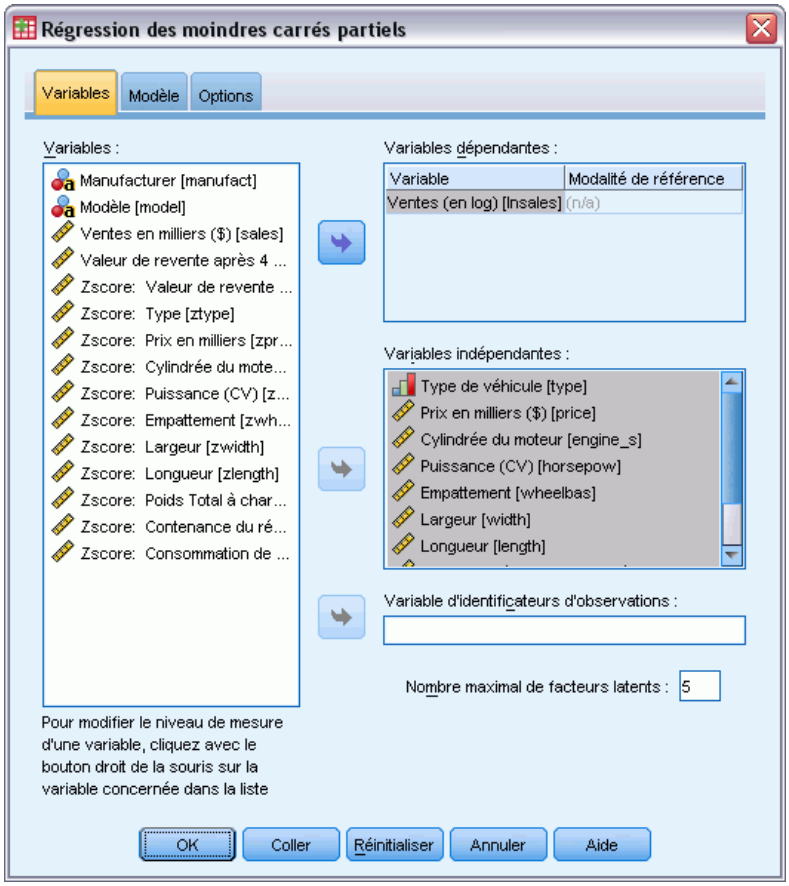
Rééchelonnement. Tous les variables de modèle sont centrées et standardisées, dont les variables d'indicateur représentant les variables qualitatives.

Pour obtenir la régression des moindres carrés partiels

A partir des menus, sélectionnez :

Analyse > Régression > Moindres carrés partiels...

Figure 19-1
Régression des moindres carrés partiels - Onglet Variables



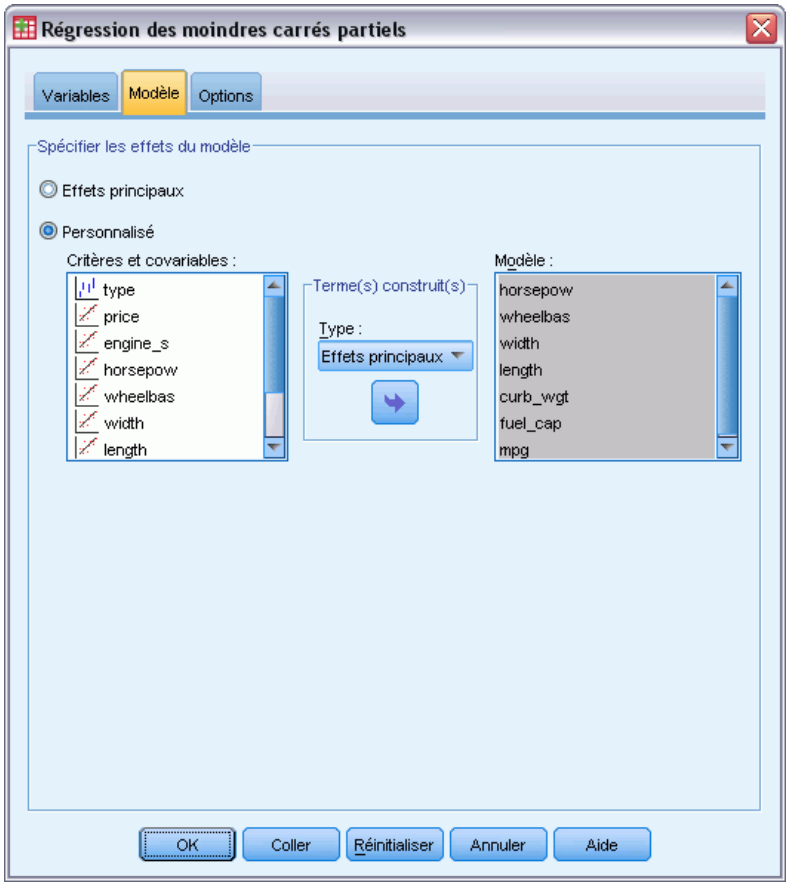
- ▶ Sélectionnez au moins une variable dépendante.
- ▶ Sélectionnez au moins une variable indépendante.

Sinon, vous pouvez :

- Indiquez une modalité de référence pour les variables qualitatives dépendantes (nominale ou ordinale).
- Indiquez une variable à utiliser comme identificateur unique pour les ensembles de données enregistrés et les résultats par observation.
- Indiquez une limite supérieure sur le nombre de facteurs latents à extraire.

Modèle

Figure 19-2
Régression des moindres carrés partiels, onglet Modèle



Spécifier les effets du modèle. Un modèle comportant des effets principaux contient tous les effets principaux de covariable et de facteur. Sélectionnez Personnalisé pour préciser les interactions. Vous devez indiquer tous les termes à inclure dans le modèle.

Critères et covariables : Les facteurs et les covariables sont répertoriés.

Modèle : Le modèle dépend de la nature de vos données. Après avoir sélectionné Autre, vous pouvez choisir les effets principaux et les interactions qui présentent un intérêt pour votre analyse.

Termes construits

Pour les facteurs et covariables sélectionnés :

Interaction : Crée le terme d'interaction du plus haut niveau de toutes les variables sélectionnées. Il s'agit de la valeur par défaut.

Effets principaux : Crée un terme d'effet principal pour chaque variable sélectionnée.

Toutes d'ordre 2 : Crée toutes les interactions d'ordre 2 possibles des variables sélectionnées.

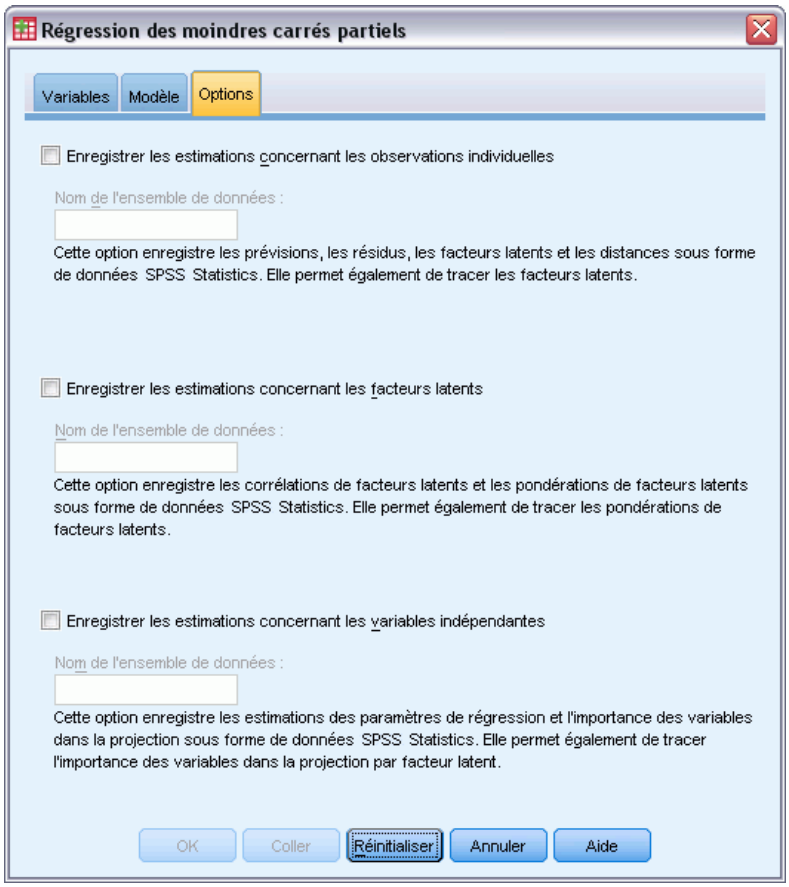
Toutes d'ordre 3 : Crée toutes les interactions d'ordre 3 possibles des variables sélectionnées.

Toutes d'ordre 4 : Crée toutes les interactions d'ordre 4 possibles des variables sélectionnées.

Toutes d'ordre 5 : Crée toutes les interactions d'ordre 5 possibles des variables sélectionnées.

Options

Figure 19-3
Régression des moindres carrés partiels, onglet Options



L'onglet Options permet d'enregistrer et de tracer des estimations de modèles pour des observations individuelles, des facteurs latents, et des variables prédites.

Pour chaque type de données, indiquez le nom d'un ensemble de données. Les noms d'ensemble de données doivent être uniques. Si vous spécifiez le nom d'un ensemble de données existant, son contenu est remplacé ; sinon, un ensemble de données est créé.

- **Enregistrez les estimations concernant les observations individuelles.** Enregistre les estimations de modèle par observation : les prévisions, les résidus, la distance au modèle de facteur latent, et les facteurs latents. Elle permet également de tracer les facteurs latents.

- **Enregistrez les estimations concernant les facteurs latents.** Enregistre les corrélations de facteurs latents et les pondérations de facteurs latents. Elle permet également de tracer les pondérations de facteurs latents.
- **Enregistrez les estimations concernant les variables indépendantes** Enregistre les estimations des paramètres de régression et l'importance des variables dans la projection (VIP). Elle permet également de tracer l'importance des variables dans la projection par facteur latent.

Chapitre
20

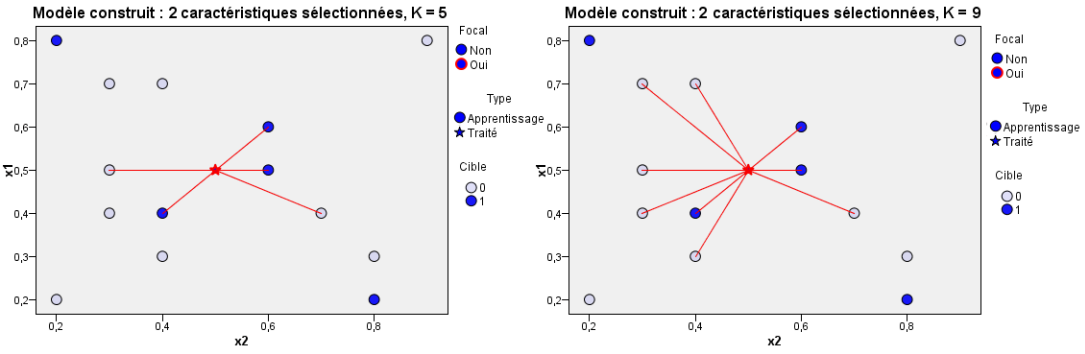
Analyse du voisin le plus proche

L'analyse du voisin le plus proche est une méthode de classification d'observations en fonction de leur similarité avec les autres observations. En apprentissage automatique, elle a été développée comme une façon de reconnaître les configurations de données sans avoir à recourir à une correspondance exacte avec d'autres configurations ou observations stockées. Les observations semblables sont proches l'une de l'autre et les observations dissemblables sont éloignées l'une de l'autre. Par conséquent, la distance entre deux observations est une mesure de leur dissemblance.

Les observations proches l'une de l'autre sont "voisines." Lorsqu'une observation est présentée (traitée), sa distance de chacune des observations du modèle est calculée. Les classifications des observations les plus semblables – les voisins les plus proches – sont comptées et la nouvelle observation est placée dans la catégorie qui contient le plus grand nombre de voisins les plus proches.

Vous pouvez spécifier le nombre de voisins les plus proches à examiner ; cette valeur est nommée *k*. Ces images indiquent comment une nouvelle observation serait répertoriée à l'aide de deux valeurs différentes de *k*. Lorsque *k* = 5, la nouvelle observation est placée dans la catégorie *1* parce qu'une majorité des voisins les plus proches appartient à la catégorie *1*. Lorsque *k* = 9, la nouvelle observation est placée dans la catégorie *0* parce qu'une majorité des voisins les plus proches appartient à la catégorie *0*.

Figure 20-1
Les effets de la modification de k sur la classification



L'analyse du voisin le plus proche peut également être utilisée pour calculer des valeurs pour une cible continue. Dans cette situation, la valeur cible de la médiane ou de la moyenne des voisins les plus proches est utilisée pour obtenir la valeur prédite de la nouvelle observation.

Cible et descriptives. La cible et les descriptives peuvent être :

- **Nominal.** Une variable peut être traitée comme étant nominale si ses valeurs représentent des modalités sans classement intrinsèque (par exemple, le service de la société dans lequel travaille un employé). La région, le code postal ou l'appartenance religieuse sont des exemples de variables nominales.

- **Ordinal.** Une variable peut être traitée comme étant ordinale si ses valeurs représentent des modalités associées à un classement intrinsèque (par exemple, des niveaux de satisfaction allant de Très mécontent à Très satisfait). Exemples de variable ordinale : des scores d'attitude représentant le degré de satisfaction ou de confiance, et des scores de classement des préférences.
- **Echelle.** Une variable peut être traitée comme une variable d'échelle (continue) si ses valeurs représentent des modalités ordonnées avec une mesure significative, de sorte que les comparaisons de distance entre les valeurs soient adéquates. L'âge en années et le revenu en milliers de dollars sont des exemples de variable d'échelle.

  Les variables qualitatives et ordinales sont traitées de manière équivalente par l'analyse du voisin le plus proche. La procédure considère que le niveau de mesure approprié a été assigné à chaque variable, bien que vous puissiez changer provisoirement le niveau de mesure d'une variable en cliquant avec le bouton droit de la souris sur la variable dans la liste des variables sources, puis en sélectionnant un niveau de mesure dans le menu contextuel.

Dans la liste des variables, une icône indique le niveau de mesure et le type de données :

	Numérique	**Chaîne**	**Date**	**Heure**
Echelle (continue).		n/a		
Ordinal				
Nominal				

Codage des variables indicatrices. La procédure recode provisoirement les variables prédites qualitatives et les variables dépendantes via le codage un-de-*c* pour la durée de la procédure. S'il existe des modalités *c* d'une variable, la variable est stockée comme vecteurs c, la première modalité étant identifiée par (1,0,...,0), la suivante par (0,1,0,...,0), ... et la dernière par (0,0,...,0,1).

Ce système de codage augmente le nombre de dimensions de l'espace des descriptives. Plus particulièrement, le nombre total de dimensions correspond au nombre de variables indépendantes d'échelle plus le nombre de modalités sur l'ensemble des variables prédites qualitatives. En conséquence, ce système de codage peut provoquer un ralentissement de la formation. Si votre formation des voisins les plus proches s'effectue très lentement, vous pouvez essayer de réduire le nombre de modalités dans vos variables prédites qualitatives en combinant des modalités similaires ou en supprimant les observations comportant des modalités extrêmement rares avant de lancer la procédure.

Tout codage un-de-*c* repose sur les données de formation, même si un échantillon traité est défini (reportez-vous à Partitions). Ainsi, si l'échantillon traité contient des observations avec des modalités de variables prédites absentes des données de formation, ces observations ne seront pas évaluées. Si l'échantillon traité contient des observations avec des modalités de variable dépendantes absentes des données de formation, ces observations seront évaluées.

Rééchelonnement. Les descriptives d'échelle sont normalisées par défaut. Le rééchelonnement repose entièrement sur les données de formation, même si un échantillon traité est défini (reportez-vous à Partitions sur p. 136). Si vous spécifiez une variable pour définir des partitions, il est important que ces descriptives présentent des distributions similaires à travers les échantillons de formation et les échantillons traités. Par exemple, utilisez la procédure Explorer pour examiner les distributions à travers les partitions.

Pondérations d'effectif. Cette procédure ignore les pondérations d'effectif.

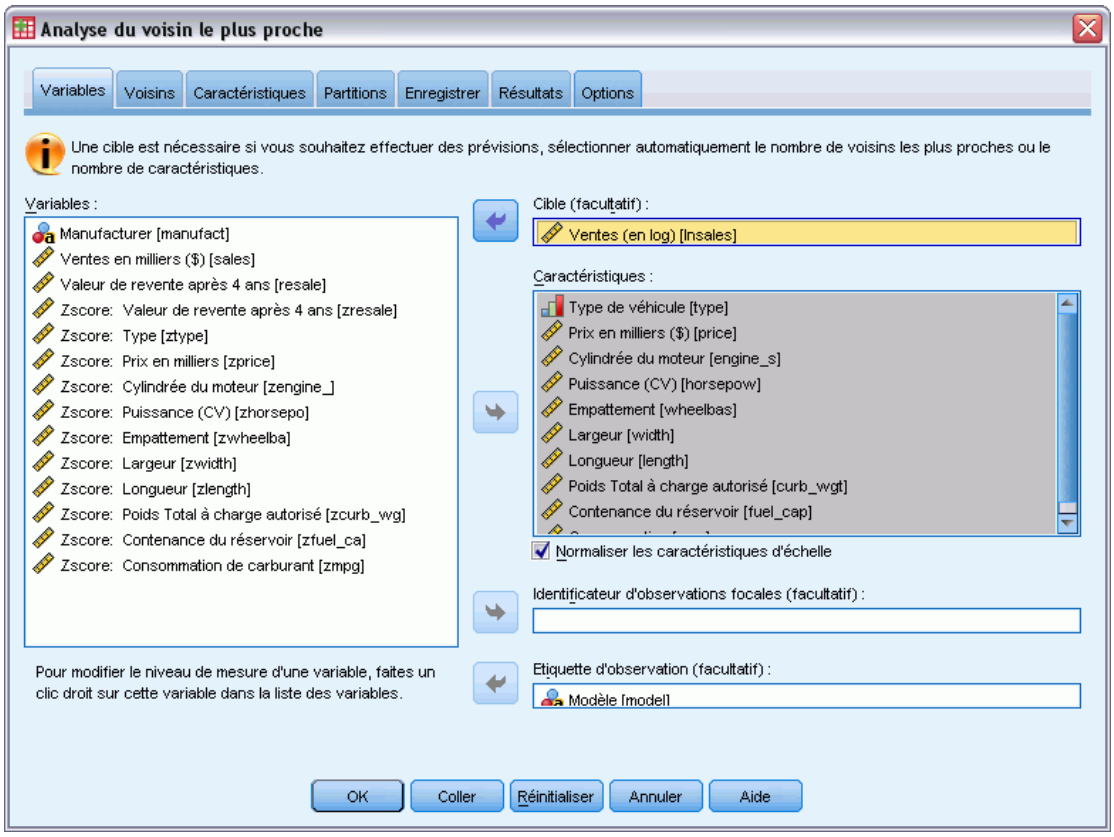
Réplication de résultats. La procédure utilise la génération de nombres aléatoires pendant l'attribution aléatoire des partitions et les niveaux de validation croisée. Si vous souhaitez répliquer vos résultats exactement, en plus d'utiliser les mêmes paramètres de procédure, définissez un générateur pour le Mersenne Twister (reportez-vous à Partitions sur p. 136), ou utilisez des variables pour définir les partitions et les niveaux de validation croisée.

Pour obtenir une analyse du voisin le plus proche

A partir des menus, sélectionnez :

Analyse > Classification > Voisin le plus proche…

Figure 20-2
Onglet Variables d'analyse du voisin le plus proche



- Spécifiez une ou plusieurs descriptives, qui peuvent être considérées comme des variables indépendantes si une cible existe.

Cible (facultative). Si aucune cible (variable dépendante ou réponse) n'est spécifiée, la procédure trouve alors les *k* voisins les plus proches seulement : aucun classement ou prévision ne sera exécuté.

Normaliser les descriptives d'échelle. Les descriptives normalisées possèdent le même intervalle de valeurs, ce qui permet d'améliorer les performances de l'algorithme d'estimation. La normalisation ajustée, [2*(x−min)/(max−min)]−1, est utilisée. Les valeurs normalisées ajustées sont comprises entre −1 et 1.

Identificateur d'observations focales (facultatif). Cela vous permet de marquer les observations présentant un intérêt particulier. Par exemple, un chercheur veut déterminer si les résultats d'un examen scolaire d'un certain district (l'observation focale) sont comparables à ceux de districts similaires. Il utilise l'analyse du voisin le plus proche pour connaître les districts scolaires les plus identiques selon un ensemble de descriptives donné. Il compare ensuite les résultats de l'examen du district focal à ceux des voisins les plus proches.

Les observations focales pourraient être également appliquées à des études cliniques pour sélectionner les observations de contrôle similaires aux observations cliniques. Les observations focales sont affichées dans le tableau des *k* voisins les plus proches et des distances, sur le graphique de l'espace des descriptives, dans le diagramme des pairs et sur la carte des quadrants. Les informations sur les observations locales sont enregistrées dans les fichiers spécifiés sur l'onglet Résultats.

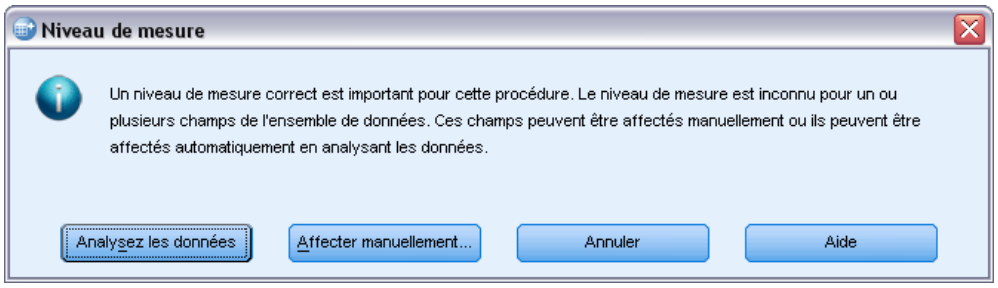
Les observations à valeur positive sur la variable spécifiée sont traitées comme des observations focales. Spécifier une variable sans valeur positive n'est pas valide.

Etiquette d'observation (facultative). Les observations sont étiquetées à l'aide de ces valeurs sur le graphique de l'espace des descriptives, dans le diagramme des pairs et sur la carte des quadrants.

Champs avec un niveau de mesure inconnu

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Figure 20-3
Alerte du niveau de mesure

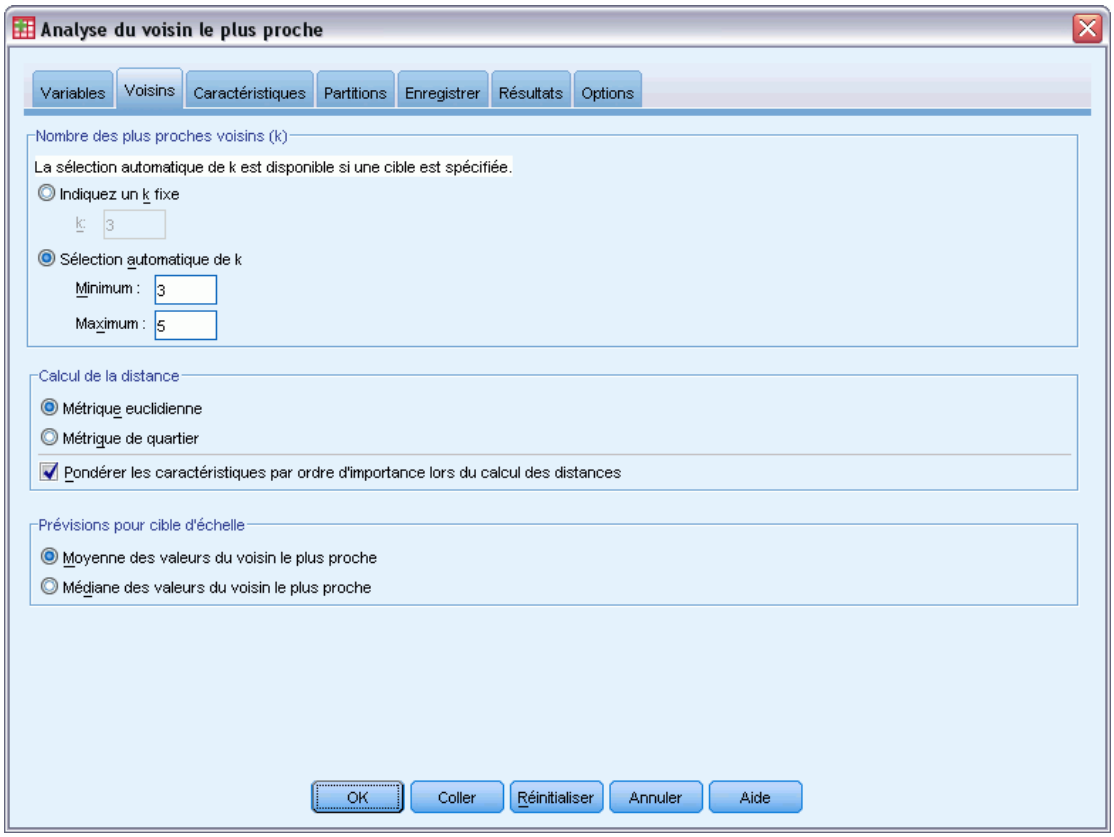


- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Voisins

Figure 20-4
Onglet Analyse du voisin le plus proche



Nombre de voisins les plus proches (k). Spécifiez le nombre de voisins les plus proches. Remarque : l'utilisation d'un nombre élevé de voisins ne garantit pas forcément un modèle plus précis.

Si une cible est spécifiée sur l'onglet Variables, vous pouvez également indiquer un intervalle de valeurs et permettre à la procédure de choisir le nombre « optimal » de voisins au sein de cet intervalle. La méthode pour déterminer le nombre de voisins les plus proches dépend si la sélection des descriptives est requise par l'onglet Descriptives ou non.

- Si oui, la sélection des descriptives sera alors exécutée pour chaque valeur de *k* dans l'intervalle requis, et le *k* ainsi que l'ensemble des descriptives l'accompagneant, avec le taux d'erreur le plus faible (ou l'erreur de la somme des carrés la plus faible si la cible est une échelle), seront sélectionnés.
- Si la séléction des descriptives n'est pas activée, alors la validation croisée de niveau *V* sera utilisée pour sélectionner le nombre de voisins "optimal". Reportez-vous à l'onglet Partitions pour contrôler l'attribution de niveaux.

Calcul de la distance. Il s'agit de la métrique employée pour spécifier la distance métrique utilisée dans la mesure de la similarité des observations.

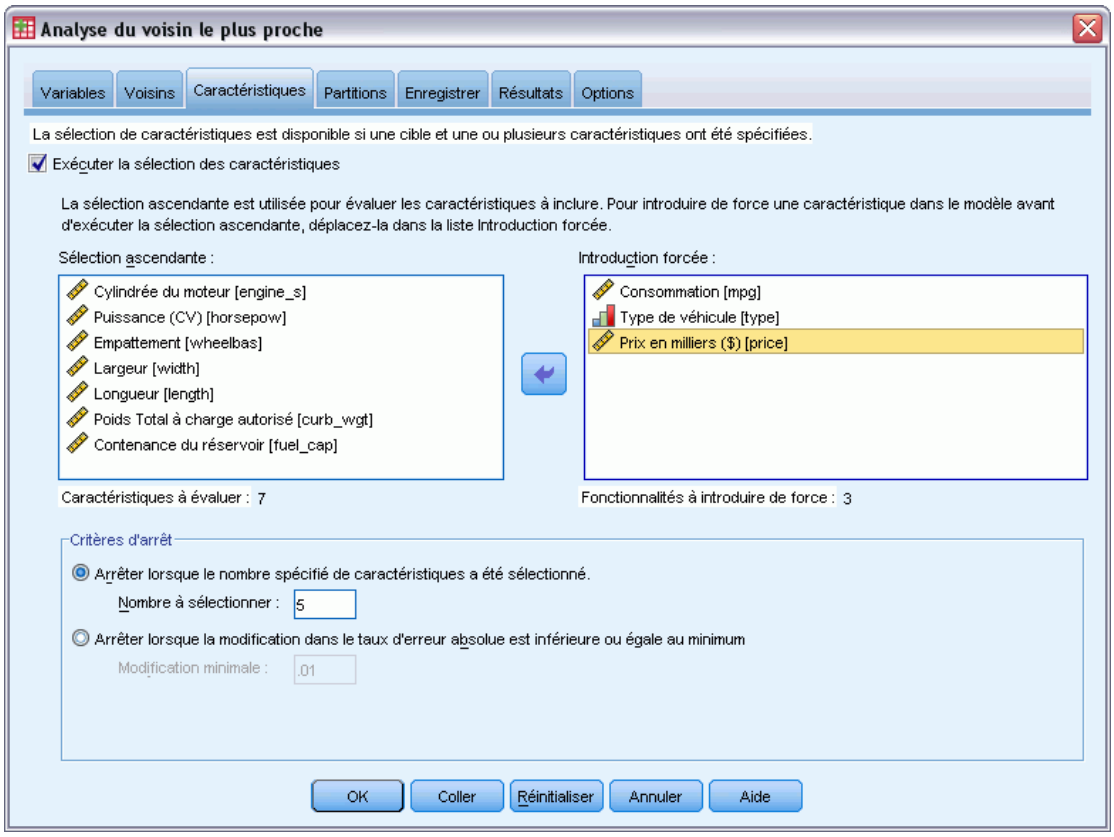
- **Métrique euclidienne.** La distance entre deux observations, x et y, est la racine carrée de la somme, sur toutes les dimensions, des carrés des différences entre les valeurs de ces observations.
- **Mesure de la distance de Manhattan.** La distance entre deux observations est la somme, sur toutes les dimensions, des différences absolues entre les valeurs de ces observations. Appelée également distance City Block.

Si une cible est spécifiée dans l'onglet Variables, vous pouvez également choisir de pondérer les descriptives selon leur importance normalisée lors du calcul des distances. L'importance des descriptives pour une variable prédite est calculée par le rapport du taux d'erreur ou l'erreur de la somme des carrés du modèle avec la valeur indépendante supprimée du modèle vers le taux d'erreur ou l'erreur de la somme des carrés pour le modèle entier. L'importance normalisée est calculée par nouvelle pondération des valeurs d'importance des descriptives de sorte que leur somme soit égale à 1.

Prévisions pour cible d'échelle. Lorsqu'une cible d'échelle est spécifiée sur l'onglet Variables, elle détermine si la valeur prévue est calculée à partir de la valeur moyenne ou la médiane des voisins les plus proches ou non.

Descriptives

Figure 20-5
Onglet Descriptives de l'analyse du voisin le plus proche



Cet onglet Descriptives vous permet de demander et de spécifier des options pour la sélection des descriptives lorsqu'une cible est spécifiée dans l'onglet Variables. Par défaut, toutes les descriptives sont prises en compte pour la sélection de descriptives, mais vous pouvez également sélectionner un sous-ensemble de descriptives à introduire de force dans le modèle.

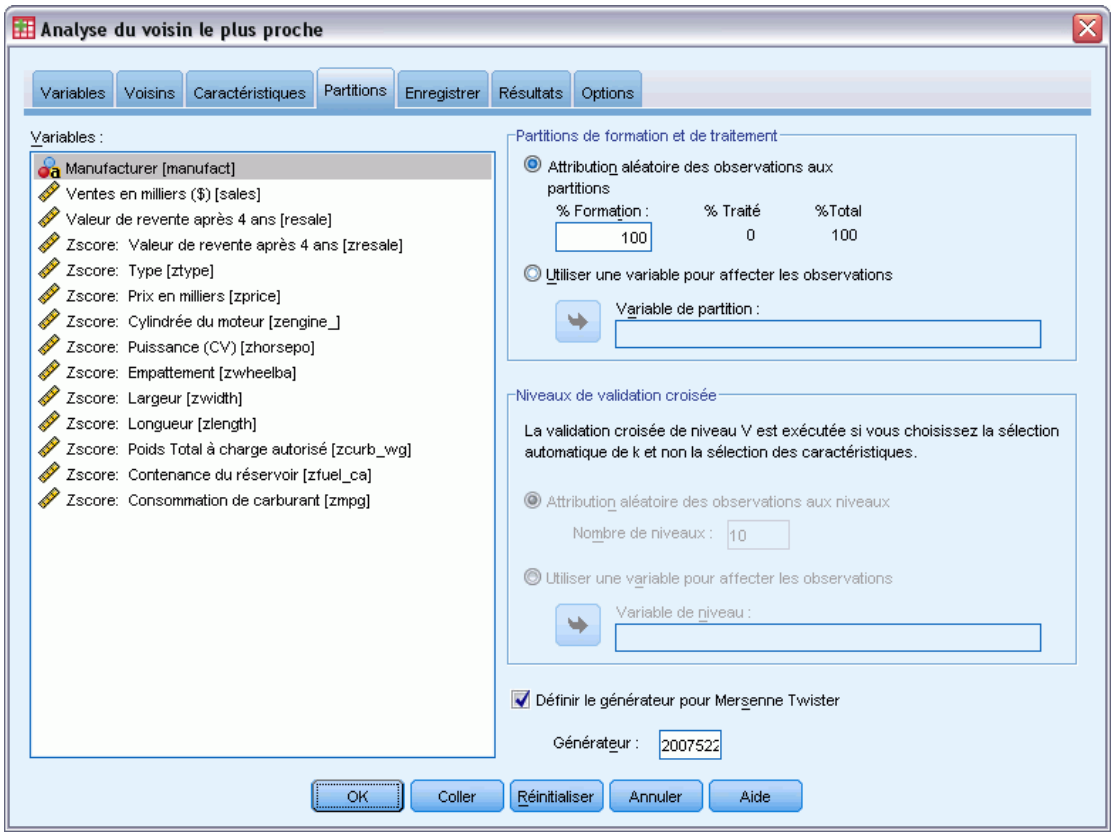
Critère d'arrêt. À chaque étape, la descriptive dont l'addition au modèle entraîne l'erreur la plus faible (calculée comme le taux d'erreur pour une cible qualitative et l'erreur de la somme des carrés pour une cible d'échelle) est prise en compte afin d'être incluse dans l'ensemble de modèle. La sélection ascendante se poursuit jusqu'à la rencontre de la condition spécifiée.

- **Nombre de descriptives spécifié.** L'algorithme ajoute un nombre fixe de descriptives en plus de celles introduites de force dans le modèle. Spécifiez un nombre entier positif. La diminution des valeurs du nombre à sélectionner produit un modèle plus réduit, au risque d'un manque de descriptives importantes. L'augmentation des valeurs du nombre à sélectionner capturera toutes les descriptives importantes, au risque d'ajouter des descriptives qui en réalité alimentent l'erreur du modèle.
- **Changement minimal dans le Ratio d'erreur absolue.** L'algorithme prend fin lorsque le changement dans le ratio d'erreur absolue indique que le modèle ne peut pas être davantage amélioré par l'ajout de nouvelles descriptives. Indiquez un nombre positif. La diminution

des valeurs pour le changement minimal aura tendance à inclure davantage de descriptives, au risque d'en inclure certaines qui n'apportent pas beaucoup de valeur au modèle. L'augmentation de la valeur du changement minimal aura tendance à exclure davantage de descriptives, au risque de perdre des descriptives importantes pour le modèle. La valeur "optimale" du changement minimal dépendra de vos données et de l'application. Reportez-vous au Journal d'erreur de sélection des descriptives pour pouvoir déterminer quelles sont les descriptives les plus importantes. Pour plus d'informations, reportez-vous à la section Journal d'erreur de sélection des descriptives sur p. 149.

Partitions

Figure 20-6
Onglet Partitions de l'analyse du voisin le plus proche



L'onglet Partitions vous permet de diviser l'ensemble de données en un ensemble d'apprentissage et un ensemble traité, et lorsque cela s'applique, il vous permet d'affecter des observations aux niveaux de validation croisée.

Partition d'apprentissage et partition traitée Ce groupe indique la méthode de partitionnement de l'ensemble de données actif en échantillons d'apprentissage et traité. L'**échantillon d'apprentissage** comprend les enregistrements de données utilisés pour former le modèle Voisin le plus proche. Un certain pourcentage d'observations contenues dans l'ensemble de données doit être affecté à l'échantillon d'apprentissage pour l'obtention d'un modèle. L'**échantillon traité**

est un ensemble indépendant d'enregistrements de données utilisé pour évaluer le modèle final ; l'erreur pour l'échantillon traité donne une estimation « honnête » de la capacité de prévision du modèle parce que les observations traitées n'ont pas été utilisées pour construire le modèle.

- **Affecter aléatoirement des observations aux partitions.** Spécifier le pourcentage d'observations à affecter à l'échantillon d'apprentissage. Le reste est affecté à l'échantillon traité.
- **Utiliser une variable pour affecter des observations.** Indiquer une variable numérique qui affecte chaque observation de l'ensemble de données actif à l'échantillon d'apprentissage et traité. Les observations contenant une valeur positive sur la variable sont affectées à l'échantillon d'apprentissage, celles contenant une valeur égale à 0 ou une valeur négative sont affectées à l'échantillon traité. Les observations contenant des valeurs manquantes sont exclues de l'analyse. Les valeurs manquantes spécifiées par l'utilisateur pour la variable de partitionnement sont toujours considérées comme étant valides.

Niveaux de validation croisée. Le Niveau *V* de validation croisée est utilisé pour déterminer le "meilleur" nombre de voisins. Il n'est pas disponible en association avec la sélection de descriptives pour des raisons de performance.

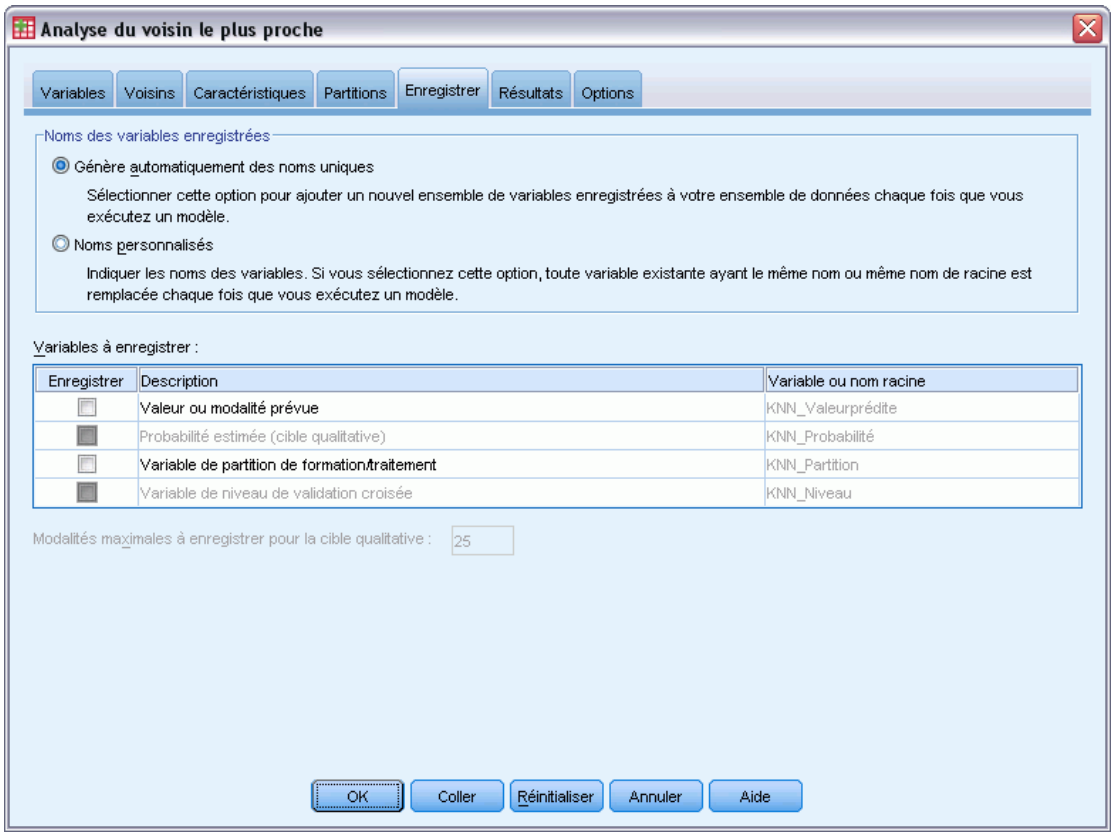
La validation croisée divise l'échantillon en plusieurs sous échantillons, ou niveaux. Les modèles du voisin le plus proche sont générés en excluant à tour de rôle les données de chaque sous-échantillon. Le premier modèle est basé sur toutes les observations à l'exception de celles du premier sous-échantillon, le deuxième modèle est basé sur toutes les observations à l'exception de celles du deuxième sous-échantillon, etc. L'erreur est estimée pour chaque modèle en appliquant le modèle au sous-échantillon exclu lors de la génération du modèle. Le "meilleur" nombre des voisins les plus proches est celui qui produit l'erreur la plus faible sur les sous-échantillons.

- **Affecter aléatoirement des observations aux niveaux.** Spécifier le nombre de niveaux à utiliser pour la validation croisée. Cette procédure affecte aléatoirement des observations aux sous-échantillons, numérotés de 1 à *V*, le nombre de sous-échantillons.
- **Utiliser une variable pour affecter des observations.** Indiquer une variable numérique qui affecte chaque observation de l'ensemble de données actif à un niveau. Cette variable doit être numérique et d'une valeur comprise entre 1 et *V*. Si une valeur manque dans cet intervalle, et que sur toutes les scissions les fichiers scindés sont activés, cela provoqusplitra une erreur.

Définissez un générateur pour le Mersenne Twister . Définir un générateur vous permet de reproduire les analyses. L'utilisation de cette commande revient à définir le Mersenne Twister comme le générateur actif et à spécifier un point de départ fixe dans la boîte de dialogue Générateurs de nombres aléatoires. La différence notoire est que la définition du générateur dans cette boîte de dialogue conserve l'état actuel du générateur de nombres aléatoires et restaure cet état une fois l'analyse terminée.

Enregistrer

Figure 20-7
Onglet Enregistrer l'analyse du voisin le plus proche



Noms des variables enregistrées. Grâce à la génération automatique de nom, vous conservez l'ensemble de votre travail. Les noms personnalisés vous permettent de supprimer/remplacer les résultats d'exécutions précédentes sans supprimer d'abord les variables enregistrées dans l'éditeur de données.

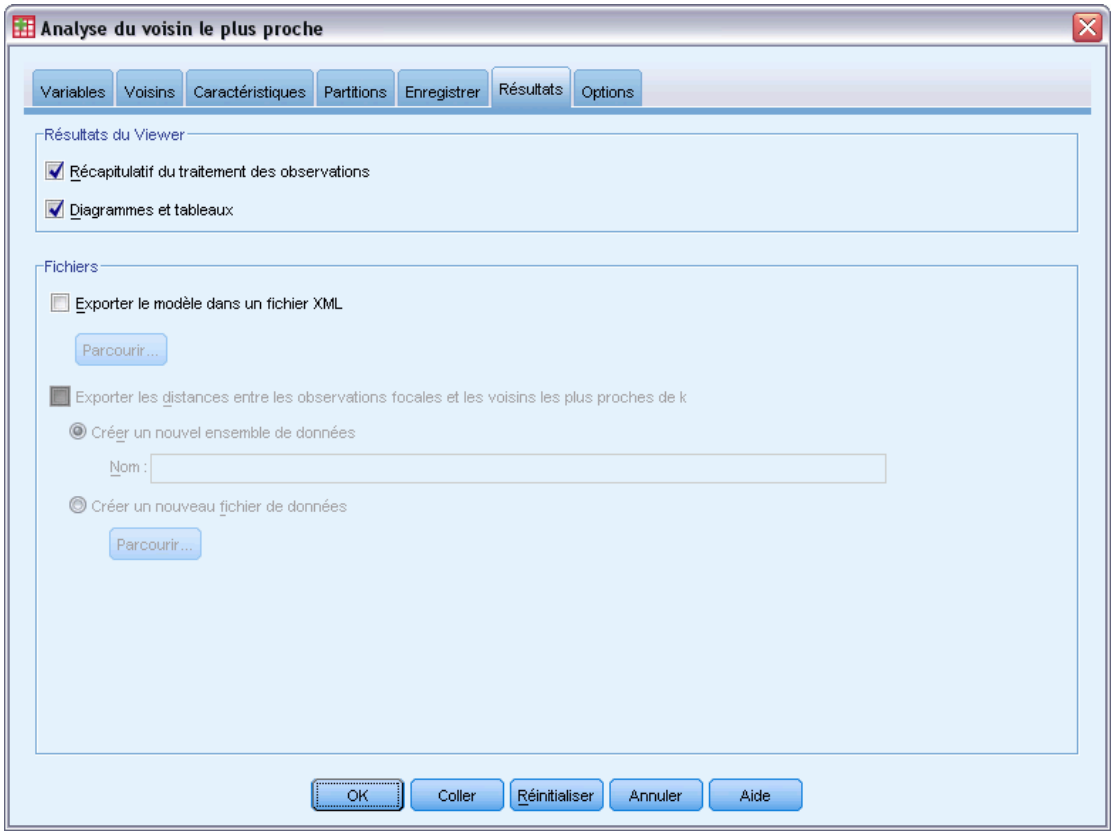
Variables à enregistrer

- **Valeur ou modalité prévue.** Cette option enregistre la valeur prévue pour une cible d'échelle ou la modalité prévue pour une cible qualitative.
- **Probabilité prévue.** Enregistre les probabilités prévues pour une cible qualitative. Une variable distincte est enregistrée pour chacune des *n* premières modalités, *n* étant spécifié dans le contrôle Modalités maximales à enregistrer pour la cible qualitative.

- **Variables de partition d'apprentissage/traitée.** Si des observations sont affectées aléatoirement aux échantillons d'apprentissage et aux échantillons traités dans l'onglet Partitions, cela enregistre la valeur de la partition (d'apprentissage ou traitée) à laquelle l'observation a été affectée.
- **Variable du niveau de validation croisée.** Si des observations ont été affectées aléatoirement à des niveaux de validation croisée dans l'onglet Partitions, cela enregistre la valeur du niveau auquel l'observation a été affectée.

Résultats

Figure 20-8
Onglet Résultats de l'analyse du voisin le plus proche



Résultats du Viewer

- **Récapitulatif du traitement des observations.** Affiche le tableau récapitulatif de traitement des observations, qui récapitule le nombre d'observations incluses et exclues de l'analyse, au total et par échantillon de formation et traité.
- **Diagrammes et tableaux.** Affiche les résultats liés au modèle, y compris les tableaux et les diagrammes. Les tables du modèle incluent les *k* voisins les plus proches et les distances pour observations focales, les variables de classement de réponse qualitative, ainsi qu'un récapitulatif d'erreur. Les résultats graphiques dans l'affichage du modèle incluent un journal d'erreur de sélection, un diagramme d'importance des descriptives, un diagramme d'espace

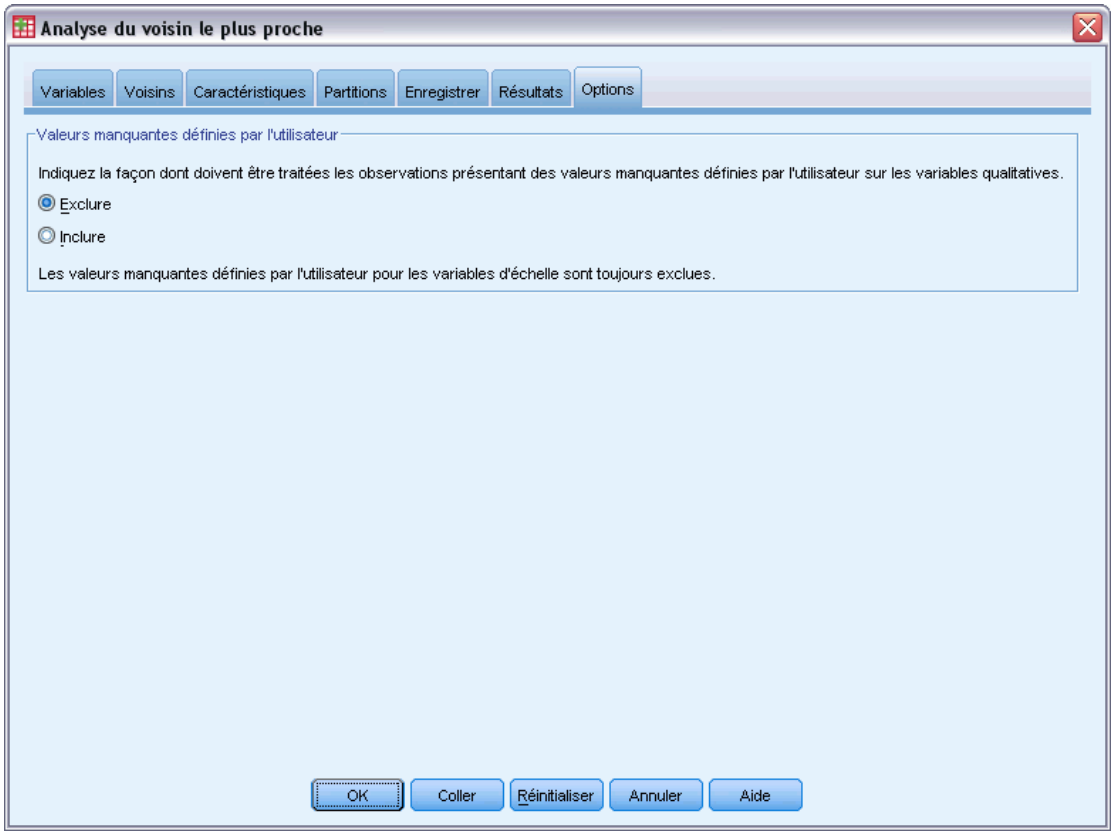
des descriptives, un diagramme des pairs et une carte des quadrants. Pour plus d'informations, reportez-vous à la section Vue du modèle sur p. 141.

Fichiers

- **Exporter le modèle vers un fichier XML.** SmartScore et IBM® SPSS® Statistics Server (produit séparé) peuvent utiliser ce fichier de modèle pour appliquer les informations du modèle à d'autres fichiers de données à des fins d'analyse. Cette option n'est pas disponible si des fichiers scindés ont été définis.
- **Exporter les distances entre les observations focales et les k voisins les plus proches.** Pour chaque observation focale, une variable distincte est créée pour chacun des *k* voisins les plus proches des observations focales (à partir de l'échantillon d'apprentissage et les *k* distances les plus proches correspondantes.

Options

Figure 20-9
Onglet Options de l'analyse du voisin le plus proche

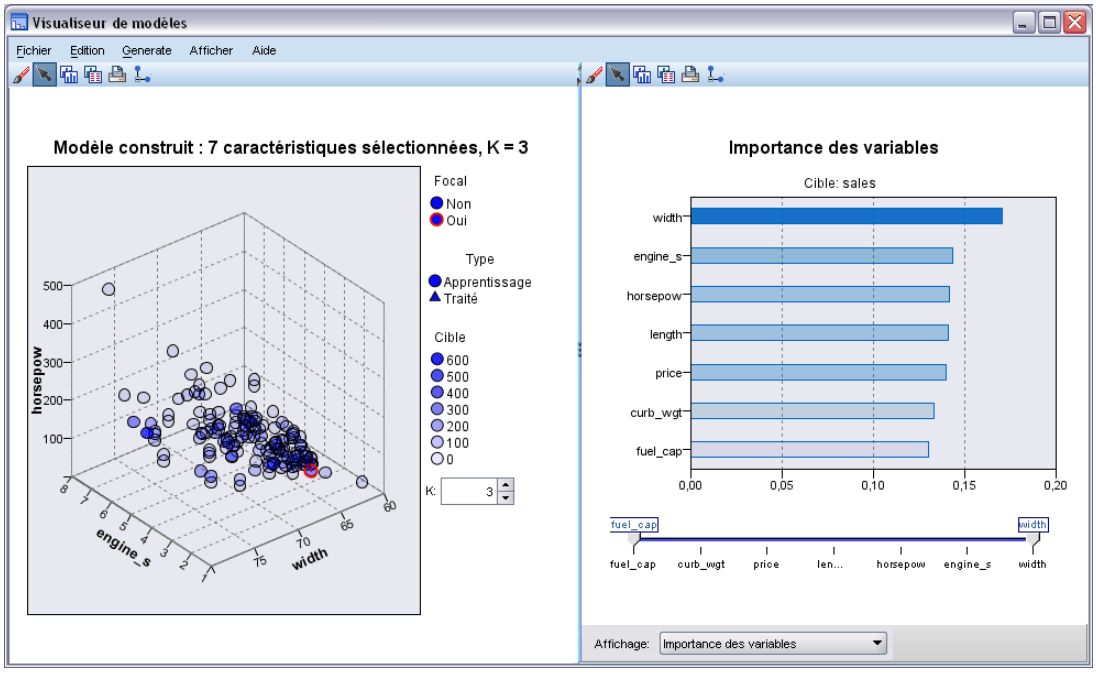


Valeurs manquantes spécifiées. Les variables qualitatives doivent avoir des valeurs valides pour qu'une observation puisse être incluse dans l'analyse. Ces commandes vous permettent d'indiquer si les valeurs manquantes spécifiées sont considérées comme valides parmi les variables qualtiatives.

Les valeurs manquantes par défaut et les valeurs manquantes pour les variables d'échelle sont toujours considérées comme non valides.

Vue du modèle

Figure 20-10
Vue du modèle de l'analyse du voisin le plus proche



Lorsque vous sélectionnez Diagrammes et tableaux dans l'onglet Résultats, la procédure produit un objet de Voisin le plus proche dans le Viewer. En activant cet objet par un double-clic, vous obtenez une vue interactive du modèle. Le modèle présente une fenêtre à double panels :

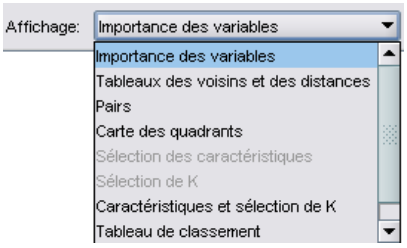
- Le premier affiche une présentation du modèle, appelée vue principale.
- Le second affiche un des deux types de vues :

  Une vue de modèle auxiliaire affiche davantage d'informations sur le modèle, mais n'est pas focalisée sur le modèle lui-même.

  Un vue liée est un affichage montrant les détails d'une descriptive du modèle lorsque l'utilisateur fait défiler une partie de la vue principale.

Par défaut, le premier panel affiche l'espace des descriptives et le second le diagramme d'importance de variable. Si ce dernier n'est pas disponible, c'est-à-dire lorsque Pondérer les descriptives par importance n'a pas été sélectionné dans l'onglet Descriptives, la première vue disponible dans la la vue déroulante est affichée.

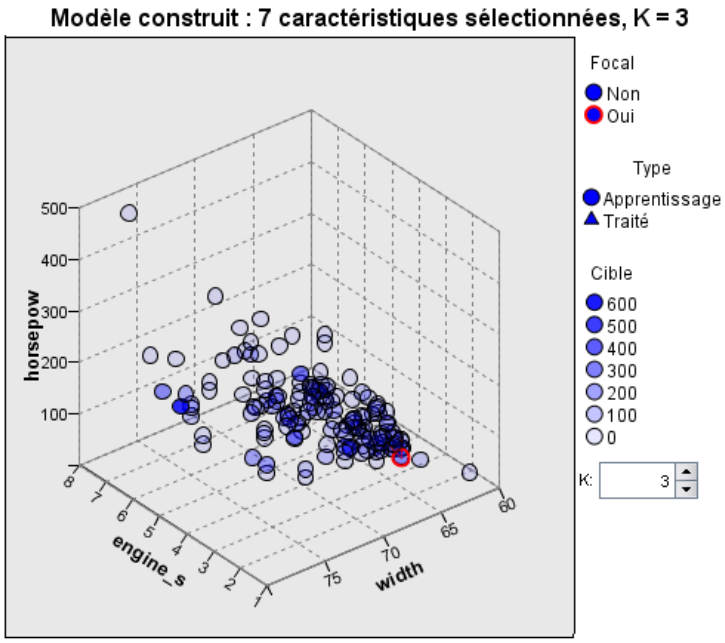
Figure 20-11
Vue déroulante de l'analyse du voisin le plus proche



Lorsqu'une vue n'a aucune information disponible, son élément texte dans la vue déroulante est désactivé.

Espace des descriptives

Figure 20-12
Espace des descriptives



Le diagramme d'espace des descriptives est un diagramme interactif de l'espace des descriptives (ou un sous-espace, s'il existe plus de 3 descriptives). Chaque axe représente une descriptive du modèle, et l'emplacement des points dans le diagramme montre la valeur de ces descriptives pour des observations dans les partitions de formation et traitée.

Clés. En plus des valeurs de descriptives, les points du diagramme contiennent d'autres informations.

- La forme indique la partition à laquelle appartient un point, Formation ou Traité.

- La couleur/ombrage d'un point indique la valeur de la cible pour cette observation, avec les valeurs de couleur distinctes correspondant aux modalités d'une cible qualitative, et les ombres indiquant l'intervalle des valeurs d'une cible continue. La valeur indiquée pour la partition de formation est la valeur observée. Pour la partition traitée, il s'agit de la valeur prévue. Si aucune cible n'est spécifiée, la clé ne s'affiche pas.
- Les contours plus épais indiquent que l'observation est focale. Les observations focales sont affichées avec un lien vers les *k* voisins les plus proches.

Commandes et intéractivité. Un certain nombre de commandes dans le graphique vous permettent d'explorer l'espace des descriptives.

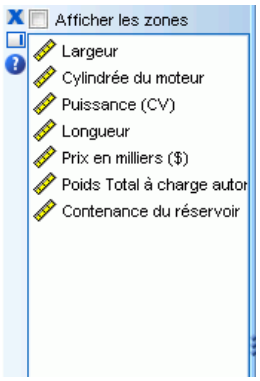
- Vous pouvez choisir quel sous-ensemble de descriptives vous souhaitez afficher dans le diagramme et modifier les descriptives à représenter dans les dimensions.
- Les "Observations focales" sont tout simplement des points sélectionnés dans le diagramme de l'espace des descriptives. Si vous avez spécifié une variable d'observation focale, les points représentant les observations focales seront sélectionnées dès le début. Cependant, tous les points peuvent devenir temporairement une observation focale si vous les sélectionnez. Les commandes "habituelles" pour la sélection des points sont appliquées. Cliquer sur un point permet de sélectionner ce point et de desélectionner tous les autres. Cliquer sur un point avec la touche Ctrl enfoncée ajoute ce point à l'ensemble des points sélectionnés. Les vues liées, tels que le diagramme des pairs, sont automatiquement mis à jour en fonction des observations sélectionnées dans l'espace des descriptives.
- Vous pouvez modifier le nombre de voisins les plus proches (*k*) à afficher pour les observations focales.
- Positionner le curseur sur un point du diagramme affiche une note d'aide avec la valeur de l'étiquette d'observation, ou le nombre d'observations si les étiquettes d'observation ne sont pas définies, et les valeurs des cibles observées et prévues.
- Un bouton "Réinitialiser" vous permet de revenir à l'Espace des descriptives à son état d'origine.

Ajout et suppression des champs et des variables

Vous pouvez ajouter de nouveaux champs ou de nouvelles variables à l'espace des descriptives, ou supprimer ceux qui y sont déjà affichés.

Palette des variables

Figure 20-13
Palette des variables



La palette des variables doit être affichée afin de pouvoir ajouter ou supprimer des variables. Pour faire apparaître la palette des variables, le Viewer de modèlesdoit être en mode Edition et une observation doit être sélectionnée dans l'espace des descriptives.

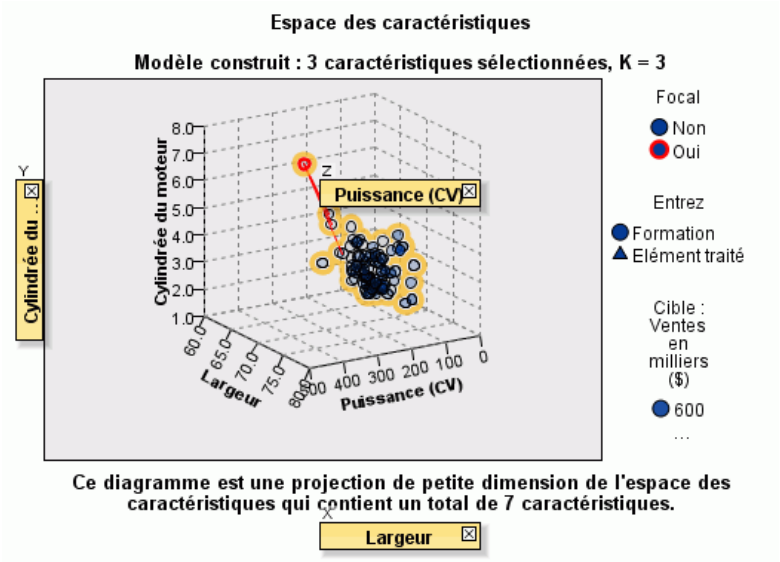
- Pour mettre le Viewer de modèle en mode Edition, sélectionnez à partir des menus :
  Affichage > Mode Edition
- Une fois le mode Edition sélectionné, cliquez sur l'une des observations de l'espace des descriptives.
- Pour afficher la palette des variables, dans les menus choisissez :
  Affichage > Palettes > Variables

  La palette des variables répertorie toutes les variables de l'espace des descriptives. L'icone en regard du nom de la variable indique le niveau de mesure de celle-ci.
- Vous pouvez modifier le niveau de mesure d'une variable de façon temporaire. Pour cela, cliquez sur la variable avec le bouton droit de la souris dans la palette des variables et choisissez une option.

Zones des variables

Les variables sont ajoutées à des zones dans l'espace des descriptives. Pour afficher les zones, faites glisser une variable depuis la palette des variables ou sélectionnez Afficher les zones.

Figure 20-14
Zones des variables



L'espace des descriptives comprend des zones pour les axes *x*, *y*, et *z*.

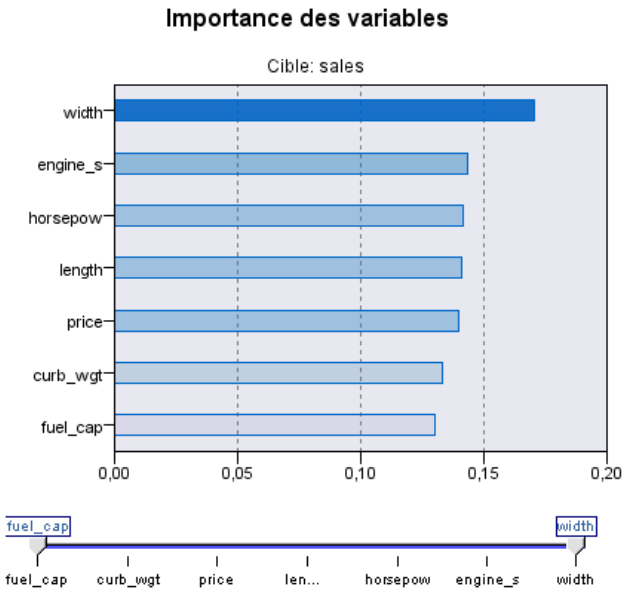
Déplacement des variables vers les zones

Voici quelques règles et conseils généraux permettant de déplacer les variables vers les zones :

- Pour déplacer une variable vers une zone, cliquez dessus et faites-la glisser de la palette des variables à la zone. Si vous sélectionnez Afficher les zones, vous pouvez cliquer avec le bouton droit sur une zone puis choisir la variable à ajouter à la zone.
- Si vous faites glisser une variable depuis la palette variables vers une zone déjà occupée par une autre variable, l'ancienne variable est remplacée par la nouvelle.
- Si vous faites glisser une variable depuis une zone vers une autre zone déjà occupée par une autre variable, les variables échangent leurs positions.
- En cliquant sur le signe X d'une zone, vous supprimez la variable située dans cette zone.
- Si la visualisation comporte plusieurs éléments graphiques, chaque élément peut posséder ses propres zones de variables associées. Sélectionnez d'abord l'élément graphique souhaité.

Importance des variables

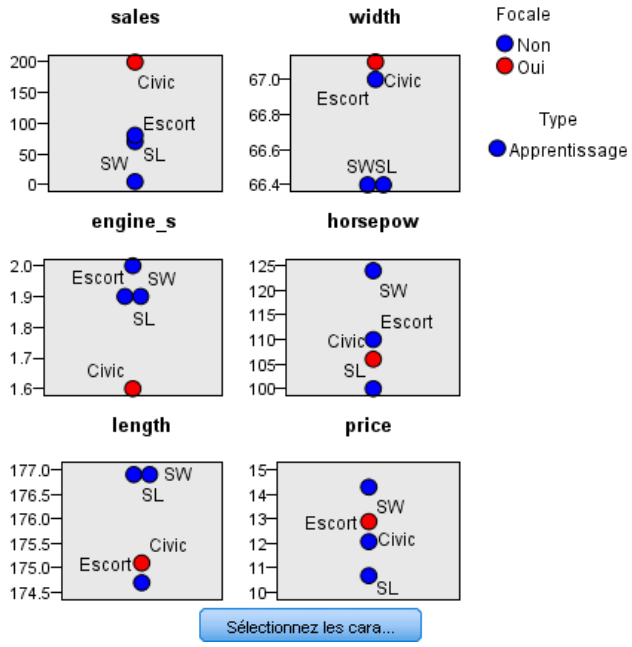
Figure 20-15
Importance des variables



Généralement, vous souhaitez concentrer vos efforts de modélisation sur les variables les plus importantes et vous envisagez d'exclure et d'ignorer les moins importantes. Le diagramme d'importance des variables peut vous y aider en indiquant l'importance relative de chaque variable en estimant le modèle. Etant donné que les valeurs sont relatives, la somme des valeurs pour l'ensemble des variables affichée est 1.0. L'importance des variables n'a aucun rapport avec la précision du modèle. Elle est juste rattachée à l'importance de chaque variable pour la réalisation d'une prévision, peu importe si cette dernière est précise ou non.

Pairs

Figure 20-16
Diagramme des pairs



Ce diagramme affiche les observations focales et leurs *k* voisins les plus proches sur chaque descriptive et sur la cible. Il est disponible si une observation focale est sélectionnée dans l'espace des descriptives.

Comportement de lien. Le diagramme des pairs est relié à l'espace des descriptives de deux manières différentes.

- Les observations sélectionnées (focales) dans l'espace des descriptives sont affichées dans le diagramme des pairs, ainsi que leurs *k* voisins les plus proches.
- La valeur de *k* sélectionnée dans l'espace des descriptives est utilisée dans le diagramme des pairs.

Distances du voisin le plus proche

Figure 20-17
Distances du voisin le plus proche

Observation focale	Voisins les plus proches			Distances les plus proches		
	1	2	3	1	2	3
Civic	SL	Escort	SW	0.053	0.059	0.064

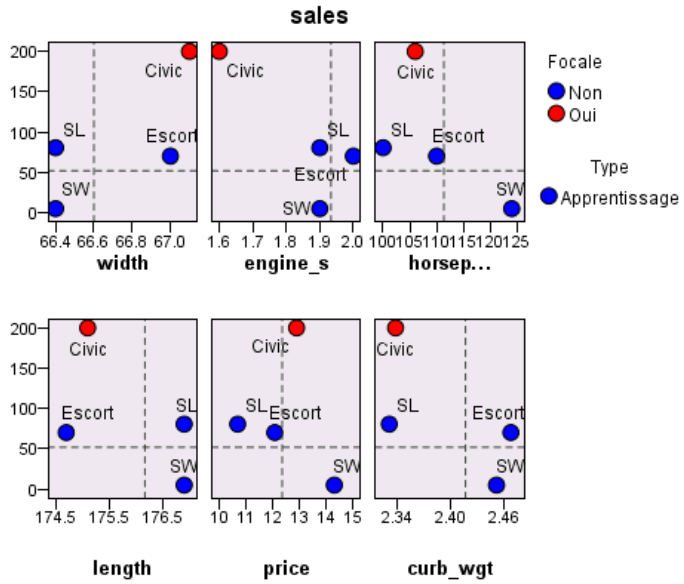
Ce tableau affiche les *k* voisins les plus proches et les distances pour les observations focales uniquement. Il est disponible si un identificateur d'observations focale est spécifié dans l'onglet Variable, et il n'affiche que les observations focales identifiées par cette variable.

Chaque ligne de :

- La colonne Observation focale contient la valeur de la variable d'étiquetage des observations pour l'observation focale. Si les étiquettes d'observations ne sont pas définies, cette colonne contient le nombre d'observations de l'observation focale.
- La *i*ème colonne sous le groupe des voisins les plus proches contient la valeur de la variable d'étiquetage d'observation pour le *i*ème voisin le plus proche de l'observation focale. Si les étiquettes d'observations ne sont pas définies, cette colonne contient le nombre d'observation du *i*ème voisin le plus proche de l'observation focale.
- La *i*ème colonne sous le groupe Distances les plus proches contient la distance du *i*ème voisin le plus proche de l'observation focale.

Carte des quadrants

Figure 20-18
Carte des quadrants

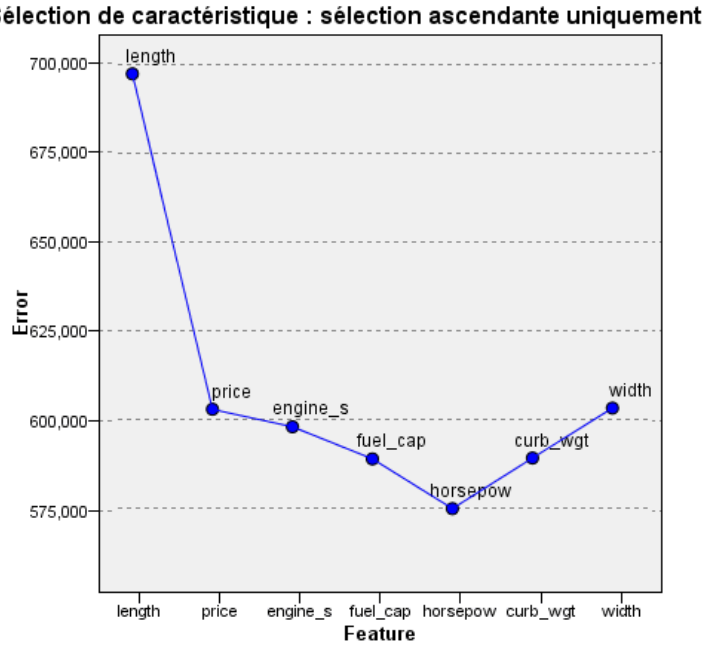


Ce diagramme affiche les observations focales et leur *k* voisins les plus proches sur un diagramme de dispersion (ou nuage de points, selon le niveau de mesure de la cible) avec la cible sur l'axe *y* et une descriptive d'échelle sur l'axe *x*, affichés sous forme de panel par descriptive. Il est disponible si une cible existe et si une observation focale est sélectionnée dans l'espace des descriptives.

- Les lignes de référence sont tracées pour des variables continues, aux moyennes variables dans la partition de formation.

Journal d'erreur de sélection des descriptives

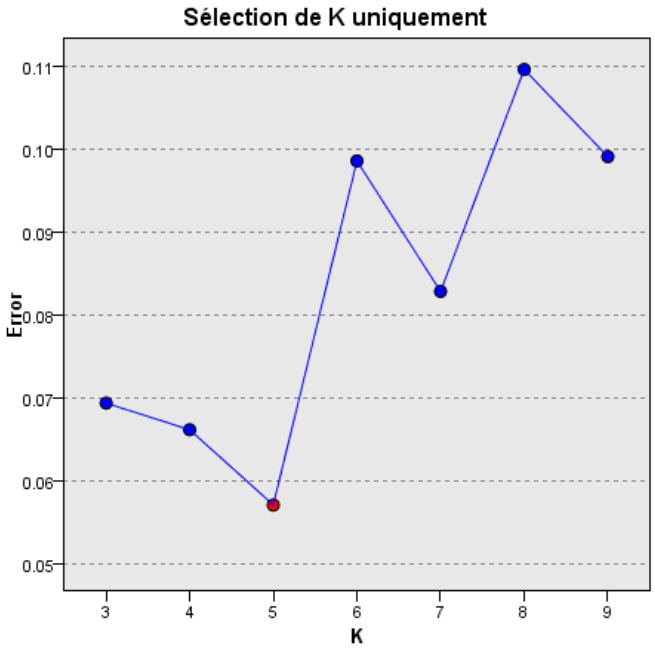
Figure 20-19
Sélection des caractérisques



Les points du diagramme affichent l'erreur (le ratio du taux d'erreur ou l'erreur de la somme des carrés, selon le niveau de mesure de la cible) sur l'axe *y* pour le modèle avec la descriptive sur l'axe *x* (plus toutes les descriptives à gauche sur l'axe *x*). Ce diagramme est disponible si une cible existe et si la sélection des descriptives est activée.

Journal d'erreur de la sélection de k

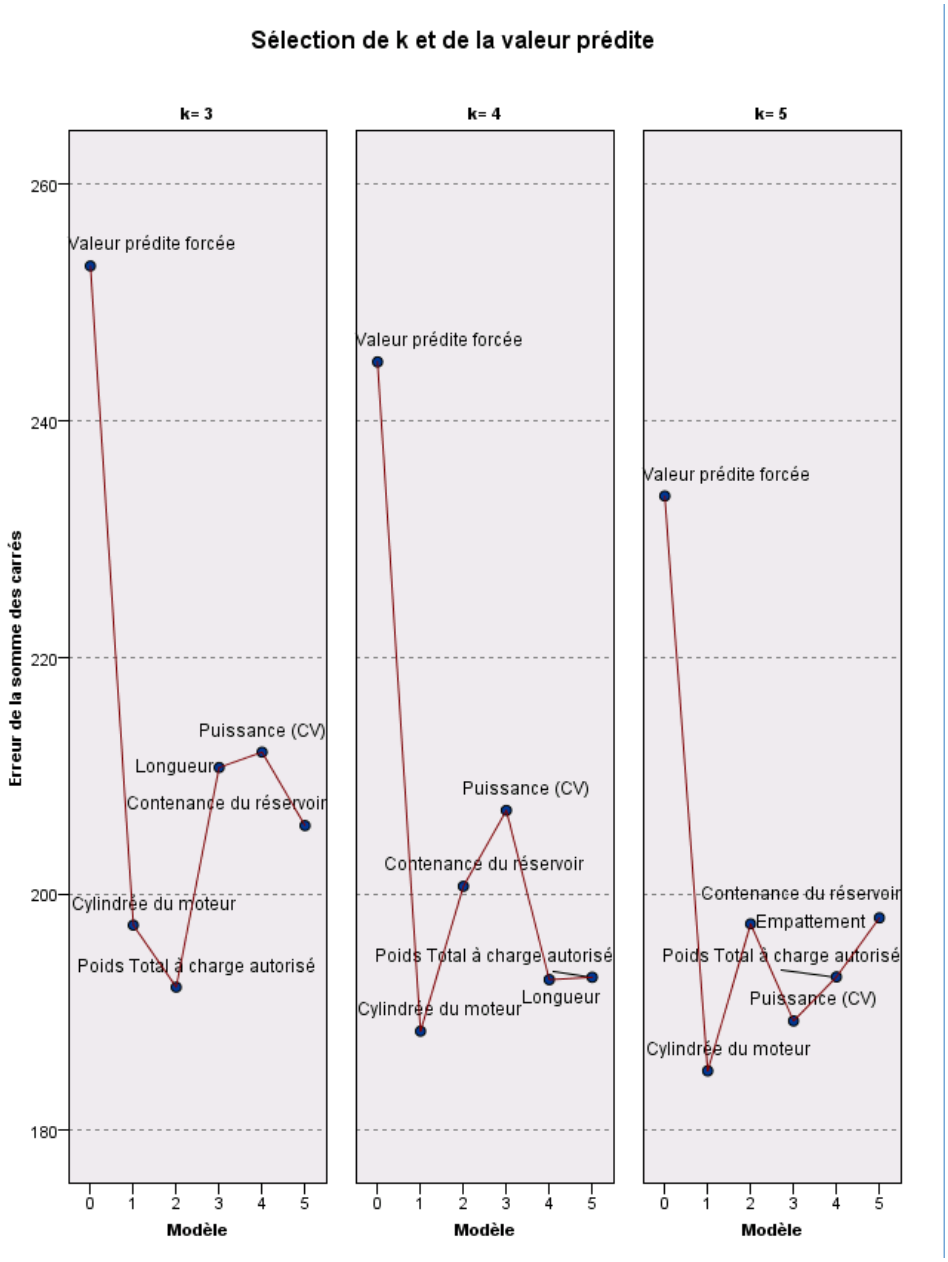
Figure 20-20
Sélection de k



Les points du diagramme affichent l'erreur (le ratio du taux d'erreur ou l'erreur de la somme des carrés, selon le niveau de mesure de la cible) sur l'axe *y* pour le modèle avec le nombre de voisins les plus proches (*k*) sur l'axe *x*. Ce diagramme est disponible si une cible existe et si la sélection de *k* est activée.

Journal d'erreur de sélection de k et des descriptives

Figure 20-21
Sélection de k et des descriptives



Ce sont des diagrammes de sélection des descriptives (reportez-vous à Journal d'erreur de sélection des descriptives sur p. 149), affiché par *k*. Ce diagramme est disponible si une cible existe et si les sélections de *k* et des descriptives sont toutes les deux activées.

Le tableau de classification

Figure 20-22
Tableau de classification

Partition		Prévisions		
		0	1	Pourcentage correct
Apprentissage	0	111	1	99.11%
	1	7	33	82.50%
	Pourcentage global	77.64%	22.37%	94.74%

Ce tableau affiche par partition la classification croisée des valeurs prévues de la cible observées contre celles prévues. Il est disponible si une cible existe et si elle est qualitative.

- La ligne (Manquante) dans la partition traitée contient des observations traitées contenant des valeurs manquantes sur la cible. Ces observations contribuent à l'échantillon traité : Les valeurs de pourcentage global mais pas les valeurs de pourcentage correct.

Récapitulatif d'erreur

Figure 20-23
Récapitulatif d'erreur

Partition	Sum-of-Squares Error
Apprentissage	622043

Ce tableau est disponible si une variable cible existe. Il affiche l'erreur associée au modèle, la somme des carrés pour une cible cible continue et le taux d'erreur (100% − pourcentage général correct) pour une cible qualitative.

Chapitre
21

Analyse discriminante

L'analyse discriminante crée un modèle de prévision de groupe d'affectation. Le modèle est composé d'une fonction discriminante (ou, pour plus de deux groupes, un ensemble de fonctions discriminantes) basée sur les combinaisons linéaires des variables explicatives qui donnent la meilleure discrimination entre groupes. Les fonctions sont générées à partir d'un échantillon d'observations pour lesquelles le groupe d'affectation est connu. Les fonctions peuvent alors être appliquées aux nouvelles observations avec des mesures de variables explicatives, mais de groupe d'affectation inconnu.

Remarque : la variable de groupe peut avoir plus de deux valeurs. Les codes de la variable de regroupement doivent cependant être des nombres entiers, et vous devez spécifier leur valeur minimale et maximale. Les observations dont les valeurs se situent hors des limites sont exclues de l'analyse.

Exemple : En moyenne, les habitants des pays des zones tempérées consomment plus de calories par jour que ceux des tropiques, et une plus grande proportion de ces habitants vit en ville. Un chercheur veut combiner ces informations en une fonction pour déterminer comment un individu peut être différencié selon les deux groupes de pays. Le chercheur pense que la taille de la population et des informations économiques peuvent aussi être importantes. L'analyse discriminante vous permet d'estimer les coefficients de la fonction discriminante linéaire, qui ressemble à la partie droite d'une équation de régression linéaire multiple. Ainsi, en utilisant les coefficients *a*, *b*, *c* et *d*, la fonction est :

D = a * climat + b * urbain + c * population + d * Produit National Brut par habitant

Si ces variables sont utiles pour établir la différence entre les deux zones climatiques, les valeurs de *D* seront différentes pour les pays tempérés et les pays tropicaux. Si vous utilisez une méthode de sélection des variables pas à pas, vous pouvez découvrir que vous n'avez pas forcément besoin d'inclure les quatre variables dans la fonction.

Statistiques : Pour chaque variable, on a les éléments suivants : moyenne, écarts-types, ANOVA à un facteur. Pour chaque analyse : *Test* de Box, matrice de corrélation intra-classe, matrice de covariance intra-classe, matrice de covariance de chaque classe, matrice de covariance totale. Pour chaque fonction discriminante canonique : valeur propre, pourcentage de la variance, corrélation canonique, lambda de Wilks, Khi-deux. Pour chaque pas : probabilités à priori, coefficients de fonction de Fisher, coefficients de fonction non standardisés, lambda de Wilks pour chaque fonction canonique.

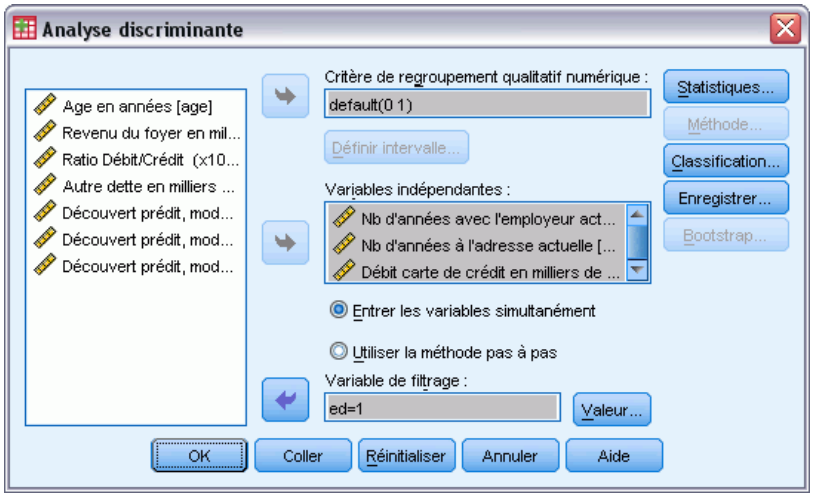
Données. La variable de regroupement doit avoir un nombre limité de modalités distinctes, codifiées sous forme de nombres entiers. Les variables indépendantes nominales doivent être recodées en variables muettes ou de contraste.

Hypothèses : Les observations doivent être indépendantes. Les variables prédites doivent avoir une distribution gaussienne multivariée, et les matrices de variance-covariance intra-groupes doivent être égales entre groupes. On part de l'hypothèse que les groupes d'affectation sont mutuellement exclusifs (c'est-à-dire qu'aucune observation n'est affectée à plus d'un groupe) et collectivement exhaustifs (c'est-à-dire que toutes les observations sont affectées à un groupe). La procédure est la plus efficace lorsque l'affectation à un groupe est une variable réellement qualitative. Si l'affectation à un groupe est basée sur les valeurs d'une variable continue (par exemple, QI élevé contre QI bas), vous devez envisager d'utiliser la régression linéaire pour exploiter les informations plus riches données par la variable continue elle-même.

Pour obtenir une analyse discriminante

▶ A partir des menus, sélectionnez :
Analyse > Classification > Analyse discriminante

Figure 21-1
Boîte de dialogue Analyse discriminante : Classement



▶ Sélectionnez une variable de regroupement à valeur entière et cliquez sur Définir intervalle pour spécifier les modalités à considérer.

▶ Sélectionnez les variables prédites (ou explicatives). (Si votre variable de regroupement n'a pas de valeurs entières, la procédure de recodification automatique du menu Transformer permettra d'en créer un avec des valeurs entières.)

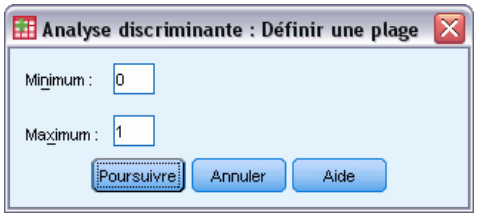
▶ Sélectionnez la méthode de saisie des variables indépendantes.

- **Entrer les variables simultanément** Entre simultanément toutes les variables indépendantes qui satisfont aux critères de tolérance.
- **Utiliser la méthode pas à pas :** Utilise la méthode pas à pas pour contrôler l'entrée et la suppression de variables.

▶ Vous pouvez également sélectionner les observations avec une variable de sélection.

Définition d'intervalles pour l'analyse discriminante

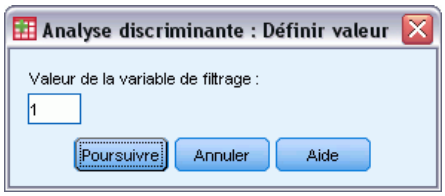
Figure 21-2
Boîte de dialogue Analyse discriminante : Définir intervalle



Spécifiez la valeur minimum et maximum de la variable de regroupement pour l'analyse. Les observations avec des valeurs hors de cet intervalle ne sont pas utilisées dans l'analyse discriminante mais elles sont classées dans un des groupes existants en fonction des résultats de l'analyse. Les valeurs minimum et maximum doivent être des entiers.

Sélection des observations pour l'analyse discriminante

Figure 21-3
Boîte de dialogue Analyse discriminante : Enregistrer



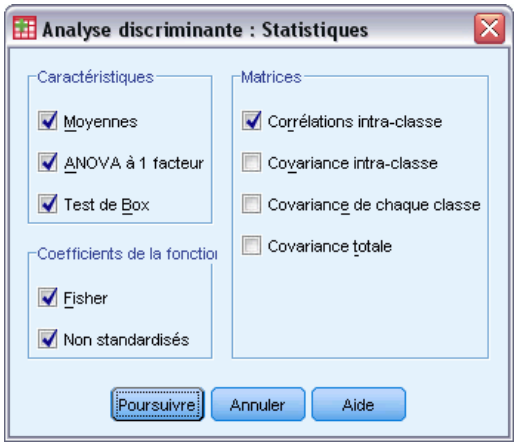
Pour sélectionner les observations pour votre analyse :

- Dans la boîte de dialogue Analyse discriminante, sélectionnez une variable de sélection.
- Cliquez sur Valeur pour entrer un entier comme valeur de sélection.

Seules les observations avec la valeur spécifiée pour la variable de sélection sont utilisées pour dériver les fonctions discriminantes. Les résultats des statistiques et de classification sont générés pour les observations sélectionnées et celles qui ne le sont pas. Ce processus fournit une méthode de classification des nouvelles observations reposant sur des données existantes ou de partitionnement de vos données dans un sous-ensemble de test ou de formation en vue d'effectuer une validation sur le modèle créé.

Statistiques de l'analyse discriminante

Figure 21-4
Boîte de dialogue Analyse discriminante: Statistiques



Descriptives. Les options disponibles sont moyennes (y compris écarts-types), ANOVA à 1 facteur et Test *M* de Box.

- **Moyennes.** Affiche le total et la moyenne de chaque groupe ainsi que l'écart-type des variables explicatives.
- **ANOVA à 1 facteur.** Effectue pour chacune des variables indépendantes une analyse de variance à 1 facteur pour tester l'égalité des moyennes de groupe.
- **M de Box.** Test d'égalité des matrices de covariance des classes. Pour les échantillons de taille suffisamment importante, une valeur p non significative indique qu'il n'est pas démontré que les matrices diffèrent. Ce test est sensible aux déviations par rapport à la distribution gaussienne multivariée.

Coefficients de la fonction : Les options disponibles sont les coefficients de la classification de Fisher et les coefficients non standardisés.

- **Fisher.** Affiche les coefficients de la fonction de classification de Fisher qui peuvent être directement utilisés pour la classification. Un groupe séparé de coefficients de fonctions de classification est obtenu pour chaque groupe et une observation est affectée au groupe qui a le plus grand score discriminant (valeur de fonction de classification).
- **Non standardisés.** Affiche les coefficients non normalisés de la fonction discriminante.

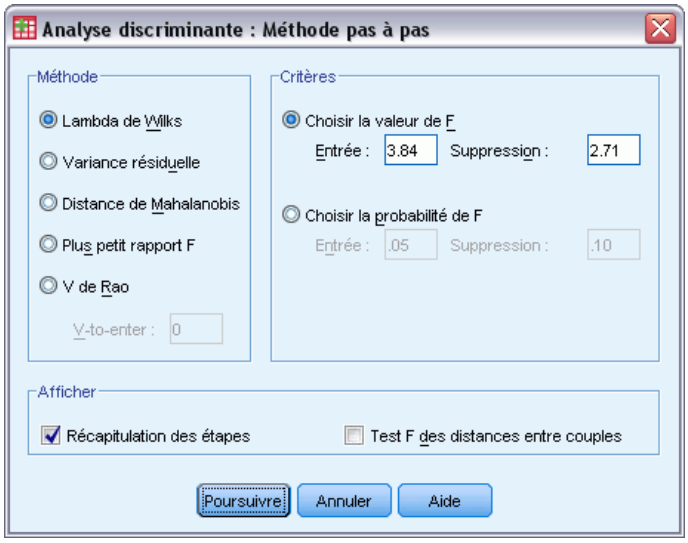
Matrices : Les matrices de coefficients pour variables indépendantes disponibles sont la matrice de corrélation intra-classe, la matrice de covariance intra-classe, la matrice de covariance de chaque classe et la matrice de covariance totale.

- **Corrélations intra-classe.** Affiche une matrice de corrélations intra-classes globale, en calculant la moyenne des matrices de covariance distinctes pour tous les groupes avant de calculer les corrélations.
- **Covariance intra-classe.** Affiche une matrice de covariances intra-classes globale, qui peut différer de la matrice de covariance totale. Cette matrice est obtenue en calculant la moyenne des matrices de covariances distinctes de tous les groupes.

- **Covariance de chaque classe.** Affiche des matrices de covariances distinctes pour chaque groupe.
- **Covariance totale.** Affiche la matrice de covariance de toutes les observations comme si elles provenaient d'un seul échantillon.

Méthode pas à pas de l'analyse discriminante

Figure 21-5
Boîte de dialogue Analyse discriminante : Méthode pas à pas



Méthode : Sélectionnez la statistique à utiliser pour ajouter ou supprimer de nouvelles variables. Les options possibles sont le lambda de Wilks, la variance résiduelle, la distance de Mahalanobis, le plus petit rapport *F* et le *V* de Rao. Avec le *V* de Rao, vous pouvez spécifier l'augmentation minimum de *V* pour entrer une variable.

- **Lambda de Wilks.** Méthode de sélection des variables pour une analyse discriminante pas à pas qui sélectionne les variables à entrer dans l'équation d'après leur capacité à faire baisser le lambda de Wilks. A chaque étape, les variables sont entrées dans l'analyse d'après leur capacité à faire baisser le lambda de Wilks.
- **Variance résiduelle.** A chaque étape, la variable qui minimise la somme des variations résiduelles entre les groupes est saisie.
- **Distance de Mahalanobis.** Mesure de la distance entre les valeurs d'une observation et la moyenne de toutes les observations sur les variables indépendantes. Une distance de Mahalanobis importante identifie une observation qui a des valeurs extrêmes pour des variables indépendantes.
- **Plus petit rapport F.** Méthode de sélection des variables en analyse pas à pas, fondée sur la maximisation d'un rapport F calculé à partir de la distance de Mahalanobis entre des groupes.
- **V de Rao.** Mesure des différences entre des moyennes de groupes. Egalement appelée trace de Lawley-Hotelling. A chaque étape, la variable qui maximise l'augmentation du V de RAO est entrée. Après avoir sélectionné cette option, entrez la valeur minimale que doit avoir une variable pour entrer dans l'analyse.

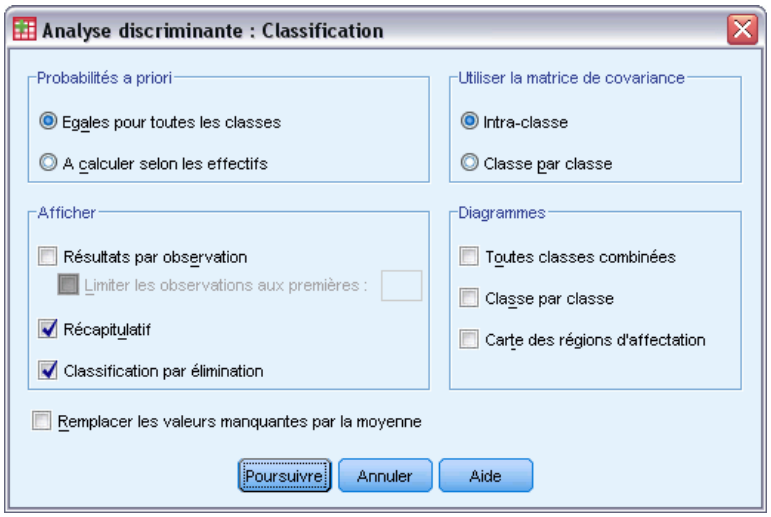
Critères : Les options disponibles sont Choisir la valeur de F et Choisir la probabilité de F. Entrez des valeurs pour ajouter et supprimer des variables.

- **Choisir la valeur de F.** Une variable est introduite dans un modèle si sa valeur F est supérieure à la valeur Entrée et elle est éliminée si la valeur F est inférieure à la valeur Elimination. La valeur Entrée doit être supérieure à la valeur Elimination et toutes deux doivent être positives. Pour introduire davantage de variables dans le modèle, réduisez la valeur du champ Entrée. Pour éliminer davantage de variables dans le modèle, augmentez la valeur du champ Elimination.
- **Choisir la probabilité de F.** Une variable est entrée dans le modèle si le seuil de signification de la valeur F est inférieur à la valeur Entrée ; la variable est éliminée si ce seuil est supérieur à la valeur Elimination. La valeur Entrée doit être inférieure à la valeur Elimination et toutes deux doivent être positives. Pour introduire davantage de variables dans le modèle, diminuez la valeur Entrée. Pour éliminer davantage de variables du modèle, réduisez la valeur Elimination.

Afficher : L'option Récapitulation des étapes affiche les statistiques de toutes les variables après chaque étape. L'option Test F des distances entre couples affiche une matrice de rapports *F* appariés pour chaque paire de groupes.

Analyse discriminante : Classement

Figure 21-6
Boîte de dialogue Classement de l'analyse discriminante



Probabilités à priori : Cette option permet de déterminer si les coefficients de classification sont ajustés pour une connaissance à priori de l'appartenance à une classe.

- **Egales pour toutes les classes.** Des probabilités à priori égales sont supposées pour toutes les classes; ceci n'a aucun incident sur les coefficients.
- **A calculer selon les effectifs.** Les tailles de classe observées dans votre échantillon déterminent les probabilités à priori de la classe d'appartenance. Par exemple, si 50% des observations comprises dans l'analyse appartiennent à la première classe, 25 % à la deuxième, et 25 % à la troisième, les coefficients de classification sont ajustés pour accroître la probabilité d'affectation de la première classe par rapport aux deux autres.

Afficher : Les options d'affichage disponibles sont les résultats par observation, le récapitulatif et la classification par élimination.

- **Résultats par observation.** Les codes du groupe actuel, du groupe prévu, des probabilités a posteriori et des scores discriminants sont affichés pour chaque observation.
- **Récapitulatif.** Nombre d'observations correctement et incorrectement affectées à chacune des classes sur la base de l'analyse discriminante. Parfois appelés « matrice de confusion ».
- **Classification par élimination.** Classement de chaque observation de l'analyse par les fonctions dérivées de l'ensemble des observations autres que cette observation. Cette classification est également appelée « méthode U ».

Remplacer les valeurs manquantes par la moyenne : Sélectionnez cette option pour remplacer la valeur manquante d'une variable indépendante par la moyenne de cette variable, mais seulement durant la phase de classification.

Utiliser la matrice d'inertie : Vous pouvez choisir de classer les observations en utilisant une matrice de covariance intra-classe ou une matrice de covariance pour chaque classe.

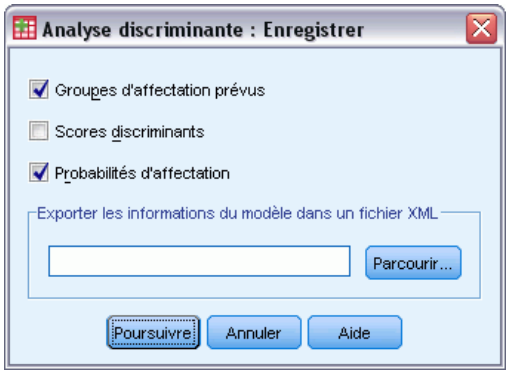
- **Intra-classe.** La matrice de covariances intra-classes globale est utilisée pour classifier les observations.
- **Classe par classe :.** Les matrices de covariances de chaque groupe sont utilisées pour la classification. Comme la classification repose sur les fonctions discriminantes et pas sur les variables d'origine, cette option n'est pas toujours équivalente à la discrimination quadratique.

Diagrammes : Les options de diagramme disponibles sont toutes classes combinées, classe par classe, et carte des régions d'affectation.

- **Toutes classes combinées.** Crée un diagramme de dispersion de tous les groupes, des valeurs des deux premières fonctions discriminantes. S'il n'y a qu'une seule fonction, un histogramme est tracé à la place.
- **Classe par classe.** Crée des diagrammes de dispersion classe par classe pour les deux premières valeurs de fonction discriminante. Lorsqu'il n'y a qu'une seule fonction, des histogrammes sont affichés à la place.
- **Carte des régions d'affectation.** Diagramme des limites servant à classifier les observations en fonction de valeurs de fonction. Les numéros correspondent aux groupes auxquels les observations ont été affectées. La moyenne de chaque groupe est indiquée par un astérisque à l'intérieur de ses limites. La carte n'est pas affichée s'il n'existe qu'une seule fonction discriminante.

Enregistrement de l'analyse discriminante

Figure 21-7
Boîte de dialogue Analyse discriminante : Enregistrer



Vous pouvez ajouter de nouvelles variables à votre fichier de données actif. Les options disponibles sont classe(s) d'affectation (une seule variable), valeurs du facteur discriminant (une variable pour chaque fonction discriminante dans la solution), et probabilités d'affectation à un groupe en fonction des valeurs du facteur discriminant (une variable pour chaque groupe).

Vous pouvez également exporter les informations du modèle vers le fichier spécifié au format XML (PMML). Vous pouvez utiliser ce fichier de modèle pour appliquer les informations du modèle aux autres fichiers de données à des fins d'évaluation.

Fonctionnalités supplémentaires de la commande DISCRIMINANT

Le langage de syntaxe de commande vous permet aussi de :

- effectuer plusieurs analyses discriminantes avec une seule commande et contrôler l'ordre d'entrée des variables (au moyen de la sous-commande `ANALYSIS`).
- spécifier des probabilités à priori pour la classification (au moyen de la sous-commande `PRIORS`).
- afficher les matrices des coordonnées factorielles et les matrices des corrélations structurelles après rotation (au moyen de la sous-commande `ROTATE`).
- limiter le nombre de fonctions discriminantes extraites (au moyen de la sous-commande `FUNCTIONS`).
- restreindre la classification aux observations sélectionnées (ou non sélectionnées) pour l'analyse (au moyen de la sous-commande `SELECT`).
- lire et analyser une matrice de corrélation (au moyen de la sous-commande `MATRIX`).
- créer une matrice de corrélation pour une analyse ultérieure (au moyen de la sous-commande `MATRIX`).

Reportez-vous à la *Référence de syntaxe de commande* pour une information complète concernant la syntaxe.

Chapitre
22

Analyse factorielle

L'analyse factorielle essaie d'identifier des variables sous-jacentes, ou **facteurs**, qui permettent d'expliquer le patron des corrélations à l'intérieur d'un ensemble de variables observées. L'analyse factorielle est souvent utilisée pour réduire un ensemble de données. L'analyse factorielle est souvent utilisée dans la factorisation, en identifiant un petit nombre de facteurs qui expliquent la plupart des variances observées dans le plus grand nombre de variables manifestes. On peut également utiliser l'analyse factorielle pour générer des hypothèses concernant des mécanismes de causalité ou pour afficher des variables pour une analyse ultérieure (par exemple, pour identifier la colinéarité avant une analyse de régression linéaire).

La procédure d'analyse factorielle offre une très grande flexibilité :

- Il existe sept méthodes d'extraction de facteur.
- Il existe cinq méthodes de rotation, dont directe Oblimin et Promax pour les rotations non orthogonales.
- Il existe trois méthodes pour calculer les facteurs, et ces facteurs peuvent être enregistrés en tant que variables pour des analyses ultérieures.

Exemple : Quelle est l'attitude sous-jacente qui pousse les personnes à répondre d'une certaine manière aux questions concernant un sondage politique ? L'examen des corrélations parmi les éléments d'un sondage révèle qu'il y a des recouvrements significatifs parmi divers sous-groupes d'éléments. Les questions sur les impôts ont tendance à être en corrélation, de même que les questions à thèmes militaires, etc. Avec l'analyse factorielle, vous pouvez enquêter sur le nombre de facteurs sous-jacents, et, dans de nombreux cas, vous pouvez identifier le concept représenté par ces facteurs. De plus, vous pouvez calculer les facteurs pour chaque répondant, facteurs que vous pouvez utiliser pour des analyses ultérieures. Par exemple, sur la base des facteurs, vous pouvez développer un modèle logistique de régression pour prévoir le comportement de vote.

Statistiques : Pour chaque variable, on a les éléments suivants : nombre d'observations valides, moyenne et écart-type. Pour chaque analyse factorielle : matrice de corrélation des variables, incluant des seuils de signification, déterminant, inverse ; les matrices des corrélations reconstituées, incluant l'anti-image ; les solutions initiales (qualités de représentation, valeurs propres et pourcentage de variance expliqué) ; mesure d'adéquation d'échantillonnage de Kaiser-Meyer-Olkin et le test de sphéricité de Bartlett ; structure avant rotation, incluant les saturations sur les facteurs, la qualité de représentation, et les valeurs propres; structure après rotation, incluant une matrice de forme après rotation et une matrice de transformation. Pour rotations obliques : type et matrices de structure après rotation ; matrice factorielle de coefficient de facteur et matrice factorielle de facteur de covariance. Diagrammes : Diagramme de valeurs propres et carte factorielle du premier, du deuxième et du troisième facteur.

Données. Les variables doivent être quantitatives au niveau de l'**intervalle** ou du **rapport**. Les données qualitatives (comme la religion ou le pays d'origine) ne conviennent pas pour l'analyse factorielle. Les données pour lesquelles la corrélation de Pearson calculée a un sens conviennent pour l'analyse factorielle.

Hypothèses : Les données doivent posséder une distribution gaussienne bivariée pour chaque paire de variables et les observations doivent être indépendantes. Le modèle d'analyse factorielle spécifie que les variables sont déterminées par des facteurs communs (les facteurs estimés par le modèle) et des facteurs uniques (qui ne sont pas corrélés entre les variables observées); les estimations calculées se basent sur l'hypothèse que tous les facteurs uniques ne sont pas en corrélation entre eux ainsi qu'avec les facteurs communs.

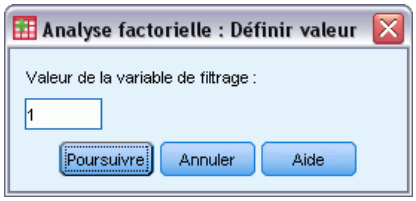
Pour obtenir une analyse factorielle

- A partir des menus, sélectionnez :
  Analyse > Réduction des dimensions > Analyse factorielle
- Sélectionnez les variables pour l'analyse factorielle.

Figure 22-1
Boîte de dialogue Analyse factorielle

Sélection des observations pour l'analyse factorielle

Figure 22-2
Sélectionnez la boîte de dialogue Analyse factorielle



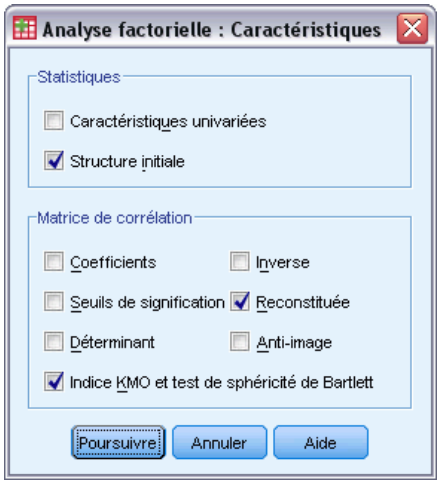
Pour sélectionner les observations pour votre analyse :

- Sélectionnez une variable de sélection.
- Cliquez sur Valeur pour entrer un entier comme valeur de sélection.

  Seules les observations ayant cette valeur pour la variable de sélection sont utilisées dans l'analyse factorielle.

Descriptives d'analyse factorielle

Figure 22-3
Boîte de dialogue Analyse Factorielle : Descriptives...



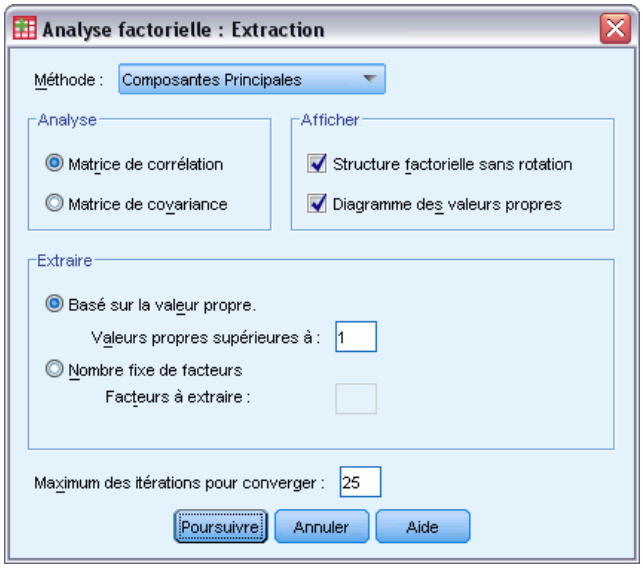
Statistiques : Les Descriptives univariées incluent la moyenne, l'écart-type et le nombre d'observations valides pour chaque variable. La structure initiale affiche la qualité de représentation initiale, les valeurs propres et le pourcentage de variance expliqué.

Matrice de corrélation : Les options disponibles sont les coefficients, les seuils de signification, les déterminants, les inverses, les reproduits, l'anti-image et l'indice KMO et le test de sphéricité de Bartlett.

- **Indice KMO et test de sphéricité de Bartlett.** Mesure de l'adéquation de l'échantillonnage de Kaiser-Meyer-Olkin qui teste si les corrélations partielles entre les variables sont faibles. Le test de sphéricité de Bartlett teste si la matrice des corrélations est une matrice d'identité, ce qui indiquerait que le modèle de facteur n'est pas adapté.
- **Reconstituée.** Matrice des corrélations estimée à partir de la solution factorielle. Les résidus (différence entre les corrélations estimées et observées) sont également affichés.
- **Anti-image.** La matrice de corrélation des anti-images contient les opposés des coefficients de corrélation partielle ; la matrice de covariance des anti-images contient les opposés des covariances partielles. Dans un bon modèle factoriel, la plupart des éléments hors diagonale doivent être petits. La mesure d'adéquation d'échantillonnage pour une variable est affichée sur la diagonale de la matrice de corrélation des anti-images.

Extraction d'analyse factorielle

Figure 22-4
Boîte de dialogue Analyse Factorielle : Extraction



Méthode. Vous permet de spécifier la méthode d'extraction de facteur. Les méthodes disponibles sont les Composantes principales, les Moindres carrés non pondérés, les Moindres carrés généralisés, le Maximum de vraisemblance, la Factorisation en axes principaux, l'Alpha-maximisation et la Factorisation en projections.

- **Analyse en composantes principales.** Méthode d'extraction de facteur utilisée pour former des combinaisons linéaires non corrélées des variables observées. La première composante principale a une variance maximale. Les autres composantes expliquent progressivement des portions plus petites de la variance sans être corrélées les unes aux autres. L'analyse des composantes principales est utilisée pour obtenir la solution factorielle initiale. Elle peut être utilisée quand la matrice des corrélations est singulière.
- **Méthode des moindres carrés non pondérés.** Méthode d'extraction de facteur qui minimise la somme des carrés des différences entre les matrices de corrélations observées et reconstituées, en ignorant les diagonales.
- **Méthode des Moindres carrés généralisés.** Méthode d'extraction de facteur qui minimise la somme des carrés des différences entre les matrices de corrélations observées et reconstituées. Les corrélations sont pondérées par l'inverse de leur unicité, de façon à ce que les variables présentant une forte unicité reçoivent une pondération inférieure à celles présentant une faible unicité.
- **Méthode du maximum de vraisemblance.** Méthode d'extraction de facteur qui fournit les estimations de paramètres les plus susceptibles d'avoir généré la matrice de corrélations observée si l'échantillon est issu d'une distribution gaussienne multivariée. Les corrélations sont pondérées par l'inverse de l'unicité des variables et un algorithme itératif est utilisé.
- **Factorisation en axes principaux.** Méthode d'extraction de facteurs à partir de la matrice des corrélations initiales où les coefficients de corrélation multiple au carré sont placés sur la diagonale comme estimation initiale des qualités. Ces cartes factorielles sont utilisées pour

une nouvelle estimation des qualités de représentation qui remplace alors l'ancienne sur la diagonale. Les itérations se poursuivent jusqu'à ce que les variations des qualités de représentation d'une itération à l'autre satisfassent le critère de convergence de l'extraction.

- **Alpha.** Méthode d'extraction de facteur qui considère les variables dans l'analyse comme un échantillon issu de la population des variables potentielles. Cette méthode maximise l'alpha de Cronbach des facteurs.
- **Factorisation en projections.** Méthode d'extraction de facteur développée par Guttman et basée sur la théorie d'une image. La partie commune de la variable, appelée image partielle, est définie comme sa régression linéaire sur les autres variables, plutôt qu'une fonction de facteurs hypothétiques.

Analyser : Vous permet de spécifier si l'analyse porte sur une matrice de corrélation ou sur une matrice de covariance.

- **Matrice de corrélation.** Utile si les variables de votre analyse sont mesurées selon des échelles différentes.
- **Matrice de covariance.** Utile lorsque vous souhaitez appliquer l'analyse factorielle à plusieurs groupes avec des variances différentes pour chaque variable.

Extraire : Vous pouvez retenir tous les facteurs dont les valeurs propres dépassent une valeur spécifique ou retenir un nombre spécifique de facteurs.

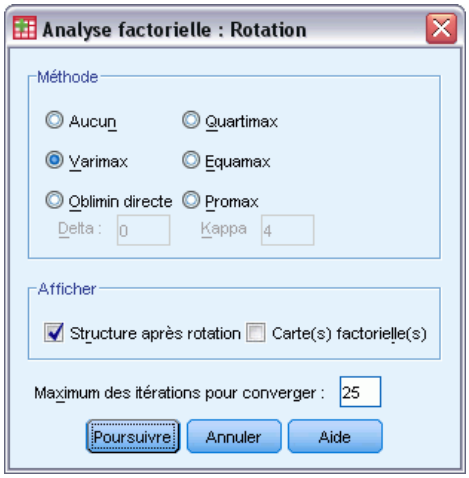
Afficher : Vous permet de demander la solution factorielle avant rotation et un diagramme des valeurs propres.

- **Solution factorielle sans rotation.** Affiche les corrélations factorielles sans rotation (matrice de projections factorielles), les qualités de représentation et les valeurs propres de la solution factorielle.
- **Diagramme des valeurs propres.** Diagramme représentant la variance associée à chaque facteur. Permet de déterminer le nombre de facteurs à conserver. Généralement, le diagramme montre une coupure franche entre la forte pente des facteurs élevés et la traîne graduelle du reste (valeurs propres).

Maximum des itérations pour converger : Vous permet de spécifier le nombre maximum de pas que l'algorithme peut utiliser pour estimer la solution.

Rotation d'analyse factorielle

Figure 22-5
Boîte de dialogue Analyse factorielle : Rotation



Méthode : Vous permet de sélectionner la méthode de rotation des facteurs. Les méthodes disponibles sont Varimax, Oblimin directe, Quartimax, Equamax ou Promax.

- **Méthode varimax.** Méthode de rotation orthogonale qui minimise le nombre de variables ayant de fortes corrélations sur chaque facteur. Simplifie l'interprétation des facteurs.
- **Critère oblimin direct.** Méthode de rotation oblique (non orthogonale). Lorsque delta est nul (valeur par défaut), les solutions sont les plus obliques. Plus la valeur de delta est négative, moins les facteurs sont obliques. Pour remplacer la valeur nulle par défaut de delta, entrez un nombre inférieur ou égal à 0,8.
- **Méthode quartimax.** Méthode de rotation qui réduit le nombre de facteurs requis pour expliquer chaque variable. Simplifie l'interprétation des variables observées.
- **Equamax.** Méthode de rotation qui est une combinaison de la méthode Varimax (qui simplifie les facteurs) et de la méthode Quartimax (qui simplifie les variables). Le nombre de variables pesant sur un facteur et le nombre de facteurs nécessaires pour expliquer une variable sont minimisés.
- **Rotation Promax.** Rotation oblique qui permet aux facteurs d'être corrélés. Peut être calculée plus rapidement qu'une rotation oblimin directe, aussi est-elle utile pour les vastes ensembles de données.

Afficher : Vous permet d'inclure le résultat de la structure après rotation, et également d'afficher les cartes factorielles sur le premier, le second et le troisième facteur (Cartes factorielles).

- **Structure après rotation.** Vous devez sélectionner une méthode de rotation pour obtenir une structure après rotation. Pour les rotations orthogonales, la matrice de forme après rotation et la matrice de transformation factorielle sont affichées. Pour les rotations obliques, le

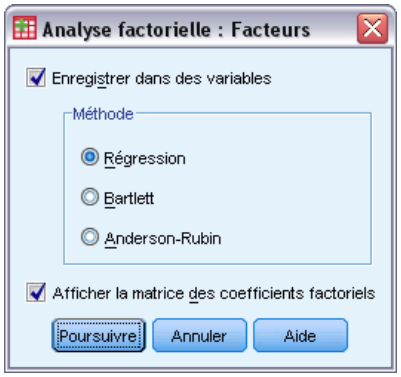
programme affiche la matrice des projections factorielles, la matrice de structure et la matrice des corrélations de facteurs.

- **Diagramme des Contributions des Facteurs.** Diagramme en trois dimensions des contributions des trois premiers facteurs. Pour une solution à deux facteurs, un diagramme en deux dimensions est affiché. Le diagramme n'est pas affiché si un seul facteur est extrait. Les diagrammes affichent des solutions ayant subi une rotation si cette dernière est demandée.

Maximum des itérations pour converger : Vous permet de spécifier le nombre maximum de pas que l'algorithme peut utiliser pour réaliser la rotation.

Scores d'analyse factorielle

Figure 22-6
Boîte de dialogue Scores de l'Analyse Factorielle



Enregistrer dans des variables : Vous permet de créer une nouvelle variable pour chaque facteur selon la structure finale.

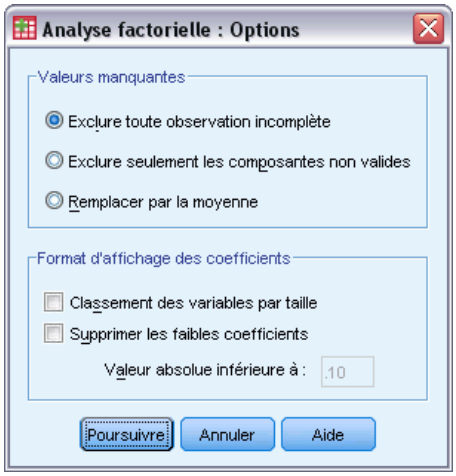
Méthode : Les méthodes alternatives pour calculer les facteurs sont la Régression, Bartlett, et Anderson-Rubin.

- **Méthode de régression.** Méthode d'estimation des coefficients factoriels. Les écarts obtenus ont une moyenne de 0 et une variance égale au carré de la corrélation multiple entre les coefficients factoriels estimés et les vraies valeurs du facteur. Les écarts peuvent être corrélés même lorsque les facteurs sont orthogonaux.
- **Facteurs de Bartlett.** Méthode d'estimation des coefficients factoriels. Les résultats ont une moyenne de 0. La somme des carrés des facteurs uniques dans la plage de variables est minimisée.
- **Méthode d'Anderson-Rubin.** Méthode d'estimation des coefficients factoriels ; variante de la méthode de Bartlett qui garantit l'orthogonalité des facteurs estimés. Les scores obtenus affichent une moyenne de 0 et un écart-type de 1 et ne sont pas corrélés.

Afficher la matrice des coefficients factoriels : Vous permet de montrer les coefficients par lesquels les variables sont multipliées pour obtenir les facteurs. Cela permet également de montrer les corrélations entre les facteurs.

Options d'analyse factorielle

Figure 22-7
Boîte de dialogue des Options de l'Analyse Factorielle



Valeurs manquantes : Vous permet de spécifier comment traiter les valeurs manquantes. Les options disponibles sont d'exclure toute observation **incomplète**, d'exclure seulement les composantes **non valides**, ou de les remplacer par la moyenne.

Affichage des projections : Vous permet de contrôler le format des matrices de résultat. Triez les coefficients par leur taille et supprimez les coefficients dont la valeur absolue est inférieure à la valeur spécifiée.

Fonctionnalités supplémentaires de la commande FACTOR

Le langage de syntaxe de commande vous permet aussi de :

- spécifier des critères de convergence pour les itérations lors de l'extraction et de la rotation ;
- définir des diagrammes factoriels individuels après rotation ;
- indiquer le nombre de facteurs à enregistrer ;
- spécifier les valeurs des diagonales pour la méthode de factorisation en axes principaux ;
- créer des matrices de corrélation ou des matrices de projections factorielles sur un disque pour une analyse ultérieure ;
- lire et analyser des matrices de corrélation ou des matrices de projections factorielles.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
23

Choix d'une procédure de classification

Vous pouvez effectuer des analyses de classes à l'aide de la procédure TwoStep, de la classification hiérarchique ou des nuées dynamiques. Chaque procédure utilise un algorithme différent pour la création des classes, et chacune d'elles comporte des options qui ne sont pas disponibles dans les autres procédures.

Analyse du composant Classe TwoStep. La procédure Analyse du composant Classe TwoStep est la méthode privilégiée pour de nombreuses applications. Elle offre les fonctionnalités spécifiques suivantes :

- Sélection automatique du meilleur nombre de classes, en plus des mesures de sélection parmi des modèles de classe.
- Possibilité de créer simultanément des modèles de classe sur la base de variables qualitatives et continues.
- Possibilité d'enregistrer le modèle de classe dans un fichier XML externe, puis de lire ce fichier et de mettre à jour le modèle de classe à l'aide des données les plus récentes.

En outre, la procédure Analyse du composant Classe TwoStep permet d'analyser des fichiers de données volumineux.

Classification hiérarchique. La procédure Classification hiérarchique est limitée à des fichiers de données plus petits (centaines d'objets à classer), mais offre les fonctionnalités spécifiques suivantes :

- Possibilité de classer des observations ou des variables.
- Possibilité de calculer plusieurs solutions possibles et d'enregistrer des classes d'affectation pour chacune de ces solutions.
- Plusieurs méthodes de formation de classes, de transformation de variables et de mesure de la dissimilarité entre les classes.

Tant que toutes les variables sont du même type, la procédure Classification hiérarchique peut analyser des variables d'intervalle (continues), d'effectif ou binaires.

Nuées dynamiques. La procédure Nuées dynamiques est limitée aux données continues et exige que vous indiquiez au préalable le nombre de classes. Elle offre néanmoins les fonctionnalités spécifiques suivantes :

- Possibilité d'enregistrer les distances à partir des centres de classes pour chaque objet.
- Possibilité de lire les centres de classes initiaux à partir d'un fichier IBM® SPSS® Statistics et d'enregistrer les centres de classes finaux dans un fichier SPSS Statistics externe .

En outre, la procédure Nuées dynamiques permet d'analyser des fichiers de données volumineux.

Chapitre
24

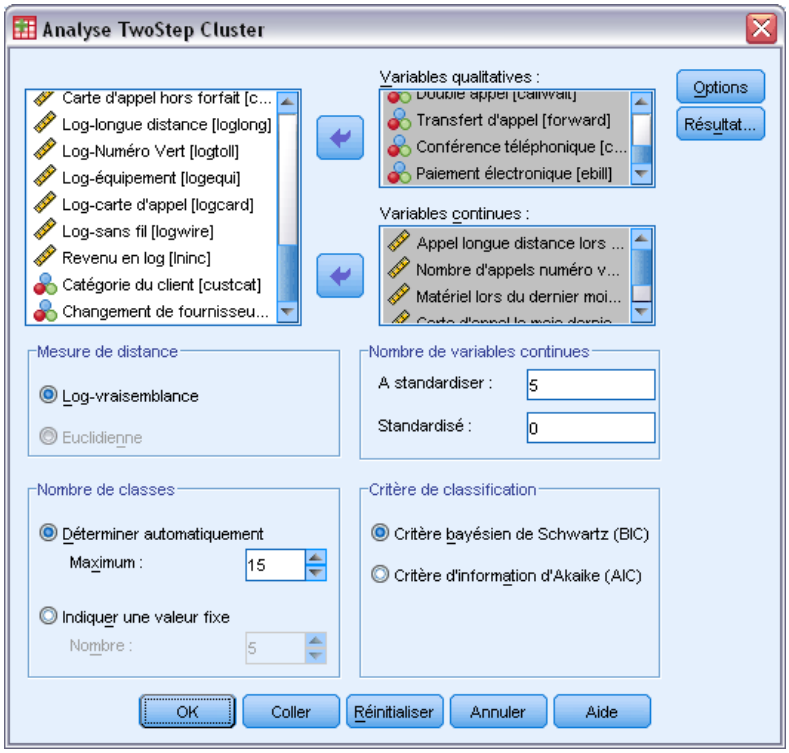
Analyse TwoStep Cluster

La procédure d'analyse TwoStep Cluster est un outil d'exploration conçu pour révéler des groupements naturels (ou classes) au sein d'un fichier de données. L'algorithme utilisé par cette procédure possède plusieurs fonctionnalités qui le distinguent des techniques de classification standard :

- **Gestion des données qualitatives et continues :** En supposant que les variables soient indépendantes, une distribution jointe multinomiale-normale peut être placée sur des variables qualitatives et continues.
- **Sélection automatique du nombre de classes :** En comparant les valeurs d'un critère de modèle-choix dans différentes solutions de classification, la procédure peut déterminer automatiquement le nombre optimal de classes.
- **Evolutivité :** En construisant une arborescence de fonctionnalités de classe (CF) qui récapitule les enregistrements, l'algorithme TwoStep vous permet d'analyser des fichiers de données volumineux.

Exemple : Les entreprises du domaine des produits de consommation et du commerce de détail utilisent régulièrement des techniques de classification des données qui décrivent les habitudes d'achat, le sexe, l'âge, le niveau de revenu, etc. de leurs clients. Ces sociétés adaptent leurs stratégies de marketing et de développement produit à chaque groupe de consommation afin d'augmenter les ventes et de développer la fidélité à la marque.

Figure 24-1
Boîte de dialogue Analyse TwoStep Cluster



Mesure de distance : Cette sélection détermine la façon dont la similarité entre deux classes est calculée.

- **Log-vraisemblance :** La mesure de vraisemblance place une distribution de probabilité sur les variables. Les variables continues sont considérées comme étant distribuées normalement alors que les variables qualitatives sont considérées comme étant multinomiales. Toutes les variables sont considérées comme étant indépendantes.
- **Euclidienne :** La mesure euclidienne est la distance « en ligne droite » entre deux classes. Elle peut être utilisée uniquement lorsque toutes les variables sont continues.

Nombre de classes : Cette sélection vous permet d'indiquer la façon dont le nombre de classes doit être déterminé.

- **Déterminer automatiquement.** Cette procédure déterminera automatiquement le « meilleur » nombre de classes en utilisant le critère défini dans le groupe Critère de classification. Vous pouvez également entrer un nombre entier positif qui définit le nombre maximal de classes que la procédure doit prendre en compte.
- **Indiquer une valeur fixe.** Vous permet d'indiquer le nombre de classes (valeur fixe) dans la solution. Entrez un entier positif.

Nombre de variables continues : Ce groupe fournit un récapitulatif des spécifications de standardisation des variables continues qui sont définies dans la boîte de dialogue Options. Pour plus d'informations, reportez-vous à la section Options de la procédure d'analyse TwoStep Cluster sur p. 173.

Critère de classification : Cette sélection détermine la façon dont l'algorithme de classification automatique détermine le nombre de classes. Vous pouvez spécifier le critère d'information bayésien (BIC) ou le critère d'information d'Akaike (AIC).

Données. Cette procédure fonctionne avec des variables continues et qualitatives. Les observations représentent les objets à classer et les variables représentent les attributs sur lesquels est basée la classification.

Tri par observation. Remarque : l'arborescence des fonctionnalités de classe et la solution finale peuvent dépendre de l'ordre des observations. Pour réduire les effets de tri, classez les observations de manière aléatoire. Vous pouvez obtenir différentes solutions pour lesquelles les observations ont été triées de manière aléatoire, afin de vérifier la stabilité d'une solution donnée. Lorsque cela s'avère difficile en raison de fichiers très volumineux, vous pouvez effectuer plusieurs fois l'opération sur un échantillon des observations triées de différentes manières aléatoires.

Hypothèses : La mesure de la distance de vraisemblance considère que les variables du modèle de classe sont indépendantes. De plus, chaque variable continue est considérée comme ayant une distribution normale (gaussienne) et chaque variable qualitative comme ayant une distribution multinomiale. Des tests internes empiriques indiquent que la procédure est assez résistante aux violations de l'hypothèse d'indépendance et des hypothèses de distribution, mais vous devez savoir comment ces hypothèses sont vérifiées.

Utilisez la procédure Corrélations bivariées pour tester l'indépendance de deux variables continues. Utilisez la procédure Tableaux croisés pour tester l'indépendance de deux variables qualitatives. Utilisez la procédure Moyennes pour tester l'indépendance entre une variable continue et une variable qualitative. Utilisez la procédure Explorer pour tester la normalité d'une variable continue. Utilisez la procédure Test du Khi-deux pour tester si une variable qualitative a une distribution multinomiale spécifiée.

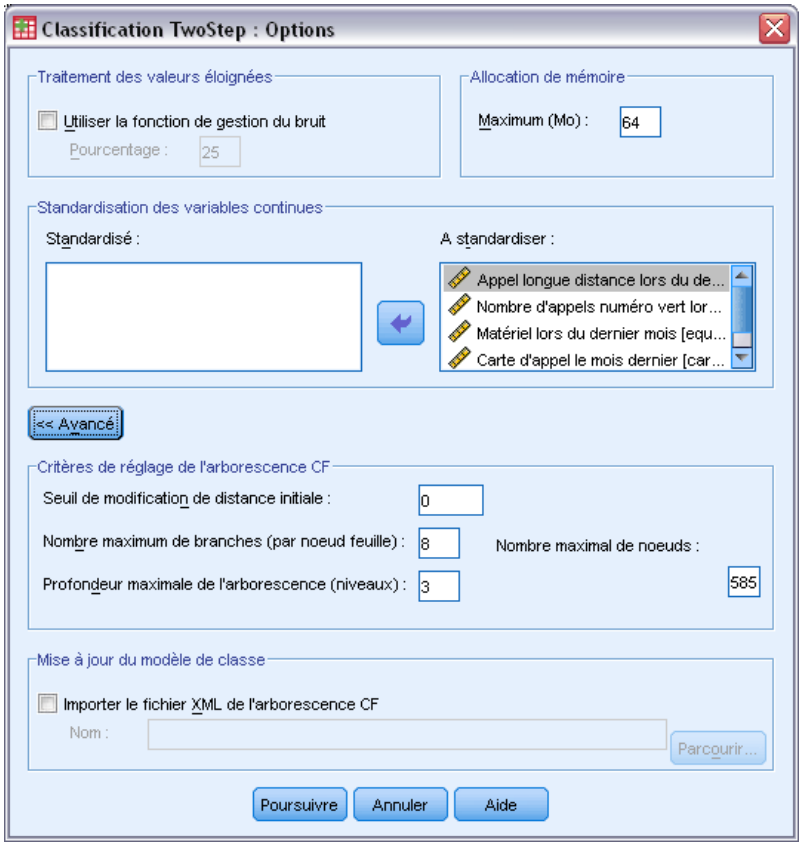
Pour effectuer une procédure d'analyse TwoStep Cluster

- A partir des menus, sélectionnez :
  Analyse > Classification > TwoStep Cluster...
- Sélectionnez une ou plusieurs variables qualitatives ou continues.

  Sinon, vous pouvez :
  - Ajuster les critères sur lesquels est basée la construction des classes.
  - Sélectionner les paramètres de gestion du bruit, d'affectation de mémoire, de standardisation de variable et d'entrée de modèle de classe.
  - Demander les résultats du Visualiseur de modèles.
  - Enregistrer les résultats de modèle dans le fichier de travail ou dans un fichier XML externe.

Options de la procédure d'analyse TwoStep Cluster

Figure 24-2
Boîte de dialogue Options TwoStep Cluster



Traitement des valeurs éloignées : Ce groupe permet de traiter les valeurs éloignées, notamment lors de la classification, si l'arborescence des fonctionnalités de classe (CF) est saturée. L'arborescence CF est saturée si elle ne peut plus accepter d'autres observations dans un noeud feuille et qu'aucun noeud feuille ne peut être divisé.

- Si vous sélectionnez la gestion du bruit et que l'arborescence CF est saturée, l'arborescence est reconstruite lorsque vous placez des observations de feuilles éclatées dans une feuille « bruit ». Une feuille est éclatée si elle contient un pourcentage inférieur au pourcentage d'observations correspondant à la taille maximale de la feuille. Une fois que l'arborescence est reconstruite, les valeurs éloignées sont placées dans l'arborescence CF si cela est possible. Sinon, les valeurs éloignées sont supprimées.
- Si vous ne sélectionnez pas la gestion du bruit et que l'arborescence CF est saturée, elle sera reconstruite à l'aide d'un seuil de changement de distance supérieur. Après la classification finale, les valeurs ne pouvant être affectées à une classe deviennent des valeurs éloignées étiquetées. Un numéro d'identification de –1 est affecté à la classe de valeur éloignée et cette dernière n'est pas prise en compte dans le nombre de classes.

Allocation de mémoire : Ce groupe vous permet d'indiquer la quantité de mémoire maximale en mégaoctets (Mo) que l'algorithme de classe doit utiliser. Si la procédure dépasse cette limite, elle utilisera le disque pour stocker les informations ne pouvant pas être enregistrées en mémoire. Spécifiez une valeur supérieure ou égale à 4.

- Consultez l'administrateur système pour connaître la plus grande valeur que vous pouvez spécifier sur votre système.
- L'algorithme risque de ne pas trouver le nombre correct ou souhaité de classes si cette valeur est trop basse.

Standardisation de variable : L'algorithme de classification fonctionne avec des variables continues standardisées. Les variables continues non standardisées doivent être laissées en tant que variables dans la liste À standardiser . Pour gagner du temps et éviter trop de calculs, vous pouvez sélectionner une variable continue déjà standardisée comme variable dans la liste Standardisée.

Options avancées

Critères de réglage de l'arborescence CF : Les paramètres d'algorithme de classification suivants s'appliquent de façon spécifique à l'arborescence CF et doivent être modifiés avec le plus grand soin :

- **Seuil de modification de distance initiale :** Il s'agit du seuil initial utilisé pour construire l'arborescence CF. Si l'insertion d'une observation dans une feuille de l'arborescence CF provoque une étroitesse inférieure au seuil, la feuille n'est pas divisée. Si l'étroitesse dépasse le seuil, la feuille est divisée.
- **Nombre maximum de branches (par noeud feuille) :** Nombre maximum de noeuds enfant qu'un noeud feuille peut contenir.
- **Profondeur maximum de l'arborescence :** Nombre maximum de niveaux que l'arborescence CF peut contenir.
- **Nombre maximal de noeuds :** Ceci indique le nombre maximal de noeuds de l'arbre CF pouvant être générés par la procédure, d'après la fonction $(b^{d+1} - 1) / (b - 1)$, b étant le nombre maximal de branches et d, la profondeur maximale de l'arbre. Notez qu'une arborescence CF trop volumineuse risque d'épuiser les ressources du système et d'avoir des effets défavorables sur les performances de la procédure. Chaque noeud exige au moins 16 octets.

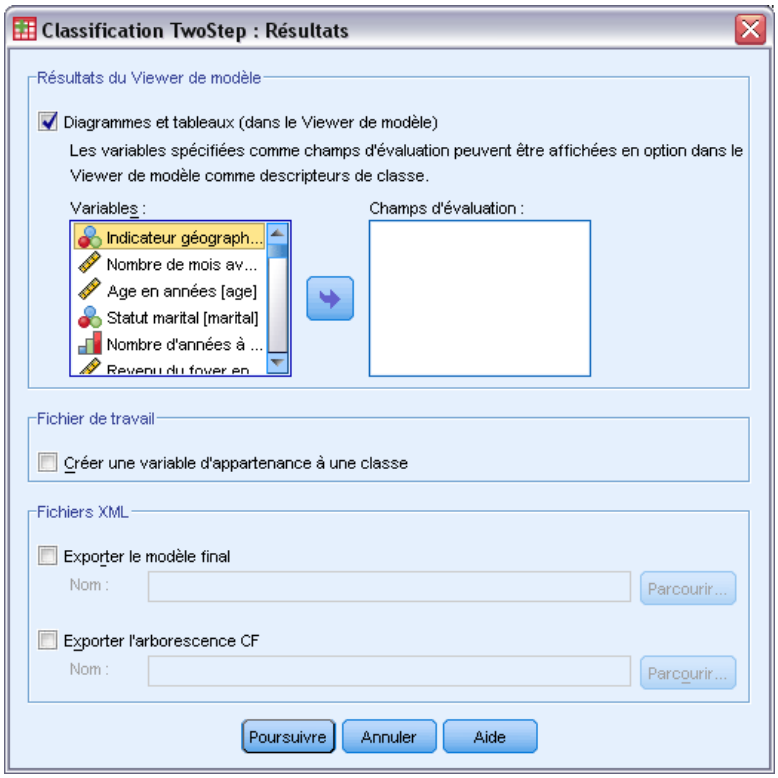
Mettre à jour le modèle de classe : Ce groupe vous permet d'importer et de mettre à jour un modèle de classe généré dans une analyse précédente. Le fichier d'entrée contient l'arborescence CF au format XML. Le modèle est ensuite mis à jour avec les données du fichier actif. Vous devez sélectionner les noms des variables dans la boîte de dialogue principale dans le même ordre que celui dans lequel ils ont été spécifiés dans l'analyse précédente. Le fichier XML demeure inchangé, sauf si vous enregistrez les informations du nouveau modèle en utilisant le même nom de fichier. Pour plus d'informations, reportez-vous à la section Résultats de l'analyse TwoStep Cluster sur p. 175.

Si vous avez indiqué la mise à jour d'un modèle de classe, les options relatives à la génération de l'arborescence CF spécifiées pour le modèle d'origine sont utilisées. Plus précisément, les paramètres de mesure de la distance, de gestion du bruit, d'affectation de mémoire ou les critères de réglage de l'arborescence CF pour le modèle enregistré sont utilisés et tous les paramètres de ces options dans les boîtes de dialogue sont ignorés.

Remarque : Lorsque vous effectuez une mise à jour d'un modèle de classe, la procédure considère qu'aucune des observations sélectionnées dans l'ensemble de données actif n'a été utilisée pour créer le modèle de classe d'origine. La procédure considère également que les observations utilisées dans la mise à jour du modèle sont issues de la même population d'observations utilisée pour créer le modèle d'origine. En d'autres termes, les moyennes et les variances des variables continues et les niveaux des variables qualitatives sont considérés comme étant identiques dans les deux groupes d'observations. Si votre « nouveau » groupe d'observations ne provient pas de la même population que votre « ancien » groupe, exécutez la procédure Analyse TwoStep Cluster sur les groupes d'observations combinés pour obtenir de meilleurs résultats.

Résultats de l'analyse TwoStep Cluster

Figure 24-3
Boîte de dialogue Résultats de l'analyse TwoStep Cluster



Résultats du Visualiseur de modèles Ce groupe fournit des options d'affichage pour les résultats de la classification.

- **Diagrammes et tableaux.** Affiche les résultats liés au modèle, y compris les tableaux et les diagrammes. Les tableaux de la vue du modèle comprennent un récapitulatif du modèle et une grille de caractéristiques par classe. Les résultats graphiques dans l'affichage du

modèle incluent un diagramme de qualité des classes, les tailles des classes, l'importance des variables, une grille de comparaison des classes et des informations sur les cellules.

- **Champs d'évaluation.** Calcule les données de classe pour les variables non utilisées dans la création des classes. Les champs d'évaluation peuvent être affichés en même temps que les caractéristiques d'entrée du Visualiseur de modèles en les sélectionnant dans la sous-boîte de dialogue Affichage. Les champs avec valeurs manquantes sont ignorés.

Fichier de travail : Ce groupe vous permet d'enregistrer des variables dans l'ensemble de données actif.

- **Créer une variable d'appartenance à une classe ;** Cette variable contient un numéro d'identification de classe pour chaque observation. Le nom de cette variable est *tsc_n*, *n* étant un nombre entier positif qui indique l'ordinal de l'opération d'enregistrement de l'ensemble de données actif effectuée par cette procédure au cours d'une session.

Fichiers XML : Le modèle de classe final et l'arborescence CF représentent deux types de fichiers de résultats pouvant être exportés au format XML.

- **Exporter le modèle final :** Le modèle de classe final est exporté vers le fichier indiqué au format XML (PMML). Vous pouvez utiliser ce fichier de modèle pour appliquer les informations du modèle aux autres fichiers de données à des fins d'évaluation.
- **Exporter l'arborescence CF :** Cette option vous permet d'enregistrer l'état actuel de l'arborescence de classe et de le mettre à jour ultérieurement en utilisant des données plus récentes.

Le viewer de classes

Les modèles de classe sont généralement utilisés pour trouver des groupes (ou des classes) d'enregistrements similaires en fonction des variables examinées, où la similarité entre les membres d'un même groupe est élevée et où la similarité entre les membres de différents groupes est faible. Les résultats peuvent être utilisés pour identifier des associations qui ne seraient pas évidentes autrement. Par exemple, grâce à la classification des préférences des clients, du niveau de revenu et des habitudes d'achat, il peut être possible d'identifier les types de clients les plus susceptibles de répondre à une campagne de marketing particulière.

Il existe deux approches pour interpréter les résultats d'un affichage de classe :

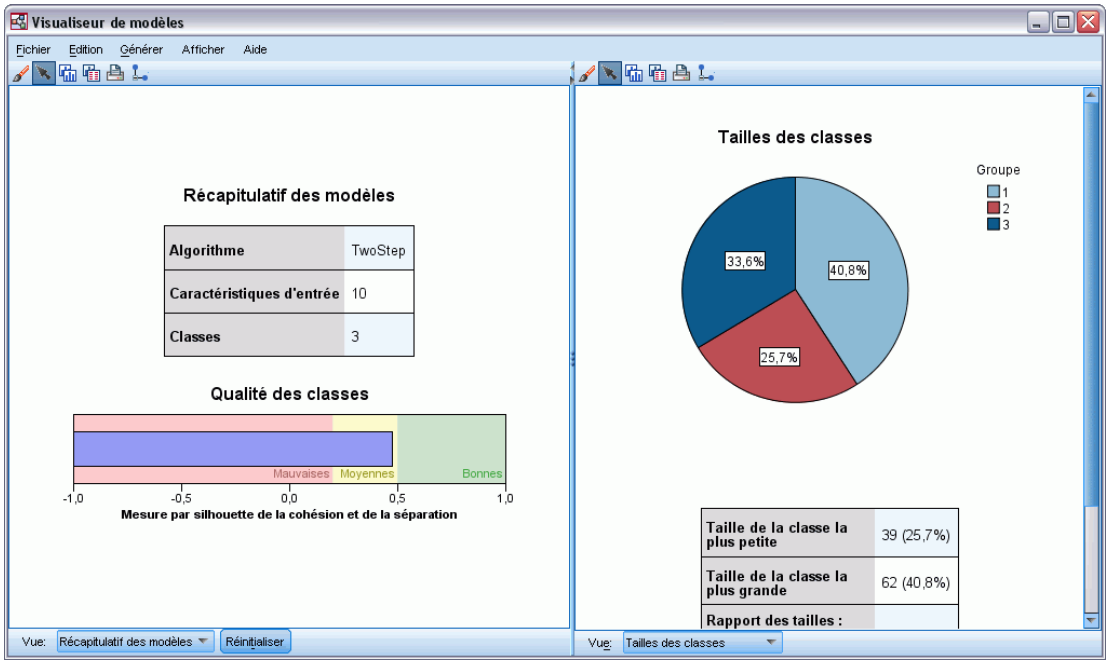
- Examiner les classes afin de déterminer les caractéristiques uniques de cette classe. *Est-ce qu'une classe contient tous les emprunteurs à revenus élevés ? Est-ce que cette classe contient davantage d'enregistrements que les autres ?*
- Examiner les champs des classes afin de déterminer comment les valeurs sont distribuées parmi les classes. *Est-ce que le niveau d'éducation détermine l'appartenance à une classe ? Est-ce qu'une cote de solvabilité élevée permet de distinguer l'appartenance à une classe spécifique ?*

Grâce à l'utilisation des vues principales et des différentes vues liées dans le viewer de classes, vous pouvez avoir un bon aperçu qui vous aidera à répondre à ces questions.

Pour consulter les informations concernant un modèle de classe, activez (double-cliquez) sur l'objet Viewer de modèle dans le Viewer.

Viewer de classes

Figure 24-4
Viewer de modèle avec affichage par défaut



Le Viewer de modèle est constitué de deux panneaux, l'affichage principal à gauche et la vue liée, ou auxiliaire, à droite. Il existe deux vues principales :

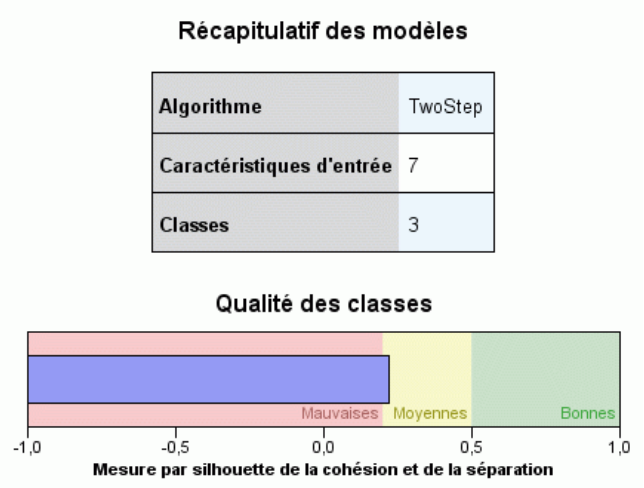
- Récapitulatif du modèle (par défaut). Pour plus d'informations, reportez-vous à la section Vue récapitulative du modèle sur p. 178.
- Classes. Pour plus d'informations, reportez-vous à la section Vue des classes sur p. 179.

Il existe quatre vues liées/auxiliaires :

- Importance des valeurs prédites. Pour plus d'informations, reportez-vous à la section Vue de l'importance des variables prédites de classe sur p. 182.
- Taille des classes (par défaut). Pour plus d'informations, reportez-vous à la section Vue de la taille des classes sur p. 183.
- Distribution des cellules. Pour plus d'informations, reportez-vous à la section Vue de la distribution des cellules sur p. 184.
- Comparaison des classes. Pour plus d'informations, reportez-vous à la section Vue de comparaison des classes sur p. 185.

Vue récapitulative du modèle

Figure 24-5
Vue récapitulative du modèle du panneau principal



La vue récapitulative du modèle affiche un instantané, ou un récapitulatif, du modèle de classe, y compris une mesure par silhouette de la cohésion et de la séparation des classes qui est ombrée pour indiquer des résultats faibles, moyens ou bons. Cet instantané vous permet de vérifier rapidement si la qualité est faible, auquel cas vous pouvez décider de revenir au noeud de modélisation afin de corriger les paramètres du modèle de classe pour obtenir un meilleur résultat.

Les résultats faibles, moyens et bons sont basés sur le travail de Kaufman et Rousseeuw (1990) concernant l'interprétation des structures de classe. Dans la vue récapitulative du modèle, d'après l'évaluation de Kaufman et Rousseeuw, un bon résultat équivaut à des données qui indiquent une preuve raisonnable ou forte de la structure de la classe, un résultat moyen signifie une preuve faible et un résultat mauvais reflète une absence de preuve significative.

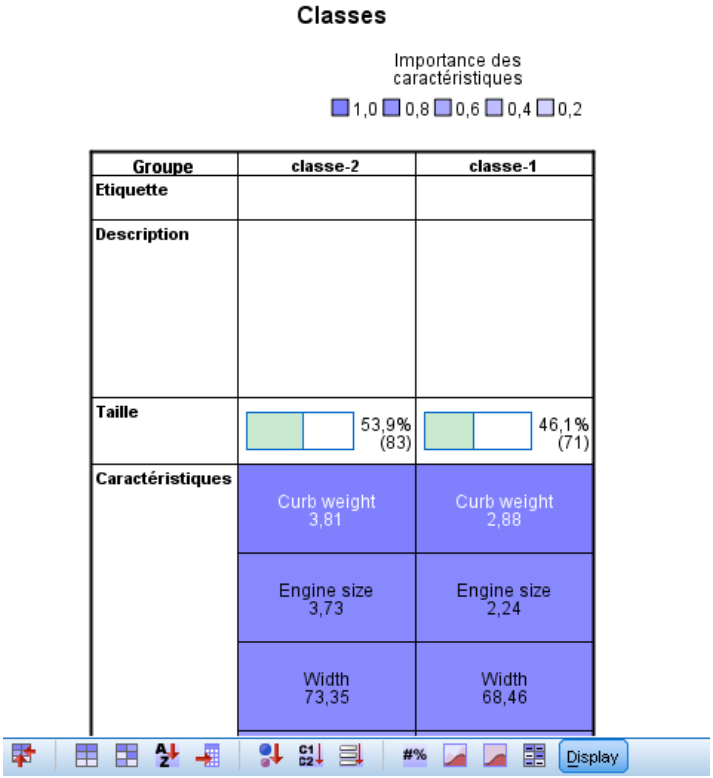
La mesure par silhouette établit la moyenne, de tous les enregistrements, (B−A) / max(A,B), où A correspond à la distance de l'enregistrement au centre de sa classe et B correspond à la distance de l'enregistrement au centre de la classe la plus proche à laquelle il n'appartient pas. Un coefficient de silhouette de 1 signifie que toutes les observations sont situées directement au centre de leurs classes. Une valeur de −1 signifie que toutes les observations sont situées au centre de classe d'autres classes. Une valeur de 0 signifie, en moyenne, que les observations sont équidistantes du centre de leur propre classe et du centre de l'autre classe la plus proche.

Le récapitulatif inclut un tableau qui contient les informations suivantes :

- **Algorithme.** L'algorithme de classification utilisé, par exemple "TwoStep".
- **Caractéristiques d'entrée.** Le nombre de champs, ou d'**entrées** ou de **variables prédites**.
- **Classes.** Le nombre de classes dans la solution.

Vue des classes

Figure 24-6
Vue des centres de la classe du panneau principal



La vue des classes contient une grille présentant les classes par caractéristiques, qui inclut les noms des classes, leurs tailles et les profils de chaque classe.

Les colonnes de la grille contiennent les informations suivantes :

- **Classe.** Les nombres de classes créées par l'algorithme.
- **Etiquette.** Toutes les étiquettes appliquées à chaque classe (celle-ci est vierge par défaut). Double-cliquez dans la cellule pour saisir une étiquette qui décrit les contenus de classe ; par exemple "Acheteurs de voitures de luxe".
- **Description :** Toutes les descriptions du contenu de classe (celle-ci est vierge par défaut). Double-cliquez dans la cellule pour saisir une description de la classe ; par exemple "professionnels, plus de 55 ans, gagnant plus de 100 000 $".
- **Taille :** La taille de chaque classe sous la forme d'un pourcentage de l'échantillon général des classes. Chaque cellule de taille de la grille affiche une barre verticale qui indique le pourcentage de taille au sein de la classe, un pourcentage de taille au format numérique et les effectifs d'observation de la classe.
- **Caractéristiques.** Les entrées ou variables prédites individuelles, triées par importance générale par défaut. Si des colonnes ont la même taille, elles sont affichées par ordre croissant des numéros de classe.

L'importance des caractéristiques générales est indiquée par la couleur d'ombrage de l'arrière-plan de la cellule ; la caractéristique la plus importante étant plus sombre et la moins importante n'étant pas ombrée. Un guide situé au-dessus du tableau indique l'importance correspondant à chaque couleur de cellule de caractéristique.

Lorsque vous passez la souris sur une cellule, le nom complet/l'étiquette de la caractéristique et la valeur de l'importance de la cellule s'affichent. De plus amples informations peuvent être affichées selon le type de vue et de caractéristique. Dans la vue Centres des classes, cela inclut les statistiques de la cellule et la valeur de la cellule ; par exemple : "Moyenne : 4.32". Pour les caractéristiques qualitatives, la cellule affiche le nom de la catégorie la plus fréquente (modale) et son pourcentage.

Dans la vue des classes, vous pouvez sélectionner plusieurs manières d'afficher les informations des classes :

- Transposer les classes et les caractéristiques. Pour plus d'informations, reportez-vous à la section Transposer les classes et les caractéristiques sur p. 180.
- Trier les caractéristiques. Pour plus d'informations, reportez-vous à la section Trier les caractéristiques sur p. 181.
- Trier les classes. Pour plus d'informations, reportez-vous à la section Trier les classes sur p. 181.
- Sélectionner le contenu des cellules. Pour plus d'informations, reportez-vous à la section Contenu des cellules sur p. 181.

Transposer les classes et les caractéristiques

Par défaut, les classes sont affichées en tant que colonnes et les caractéristiques sont affichées en tant que lignes. Pour inverser cet affichage, cliquez sur le bouton Transposer les classes et les caractéristiques à gauche des boutons Trier les caractéristiques par. Par exemple, vous pouvez réaliser ceci lorsque de nombreuses classes sont affichées, afin de réduire le défilement horizontal nécessaire pour visualiser les données.

Figure 24-7
Classes transposées dans le panneau principal

Classe	Etiquette	Description	Taille	
cluster-1			45,0% (91)	BP HIGH (41,8%)
cluster-3			35,0% (70)	BP NORMAL (51,4%)
cluster-2			19,0% (39)	BP HIGH (100,0%)

Trier les caractéristiques

Le bouton Trier les caractéristiques par vous permet de sélectionner la façon dont les cellules de caractéristiques sont affichées :

- **Importance générale.** Il s'agit de l'ordre de tri par défaut. Les caractéristiques sont triées en ordre décroissant de l'importance générale, et l'ordre de tri est le même pour toutes les classes. Si des caractéristiques ont des valeurs d'importance liées, les caractéristiques liées sont répertoriées par ordre croissant des noms de caractéristique.
- **Importance intra-classe.** Les caractéristiques sont triées par rapport à leur importance pour chaque classe. Si des caractéristiques ont des valeurs d'importance liées, les caractéristiques liées sont répertoriées par ordre croissant des noms de caractéristique. Lorsque cette option est sélectionnée, l'ordre de tri varie généralement d'une classe à l'autre.
- **Nom.** Les caractéristiques sont triées par nom dans l'ordre alphabétique.
- **Ordre des données.** Les caractéristiques sont triées selon leur ordre dans l'ensemble de données.

Trier les classes

Par défaut, les classes sont triées en ordre de taille décroissante. Les boutons Trier les classes par vous permettent de les trier par nom dans l'ordre alphabétique, ou, si vous avez créé des étiquettes uniques, par ordre alphabétique des étiquettes.

Les caractéristiques ayant la même étiquette sont triées selon le nom de classe. Si des classes sont triées par étiquette et que vous modifiez l'étiquette d'une classe, l'ordre de tri est automatiquement mis à jour.

Contenu des cellules

Les boutons Cellules vous permettent de modifier l'affichage du contenu des cellules pour les champs de caractéristiques et d'évaluation.

- **Centre de classes.** Par défaut, les cellules affichent les noms/étiquettes des caractéristiques et la tendance centrale pour chaque combinaison de classe/caractéristique. La moyenne est affichée pour des champs continus et le mode (de la catégorie se présentant le plus fréquemment) avec le pourcentage de catégorie des champs qualitatifs.
- **Distributions absolues.** Affiche des noms/étiquettes de caractéristiques et des distributions absolues des caractéristiques dans chaque classe. Pour les caractéristiques qualitatives, l'affichage montre des diagrammes en bâtons où sont superposées des catégories en ordre croissant de la valeur des données. Pour les caractéristiques continues, l'affichage montre un diagramme de densité lissé qui utilise les mêmes extrema et intervalles pour chaque classe.

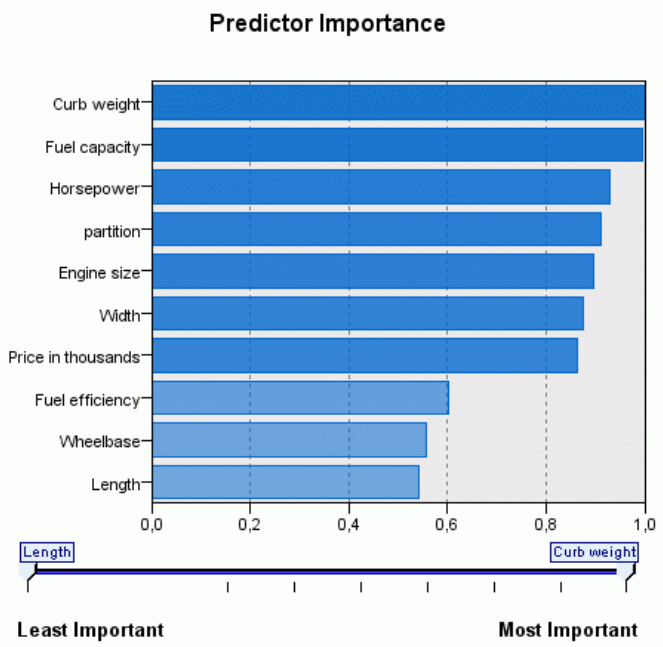
  L'affichage en rouge uni montre la distribution des classes, alors qu'un affichage plus pâle représente les données générales.
- **Distributions relatives.** Affiche les noms/étiquette des caractéristiques et les distributions relatives dans les cellules. En général, les affichages sont similaires à ceux des distributions absolues, excepté les distributions relatives qui sont affichées à la place.

L'affichage en rouge uni montre la distribution des classes, alors qu'un affichage plus pâle représente les données générales.

- **Vue de base.** Là où il y a beaucoup de classes, il peut être difficile de distinguer tous les détails sans procéder à un défilement. Afin de réduire le défilement, sélectionnez cette vue pour modifier l'affichage en une version plus compacte du tableau.

Vue de l'importance des variables prédites de classe

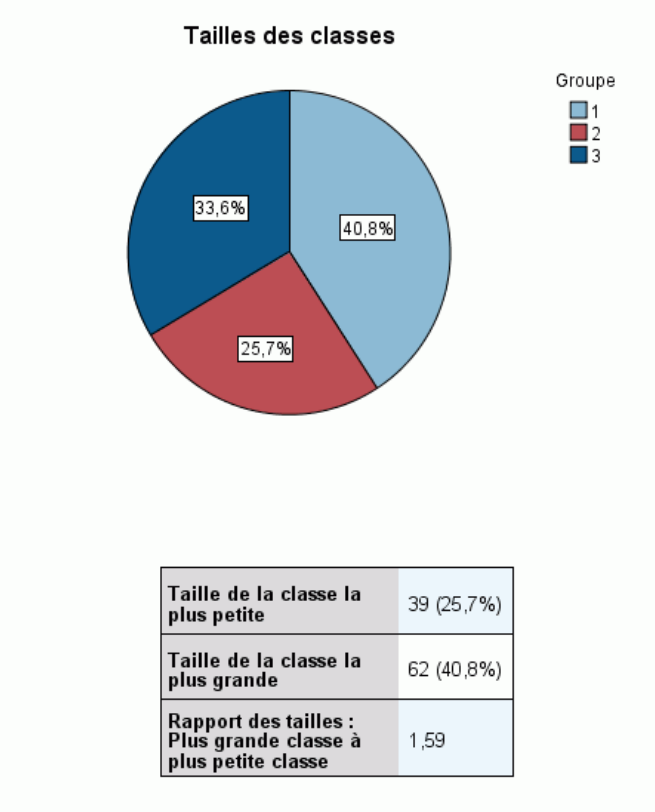
Figure 24-8
Vue de l'importance des variables prédites de classe dans le panneau des liens



La vue de l'importance des variables prédites affiche l'importance relative de chaque champ dans l'estimation du modèle.

Vue de la taille des classes

Figure 24-9
Vue de la taille des classes dans le panneau des liens



Taille de la classe la plus petite	39 (25,7%)
Taille de la classe la plus grande	62 (40,8%)
Rapport des tailles : Plus grande classe à plus petite classe	1,59

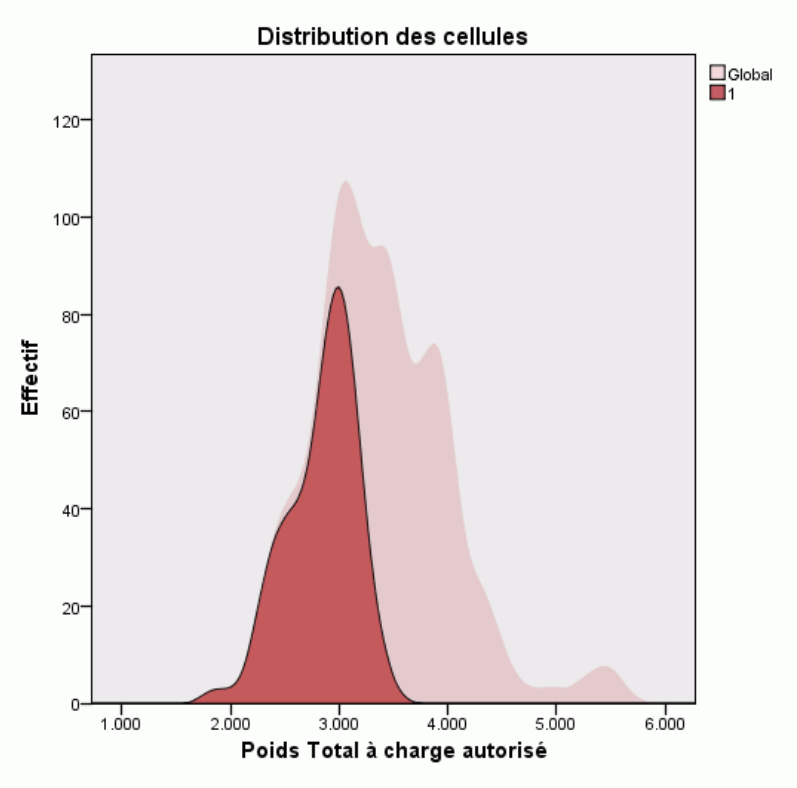
La vue de la taille des classes affiche un diagramme en secteurs qui contient chaque classe. La taille en pourcentage de chaque classe est affichée sur chaque tranche ; passez la souris sur chaque tranche pour afficher l'effectif de celle-ci.

En dessous du diagramme, un tableau répertorie les informations suivantes sur la taille :

- La taille de la classe la plus petite (en effectif et en pourcentage de l'ensemble).
- La taille de la classe la plus grande (en effectif et en pourcentage de l'ensemble).
- Le rapport de taille de la classe la plus grande par rapport à la classe la plus petite.

Vue de la distribution des cellules

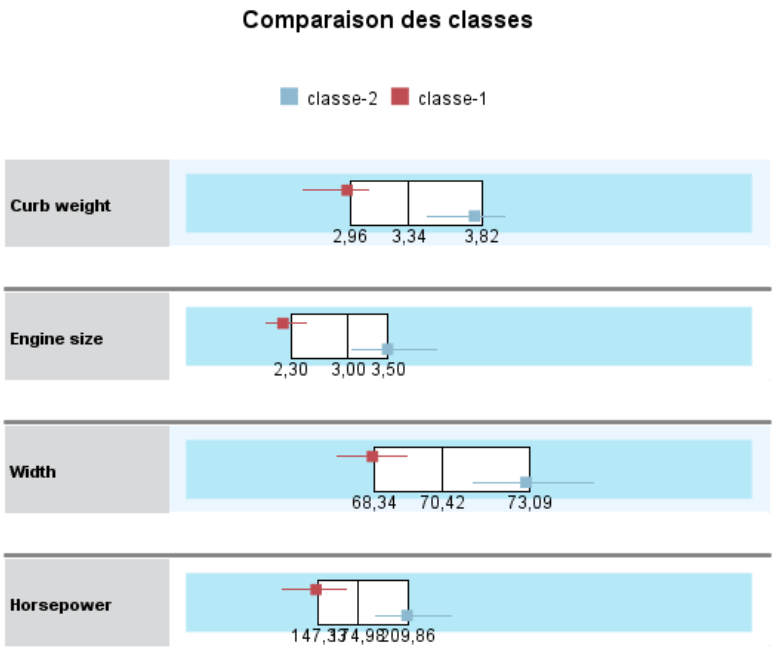
Figure 24-10
Vue de la distribution des cellules dans le panneau des liens



La vue de la distribution des cellules affiche un diagramme étendu et plus détaillé de la distribution des données pour toutes les cellules de caractéristique que vous sélectionnez dans le tableau du panneau principal des classes.

Vue de comparaison des classes

Figure 24-11
Vue de comparaison des classes dans le panneau des liens



La vue de comparaison des classes se compose d'une présentation sous forme de grille avec des caractéristiques dans les lignes et des classes sélectionnées dans les colonnes. Cette vue vous aide à mieux comprendre les facteurs qui composent les classes ; elle vous permet aussi de voir les différences entre les classes non seulement par comparaison avec les données générales mais aussi en comparant les classes les unes aux autres.

Pour sélectionner des classes à afficher, cliquez en haut de la colonne de classe sur le panneau principal des classes. Cliquez en maintenant la touche Ctrl ou Maj enfoncée pour sélectionner ou désélectionner plus d'une classe pour la comparaison.

Remarque : Vous pouvez sélectionner jusqu'à cinq classes à afficher.

Les classes sont affichées dans l'ordre où elles ont été sélectionnées, alors que l'ordre des champs est déterminé par l'option Trier les caractéristiques par. Lorsque vous sélectionnez Importance intra-classe, les champs sont toujours triés par ordre d'importance générale.

Les diagrammes d'arrière-plan affichent les distributions générales de chaque caractéristique :

- Les caractéristiques qualitatives sont affichées sous forme de tracés de points, où la taille des points indique la catégorie la plus fréquente/modale pour chaque classe (par caractéristique).
- Les caractéristiques continues seront affichées sous forme de boîtes à moustache, qui affichent les médianes générales et les intervalles interquartiles.

Des boîtes à moustache des classes sélectionnées recouvrent ces vues d'arrière-plan :

- Pour les caractéristiques continues, des marqueurs en points carrés et des lignes horizontales indiquent la médiane et l'intervalle interquartile de chaque classe.
- Chaque classe est représentée par une couleur différente, affichée en haut de la vue.

Navigation dans le viewer de classes

Le viewer de classes est un affichage interactif. Vous pouvez :

- sélectionner un champ ou une classe pour afficher davantage de détails ;
- comparer des classes afin de sélectionner des éléments intéressants ;
- modifier l'affichage ;
- transposer les axes.

Utilisation des barres d'outils

Vous contrôlez les informations affichées dans les panneaux de gauche et de droite à l'aide des options des barres d'outils. Vous pouvez modifier l'orientation de l'affichage (haut-bas, gauche-droite ou droite-gauche) à l'aide des commandes des barres d'outils. En outre, vous pouvez aussi réinitialiser le viewer à ses paramètres par défaut et ouvrir une boîte de dialogue pour spécifier le contenu de la vue des classes dans le panneau principal.

Figure 24-12
Barres d'outils pour contrôler les données affichées dans le viewer de classes

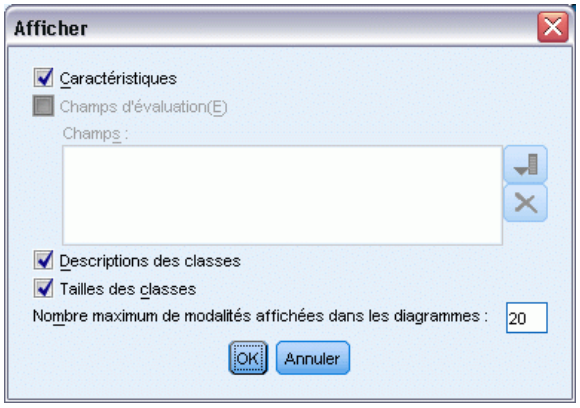
Les options Trier les caractéristiques par, Trier les classes par, Cellules et Affichage ne sont disponibles que lorsque vous sélectionnez la vue Classes du panneau principal. Pour plus d'informations, reportez-vous à la section Vue des classes sur p. 179.

	Reportez-vous à Transposer les classes et les caractéristiques sur p. 180
	Reportez-vous à Trier les caractéristiques par sur p. 181
	Reportez-vous à Trier les classes par sur p. 181
	Reportez-vous à Cellules sur p. 181

Contrôler l'affichage de la vue des classes

Pour contrôler ce qui est affiché dans la vue des classes sur le panneau principal, cliquez sur le bouton Afficher. La boîte de dialogue Afficher s'ouvre.

Figure 24-13
Viewer de classes - option d'affichage



Caractéristiques. Sélectionné par défaut. Pour masquer toutes les caractéristiques entrées, décochez la case.

Champs d'évaluation. Sélectionnez les champs d'évaluation à afficher (les champs qui ne sont pas utilisés pour créer le modèle de classe, mais envoyés au viewer de modèle pour évaluer les classes) ; par défaut, aucun n'est affiché. *Remarque* : Cette case à cocher n'est pas disponible si aucun champ d'évaluation n'est disponible.

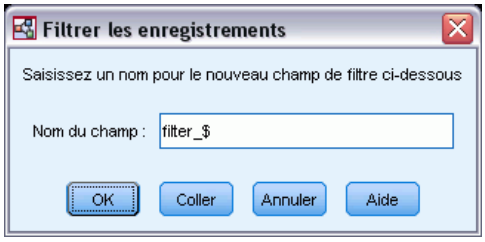
Descriptions des classes. Sélectionné par défaut. Pour masquer toutes les cellules de description des classes, décochez la case.

Tailles des classes. Sélectionné par défaut. Pour masquer toutes les cellules de taille des classes, décochez la case.

Nombre maximal de catégories. Spécifiez le nombre maximal de catégorie à afficher dans les diagrammes de caractéristiques qualitatives ; la valeur par défaut est de 20.

Filtrage des enregistrements

Figure 24-14
viewer de classes - Filtrer des observations



Si vous souhaitez en savoir plus sur les observations d'une classe particulière ou d'un groupe de classes, vous pouvez sélectionner un sous-ensemble d'enregistrements afin de poursuivre l'analyse en fonction des classes sélectionnées.

- Sélectionnez les classes dans la vue des classes du viewer de classes. Pour sélectionner plusieurs classes, cliquez tout en maintenant la touche Ctrl enfoncée.
- A partir des menus, sélectionnez :
  Générer > Filtrer des enregistrements...
- Entrez le nom d'une variable de filtre. Les enregistrements des classes sélectionnées reçoivent une valeur de 1 pour ce champ. Tous les autres enregistrements reçoivent une valeur de 0 et sont exclus des analyses ultérieures jusqu'à ce que vous modifiiez l'état du filtre.
- Cliquez sur OK.

Chapitre
25

Classification hiérarchique

Cette procédure tente d'identifier les classes d'observations (ou de variables) relativement homogènes basées sur des caractéristiques sélectionnées, en utilisant un algorithme qui débute avec chaque observation (ou variable) dans une classe séparée et qui combine les classes jusqu'à ce qu'il n'en reste qu'une. Vous pouvez analyser des variables non normées ou vous pouvez choisir parmi un assortiment de transformations standardisées. Les mesures de distance ou de similarité sont générées par la procédure Proximities (Proximités). Les statistiques s'affichent à chaque étape pour vous aider à choisir la meilleure solution.

Exemple : Y a-t-il des classes identifiables de spectacles télévisuels qui attirent des audiences similaires à l'intérieur de chaque classe ? Avec une classification hiérarchique, vous pouvez reclasser les spectacles télévisuels (observations) en classes homogènes basées sur les caractéristiques du spectateur. Cette méthode peut être utilisée pour identifier des segments à des fins commerciales. Vous pouvez aussi classer les villes (observations) en groupes homogènes pour permettre la sélection de villes comparables afin de tester diverses stratégies commerciales.

Statistiques : Chaîne des agrégations, matrice de distances (ou des similarités) et classe d'affectation pour une seule solution ou un ensemble de solutions. Diagrammes : arbres hiérarchiques et Stalactites.

Données. Les variables peuvent être des données quantitatives, binaires ou d'effectif. L'échelle des variables est un élément important : des différences d'échelle qui peuvent affecter votre (vos) solution(s) en classes hiérarchiques. Si vos variables sont d'échelles très différentes (par exemple, une variable est mesurée en dollars et l'autre est mesurée en années), vous devez envisager de les standardiser (ceci peut être fait automatiquement avec la procédure de la classification hiérarchique).

Tri par observation. Si des distances ex aequo ou des similitudes se présentent dans les données de saisie ou entre les classes mises à jour au cours de l'opération de jointure, la solution de classe qui en résulte risque de dépendre de l'ordre des observations dans le fichier. Vous pouvez obtenir différentes solutions pour lesquelles les observations ont été triées de manière aléatoire, afin de vérifier la stabilité d'une solution donnée.

Hypothèses : Les mesures de distance ou de similarité utilisées doivent convenir aux données analysées (Voir la procédure Proximities (proximités) pour plus de renseignements sur le choix des mesures de distances et de similarité). Vous devez aussi inclure toutes les variables appropriées dans votre analyse. L'omission de variables influentes peut aboutir à une solution erronée. Parce que la classification hiérarchique est une méthode d'exploration, les résultats doivent être considérés comme provisoires tant qu'ils ne sont pas confirmés avec un échantillon indépendant.

Obtenir une classification hiérarchique

- A partir des menus, sélectionnez :
  Analyse > Classification > Classification hiérarchique

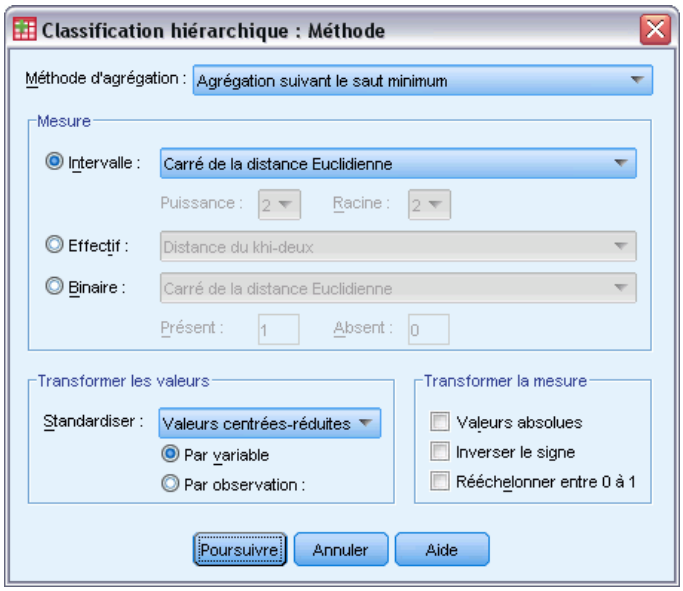
Figure 25-1
Boîte de dialogue Classification hiérarchique

- Si vous classez des observations, sélectionnez au moins une variable numérique. Si vous classez des variables, sélectionnez au moins trois variables numériques.

  Vous avez la possibilité de sélectionner une variable d'identification pour étiqueter les observations.

Méthode de classification hiérarchique

Figure 25-2
Boîte de dialogue Classification hiérarchique : Méthode



Méthode d'agrégation : Les choix disponibles sont : la Distance moyenne entre classes, la Distance moyenne dans les classes, l'Agrégation suivant le saut minimum, l'Agrégation suivant le diamètre, les Barycentres, la Médiane et la Méthode de Ward.

Mesure : Il permet de spécifier la mesure de distance ou de similarité devant être utilisée pour la classification. Sélectionnez le type de données et la mesure appropriée de distance ou de similarité :

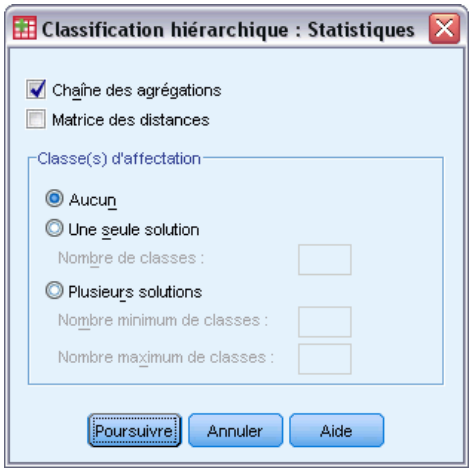
- **Intervalle :** Les choix possibles sont la Distance Euclidienne, le Carré de la distance Euclidienne, le Cosinus, la Corrélation de Pearson, la Distance de Tchebycheff, la Distance de Manhattan (bloc), la Distance de Minkowski, et Autre.
- **Effectif :** Les choix possibles sont la Distance du Khi-deux et la Distance du phi-deux.
- **Binaire :** Les choix possibles sont la Distance Euclidienne, le Carré de la distance Euclidienne, l'Ecart de taille, la Différence de motif, la Variance, la Dispersion, la Forme, l'Indice de Sokal et Michener, la Corrélation phi tétrachorique, le Lambda, le *D* d'Anderberg, Dice, Hamann, Jaccard, Kulczynski 1, Kulczynski 2, Lance et Williams, Ochiai, Rogers et Tanimoto, Russel et Rao, Sokal et Sneath 1, Sokal et Sneath 2, Sokal et Sneath 3, Sokal et Sneath 4, Sokal et Sneath 5, le *Y* de Yule, et le *Q* de Yule.

Transformer les valeurs : Vous permet de standardiser les valeurs des données pour les observations ou les valeurs avant le calcul des proximités (non disponible pour les données binaires). Les méthodes de standardisation disponibles sont *Centrer-réduire*, Entre −1 et 1, Entre 0 et 1, Maximum = 1, Moyenne = 1 ou Ecart type = 1.

mesures : Vous permet de transformer les valeurs générées par la mesure de distance. Elles sont appliquées après le calcul de la mesure d'indice. Les choix possibles sont Valeurs absolues, Inverser le signe, et Rééchelonner entre 0 et 1.

Statistiques de la classification hiérarchique

Figure 25-3
Boîte de dialogue Classification hiérarchique : Statistiques



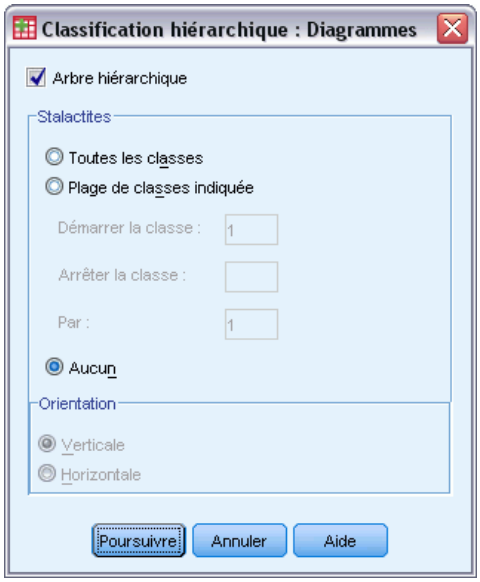
Chaîne des agrégations : Affiche les observations ou les classes combinées à chaque étape, les distances entre les observations ou les classes en cours de combinaison, et le dernier niveau de classe auquel une observation (ou une variable) a rejoint la classe.

Matrice des distances : Indique les distances ou les similarités entre éléments.

Classe(s) d'affectation : Affiche le groupe auquel chaque observation appartient lors d'une ou plusieurs étapes de la combinaison de classes. Les options disponibles sont Une seule partition, Plusieurs partitions ou Aucune.

Diagrammes (graphiques) de classification hiérarchique

Figure 25-4
Boîte de dialogue Classification hiérarchique : Graphiques (diagrammes)

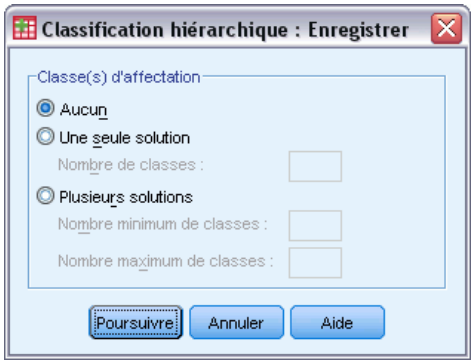


Arbre hiérarchique : Affiche un **dendrogramme**. Les arbres hiérarchiques peuvent être utilisés pour évaluer la cohésion des groupes formés et ils fournissent des renseignements sur le nombre approprié de groupes à conserver.

Stalactites : Affiche un **diagramme en stalactite**, incluant tous les groupes ou une plage de groupes spécifiée. Les diagrammes en stalactite affichent des informations sur la façon dont les observations sont regroupées à chaque itération de l'analyse. Orientation vous permet de sélectionner un diagramme vertical ou horizontal.

Sauvegarde des nouvelles variables de classification hiérarchique

Figure 25-5
Boîte de dialogue Classification hiérarchique : Enregistrer les nouvelles...



Classe(s) d'affectation : Vous permet de sauvegarder les classes d'affectation pour une ou plusieurs ou aucune partition(s). Les variables sauvegardées peuvent alors être utilisées pour des analyses ultérieures pour explorer d'autres différences entre groupes.

Fonctionnalités supplémentaires de la syntaxe de commande CLUSTER

La procédure de classification hiérarchique utilise la syntaxe de commande `CLUSTER`. Le langage de syntaxe de commande vous permet aussi de :

- Utiliser plusieurs méthodes de classification dans une seule analyse.
- Lire et analyser une matrice de proximité.
- Ecrire une matrice de proximité à analyser ultérieurement.
- Indiquer des valeurs pour la puissance et la racine dans la mesure de distance personnalisée (Puissance).
- Spécifier les noms des variables enregistrées.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
26

Nuées dynamiques

Cette procédure cherche à identifier des groupes d'observations relativement homogènes d'après des caractéristiques sélectionnées, au moyen d'un algorithme qui peut traiter de grands nombres d'observations. L'algorithme vous demande toutefois d'indiquer le nombre de classes. Vous pouvez indiquer les centres de classe initiaux si vous connaissez cette information. Vous pouvez choisir entre deux méthodes de classement des observations, soit la mise à jour des centres de classe de façon itérative, soit la classification seule. Vous pouvez enregistrer l'appartenance à une classe, les informations de distance et les centres de classes finaux. Vous pouvez éventuellement indiquer une variable dont les valeurs servent à étiqueter les résultats par observations. Vous pouvez également demander des statistiques *F* d'analyse de la variance. Bien que ces statistiques soient opportunistes (la procédure cherche à former des groupes qui diffèrent), la taille relative des statistiques fournit des informations sur la contribution de chaque variable à la séparation des groupes.

Exemple : Quels sont les groupes de programmes de télévision identifiables qui attirent des publics similaires au sein de chaque groupe ? Grâce à l'analyse des nuées dynamiques, vous pouvez classer les programmes de télévision (observations) en *k* groupes homogènes d'après les caractéristiques des téléspectateurs. Cette méthode peut être utilisée pour identifier des segments à des fins commerciales. Vous pouvez aussi classer les villes (observations) en groupes homogènes pour permettre la sélection de villes comparables afin de tester diverses stratégies commerciales.

Statistiques. Solution complète : centres de classes initiaux, tableau ANOVA. Chaque observation: information de classe, distance au centre de classe.

Données. Les variables doivent être quantitatives au niveau intervalle ou ratio. Si vos variables sont binaires ou sont des effectifs, utilisez la procédure de classification hiérarchique.

Ordre des observations et des centres de classe initiaux. L'algorithme par défaut permettant de choisir les centres de classe initiaux varie en fonction du tri par observation. L'option Utiliser les nouveaux centres de la boîte de dialogue Itérer rend la solution résultante potentiellement dépendante du tri par observation, quel que soit le mode de sélection des centres de classe initiaux. Si vous utilisez l'une de ces méthodes, vous pouvez obtenir différentes solutions pour lesquelles les observations ont été triées de manière aléatoire, afin de vérifier la stabilité d'une solution donnée. Si vous indiquez les centres de classe initiaux et que vous n'utilisez pas l'option Utiliser les nouveaux centres, vous évitez tout problème lié au tri par observation. Toutefois, le tri des centres de classe initiaux risque d'affecter la solution s'il existe des distances ex aequo entre les observations et les centres de classe. Pour évaluer la stabilité d'une solution donnée, vous pouvez comparer les résultats des analyses pour lesquelles les valeurs des centres initiaux ont été permutées de différentes manières.

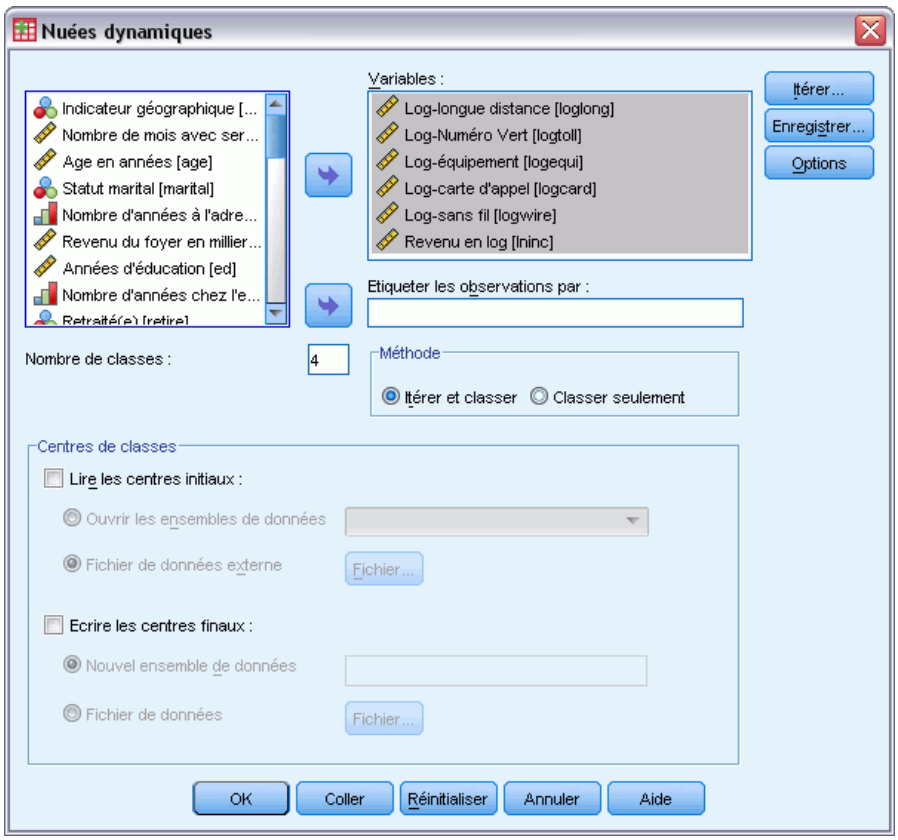
Hypothèses : Les distances sont calculées à l'aide de la distance euclidienne simple. Si vous souhaitez utiliser une autre distance ou une mesure de similarité, utilisez la procédure de classification hiérarchique. Il est important de prendre en compte la mise à l'échelle des variables. Si vos variables sont mesurées selon des échelles différentes (une variable est exprimée en dollars par exemple et une autre en années), vos résultats risquent d'être erronés. Dans ces

cas, vous pouvez envisager de standardiser vos variables avant d'effectuer l'analyse des *nuées dynamiques* (cela peut être fait dans la procédure Descriptives). La procédure suppose que vous avez sélectionné le nombre voulu de classes et que vous avez inclus toutes les variables pertinentes. Si vous avez choisi un nombre de classes inadéquat ou omis de variables importantes, vos résultats risquent d'être erronés.

Obtenir une analyse de nuées dynamiques

- A partir des menus, sélectionnez :
  Analyse > Classification > Nuées dynamiques

Figure 26-1
Boîte de dialogue Nuées dynamiques



- Sélectionnez les variables à utiliser dans l'analyse.
- Spécifiez le nombre de classes. Le nombre de classes doit être au moins de deux et ne doit pas être supérieur au nombre d'observations contenues dans le fichier de données.
- Sélectionnez soit la méthode Itérer et classer soit la méthode Classer seulement.
- Vous avez la possibilité de sélectionner une variable d'identification pour étiqueter les observations.

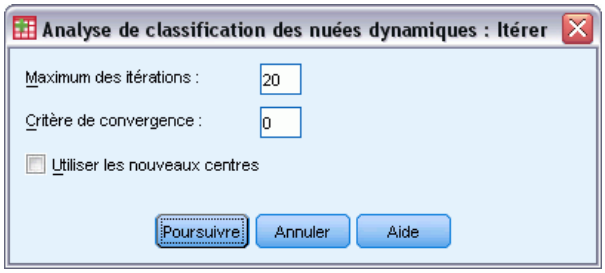
Efficacité de la classification en nuées dynamiques

La commande d'analyse des *nuées dynamiques* est efficace essentiellement parce qu'elle ne calcule pas les distances entre toutes les paires d'observations, comme c'est le cas dans de nombreux algorithmes de classification, y compris celui utilisé par la commande de classification hiérarchique.

Pour plus d'efficacité, prenez un échantillon d'observations et utilisez la méthode Itérer et classer pour déterminer les centres de classe. Sélectionnez Ecrire les centres finaux dans Fichier. Ensuite, restaurez la totalité du fichier de données et sélectionnez la méthodeClasser seulement puis sélectionnez Lire les fichiers initiaux à partir de pour classer tout le fichier en utilisant les centres qui sont estimés à partir de l'échantillon. Vous pouvez écrire et lire depuis un fichier ou un ensemble de données. Les ensembles de données sont disponibles pour utilisation ultérieure dans la même session mais ne sont pas enregistrés en tant que fichiers sauf si vous le faites explicitement avant la fin de la session. Le nom des ensembles de données doit être conforme aux règles de dénomination de variables.

Itération de la classification en nuées dynamiques

Figure 26-2
Boîte de dialogue Nuées dynamiques : Itérer



Remarque : Ces options sont disponibles uniquement si vous avez sélectionné la méthode Itérer et classer dans la boîte de dialogue Analyse de nuées dynamiques.

Maximum des itérations : Limite le nombre des itérations dans l'algorithme des nuées dynamiques. L'itération s'arrête après ce nombre d'itérations même si le critère de convergence n'est pas satisfait. Ce nombre doit être compris entre 1 et 999.

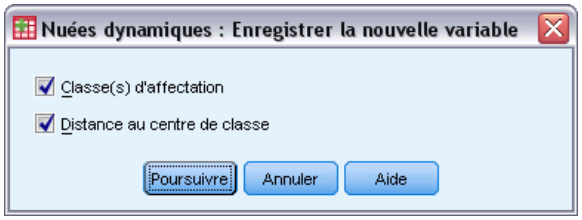
Pour reproduire l'algorithme utilisé par la commande Quick Cluster antérieure à la version 5.0, définissez Maximum des itérations sur 1.

Critère de convergence. Détermine le moment où l'itération s'arrête. Il représente une proportion de la distance minimale entre les centres de classes initiaux et doit donc être plus grand que 0 mais plus petit que 1. Si le critère est égal à 0,02 par exemple, l'itération cesse lorsqu'une itération complète ne déplace plus aucun des centres d'une distance de plus de deux pour cent de la plus petite distance entre n'importe quels centres initiaux.

Utiliser les nouveaux centres : Vous permet de demander la mise à jour des centres après l'affectation de chaque observation. Si vous ne sélectionnez pas cette option, les nouveaux centres seront calculés lorsque toutes les observations auront été affectées.

Enregistrement des analyses de classes de nuées dynamiques

Figure 26-3
Boîte de dialogue Nuées dynamiques : Enregistrer les nouvelles variables...



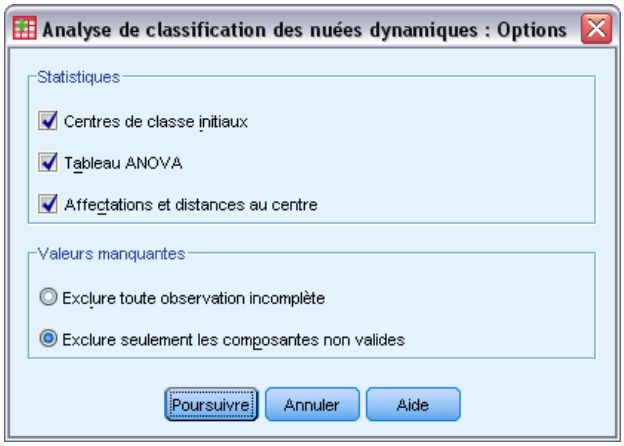
Vous pouvez enregistrer des informations sur la solution relatives aux nouvelles variables à utiliser dans les analyses ultérieures :

Classe(s) d'affectation : Crée une nouvelle variable indiquant la classe d'affectation finale de chaque observation. Les valeurs de la nouvelle variable vont de 1 au nombre de classes.

Distance au centre de classe : Crée une nouvelle variable indiquant la distance euclidienne entre chaque nouvelle variable et son centre de classification.

Options d'analyses des classes de nuées dynamiques

Figure 26-4
Boîte de dialogue Nuées dynamiques : Options



Statistiques. Vous pouvez sélectionner les statistiques suivantes : Centres de classes initiaux, Tableau ANOVA, et Affectation et distances au centre.

- **Centres de classe initiaux.** Première estimation des moyennes des variables de chacune des classes. Par défaut, le nombre d'observations assez espacées sélectionné dans les données est égal au nombre de classes. Les centres de classes initiaux sont utilisés pour une première classification et sont ensuite mis à jour.

- **Tableau ANOVA.** Affiche un tableau d'analyse de la variance, incluant les tests F univariés pour chacune des variables de la classification. Les tests F sont uniquement descriptifs et les probabilités qui en résultent ne doivent pas être interprétées. Le tableau ANOVA n'apparaît pas si toutes les observations sont affectées à une seule classe.
- **Affectations et distances au centre.** Affiche pour chaque observation l'affectation de classe finale et la distance euclidienne entre l'observation et le centre de classe utilisé pour classer l'observation. Affiche également la distance euclidienne entre les centres de classe finaux.

Valeurs manquantes : Les options disponibles sont Exclure toute observation incomplète ou Exclure seulement les composantes non valides.

- **Exclure toute observation incomplète :** Exclut de l'analyse les observations qui ont des valeurs manquantes pour une variable de grappe.
- **Exclure seulement les composantes non valides :** Affecte des observations aux classes basées sur des distances calculées à partir de toutes les variables n'ayant pas de valeur manquante.

Fonctionnalités supplémentaires de la commande QUICK CLUSTER

La procédure de nuées dynamiques utilise la syntaxe de commande `QUICK CLUSTER`. Le langage de syntaxe de commande vous permet aussi de :

- Accepter les premières observations *k* comme centres de classes initiaux, évitant ainsi le passage de données normalement utilisé pour les estimer.
- Spécifier les centres de classe initiaux directement comme faisant partie de la syntaxe de commande.
- Spécifier les noms des variables enregistrées.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
27

Tests non paramétriques

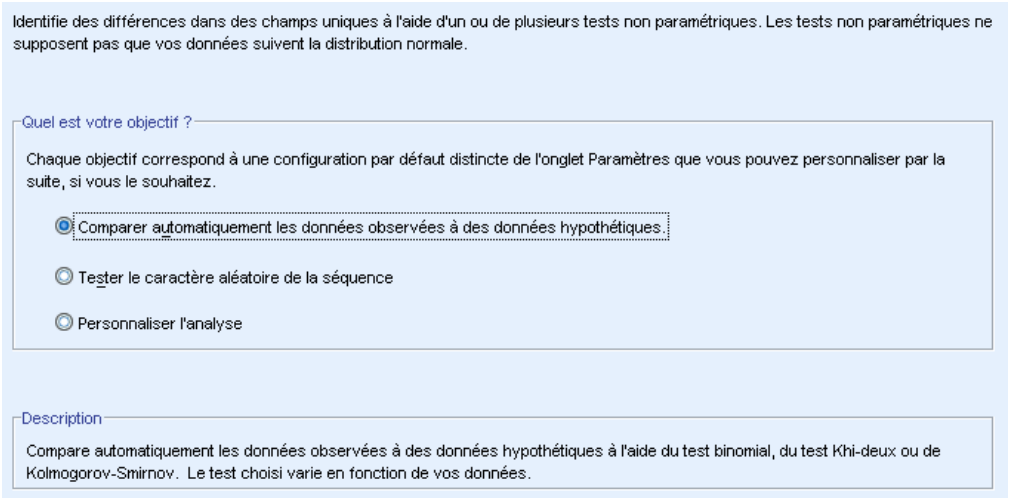
Les tests non paramétriques effectuent des suppositions minimales sur la distribution sous-jacente des données. Les tests disponibles dans ces boîtes de dialogue peuvent être groupés en trois grandes catégories basées sur la façon dont les données sont organisées :

- Un test à un échantillon analyse un champ.
- Un test pour échantillons liés compare deux champs ou plus pour le même ensemble d'observations.
- Un test pour échantillons indépendants analyse un champ qui est regroupé par modalités d'un autre champ.

Tests non paramétriques à un échantillon

Les tests non paramétriques à un échantillon identifient les différences dans les champs uniques en utilisant un ou plusieurs tests non paramétriques. Les tests non paramétriques ne supposent pas que vos données suivent la distribution normale.

Figure 27-1
Onglet Objectif des tests non paramétriques à un échantillon



Quel est votre objectif ? Les objectifs servent à spécifier rapidement des paramètres de test différents mais fréquemment utilisés.

- **Comparer automatiquement les données observées à des données hypothétiques.** Cet objectif applique le test binomial aux champs qualitatifs avec seulement deux modalités, le test du khi-deux à tous les autres champs qualitatifs et le test de Kolmogorov-Smirnov aux champs continus.

- **Tester le caractère aléatoire de la séquence.** Cet objectif utilise les suites en séquences pour tester la séquence de valeurs de données observées pour le caracère aléatoire.
- **Analyse personnalisée.** Lorsque vous souhaitez modifier manuellement les paramètres du test dans l'onglet Paramètres, sélectionnez cette option. Veuillez noter que ce paramètre est automatiquement sélectionné si vous modifiez ensuite des options dans l'onglet Paramètres qui ne sont pas compatibles avec l'objectif actuellement sélectionné.

Obtenir des tests non paramétriques à un échantillon

A partir des menus, sélectionnez :
Analyse > Tests non paramétriques > Un échantillon...

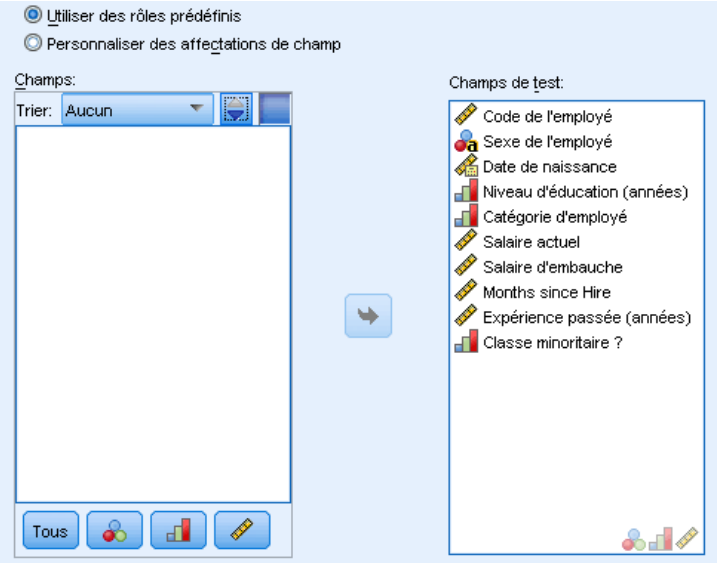
▶ Cliquez sur Exécuter.

Sinon, vous pouvez :

- Spécifiez un objectif dans l'onglet Objectif.
- spécifiez les affectations de champ dans l'onglet Champs.
- spécifiez les paramètres d'expert dans l'onglet Paramètres.

Onglet Champs

Figure 27-2
Onglet Champs des tests non paramétriques à un échantillon



L'onglet Champs indique les champs à tester.

Utiliser des rôles prédéfinis. Cette option utilise des informations sur des champs existants. Tous les champs avec un rôle prédéfini d'Entrée, Cible ou Les deux seront utilisés comme champs de test. Au moins un champ de test est requis.

Utiliser des affectations de champs personnalisées. Cette option permet de modifier les rôles des champs. Après avoir sélectionné cette option, spécifiez les champs ci-dessous :

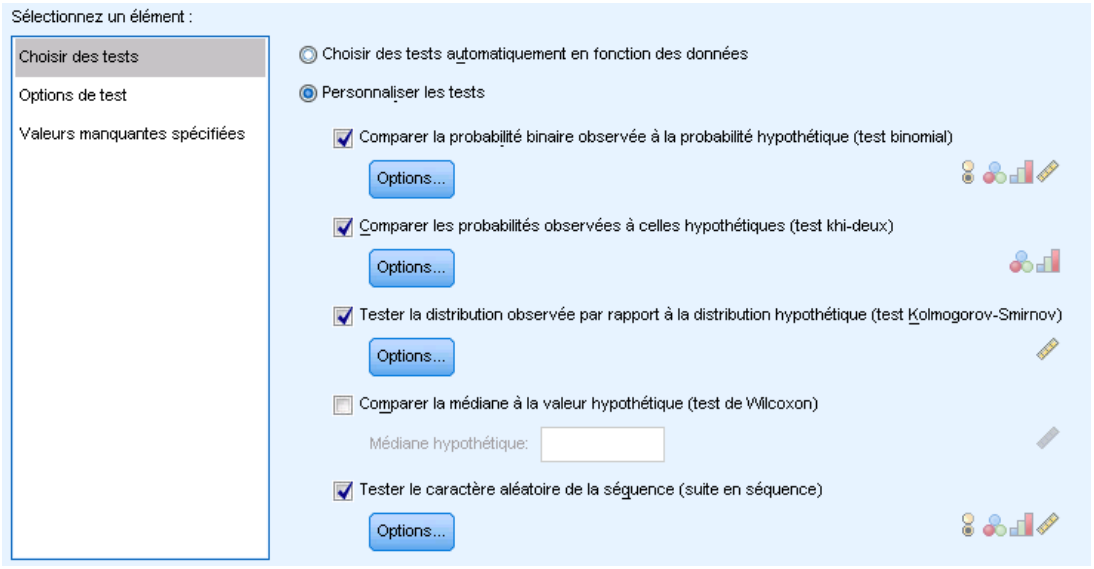
- **Champs de test.** Sélectionnez un ou plusieurs champs.

Onglet Paramètres

L'onglet Paramètres contient plusieurs groupes de paramètres différents que vous pouvez modifier pour affiner le traitement des données par l'algorithme. Si vous modifiez les paramètres par défaut et que ces modifications sont incompatibles avec l'objectif actuellement sélectionné, l'onglet Objectif est automatiquement mis à jour pour sélectionner l'option Personnaliser l'analyse.

Choisir les tests

Figure 27-3
Tests non paramétriques à un échantillon - Choisir les tests, paramètres



Ces paramètres indiquent les tests à effectuer sur les champs spécifiés dans l'onglet Champs.

Choisir automatiquement les tests en fonction des données. Ce paramètre applique le test binomial aux champs qualitatifs avec seulement deux modalités valides (non manquantes), le test du khi-deux à tous les autres champs qualitatifs et le test de Kolmogorov-Smirnov aux champs continus.

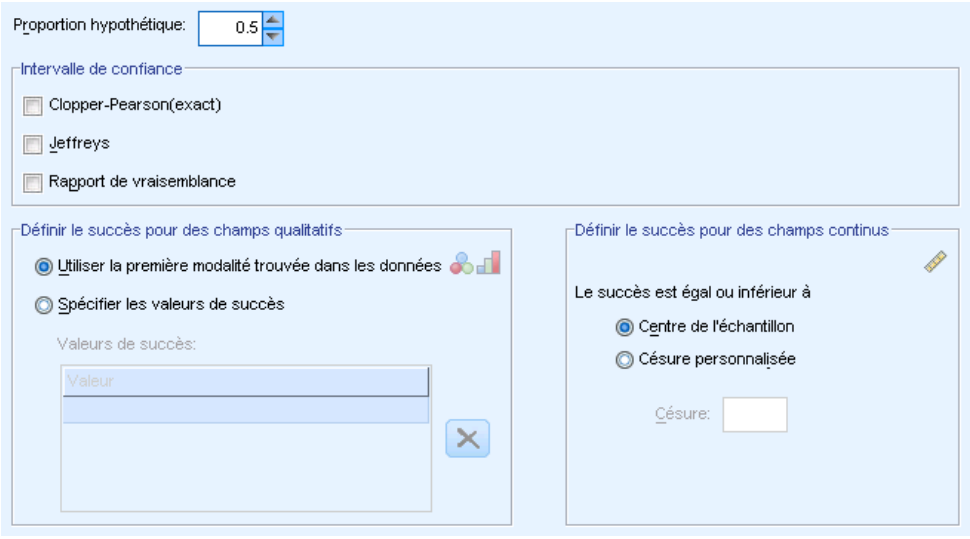
Personnaliser les tests. Ce paramètre permet de choisir des tests spécifiques à exécuter.

- **Comparer la probabilité binaire observée à la probabilité hypothétique (test binomial).** Le test binomial peut s'appliquer à tous les champs. Cela produit un test à un échantillon qui évalue si la distribution observée d'un champ booléen (un champ qualitatif avec seulement deux modalités) est semblable à ce qui est attendu d'une distribution binomiale spécifiée. De plus, vous pouvez demander des intervalles de confiance. Consultez Options du Test binomial pour obtenir des détails sur les paramètres de test.

- **Comparer les probabilités observées aux probabilités hypothétiques (test khi-deux).** Le test du Khi-deux s'applique aux champs nominaux et ordinaux. Cela produit un test à un échantillon qui calcule une statistique du Khi-deux basée sur les différences entre les fréquences observées et attendues des modalités d'un champ. Consultez Options du Test de Khi-deux pour obtenir des détails sur les paramètres de test.
- **Tester la distribution observée par rapport à la distribution hypothétique (test de Kolmogorov-Smirnov).** Le test de Kolmogorov-Smirnov s'applique aux champs continus. Cela produit un test à un échantillon évaluant si l'échantillon de la fonction de distribution cumulée d'un champ est homogène avec une distribution uniforme, normale, de Poisson ou exponentielle. Consultez Options Kolmogorov-Smirnov pour obtenir des détails sur les paramètres de test.
- **Comparer la médiane à la valeur hypothétique (test de Wilcoxon).** Le test de Wilcoxon s'applique aux champs continus. Cela produit un test à un échantillon de la valeur de la médiane d'un champ. Spécifier un nombre comme médiane hypothétique.
- **Tester le caractère aléatoire de la séquence (suite en séquence).** La suite en séquence s'applique à tous les champs. Cela produit un test à un échantillon évaluant si la séquence des valeurs d'un champ dichotomisé est aléatoire. Consultez Options Suites en séquence pour obtenir des détails sur les paramètres de test.

Options du Test binomial

Figure 27-4
Options du Test binomial des tests non paramétriques à un échantillon



Le test binomial est pour les champs booléens (champs qualitatifs avec seulement deux modalités) mais peut s'appliquer à tous les champs en utilisant des règles pour définir le « succès ».

Proportion hypothétique. Cela spécifie la proportion d'enregistrements attendue définie en tant que « succès » ou *p*. Définissez une valeur supérieure à 0 et inférieure à 1. La valeur par défaut est 0,5.

Intervalle de confiance : Les méthodes de calculs d'intervalles de confiance pour les données binaires suivantes sont disponibles :

- **Clopper-Pearson (exact).** Un intervalle exact basé sur la distribution binomiale cumulée.
- **Jeffreys.** Un intervalle bayésien basé sur la distribution postérieure de *p* utilisant la loi a priori de Jeffreys.
- **Rapport de vraisemblance.** Un intervalle basé sur la fonction de vraisemblance pour *p*.

Définir le succès pour des champs qualitatifs. Cela indique comment le « succès », la/les valeurs de données testée(s) par rapport à la proportion hypothétique, est défini pour les champs qualitatifs.

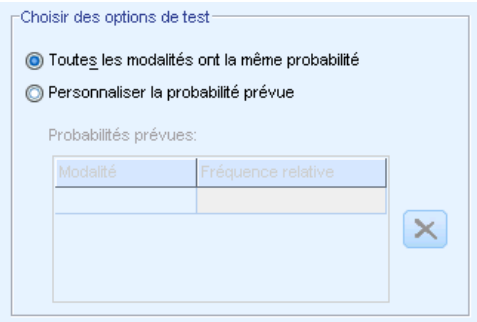
- Utiliser la première catégorie trouvée dans les données effectue le test binomial en utilisant la première valeur trouvée dans l'échantillon pour définir le « succès ». Cette option s'applique uniquement aux champs nominaux ou ordinaux avec seulement deux valeurs ; tous les autres champs qualitatifs spécifiés dans l'onglet Champs où cette option est utilisée, ne seront pas testés. Il s'agit de la valeur par défaut.
- Spécifier les valeurs de succès effectue le test binomial à l'aide d'une liste de valeurs spécifiée pour définir le « succès ». Spécifier une liste de chaîne ou des valeurs numériques. Les valeurs de la liste n'ont pas besoin d'être présentes dans l'échantillon.

Définir le succès pour des champs continus. Cela indique comment le « succès », la/les valeurs de données testée(s) par rapport à la valeur de test, est défini pour les champs continus. Le succès est défini comme des valeurs inférieures ou égales à la césure.

- Centre de l'échantillon définit la césure à la moyenne des valeurs minimales et maximales.
- Césure personnalisée permet de spécifier une valeur de césure.

Options du Test de Khi-deux

Figure 27-5
Options du Test de Khi-deux des tests non paramétriques à un échantillon



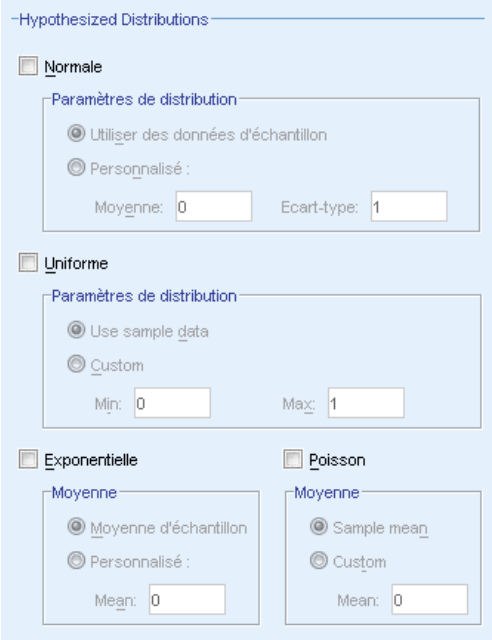
Toutes les modalités ont la même probabilité. Cela produit des fréquences égales pour toutes les modalités de l'échantillon. Il s'agit de la valeur par défaut.

Personnaliser la probabilité prévue. Cela vous permet de spécifier les fréquences inégales pour une liste de modalités spécifiée. Spécifier une liste de chaîne ou des valeurs numériques. Les valeurs de la liste n'ont pas besoin d'être présentes dans l'échantillon. Dans la colonne Modalité , spécifiez les valeurs des modalités. Dans la colonne Fréquence relative , spécifiez une valeur supérieure à 0 pour chaque modalité. Les fréquences personnalisées sont traitées comme des ratios. Par

exemple, spécifier des fréquences 1, 2 et 3 est l'équivalent de spécifier des fréquences 10, 20 et 30 et les deux spécifient que 1/6 des enregistrements doit se trouver dans la première modalité, 1/3 dans la seconde et 1/2 dans la troisième. Lorsque les probabilités attendues personnalisées sont spécifiées, les valeurs de modalités personnalisées doivent inclure toutes les valeurs de champ dans les données ; sinon, le test n'est pas exécuté pour ce champ.

Options Kolmogorov-Smirnov

Figure 27-6
Options Kolmogorov-Smirnov des tests non paramétriques à un échantillon



Cette boîte de dialogue indique les distributions à tester et les paramètres des distributions hypothétiques.

Normale. Utiliser des données d'échantillon utilise la moyenne et l'écart-type observés, Personnaliser vous permet de spécifier des valeurs.

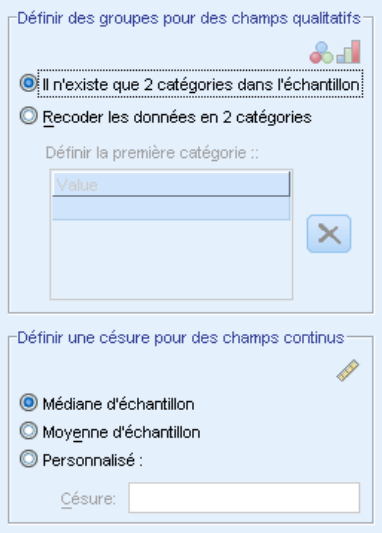
Uniforme. Utiliser des données d'échantillon utilise le minimum et le maximum observés, Personnaliser vous permet de spécifier des valeurs.

Exponentielle. Moyenne d'échantillon utilise la moyenne observée, Personnaliser vous permet de spécifier des valeurs.

Poisson. Moyenne d'échantillon utilise la moyenne observée, Personnaliser vous permet de spécifier des valeurs.

Options Suites en séquence

Figure 27-7
Options Suites en séquence des tests non paramétriques à un échantillon



Les suites en séquence sont pour les champs booléens (champs qualitatifs avec seulement deux modalités) mais peuvent être appliquées à tous les champs en utilisant des règles pour définir les groupes.

Définir des groupes pour des champs qualitatifs

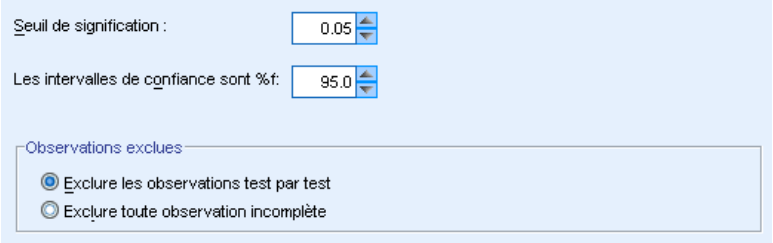
- Il n'existe que 2 catégories dans l'échantillon effectue les suites en séquence en utilisant des valeurs trouvées dans l'échantillon pour définir les groupes. Cette option s'applique uniquement aux champs nominaux ou ordinaux avec seulement deux valeurs ; tous les autres champs qualitatifs spécifiés dans l'onglet Champs où cette option est utilisée, ne seront pas testés.
- Recoder les données en 2 catégories effectue les suites en séquence à l'aide de la liste de valeurs spécifiée pour définir un des groupes. Toutes les autres valeurs de l'échantillon définissent l'autre groupe. Les valeurs de la liste n'ont pas besoin d'être présentes dans l'échantillon, mais au moins un enregistrement doit se trouver dans chaque groupe.

Définir une césure pour des champs continus Cela spécifie la façon dont les groupes sont définis pour les champs continus. Le premier groupe est défini comme comprenant des valeurs inférieures ou égales à la césure.

- Médiane d'échantillon définit la césure à la médiane d'échantillon.
- Moyenne d'échantillon définit la césure à la moyenne d'échantillon.
- Personnalisé permet de spécifier une valeur de césure.

Options de test

Figure 27-8
Tests non paramétriques à un échantillon - Option de tests, paramètres



Niveau de signification : Indique le niveau de signification (alpha) pour tous les tests. Spécifier une valeur numérique entre 0 et 1. 0,05 est la valeur par défaut.

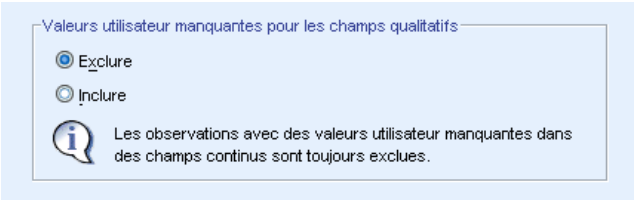
Intervalle de confiance (%). Indique le niveau de confiance pour tous les intervalles de confiance générés. Spécifier une valeur numérique entre 0 et 100. 95 est la valeur par défaut.

Observations exclues. Indique comment déterminer la base des observations pour les tests.

- Exclure toute observation incomplète signifie que les enregistrements avec des valeurs manquantes dans les champs qui sont nommés dans l'onglet Champs sont exclus de toutes les analyses.
- Exclure les observations test par test signifie que les enregistrements avec des valeurs manquantes dans un champ utilisé pour un test spécifique sont ignorés pendant ce test. Lorsque plusieurs tests sont spécifiés dans l'analyse, chaque test est évalué séparément.

Valeurs manquantes spécifiées

Figure 27-9
Tests non paramétriques à un échantillon - Valeurs manquantes spécifiées, paramètres

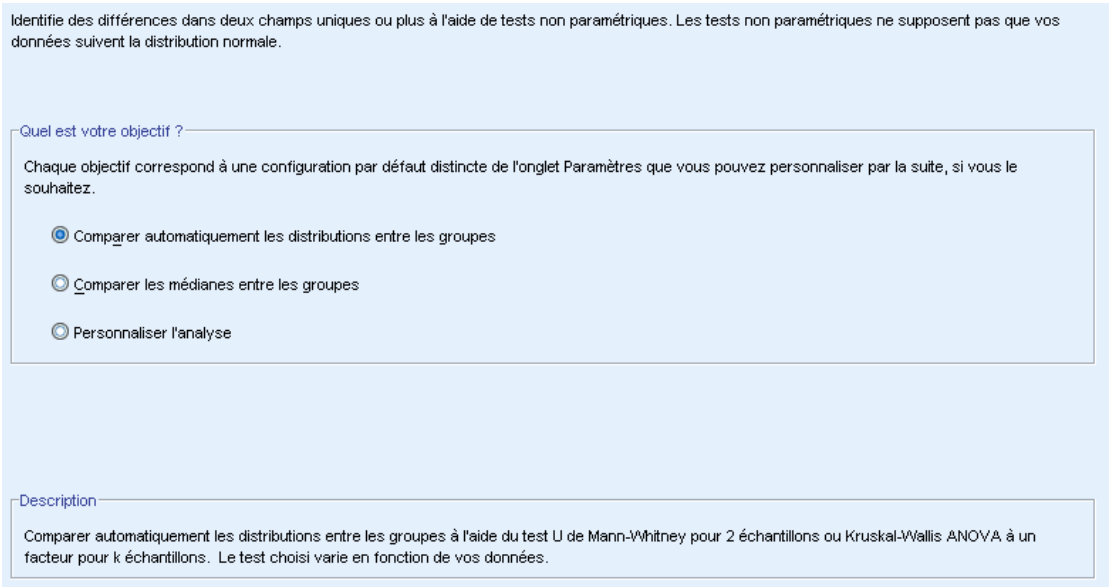


Valeurs utilisateur manquantes pour les champs qualitatifs. Les champs qualitatifs doivent avoir des valeurs valides pour qu'un enregistrement puisse être inclus dans l'analyse. Ces commandes vous permettent d'indiquer si les valeurs manquantes spécifiées sont considérées comme valides parmi les champs qualitatifs. Les valeurs manquantes par défaut et les valeurs manquantes pour les champs continus sont toujours considérées comme non valides.

Tests non paramétriques pour échantillons indépendants

Les tests non paramétriques pour échantillons indépendants identifient les différences entre deux groupes ou plus à l'aide d'un ou plusieurs tests non paramétriques. Les tests non paramétriques ne supposent pas que vos données suivent la distribution normale.

Figure 27-10
Onglet Objectif des tests non paramétriques pour échantillons indépendants



Quel est votre objectif ? Les objectifs servent à spécifier rapidement des paramètres de test différents mais fréquemment utilisés.

- **Comparer automatiquement les distributions entre les groupes.** Cet objectif applique le test U de Mann-Whitney aux données avec 2 groupes ou le Kruskal-Wallis ANOVA à un facteur aux données avec *k* groupes.
- **Comparer les médianes entre les groupes.** Cet objectif utilise le test de la médiane pour comparer les médianes observées entre les groupes.
- **Analyse personnalisée.** Lorsque vous souhaitez modifier manuellement les paramètres du test dans l'onglet Paramètres, sélectionnez cette option. Veuillez noter que ce paramètre est automatiquement sélectionné si vous modifiez ensuite des options dans l'onglet Paramètres qui ne sont pas compatibles avec l'objectif actuellement sélectionné.

Obtenir des tests non paramétriques pour échantillons indépendants

A partir des menus, sélectionnez :
Analyse > Tests non paramétriques > Echantillons indépendants...

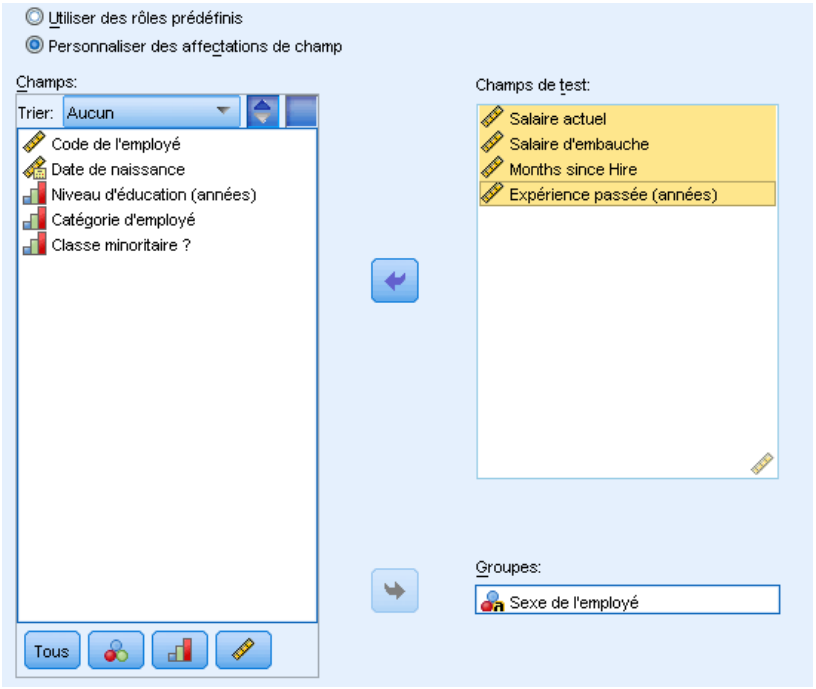
▶ Cliquez sur Exécuter.

Sinon, vous pouvez :

- Spécifiez un objectif dans l'onglet Objectif.
- spécifiez les affectations de champ dans l'onglet Champs.
- spécifiez les paramètres d'expert dans l'onglet Paramètres.

Onglet Champs

Figure 27-11
Onglet Champs des tests non paramétriques pour échantillons indépendants



L'onglet Champs indique les champs à tester et le champ utilisé pour définir les groupes.

Utiliser des rôles prédéfinis. Cette option utilise des informations sur des champs existants. Tous les champs continus avec un rôle prédéfini de Cible ou Les deux seront utilisés comme champs de test. Si un champ unique qualitatif avec un rôle prédéfini d'Entrée est disponible, il sera utilisé comme champ de regroupement. Sinon, il n'y aura pas d'utilisation par défaut de champ de regroupement et vous devrez utiliser des affectations de champs personnalisées. Au moins un champ de test et un champ de regroupement sont requis.

Utiliser des affectations de champs personnalisées. Cette option permet de modifier les rôles des champs. Après avoir sélectionné cette option, spécifiez les champs ci-dessous :

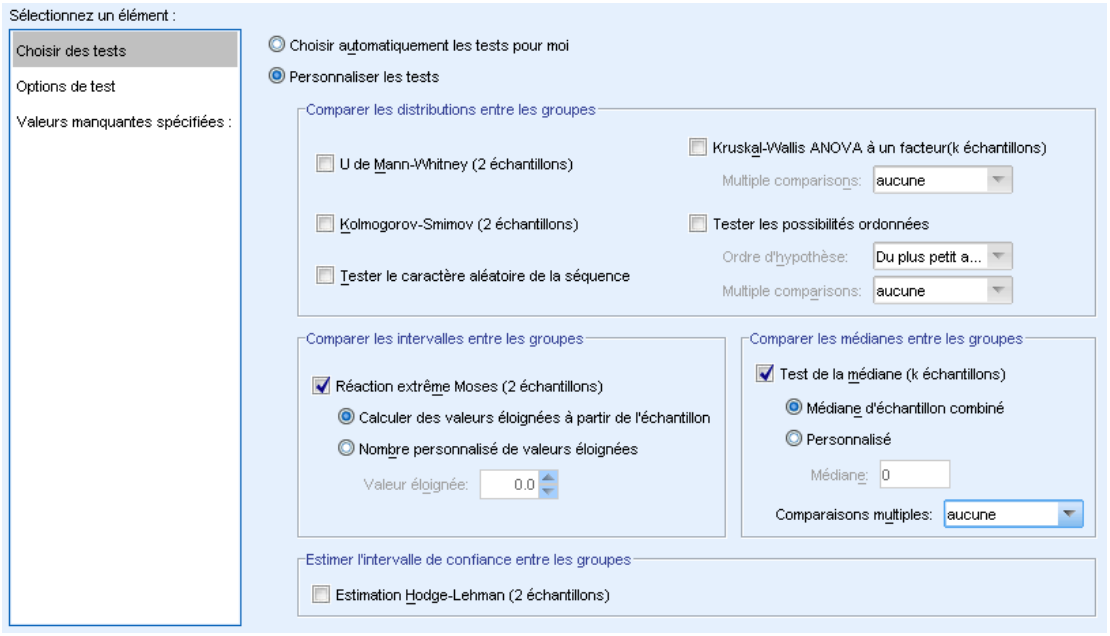
- **Champs de test.** Sélectionnez un ou plusieurs champs continus.
- **Groupes.** Sélectionnez un champ qualitatif.

Onglet Paramètres

L'onglet Paramètres contient plusieurs groupes de paramètres différents que vous pouvez modifier pour affiner le traitement des données par l'algorithme. Si vous modifiez les paramètres par défaut et que ces modifications sont incompatibles avec l'objectif actuellement sélectionné, l'onglet Objectif est automatiquement mis à jour pour sélectionner l'option Personnaliser l'analyse.

Choisir les tests

Figure 27-12
Tests non paramétriques pour échantillons indépendants - Choisir les tests, paramètres



Ces paramètres indiquent les tests à effectuer sur les champs spécifiés dans l'onglet Champs.

Choisir automatiquement les tests en fonction des données. Ce paramètre applique le test U de Mann-Whitney aux données avec 2 groupes ou le Kruskal-Wallis ANOVA à un facteur aux données avec *k* groupes.

Personnaliser les tests. Ce paramètre permet de choisir des tests spécifiques à exécuter.

- **Comparer les distributions entre les groupes.** Ces paramètres produisent des tests pour échantillons indépendants indiquant si les échantillons sont issus de la même population.

  Mann-Whitney U (2 échantillons) utilise le rang de chaque observation pour tester si les groupes sont issus de la même population. La première valeur dans l'ordre croissant du champ de regroupement définit le premier groupe et la deuxième valeur définit le deuxième groupe. Si le champ de regroupement a plus de deux valeurs, ce test n'est pas généré.

  Kolmogorov-Smirnov (2 échantillons) est sensible aux différences de médiane, de dispersion, d'asymétrie, etc. entre les deux distributions. Si le champ de regroupement a plus de deux valeurs, ce test n'est pas généré.

  Tester le caractère aléatoire de la séquence (Wald-Wolfowitz pour 2 échantillons) génère des suites en séquence avec l'appartenance au groupe comme critère. Si le champ de regroupement a plus de deux valeurs, ce test n'est pas généré.

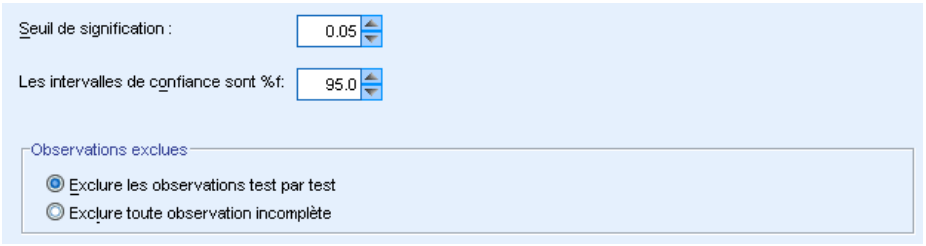
  Kruskal-Wallis ANOVA à un facteur (k échantillons) est une extension du test U de Mann-Whitney et l'équivalent non paramétrique de l'analyse de variance à un facteur. En option, vous pouvez demander des comparaisons multiples des *k* échantillons, soit toutes les comparaisons multiples par paire soit les comparaisons pas à pas descendantes .

Tester les possibilités ordonnées (Jonckheere-Terpstra pour k échantillons) est une alternative à Kruskal-Wallis plus puissante lorsque les *k* échantillons ont un ordre naturel. Par exemple, les *k* populations peuvent représenter *k* températures croissantes. L'hypothèse selon laquelle différentes températures produisent la même distribution des réponses est testée contre l'hypothèse alternative selon laquelle l'accroissement de température fait augmenter la magnitude de la réponse. Ici, l'hypothèse alternative est ordonnée ; le test de Jonckheere-Terpstra est donc le plus approprié. Spécifier l'ordre des hypothèses alternatives ; De la plus petite à la plus grande indique une hypothèse alternative stipulant que le paramètre d'emplacement du premier groupe n'est pas égal au deuxième qui lui-même n'est pas égal au troisième, etc. ; De la plus grande à la plus petite indique une hypothèse alternative stipulant que le paramètre d'emplacement du premier groupe n'est pas égal au second, qui est lui-même n'est pas égal au troisième, etc. En option, vous pouvez demander des comparaisons multiples des *k* échantillons, soit toutes les comparaisons multiples par paire soit les comparaisons pas à pas descendantes .

- **Comparer les intervalles entre les groupes.** Cela génère des tests pour échantillons indépendants indiquant si les échantillons ont le même intervalle. La Réaction extrême de Moses (2 échantillons) teste un groupe de commandes par rapport à un groupe de comparaisons. La première valeur dans l'ordre croissant du champ de regroupement définit le groupe de commandes et la deuxième valeur définit le groupe de comparaisons. Si le champ de regroupement a plus de deux valeurs, ce test n'est pas généré.
- **Comparer les médianes entre les groupes.** Cela génère des tests pour échantillons indépendants indiquant si les échantillons ont la même médiane. Le Test de la médiane (k échantillons) peut utiliser soit la médiane d'échantillon combiné (calculée à partir de tous les enregistrements de l'ensemble de données) ou une valeur personnalisée comme la médiane hypothétique. En option, vous pouvez demander des comparaisons multiples des *k* échantillons, soit toutes les comparaisons multiples par paire soit les comparaisons pas à pas descendantes .
- **Estimer les intervalles de confiance entre les groupes.** L'estimation de Hodges-Lehman (2 échantillons) génère une estimation d'échantillons indépendants et un intervalle de confiance pour la différence entre les médianes des deux groupes. Si le champ de regroupement a plus de deux valeurs, ce test n'est pas généré.

Options de test

Figure 27-13
Tests non paramétriques pour échantillons indépendants - Options de test, paramètres



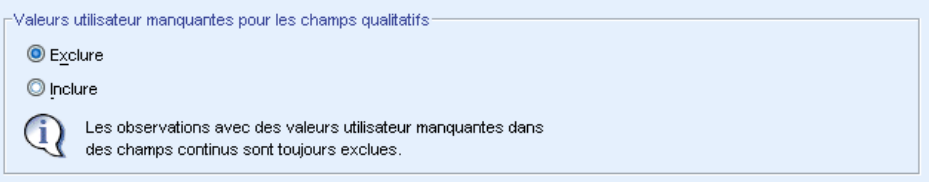
Niveau de signification : Indique le niveau de signification (alpha) pour tous les tests. Spécifier une valeur numérique entre 0 et 1. 0,05 est la valeur par défaut.

Intervalle de confiance (%). Indique le niveau de confiance pour tous les intervalles de confiance générés. Spécifier une valeur numérique entre 0 et 100. 95 est la valeur par défaut.

Observations exclues. Indique comment déterminer la base des observations pour les tests. Exclure toute observation incomplète signifie que les enregistrements avec des valeurs manquantes dans les champs nommés dans une sous-commande sont exclus de toutes les analyses. Exclure les observations test par test signifie que les enregistrements avec des valeurs manquantes dans un champ utilisé pour un test spécifique sont ignorés pendant ce test. Lorsque plusieurs tests sont spécifiés dans l'analyse, chaque test est évalué séparément.

Valeurs manquantes spécifiées

Figure 27-14
Tests non paramétriques pour échantillons indépendants - Valeurs manquantes spécifiées, paramètres



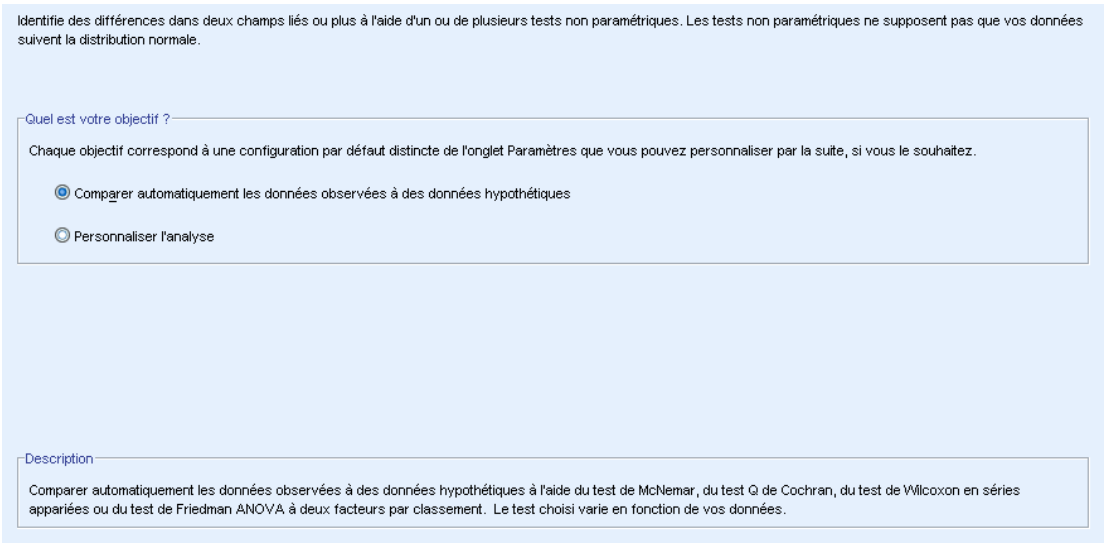
Valeurs utilisateur manquantes pour les champs qualitatifs. Les champs qualitatifs doivent avoir des valeurs valides pour qu'un enregistrement puisse être inclus dans l'analyse. Ces commandes vous permettent d'indiquer si les valeurs manquantes spécifiées sont considérées comme valides parmi les champs qualitatifs. Les valeurs manquantes par défaut et les valeurs manquantes pour les champs continus sont toujours considérées comme non valides.

Tests non paramétriques pour échantillons liés

Identifie des différences entre deux champs liés ou plus à l'aide d'un ou de plusieurs tests non paramétriques. Les tests non paramétriques ne supposent pas que vos données suivent la distribution normale.

Analyse des données. Chaque enregistrement correspond à un sujet donné pour lequel deux mesures associées ou plus sont stockées dans des champs distincts de l'ensemble de données. Par exemple, une étude concernant l'efficacité d'un régime peut être analysée à l'aide de tests d'échantillons liés non paramétriques, si le poids de chaque sujet est mesuré à intervalles réguliers et stocké dans des champs tels que *Poids avant le régime*, *Poids intermédiaire*, et *Poids après le régime*. Ces champs sont " liés ".

Figure 27-15
Onglet Objectif des tests non paramétriques pour échantillons liés



Quel est votre objectif ? Les objectifs servent à spécifier rapidement des paramètres de test différents mais fréquemment utilisés.

- **Comparer automatiquement les données observées à des données hypothétiques.** Cet objectif applique le test de McNemar aux donnés qualitatives lorsque 2 champs sont spécifiés, le Q de Cochran aux données qualitatives lorsque plus de 2 champs sont spécifiés, le test de Wilcoxon en séries appariées aux données continues lorsque 2 champs sont spécifiés et le test de Friedman ANOVA à deux facteurs par classement aux données continues lorsque plus de 2 champs sont spécifiés.
- **Analyse personnalisée.** Lorsque vous souhaitez modifier manuellement les paramètres du test dans l'onglet Paramètres, sélectionnez cette option. Veuillez noter que ce paramètre est automatiquement sélectionné si vous modifiez ensuite des options dans l'onglet Paramètres qui ne sont pas compatibles avec l'objectif actuellement sélectionné.

Lorsque des champs de niveaux de mesure différents sont spécifiés, ils sont séparés en fonction de leur niveau de mesure puis le test approprié est appliqué à chaque groupe. Par exemple, si vous choisissez Comparer automatiquement les données observées à des données hypothétiques comme objectif et spécifiez 3 champs continus et 2 champs nominaux, le test de Friedman est appliqué aux champs continus et le test de McNemar est appliqué aux champs nominaux.

Obtenir des tests non paramétriques pour échantillons liés

A partir des menus, sélectionnez :
Analyse > Tests non paramétriques > Echantillons liés...

► Cliquez sur Exécuter.

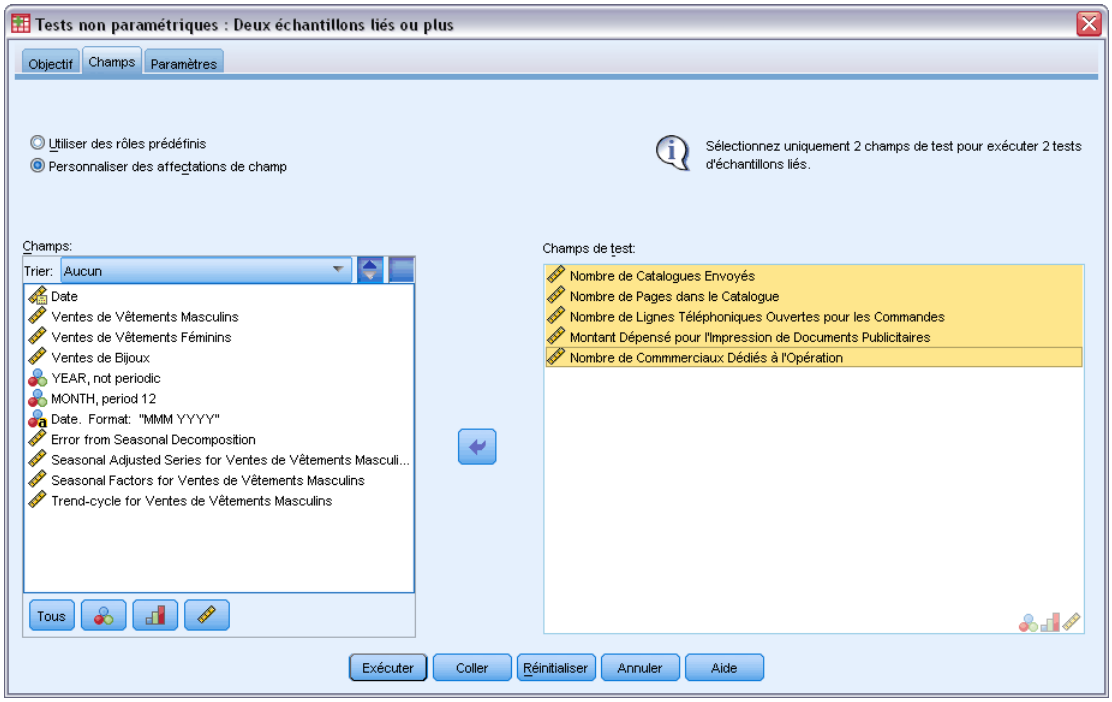
Sinon, vous pouvez :

- Spécifiez un objectif dans l'onglet Objectif.

- spécifiez les affectations de champ dans l'onglet Champs.
- spécifiez les paramètres d'expert dans l'onglet Paramètres.

Onglet Champs

Figure 27-16
Onglet Champs des tests non paramétriques pour échantillons liés



L'onglet Champs indique les champs à tester.

Utiliser des rôles prédéfinis. Cette option utilise des informations sur des champs existants. Tous les champs avec un rôle prédéfini de Cible ou Les deux seront utilisés comme champs de test. Au moins deux champs de test sont requis.

Utiliser des affectations de champs personnalisées. Cette option permet de modifier les rôles des champs. Après avoir sélectionné cette option, spécifiez les champs ci-dessous :

- **Champs de test.** Sélectionnez deux ou plusieurs champs. Chaque champ correspond à un échantillon lié séparé.

Onglet Paramètres

L'onglet Paramètres contient plusieurs groupes de paramètres différents que vous pouvez modifier pour affiner le traitement des données par la procédure. Si vous modifiez les paramètres par défaut et que ces modifications sont incompatibles avec les autres objectifs, l'onglet Objectif est automatiquement mis à jour pour sélectionner l'option Personnaliser l'analyse.

Choisir les tests

Figure 27-17
Paramètres des tests Choisir les tests non paramétriques pour échantillons liés

Ces paramètres indiquent les tests à effectuer sur les champs spécifiés dans l'onglet Champs.

Choisir automatiquement les tests en fonction des données. Ce paramètre applique le test de McNemar aux donnés qualitatives lorsque 2 champs sont spécifiés, le Q de Cochran aux données qualitatives lorsque plus de 2 champs sont spécifiés, le test de Wilcoxon en séries appariées aux données continues lorsque 2 champs sont spécifiés et le test de Friedman ANOVA à deux facteurs par classement aux données continues lorsque plus de 2 champs sont spécifiés.

Personnaliser les tests. Ce paramètre permet de choisir des tests spécifiques à exécuter.

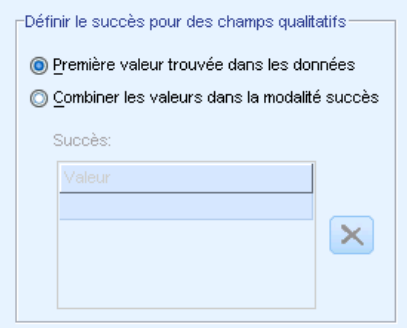
- **Tester les modifications au sein des données binaires.** Le test de McNemar (2 échantillons) peut être appliqué aux champs qualitatifs. Cela produit un test pour échantillons liés évaluant si les combinaisons de valeurs entre deux champs booléens (champs qualitatifs avec seulement deux valeurs) sont aussi probables l'une que l'autre. S'il existe plus de deux champs spécifiés dans l'onglet Champs, ce test n'est pas effectué. Consultez Test de McNemar : Définir le succès pour obtenir des détails sur les paramètres de test. Le Q de Cochran (k échantillons) peut être appliqué aux champs qualitatifs. Cela produit un test pour échantillons liés évaluant si les combinaisons de valeurs entre *k* champs booléens (champs qualitatifs avec seulement deux valeurs) sont aussi probables l'une que l'autre. En option, vous pouvez demander des comparaisons multiples des *k* échantillons, soit toutes les comparaisons multiples par paire soit les comparaisons pas à pas descendantes . Consultez Q de Cochran : Définir le succès pour obtenir des détails sur les paramètres de test.
- **Tester les modifications au sein des données multinomiales.** Le Test de l'homogénéité marginale (2 échantillons) génère un test d'échantillons liés évaluant si des combinaisons de valeurs entre deux champs ordinaux appariés sont aussi probables l'une que l'autre. Le test d'homogénéité marginale est généralement utilisé dans les cas avec des mesures répétées. Ce test est un

développement du test de McNemar d'une réponse binaire à une réponse multinomiale. S'il existe plus de deux champs spécifiés dans l'onglet Champs, ce test n'est pas effectué.

- **Comparer la différence de médiane à la différence de médiane hypothétique.** Chacun de ces tests génère un test pour échantillons liés évaluant si la différence de médiane entre deux champs continus est différente de 0. S'il y a plus de deux champs spécifiés dans l'onglet Champs, ces tests ne sont pas effectués.
- **Estimer l'intervalle de confiance.** Cela génère une estimation et un intervalle de confiance d'échantillons liés pour la différence de médiane entre deux champs continus appariés. S'il existe plus de deux champs spécifiés dans l'onglet Champs, ce test n'est pas effectué.
- **Quantifier les associations.** Le coefficient de concordance de Kendall (k échantillons) génère une mesure d'accord entre les juges ou les indicateurs, où chaque enregistrement est le classement par un juge de plusieurs éléments (champs). En option, vous pouvez demander des comparaisons multiples des *k* échantillons, soit toutes les comparaisons multiples par paire soit les comparaisons pas à pas descendantes .
- **Comparer les distributions.** Le Test de Friedman ANOVA à deux facteurs par classement (k échantillons) génère un test d'échantillons liés évaluant si *k* échantillons liés sont issus de la même population. En option, vous pouvez demander des comparaisons multiples des *k* échantillons, soit toutes les comparaisons multiples par paire soit les comparaisons pas à pas descendantes .

Test de McNemar : Définir le succès

Figure 27-18
Test de McNemar des tests non paramétriques pour échantillons liés : Paramètres Définir le succès



Le test de McNemar est pour les champs booléens (champs qualitatifs avec seulement deux catégories) mais peut s'appliquer à tous les champs qualitatifs en utilisant des règles pour définir le « succès ».

Définir le succès pour des champs qualitatifs. Cela indique comment le « succès » est défini pour les champs qualitatifs.

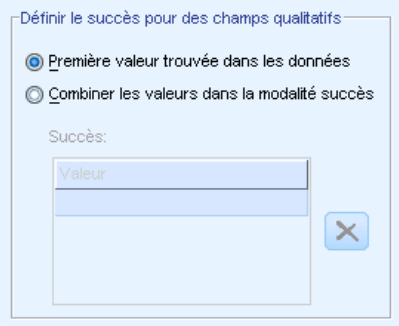
- Utiliser la première catégorie trouvée dans les données effectue le test en utilisant la première valeur trouvée dans l'échantillon pour définir le « succès ». Cette option s'applique uniquement aux champs nominaux ou ordinaux avec seulement deux valeurs ; tous les autres

champs qualitatifs spécifiés dans l'onglet Champs où cette option est utilisée, ne seront pas testés. Il s'agit de la valeur par défaut.

- Spécifier les valeurs de succès effectue le test à l'aide d'une liste de valeurs spécifiée pour définir le « succès ». Spécifier une liste de chaîne ou des valeurs numériques. Les valeurs de la liste n'ont pas besoin d'être présentes dans l'échantillon.

Q de Cochran : Définir le succès

Figure 27-19
Q de Cochran des tests non paramétriques pour échantillons liés : Définir le succès



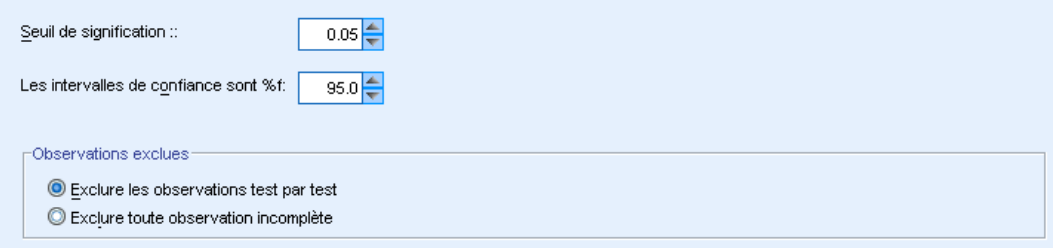
Le test Q de Coachran est pour les champs booléens (champs qualitatifs avec seulement deux catégories) mais peut s'appliquer à tous les champs qualitatifs en utilisant des règles pour définir le « succès ».

Définir le succès pour des champs qualitatifs. Cela indique comment le « succès » est défini pour les champs qualitatifs.

- Utiliser la première catégorie trouvée dans les données effectue le test en utilisant la première valeur trouvée dans l'échantillon pour définir le « succès ». Cette option s'applique uniquement aux champs nominaux ou ordinaux avec seulement deux valeurs ; tous les autres champs qualitatifs spécifiés dans l'onglet Champs où cette option est utilisée, ne seront pas testés. Il s'agit de la valeur par défaut.
- Spécifier les valeurs de succès effectue le test à l'aide d'une liste de valeurs spécifiée pour définir le « succès ». Spécifier une liste de chaîne ou des valeurs numériques. Les valeurs de la liste n'ont pas besoin d'être présentes dans l'échantillon.

Options de test

Figure 27-20
Paramètres Options de test des Tests non paramétriques pour échantillons liés



Niveau de signification : Indique le niveau de signification (alpha) pour tous les tests. Spécifier une valeur numérique entre 0 et 1. 0,05 est la valeur par défaut.

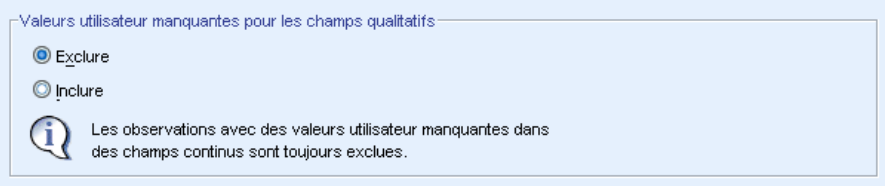
Intervalle de confiance (%). Indique le niveau de confiance pour tous les intervalles de confiance générés. Spécifier une valeur numérique entre 0 et 100. 95 est la valeur par défaut.

Observations exclues. Indique comment déterminer la base des observations pour les tests.

- Exclure toute observation incomplète signifie que les enregistrements avec des valeurs manquantes dans les champs nommés dans une sous-commande sont exclus de toutes les analyses.
- Exclure les observations test par test signifie que les enregistrements avec des valeurs manquantes dans un champ utilisé pour un test spécifique sont ignorés pendant ce test. Lorsque plusieurs tests sont spécifiés dans l'analyse, chaque test est évalué séparément.

Valeurs manquantes spécifiées

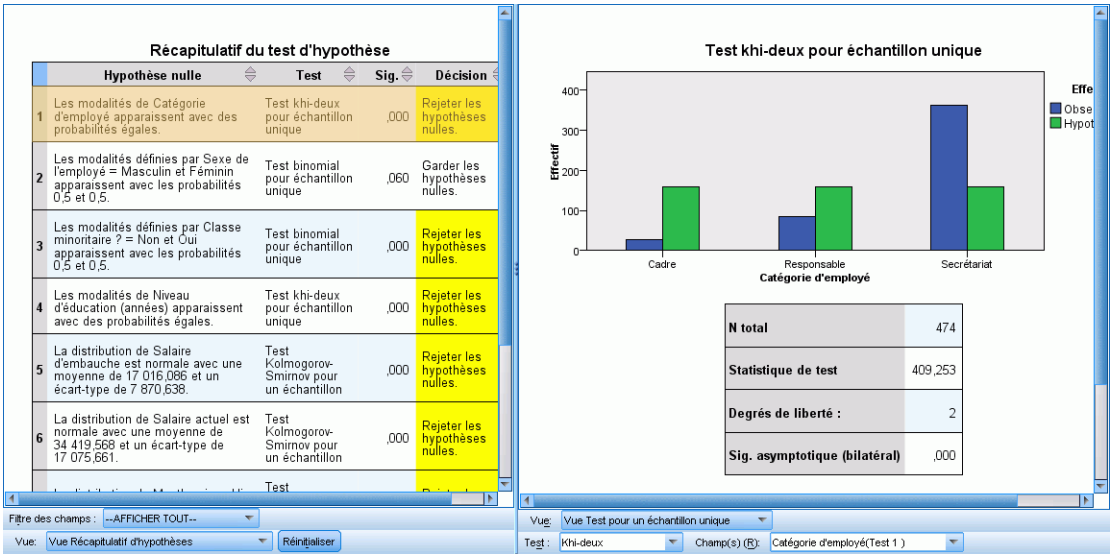
Figure 27-21
Paramètres des Valeurs manquantes spécifiées des Tests non paramétriques pour échantillons liés



Valeurs utilisateur manquantes pour les champs qualitatifs. Les champs qualitatifs doivent avoir des valeurs valides pour qu'un enregistrement puisse être inclus dans l'analyse. Ces commandes vous permettent d'indiquer si les valeurs manquantes spécifiées sont considérées comme valides parmi les champs qualitatifs. Les valeurs manquantes par défaut et les valeurs manquantes pour les champs continus sont toujours considérées comme non valides.

Vue du modèle

Figure 27-22
Vue du modèle Tests non paramétriques



La procédure crée un objet Visualiseur de modèle dans le Viewer. En activant cet objet par un double-clic, vous obtenez une vue interactive du modèle. La vue du modèle est composé d'une fenêtre à deux panneaux, la vue principale à gauche et la vue liée, ou auxiliaire, à droite.

Il existe deux vues principales :

- Récapitulatif d'hypothèses. Il s'agit de la vue par défaut.Pour plus d'informations, reportez-vous à la section Récapitulatif d'hypothèses sur p. 219.
- Récapitulatif de l'intervalle de confiance. Pour plus d'informations, reportez-vous à la section Récapitulatif de l'intervalle de confiance sur p. 220.

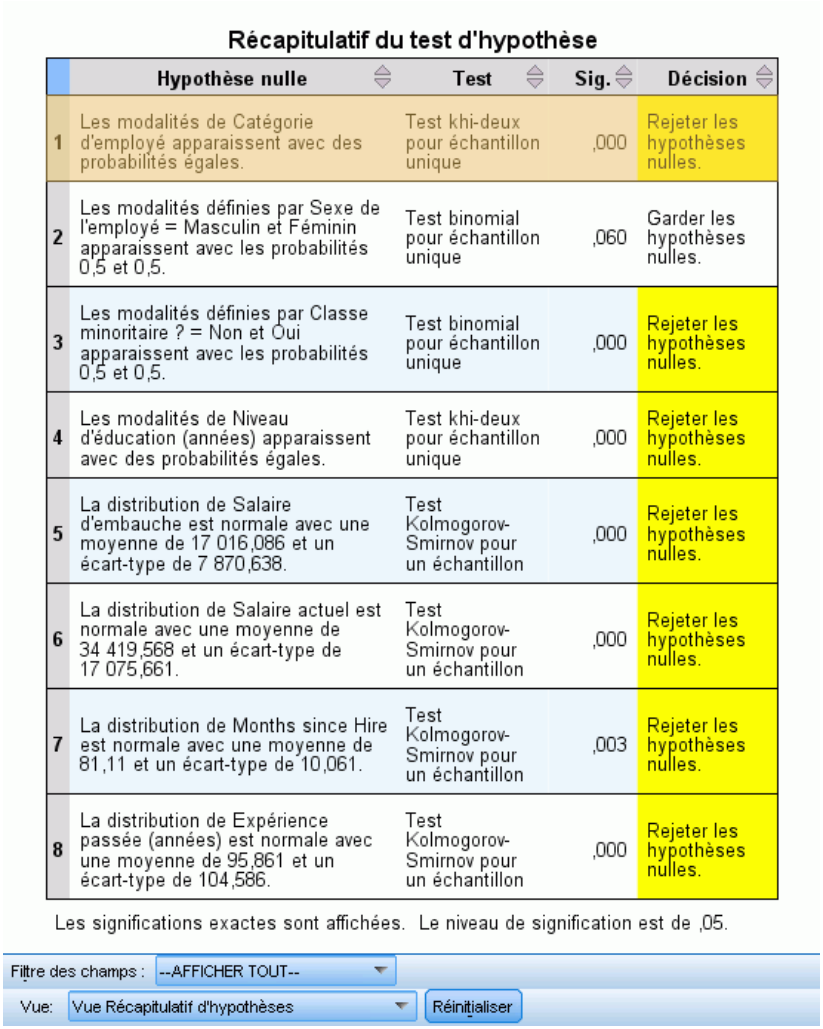
Il existe sept vues liées/auxiliaires :

- Test pour un échantillon unique. Il s'agit de la vue par défaut si des tests pour un échantillon unique sont requis. Pour plus d'informations, reportez-vous à la section Test à un échantillon sur p. 221.
- Test pour échantillons liés. Il s'agit de la vue par défaut si des tests pour échantillons liés sont requis et qu'aucun test pour un échantillon unique n'est requis. Pour plus d'informations, reportez-vous à la section Test pour échantillons liés sur p. 225.
- Test pour échantillons indépendants. Il s'agit de la vue par défaut si aucun test pour échantillons liés ou aucun test pour un échantillon unique n'est requis. Pour plus d'informations, reportez-vous à la section Test pour échantillons indépendants sur p. 232.
- Informations sur les champs qualitatifs. Pour plus d'informations, reportez-vous à la section Informations sur les champs qualitatifs sur p. 240.
- Informations sur les champs continus. Pour plus d'informations, reportez-vous à la section Informations sur les champs continus sur p. 241.

- Comparaisons par paire. Pour plus d'informations, reportez-vous à la section Comparaisons par paire sur p. 242.
- Sous-ensembles homogènes. Pour plus d'informations, reportez-vous à la section Sous-ensembles homogènes sur p. 243.

Récapitulatif d'hypothèses

Figure 27-23
Récapitulatif d'hypothèses



Récapitulatif du test d'hypothèse

	Hypothèse nulle	Test	Sig.	Décision
1	Les modalités de Catégorie d'employé apparaissent avec des probabilités égales.	Test khi-deux pour échantillon unique	,000	Rejeter les hypothèses nulles.
2	Les modalités définies par Sexe de l'employé = Masculin et Féminin apparaissent avec les probabilités 0,5 et 0,5.	Test binomial pour échantillon unique	,060	Garder les hypothèses nulles.
3	Les modalités définies par Classe minoritaire ? = Non et Oui apparaissent avec les probabilités 0,5 et 0,5.	Test binomial pour échantillon unique	,000	Rejeter les hypothèses nulles.
4	Les modalités de Niveau d'éducation (années) apparaissent avec des probabilités égales.	Test khi-deux pour échantillon unique	,000	Rejeter les hypothèses nulles.
5	La distribution de Salaire d'embauche est normale avec une moyenne de 17 016,086 et un écart-type de 7 870,638.	Test Kolmogorov-Smirnov pour un échantillon	,000	Rejeter les hypothèses nulles.
6	La distribution de Salaire actuel est normale avec une moyenne de 34 419,568 et un écart-type de 17 075,661.	Test Kolmogorov-Smirnov pour un échantillon	,000	Rejeter les hypothèses nulles.
7	La distribution de Months since Hire est normale avec une moyenne de 81,11 et un écart-type de 10,061.	Test Kolmogorov-Smirnov pour un échantillon	,003	Rejeter les hypothèses nulles.
8	La distribution de Expérience passée (années) est normale avec une moyenne de 95,861 et un écart-type de 104,586.	Test Kolmogorov-Smirnov pour un échantillon	,000	Rejeter les hypothèses nulles.

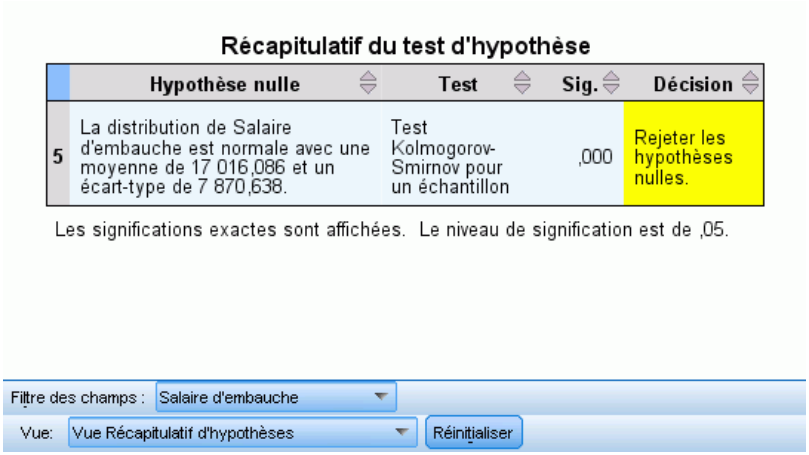
Les significations exactes sont affichées. Le niveau de signification est de ,05.

La vue Récapitulatif du modèle est un instantané, permettant de consulter en un coup d'oeil les tests non-paramétriques. Elle met en évidence les hypothèses nulles et les décisions, portant l'attention sur les valeurs *p* significatives.

- Chaque ligne correspond à un test distinct. Cliquez sur une ligne pour afficher des informations supplémentaires sur le test dans la vue liée.

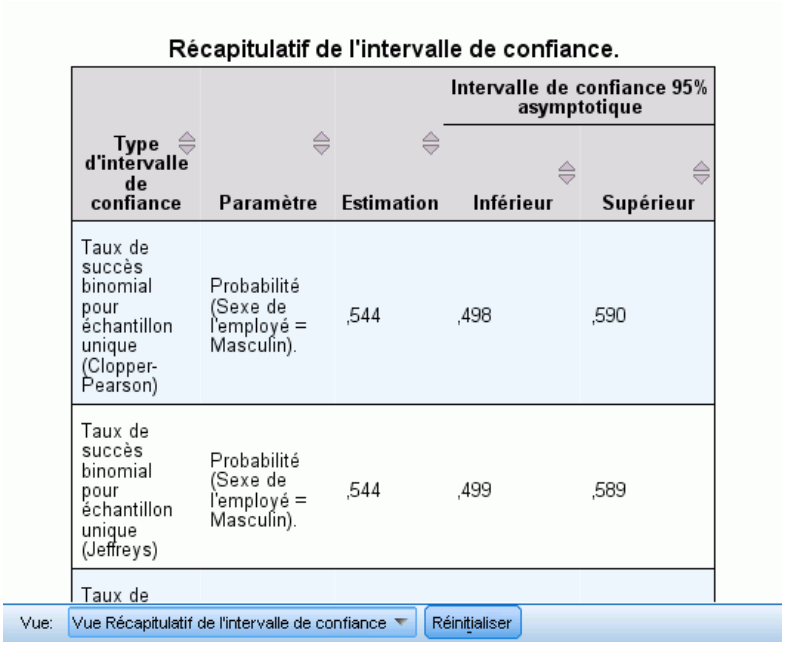
- Cliquez sur un en-tête de colonne pour trier les lignes de la colonne concernée en fonction de leurs valeurs.
- Le bouton Réinitialiser vous permet de revenir à l'état d'origine du Visualiseur de modèle.
- La liste déroulante Filtre des champs vous permet d'afficher uniquement les tests concernant le champ sélectionné. Par exemple, si *Salaire d'embauche* est sélectionné dans la liste Filtre des champs, seuls deux tests s'affichent dans le récapitulatif d'hypothèses.

Figure 27-24
Récapitulatif d'hypothèses filtré pour le Salaire d'embauche



Récapitulatif de l'intervalle de confiance

Figure 27-25
Récapitulatif de l'intervalle de confiance



Le récapitulatif de l'intervalle de confiance affiche tout intervalle de confiance produit par les tests non paramétriques.

- Chaque ligne correspond à un intervalle de confiance distinct.
- Cliquez sur un en-tête de colonne pour trier les lignes de la colonne concernée en fonction de leurs valeurs.

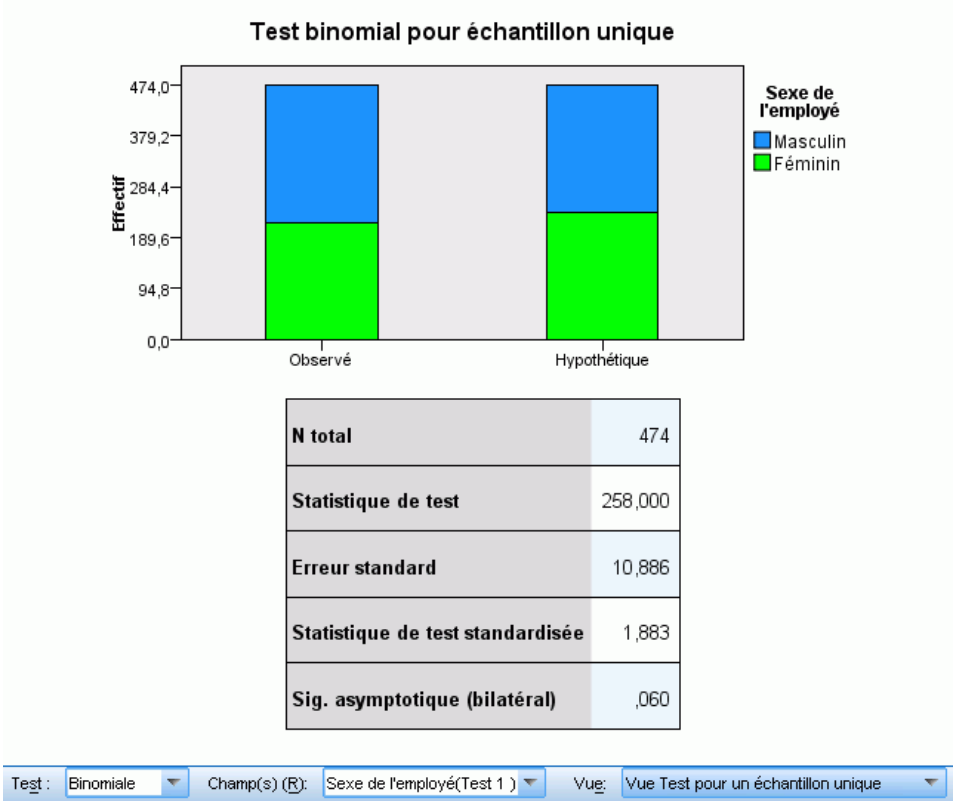
Test à un échantillon

La vue Test pour un échantillon unique affiche des informations détaillées sur tout test non paramétrique pour un échantillon unique requis. Les informations affichées dépendent du test sélectionné.

- La liste déroulante Test vous permet de sélectionner un type de test à un échantillon.
- La liste déroulante Champ(s) vous permet de sélectionner un champ testé à l'aide du test sélectionné dans la liste déroulante Test.

Test binomial

Figure 27-26
Vue Test pour un échantillon unique, test binomial

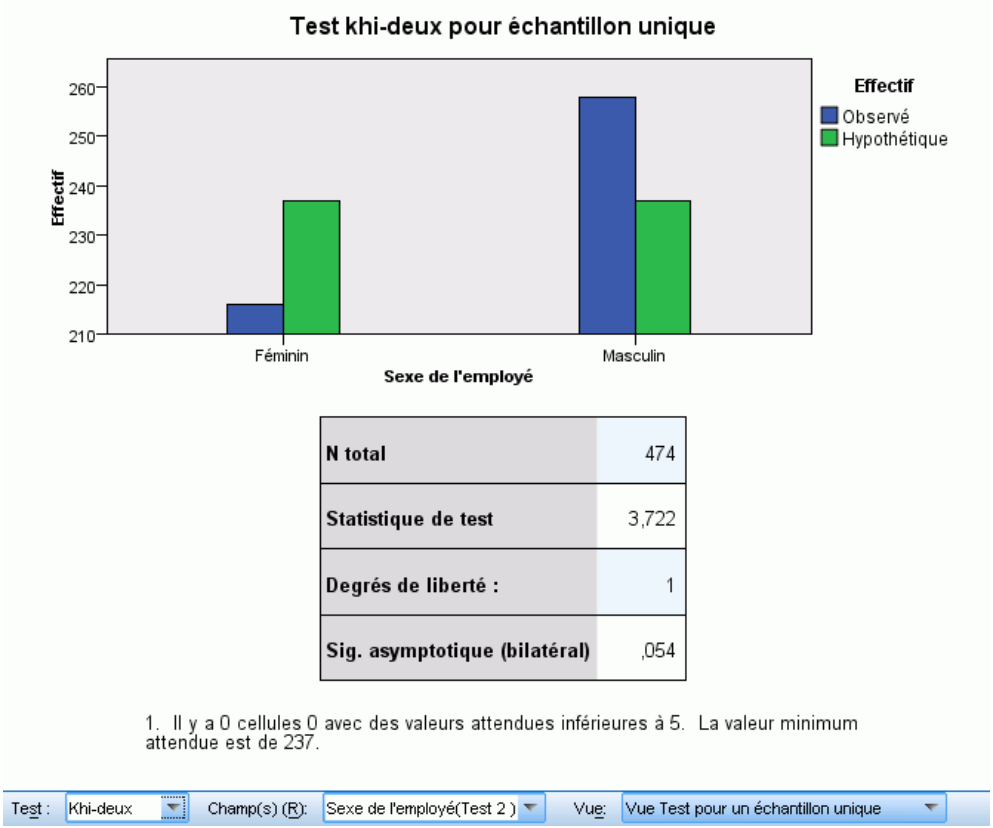


Le test binomial affiche un diagramme en bâtons empilés et un tableau des tests.

- Le diagramme en bâtons empilés affiche les fréquences observées et théoriques pour les modalités “succès” et “échec” du champ test, les “échecs” étant empilés au dessus des “succès”. Lorsque vous passez la souris sur un bâton, les pourcentages des modalités s’affichent dans une info-bulle. Des différences visibles entre les bâtons indiquent que le champ de test peut ne pas comporter la distribution binomiale hypothétique.
- Le tableau affiche des informations détaillées sur le test.

Test du Khi-deux

Figure 27-27
Vue Test pour échantillon unique, test du Khi-deux



N total	474
Statistique de test	3,722
Degrés de liberté :	1
Sig. asymptotique (bilatéral)	,054

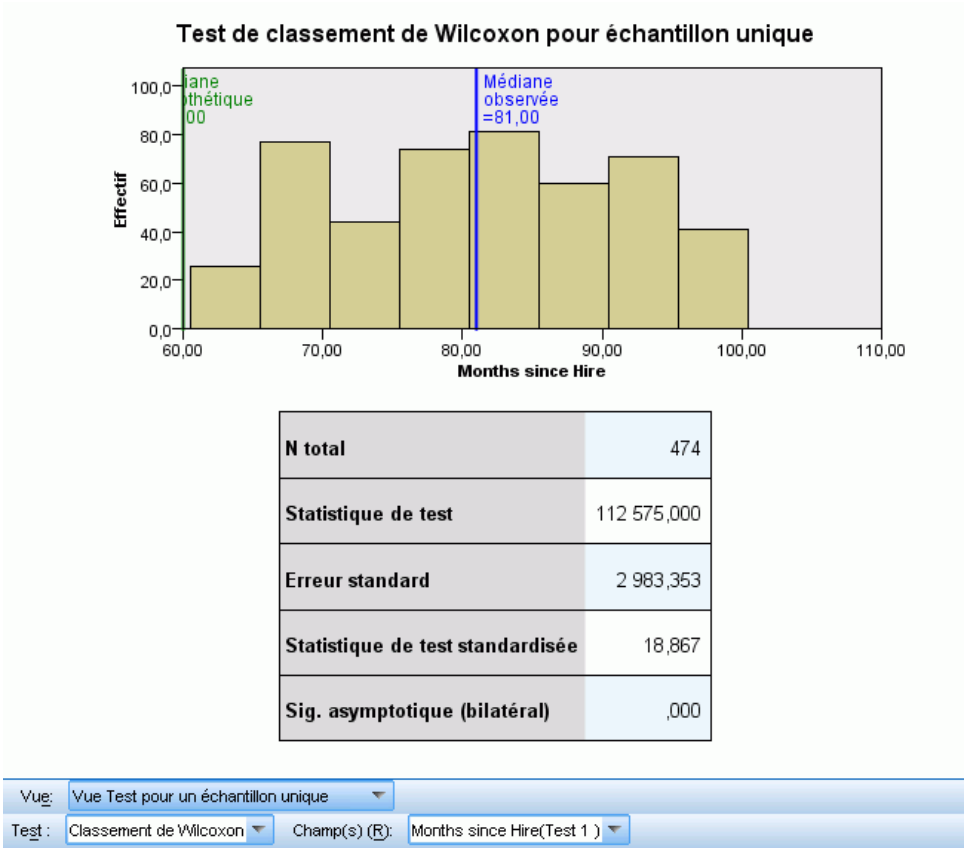
1. Il y a 0 cellules 0 avec des valeurs attendues inférieures à 5. La valeur minimum attendue est de 237.

Le test du Khi-deux affiche un diagramme en bâtons groupés et un tableau des tests.

- Le diagramme en bâtons groupés affiche les fréquences observées et théoriques pour chaque modalité du champ de test. Lorsque vous passez la souris sur un bâton, les fréquences observées et théoriques et leur différence (résidu) s’affichent dans une info-bulle. Des différences visibles entre les fréquences observées et théoriques indiquent que le champ de test peut ne pas comporter la distribution hypothétique.
- Le tableau affiche des informations détaillées sur le test.

Classement de Wilcoxon

Figure 27-28
Vue Test pour échantillon unique, test de classement de Wilcoxon



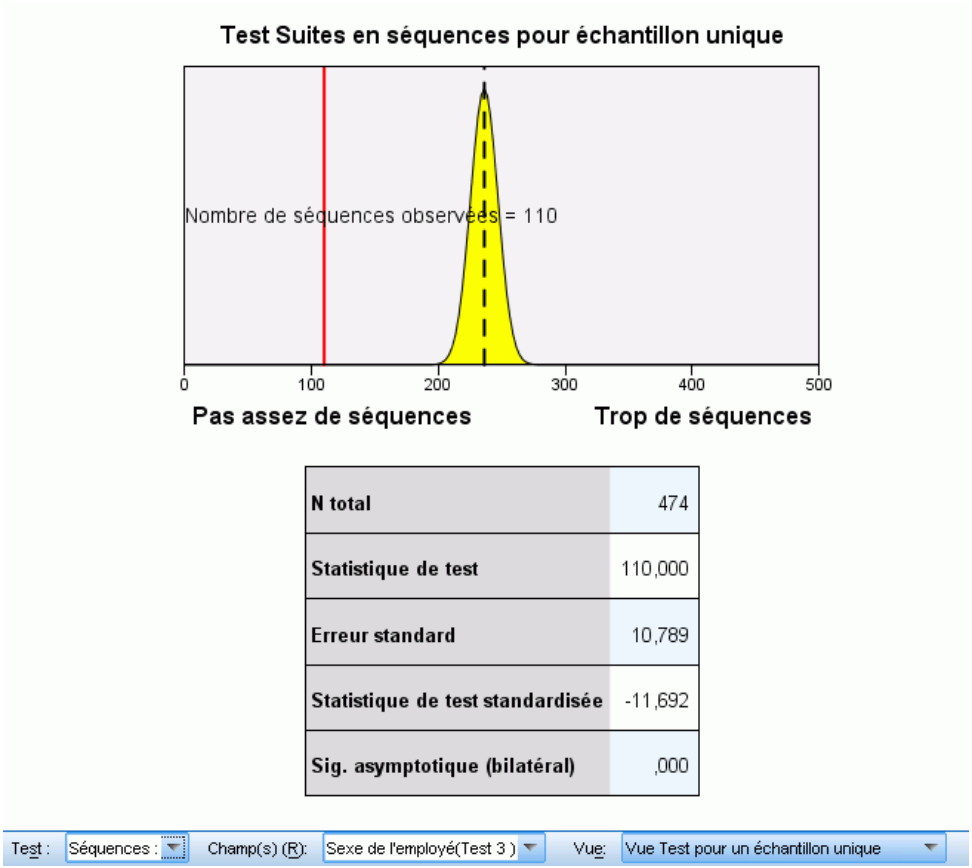
N total	474
Statistique de test	112 575,000
Erreur standard	2 983,353
Statistique de test standardisée	18,867
Sig. asymptotique (bilatéral)	,000

Le test de classement de Wilcoxon affiche un histogramme et un tableau des tests.

- L'histogramme comprend des lignes verticales qui représentent les médianes observées et théoriques.
- Le tableau affiche des informations détaillées sur le test.

Suites en séquences

Figure 27-29
Vue Test pour un échantillon unique, test de suites en séquences

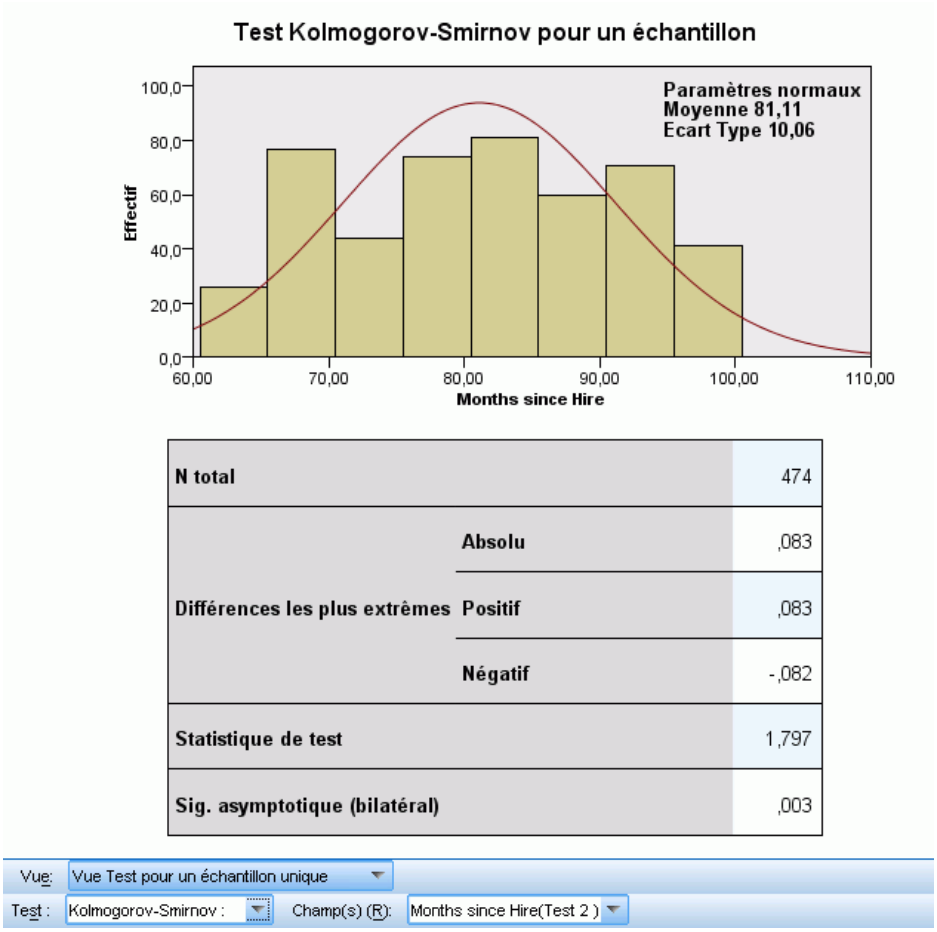


La vue Test de suites en séquences affiche un diagramme et un tableau des tests.

- Le diagramme affiche une distribution normale avec le nombre de suites en séquences observées indiqué par une ligne verticale. Remarque : lorsque le test exact est réalisé, il ne repose pas sur une distribution normale.
- Le tableau affiche des informations détaillées sur le test.

Test de Kolmogorov-Smirnov

Figure 27-30
Vue Test pour un échantillon unique, test de Kolmogorov-Smirnov



Le test de Kolmogorov-Smirnov affiche un histogramme et un tableau des tests.

- L'histogramme comprend une superposition de la fonction de densité de la probabilité pour la distribution uniforme hypothétique, normale, de Poisson ou exponentielle. Remarque : le test est basé sur les distributions cumulées et les différences les plus extrêmes rapportées dans le tableau doivent être interprétées en fonction de ces distributions cumulées.
- Le tableau affiche des informations détaillées sur le test.

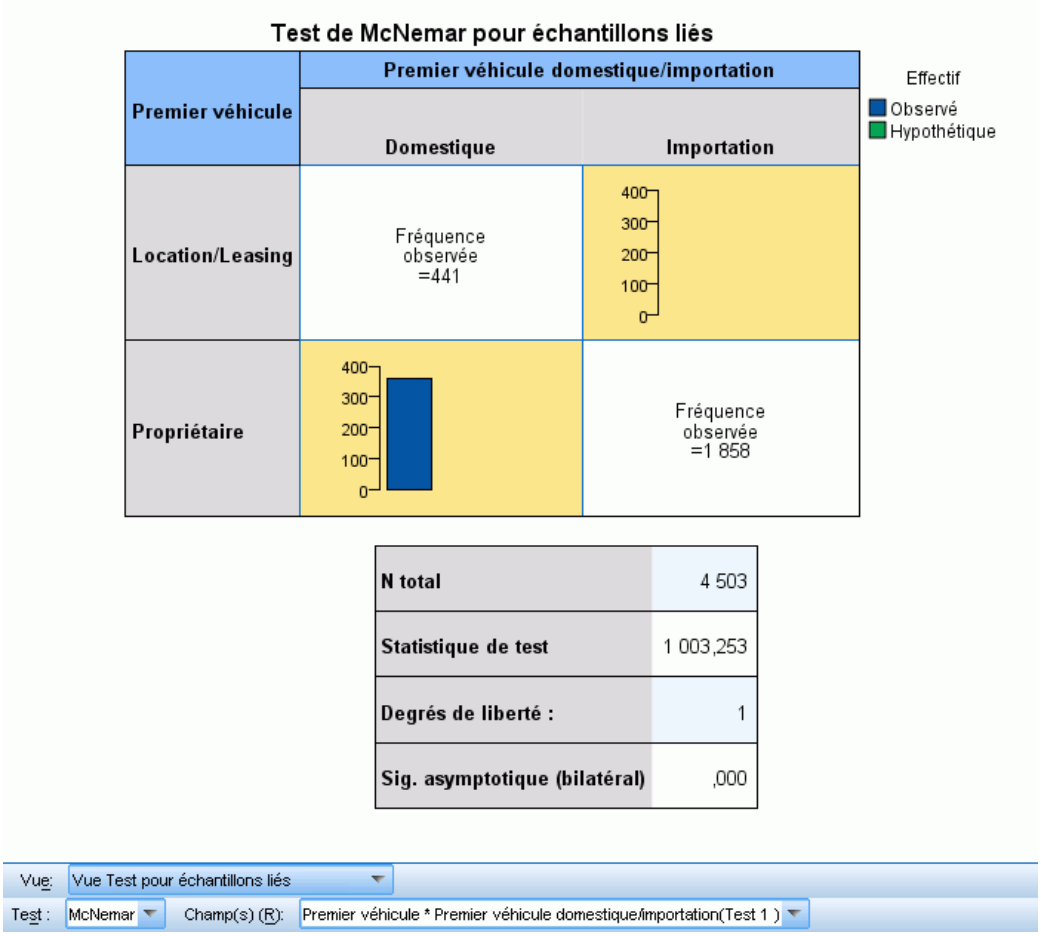
Test pour échantillons liés

La vue Test pour un échantillon unique affiche des informations détaillées sur tout test non paramétrique pour un échantillon unique requis. Les informations affichées dépendent du test sélectionné.

- La liste déroulante Test vous permet de sélectionner un type de test à un échantillon.
- La liste déroulante Champ(s) vous permet de sélectionner un champ testé à l'aide du test sélectionné dans la liste déroulante Test.

Test de McNemar :

Figure 27-31
Vue Test pour échantillons liés, test de McNemar

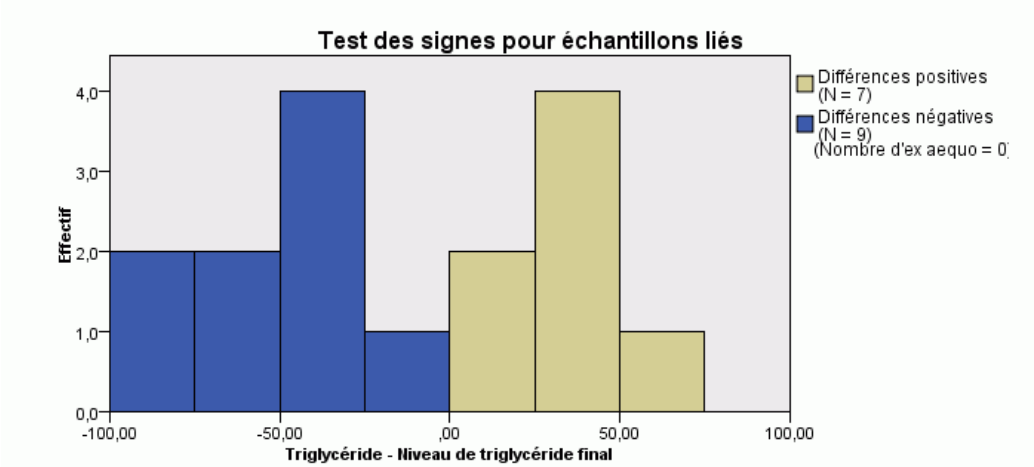


Le test de McNemar affiche un diagramme en bâtons groupés et un tableau des tests.

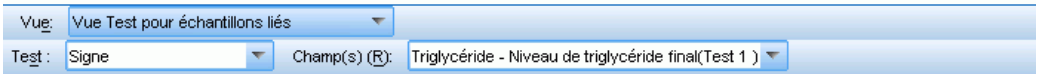
- Le diagramme en bâtons groupés affiche les fréquences observées et théoriques pour les cellules hors diagonale du tableau 2×2 défini par les champs du test.
- Le tableau affiche des informations détaillées sur le test.

Test des signes

Figure 27-32
Vue Test pour échantillons liés, test des signes



N total	16
Statistique de test	7,000
Erreur standard	2,000
Statistique de test standardisée	-,250
Sig. asymptotique (bilatéral)	,803
Sig. exacte (bilatéral)	,804

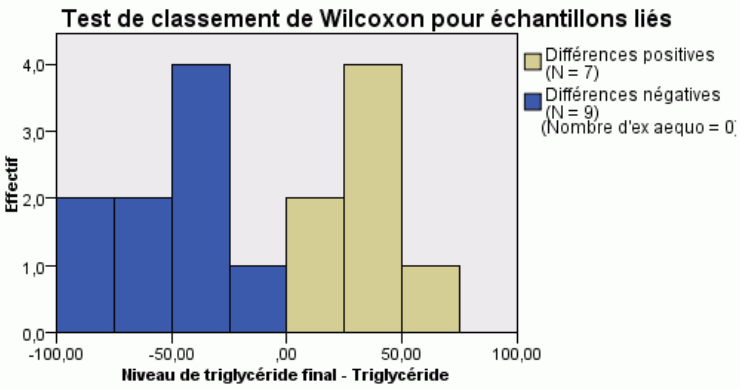


La vue Test des signes affiche un histogramme empilé et un tableau des tests.

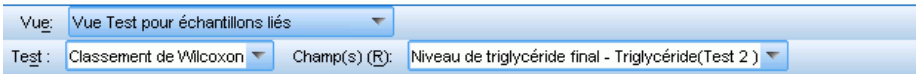
- L'histogramme empilé affiche les différences entre les champs à l'aide du signe de la différence comme champ d'empilement.
- Le tableau affiche des informations détaillées sur le test.

Test de classement de Wilcoxon

Figure 27-33
Vue Test pour échantillons liés, test de classement de Wilcoxon



N total	16
Statistique de test	45,000
Erreur standard	19,339
Statistique de test standardisée	-1,189
Sig. asymptotique (bilatéral)	,234

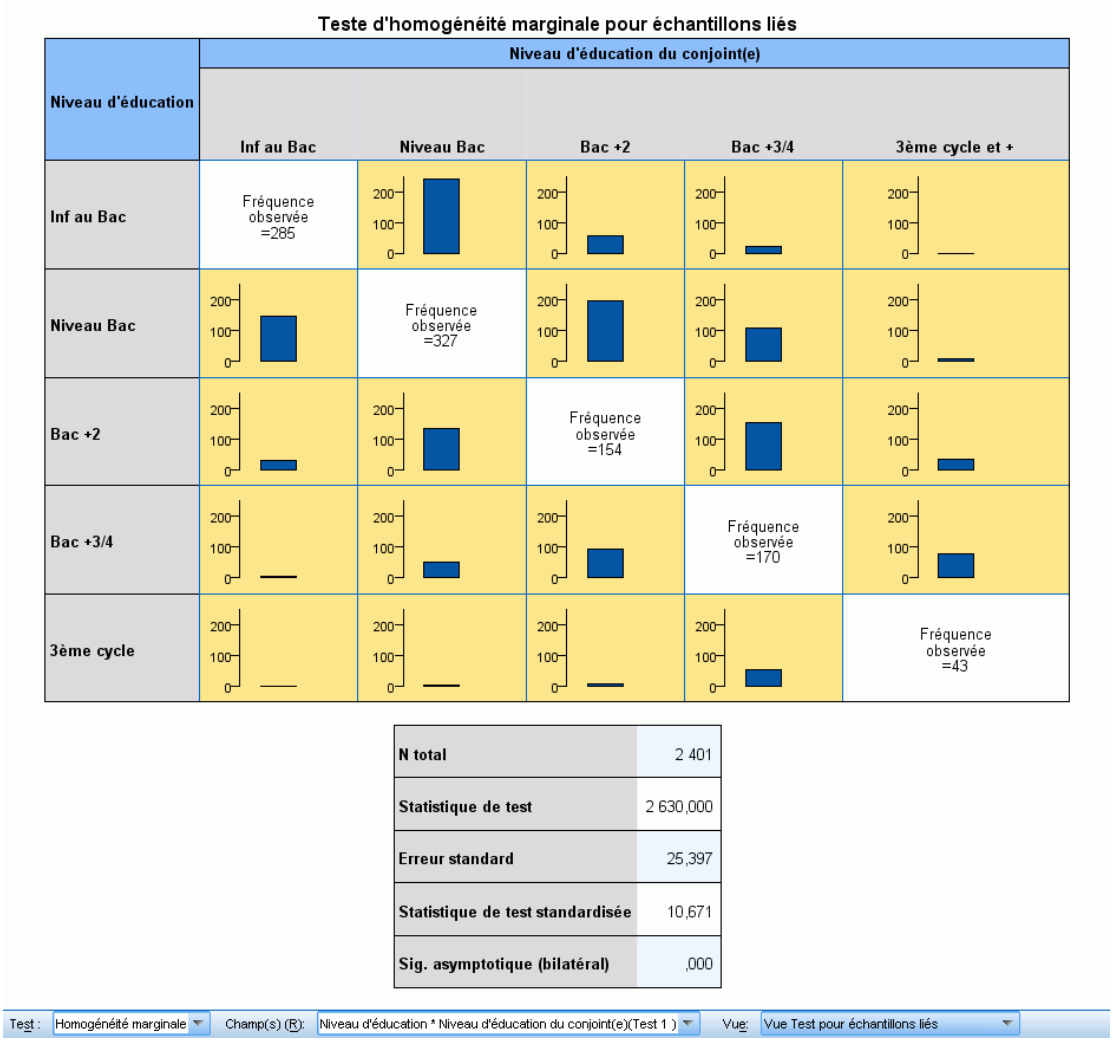


Le test de classement de Wilcoxon affiche un histogramme empilé et un tableau des tests.

- L'histogramme empilé affiche les différences entre les champs à l'aide du signe de la différence comme champ d'empilement.
- Le tableau affiche des informations détaillées sur le test.

Test d'homogénéité marginale

Figure 27-34
Vue Test pour échantillons liés, test d'homogénéité marginale

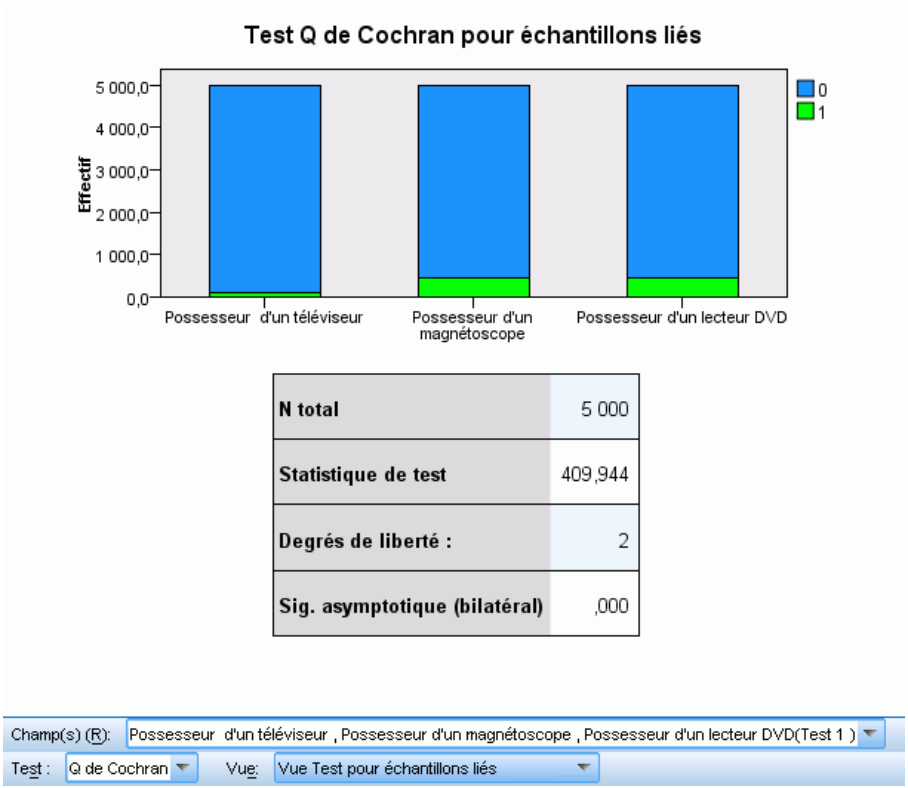


La vue Test d'homogénéité marginale affiche un diagramme en bâtons groupés et un tableau des tests.

- Le diagramme en bâtons groupés affiche les fréquences observées pour les cellules hors diagonale du tableau défini par les champs du test.
- Le tableau affiche des informations détaillées sur le test.

Test Q de Cochran

Figure 27-35
Vue Test pour échantillons liés, test Q de Cochran

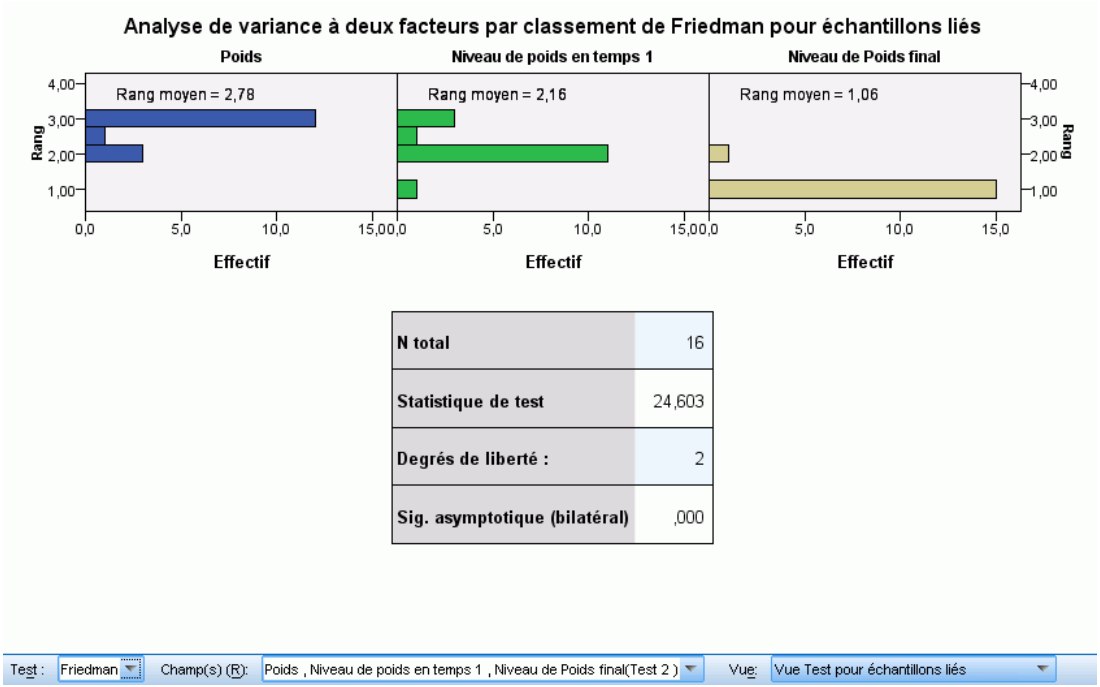


Le test Q de Cochran affiche un diagramme en bâtons empilés et un tableau des tests.

- Le diagramme en bâtons empilés affiche les fréquences observées pour les modalités “succès” et “échec” des champs du test, les “échecs” étant empilés au dessus des “succès”. Lorsque vous passez la souris sur un bâton, les pourcentages des modalités s’affichent dans une info-bulle.
- Le tableau affiche des informations détaillées sur le test.

Analyse de variance à deux facteurs par classement de Friedman

Figure 27-36
Vue Test pour échantillons liés, analyse de variance à deux facteurs par classement de Friedman

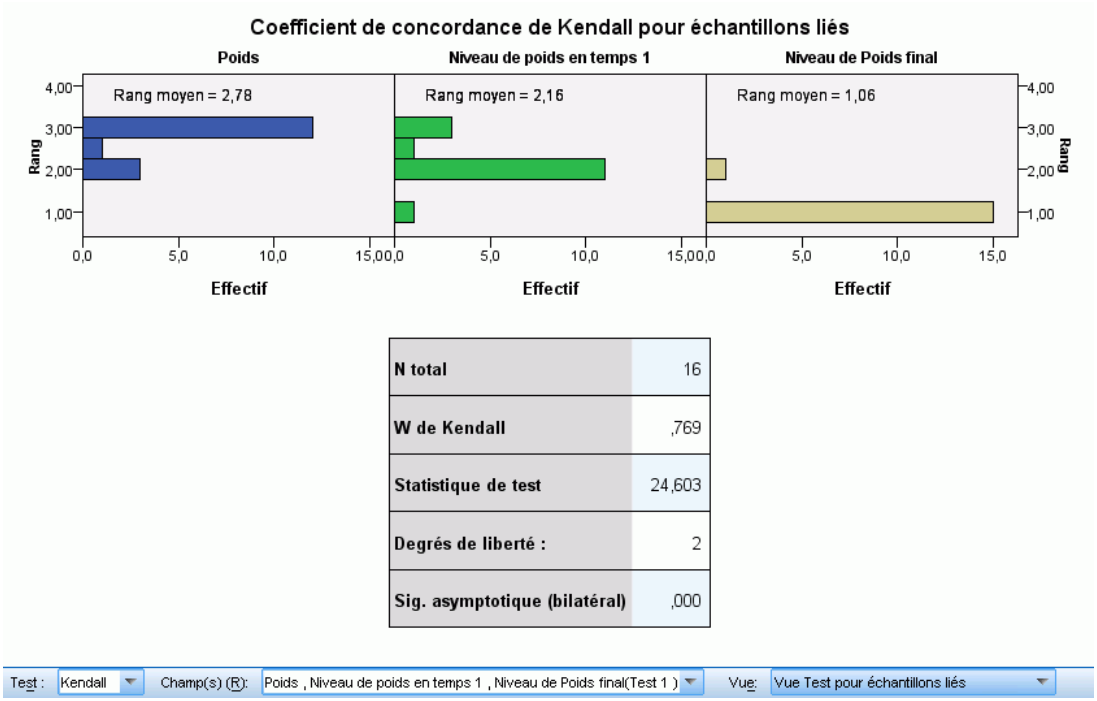


La vue Analyse de variance à deux facteurs par classement de Friedman affiche des histogrammes sous forme de panels et un tableau des tests.

- Les histogrammes affichent la distribution observée des rangs, presentée sous forme de panels par les champs du test.
- Le tableau affiche des informations détaillées sur le test.

Coefficient de concordance de Kendall

Figure 27-37
Vue Test pour échantillons liés, coefficient de concordance de Kendall



La vue Coefficient de concordance de Kendall affiche des histogrammes sous forme de panels et un tableau des tests.

- Les histogrammes affichent la distribution observée des rangs, panélisée par les champs du test.
- Le tableau affiche des informations détaillées sur le test.

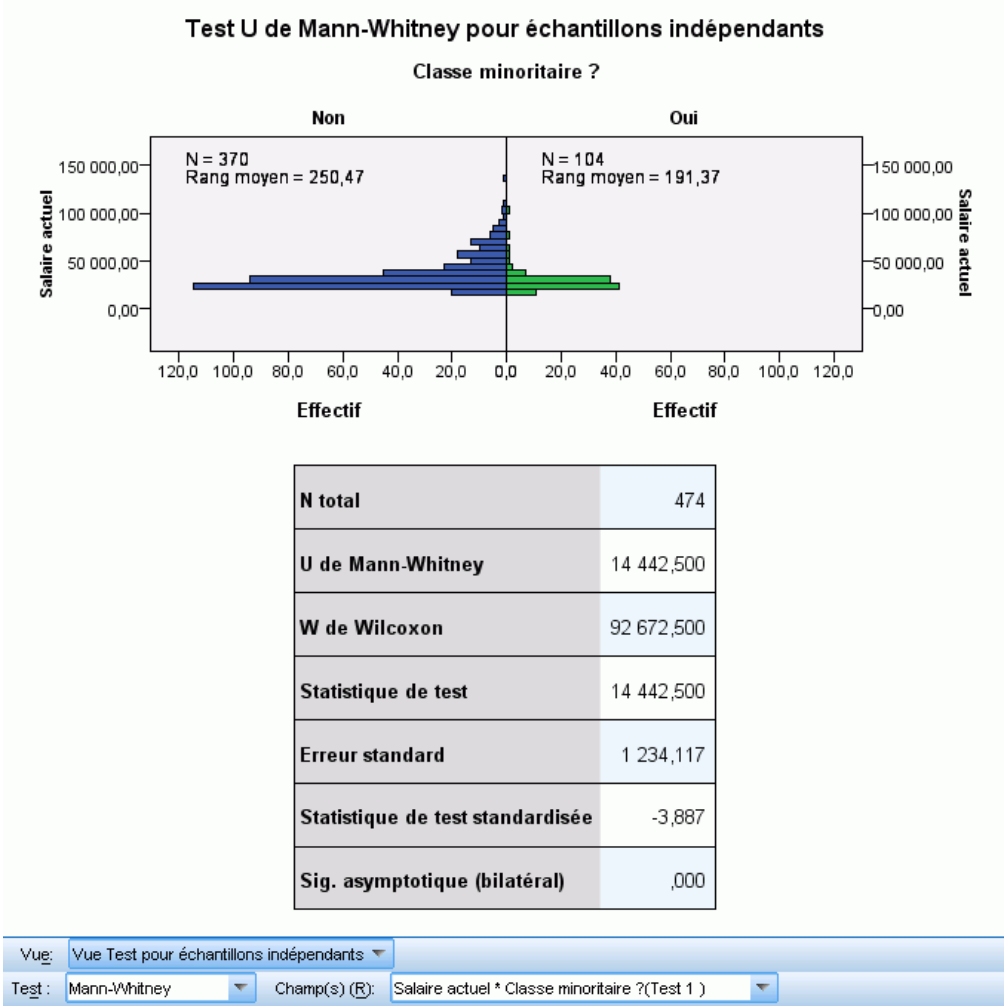
Test pour échantillons indépendants

La vue Test pour échantillons indépendants affiche des informations détaillées sur tout test non paramétrique pour échantillons indépendants requis. Les informations affichées dépendent du test sélectionné.

- La liste déroulante Test vous permet de sélectionner un type de test pour échantillons indépendants.
- La liste déroulante Champ(s) vous permet de sélectionner un test et une combinaison de champs de regroupement testés à l'aide du test sélectionné dans la liste déroulante Test.

Test de Mann-Whitney

Figure 27-38
Vue Test pour échantillons indépendants, test de Mann-Whitney



N total	474
U de Mann-Whitney	14 442,500
W de Wilcoxon	92 672,500
Statistique de test	14 442,500
Erreur standard	1 234,117
Statistique de test standardisée	-3,887
Sig. asymptotique (bilatéral)	,000

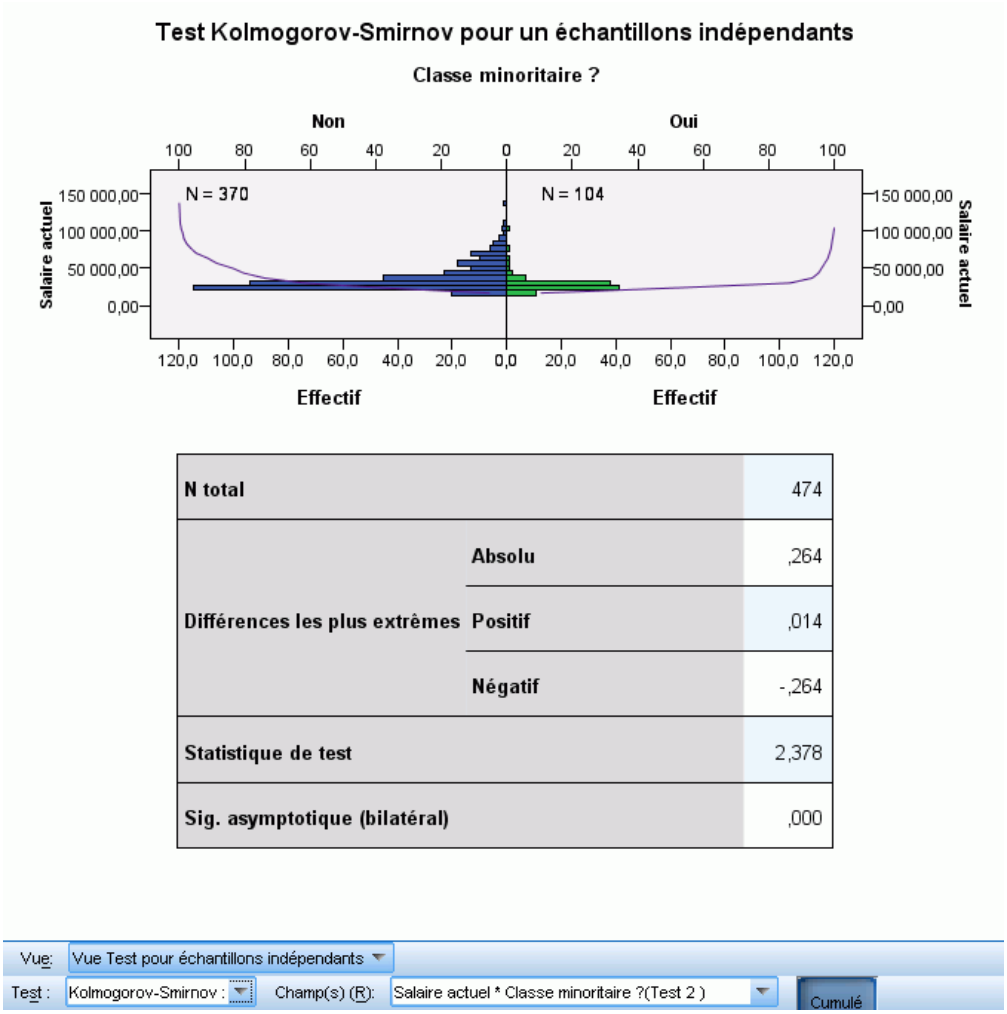
Vue: Vue Test pour échantillons indépendants
Test : Mann-Whitney Champ(s) (R): Salaire actuel * Classe minoritaire ?(Test 1)

Le test de Mann-Whitney affiche une pyramide de population et un tableau des tests.

- La pyramide de population affiche des histogrammes collés dos à dos classés par catégories du champ de regroupement et indiquant le nombre d'enregistrements dans chaque groupe et le rang moyen de chaque groupe.
- Le tableau affiche des informations détaillées sur le test.

Test de Kolmogorov-Smirnov

Figure 27-39
Vue Test pour échantillons indépendants, test de Kolmogorov-Smirnov



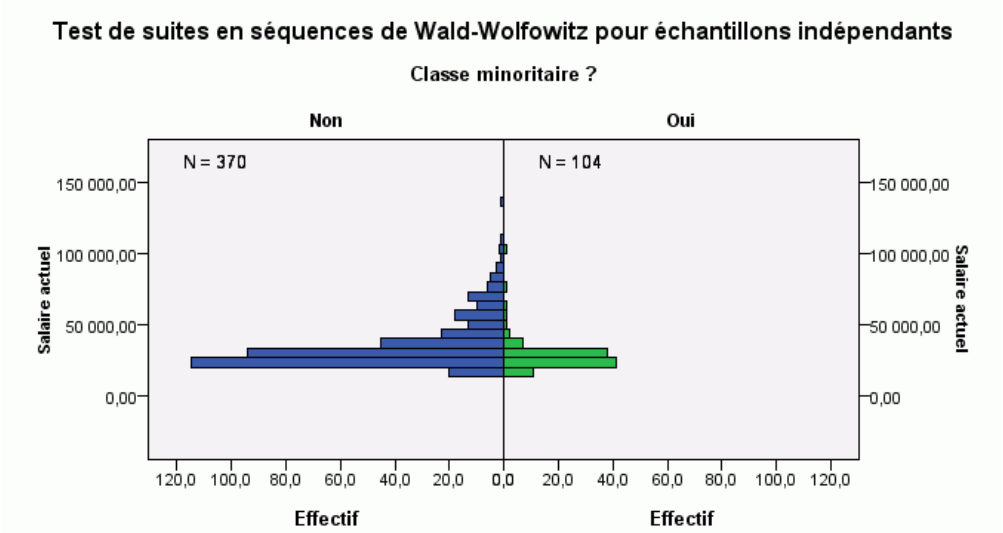
N total		474
Différences les plus extrêmes	**Absolu**	,264
	Positif	,014
	Négatif	-,264
Statistique de test		2,378
Sig. asymptotique (bilatéral)		,000

Le test de Kolmogorov-Smirnov affiche une pyramide de population et un tableau des tests.

- La pyramide de population affiche des histogrammes collés dos à dos classés par catégories du champ de regroupement et indiquant le nombre d'enregistrements dans chaque groupe. Les lignes de la distribution cumulée observée peuvent être affichées ou masquées en cliquant sur le bouton Cumulé.
- Le tableau affiche des informations détaillées sur le test.

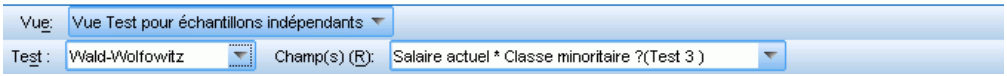
Suites en séquences de Wald-Wolfowitz :

Figure 27-40
Vue Test pour échantillons indépendants, test des suites en séquences de Wald-Wolfowitz



N total		474
Minimum possible	**Statistique de test[1]**	97,000
	Erreur standard	7,442
	Statistique de test standardisée	-8,917
	Sig. asymptotique (bilatéral)	,000
Maximum possible	**Statistique de test[1]**	199,000
	Erreur standard	7,442
	Statistique de test standardisée	4,788
	Sig. asymptotique (bilatéral)	1,000

[1]The test statistic is the number of runs.
1. There are 55 inter-group ties involving 228 records.

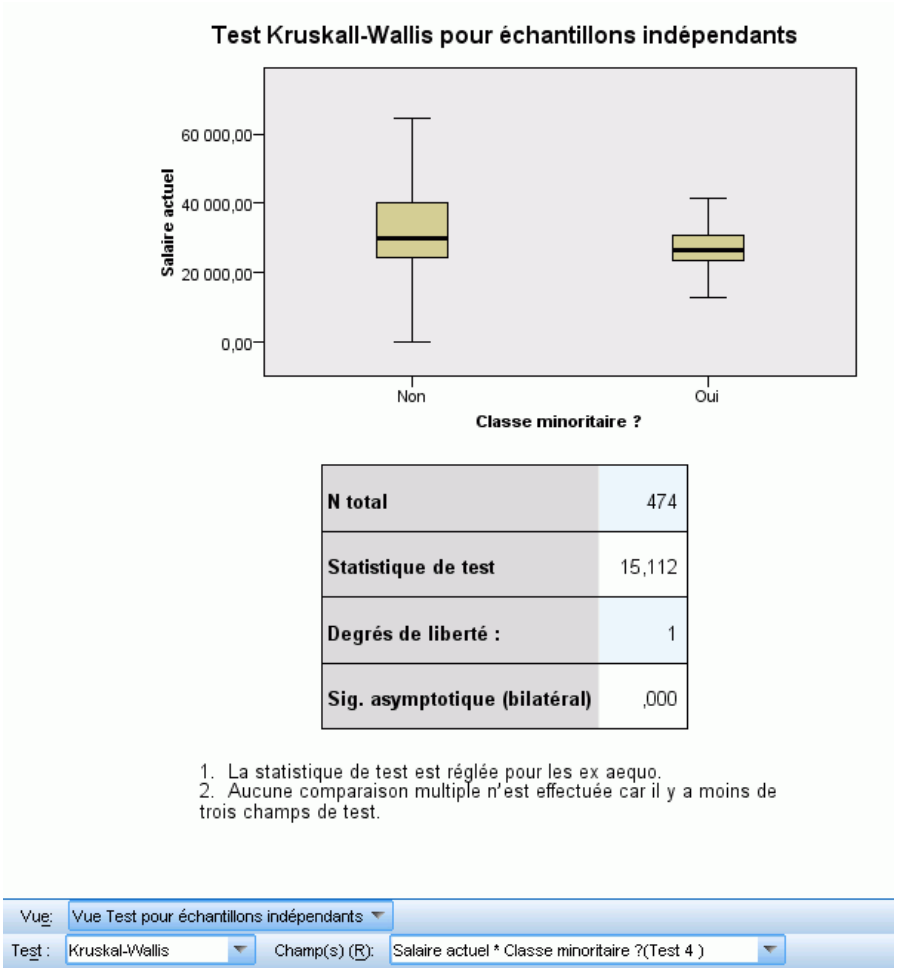


Le test des suites en séquences de Wald-Wolfowitz affiche un diagramme en bâtons empilés et un tableau des tests.

- La pyramide de population affiche des histogrammes collés dos à dos classés par catégories du champ de regroupement et indiquant le nombre d'enregistrements dans chaque groupe.
- Le tableau affiche des informations détaillées sur le test.

Test de Kruskal-Wallis

Figure 27-41
Vue Test pour échantillons indépendants, test de Kruskal-Wallis



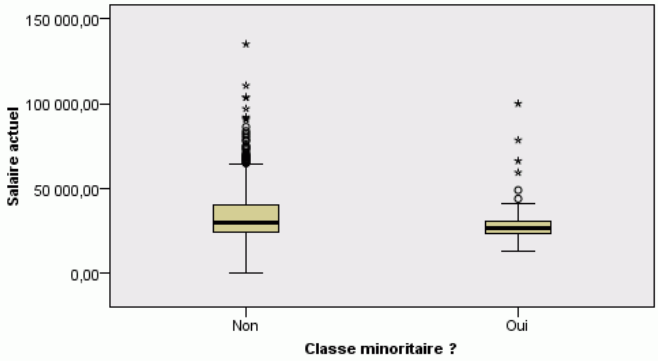
Le test de Kruskal-Wallis affiche des boîtes à moustaches et un tableau des tests.

- Des boîtes à moustaches distinctes sont affichées pour chaque modalité du champ de regroupement. Lorsque vous passez la souris sur une boîte, le rang moyen s'affiche dans une info-bulle.
- Le tableau affiche des informations détaillées sur le test.

Test de Jonckheere-Terpstra :

Figure 27-42
Vue Test pour échantillons indépendants, test de Jonckheere-Terpstra

Test des alternatives ordonnées de Jonckheere-Terpstra pour échantillons indépendants



N total	474
Statistique de test	14 442,500
Erreur standard	1 234,117
Statistique de test standardisée	-3,887
Sig. asymptotique (bilatéral)	,000

1. Aucune comparaison multiple n'est effectuée car il y a moins de trois champs de test.

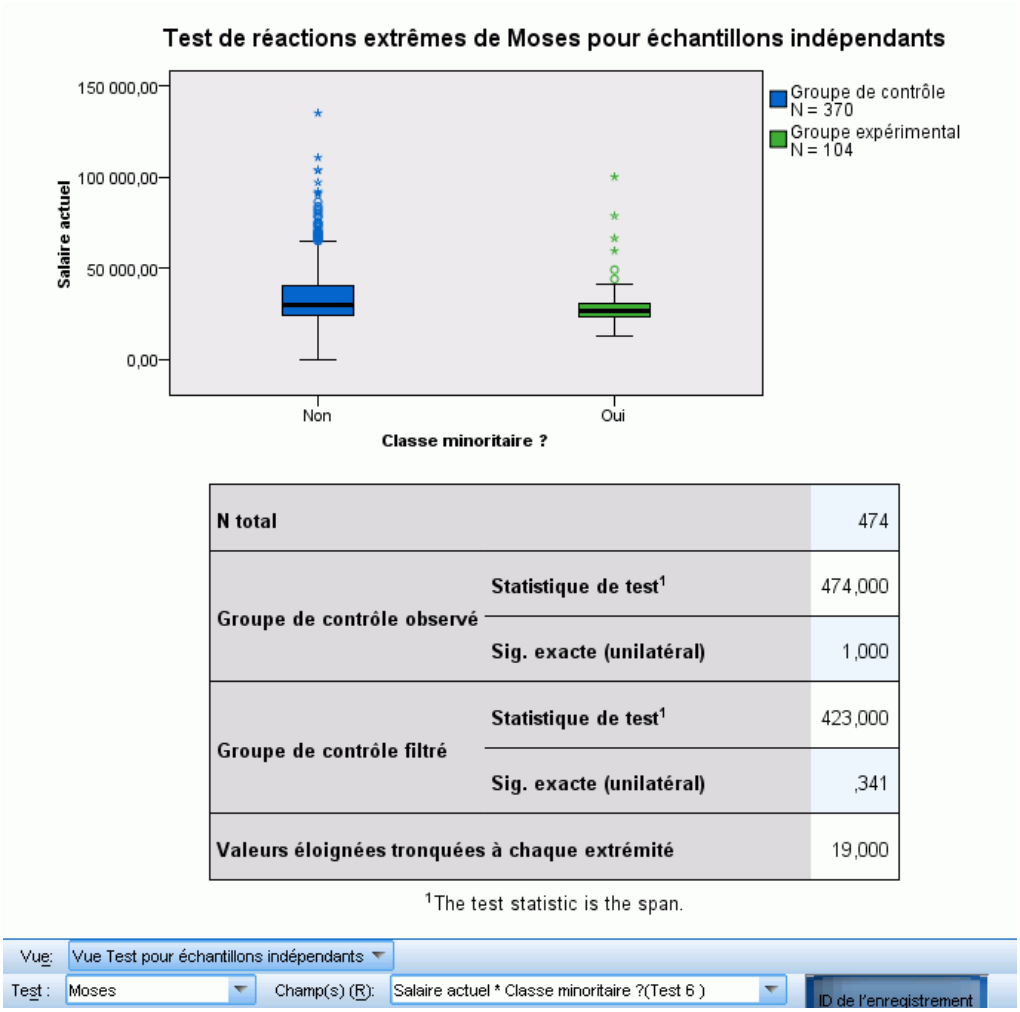
Test : Jonkheere-Terpstra | Champ(s) (R): Salaire actuel * Classe minoritaire ?(Test 5) | Vue: Vue Test pour échantillons indépendants

Le test de Jonckheere-Terpstra affiche des boîtes à moustaches et un tableau des tests.

- Des boîtes à moustaches distinctes sont affichées pour chaque modalité du champ de regroupement.
- Le tableau affiche des informations détaillées sur le test.

Test de réactions extrêmes de Moses

Figure 27-43
Vue Test pour échantillons indépendants, test de réactions extrêmes de Moses



N total		474
Groupe de contrôle observé	**Statistique de test[1]**	474,000
	Sig. exacte (unilatéral)	1,000
Groupe de contrôle filtré	**Statistique de test[1]**	423,000
	Sig. exacte (unilatéral)	,341
Valeurs éloignées tronquées à chaque extrémité		19,000

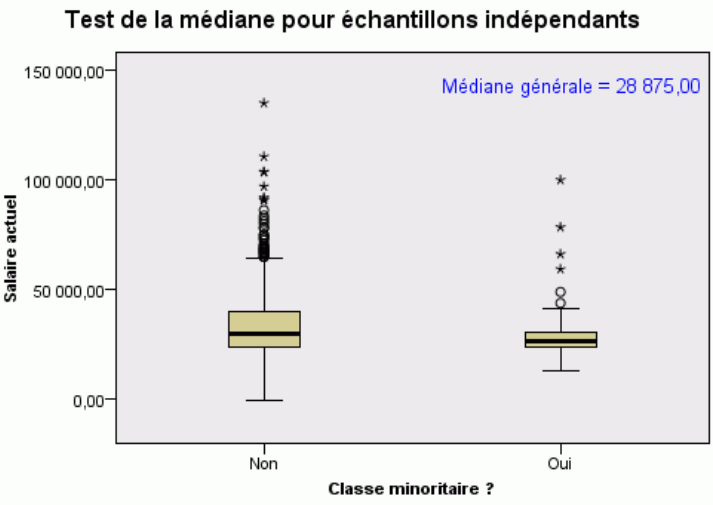
[1]The test statistic is the span.

Le test de réactions extrêmes de Moses affiche des boîtes à moustaches et un tableau des tests.

- Des boîtes à moustaches distinctes sont affichées pour chaque modalité du champ de regroupement. Les étiquettes des points peuvent être affichées ou masquées en cliquant sur le bouton ID de l'enregistrement.
- Le tableau affiche des informations détaillées sur le test.

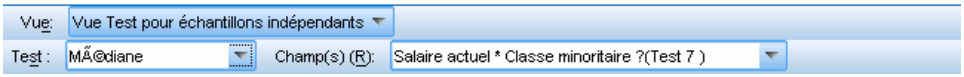
Test de la médiane

Figure 27-44
Vue Test pour échantillons indépendants, test de la médiane



N total		474
MÃ©diane		28 875,000
Statistique de test		14,240
Degrés de liberté :		1
Sig. asymptotique (bilatéral)		,000
Correction pour la continuité de Yates	**Khi-deux**	13,414
	Degrés de liberté :	1
	Sig. asymptotique (bilatéral)	,000

1. Aucune comparaison multiple n'est effectuée car il y a moins de trois champs de test.

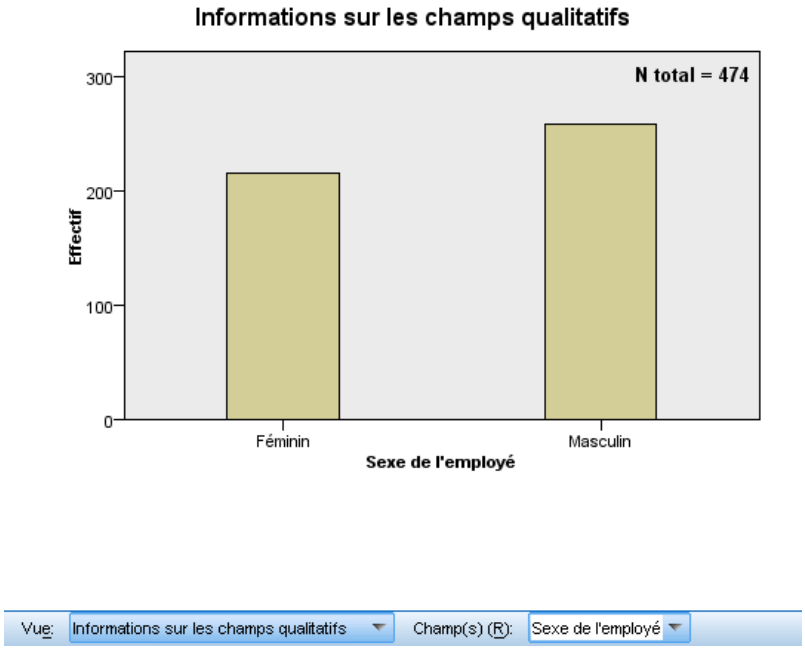


Le test de la médiane affiche des boîtes à moustaches et un tableau des tests.

- Des boîtes à moustaches distinctes sont affichées pour chaque modalité du champ de regroupement.
- Le tableau affiche des informations détaillées sur le test.

Informations sur les champs qualitatifs

Figure 27-45
Informations sur les champs qualitatifs

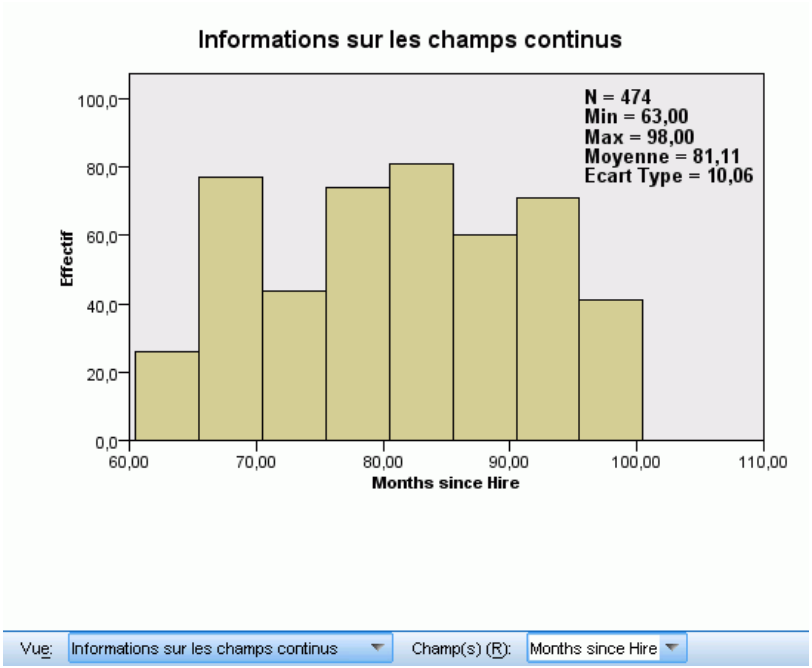


La vue Informations sur les champs qualitatifs affiche un diagramme en bâtons pour le champ qualitatif sélectionné dans la liste déroulante Champ(s). La liste des champs disponibles est limitée aux champs qualitatifs utilisés dans le test actuellement sélectionné dans la vue Récapitulatif d'hypothèses.

- Lorsque vous passez la souris sur un bâton, les pourcentages des modalités s'affichent dans une info-bulle.

Informations sur les champs continus

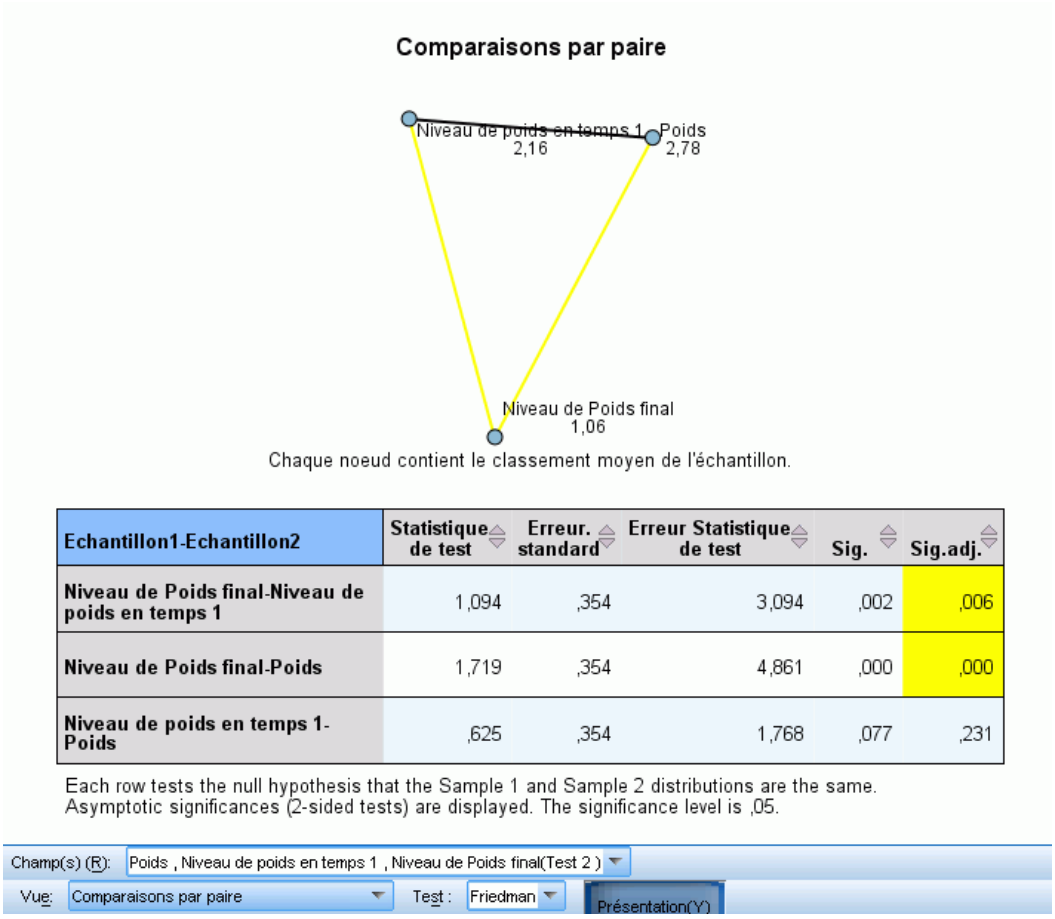
Figure 27-46
Informations sur les champs continus



La vue Informations sur les champs continus affiche un histogramme pour le champ continu sélectionné dans la liste déroulante Champ(s). La liste des champs disponibles est limitée aux champs continus utilisés dans le test actuellement sélectionné dans la vue Récapitulatif d'hypothèses.

Comparaisons par paire

Figure 27-47
Comparaisons par paire



Echantillon1-Echantillon2	Statistique de test	Erreur. standard	Erreur Statistique de test	Sig.	Sig.adj.
Niveau de Poids final-Niveau de poids en temps 1	1,094	,354	3,094	,002	,006
Niveau de Poids final-Poids	1,719	,354	4,861	,000	,000
Niveau de poids en temps 1-Poids	,625	,354	1,768	,077	,231

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is ,05.

La vue Comparaisons par paire affiche un diagramme de réseau des distances et un tableau des comparaisons produits par des tests non paramétriques à échantillon-*k*, lorsque des comparaisons par paire multiples sont requises.

- Le diagramme de réseau des distances est une représentation graphique du tableau des comparaisons dans lequel les distances entre les noeuds du réseau correspondent aux différences entre les échantillons. Les lignes jaunes correspondent aux différences statistiques significatives, alors que les lignes noires correspondent aux différences non significatives. Lorsque vous passez la souris sur une ligne du réseau, la signification ajustée de la différence entre les noeuds connectés par la ligne s'affiche dans une info-bulle.
- Le tableau des comparaisons affiche les résultats numériques de toutes les comparaisons par paire. Chaque ligne correspond à une comparaison par paire distincte. Cliquez sur un en-tête de colonne pour trier les lignes de la colonne concernée en fonction de leurs valeurs.

Sous-ensembles homogènes

Figure 27-48
Sous-ensembles homogènes

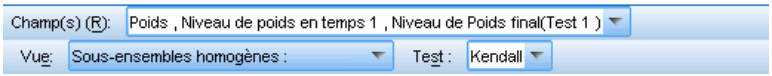
Sous-ensembles homogènes :

		Sous-ensemble		
		1	2	3
Echantillon[1]	Niveau de Poids final	1,063		
	Niveau de poids en temps 1		2,156	
	Poids			2,781
Statistique de test		.[2]	.[2]	.[2]
Sig. (2 - latéral)		.	.	.
Sig. ajustée (2 - latéral)		.	.	.

Les sous-ensembles homogènes sont basés sur les significations asymptotiques. Le niveau de signification est de ,05.

[1]Chaque cellule contient le classement moyen de l'échantillon.

[2]Unable to compute because the subset contains only one sample.



La vue Sous-ensembles homogènes affiche un tableau des comparaisons produits par des tests non paramétriques à échantillon-*k*, lorsque des comparaisons multiples pas à pas descendantes sont requises.

- Chaque ligne du groupe Echantillons correspond à un échantillon lié distinct (représenté dans les données par des champs distincts). Les échantillons qui ne diffèrent pas de manière significative d'un point de vue statistique sont groupés dans des sous-ensembles de même couleur, une colonne distincte comprenant chaque sous-ensemble identifié. Lorsque tous les échantillons diffèrent de manière significative d'un point de vue statistique, il n'y a qu'un sous-ensemble distinct pour chaque échantillon. Lorsque les échantillons ne diffèrent pas du tout de manière significative d'un point de vue statistique, il n'y a qu'un sous-ensemble unique.
- Une statistique de test, une valeur de signification et une valeur de signification ajustée sont calculées pour chaque sous-ensemble contenant plus d'un échantillon.

Fonctions supplémentaires de la commande NPTESTS

Le langage de syntaxe de commande vous permet aussi de :

- spécifier des tests à un échantillon, pour échantillon indépendants et pour échantillons liés en n'exécutant la procédure qu'une seule fois.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Boîtes de dialogue ancienne version

Il existe un certain nombre de boîtes de dialogue « ancienne version » qui effectuent également des tests non paramétriques. Ces boîtes de dialogue prennent en charge la fonctionnalité fournie par l'option Tests exacts.

Test Khi-deux : Tabule une variable en modalités et calcule une statistique khi-deux basée sur les différences entre les fréquences observées et les fréquences attendues.

Test binomial : Compare la fréquence observée dans chaque modalité d'une variable dichotomique avec les fréquences attendues de la distribution binomiale.

Suites en séquence : Teste si l'ordre d'occurrence de deux valeurs d'une variable est aléatoire.

Test Kolmogorov-Smirnov pour un échantillon : Compare la fonction de distribution cumulée observée pour une variable avec une distribution théorique spécifiée, qui peut être normale, uniforme, exponentielle ou Poisson.

Tests de deux échantillons indépendants : Compare deux groupes d'observations d'une variable. Le test *U* de Mann-Whitney, le test de Kolmogorov-Smirnov pour deux échantillons, le test de réactions extrêmes Moses et les suites en séquences Wald-Wolfowitz sont disponibles.

Tests de deux échantillons liés : Compare les distributions de deux variables. Le test de Wilcoxon, le test des signes et le test de McNemar sont disponibles.

Tests de plusieurs échantillons indépendants : Compare deux groupes d'observations ou plus d'une variable. Le test de Kruskal-Wallis, le test de la médiane et le test de Jonckheere-Terpstra sont disponibles.

Tests de plusieurs échantillons liés : Compare les distributions de deux variables ou plus. Le test de Friedman, le test *W* de Kendall et le test *Q* de Cochran sont disponibles.

Les quartiles et la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations sont disponibles pour tous les tests ci-dessus.

Test du Khi-deux

La procédure de test du Khi-deux tabule une variable en modalités et calcule une statistique Khi-deux. Ce test de qualité de l'ajustement compare les fréquences observées et attendues dans chaque modalité pour vérifier si toutes les modalités contiennent la même proportion de valeurs ou si chaque modalité contient une proportion de valeurs spécifiées par l'utilisateur.

Exemples : Le test du Khi-deux peut être utilisé pour déterminer si un sac de bonbons contient les mêmes proportions de bonbons bleus, marrons, verts, oranges, rouges et jaunes. Vous pouvez aussi tester si le sac de bonbons contient 5 % de bonbons bleus, 30 % de bonbons marrons, 10 % de bonbons verts, 20 % bonbons oranges, 15 % de bonbons rouges et 15 % de bonbons jaunes.

Statistiques : Moyenne, écart-type, minimum, maximum, et quartiles. Le nombre et le pourcentage d'observations manquantes et non manquantes, le nombre de cas observés et attendus pour chaque modalité, les résidus et la statistique du Khi-deux.

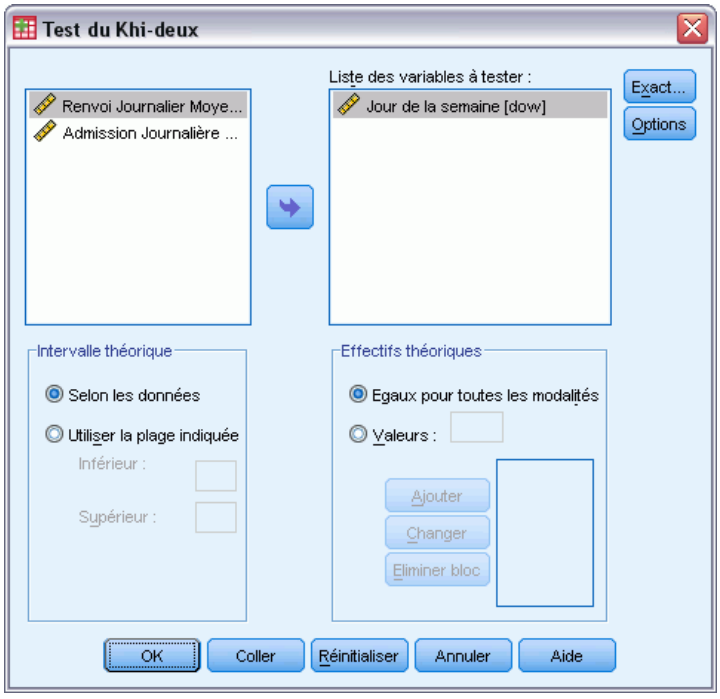
Données. Utilisez des variables qualitatives numériques ordonnées ou désordonnées (niveau de mesure ordinal ou nominal). Pour convertir des variables chaînes en variables numériques, utilisez la procédure de recodage automatique, disponible dans le menu Transformer.

Hypothèses : Les tests non paramétriques ne nécessitent pas d'hypothèses sur la forme de la distribution sous-jacente. On part du principe que les données constituent un échantillon aléatoire. Les fréquences attendues pour chaque modalité doivent être au moins égales à 1,20 % des modalités au maximum doivent avoir des fréquences inférieures à 5.

Pour obtenir un test Khi-deux

- A partir des menus, sélectionnez :
  Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > Khi-deux

Figure 27-49
Boîte de dialogue Test du khi-deux



- Sélectionnez des variables de test. Chaque variable produit un test distinct.
- Vous pouvez également cliquer sur Options pour les statistiques descriptives, les quartiles et le contrôle du traitement des données manquantes.

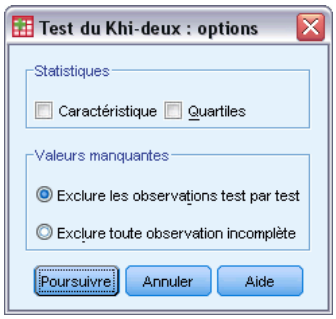
Valeurs et intervalles théoriques du test du Khi-deux

Intervalle théorique. Par défaut, chaque valeur distincte de la variable est définie comme modalité. Pour établir des modalités dans un intervalle spécifique, sélectionnez l'option Dans les limites spécifiées, et indiquez les valeurs entières pour les limites inférieure et supérieure. Des modalités sont établies pour chaque valeur entière comprise dans l'intervalle, et les observations à l'extérieur des limites sont exclues. Par exemple, si vous spécifiez une valeur de 1 pour la limite inférieure et une valeur de 4 pour la limite supérieure, seules les valeurs entières comprises entre 1 et 4 sont utilisées pour le test du Khi-deux.

Valeurs théoriques. Par défaut, toutes les modalités ont des valeurs théoriques égales. Les modalités peuvent avoir des proportions attendues définies par l'utilisateur. Sélectionnez Valeurs, indiquez une valeur supérieure à 0 pour chaque modalité de variable de test et cliquez ensuite sur Ajouter. Chaque fois que vous ajoutez une valeur, celle-ci apparaît au bas de la liste des valeurs. L'ordre des valeurs est important. Il correspond à l'ordre croissant des valeurs des modalités de la variable de test. La première valeur de la liste correspond à la valeur de groupe la plus basse de la variable de test et la dernière valeur correspond à la valeur la plus élevée. Les éléments de la liste des valeurs sont additionnés et chaque valeur est ensuite divisée par cette somme pour calculer la proportion d'observations attendues dans la modalité correspondante. Par exemple, une liste de valeurs de 3, 4, 5, 4 indique les proportions attendues de 3/16, 4/16, 5/16 et 4/16.

Test du Khi-deux : Options

Figure 27-50
Boîte de dialogue Test du Khi-deux : Options



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :** Affiche la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations non manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour l'une ou l'autre variable sont exclues de toutes les analyses.

Fonctionnalités supplémentaires de la commande NPAR TESTS (test du Khi-deux)

Le langage de syntaxe de commande vous permet aussi de :

- Spécifier des valeurs minimale et maximale différentes, ou des fréquences attendues pour différentes variables (avec la sous-commande `CHISQUARE`).
- Tester la même variable avec différentes fréquences attendues ou utiliser différentes plages (avec la sous-commande `EXPECTED`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Test binomial

La procédure de test binomial compare les fréquences observées des deux modalités d'une variable dichotomique avec les fréquences que l'on peut attendre d'une distribution binomiale avec un paramètre de probabilité spécifié. Par défaut, le paramètre de probabilité pour les deux groupes est de 0,5. Pour modifier les probabilités, vous pouvez entrer un test de proportion pour le premier groupe. La probabilité pour le second groupe sera de 1 moins la probabilité spécifiée pour le premier groupe.

Exemple : Quand vous lancez une pièce, la probabilité de tomber sur le côté face est de 1/2. Sur la base de cette hypothèse, une pièce est lancée 40 fois, et les résultats sont enregistrés (pile ou face). Du test binomial, il se peut que vous observiez que les 3/4 des lancements sont tombés sur le côté face et que le seuil de signification observé est bas (0,0027). Ces résultats indiquent qu'il est peu probable que la probabilité pour que la pièce tombe sur le côté face soit égale à 1/2. La pièce est probablement truquée.

Statistiques. Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles.

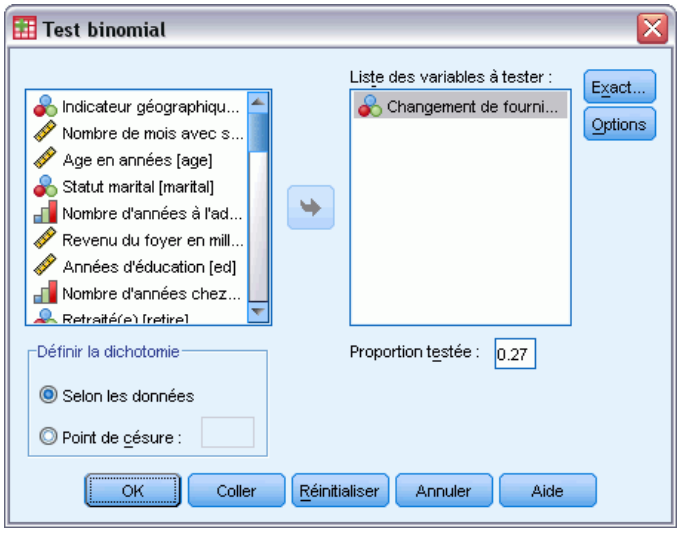
Données. Les variables testées doivent être numériques et dichotomiques. Pour convertir des variables chaînes en variables numériques, utilisez la procédure de recodage automatique, disponible dans le menu Transformer. Une **variable dichotomique** est une variable qui ne peut prendre que deux valeurs possibles : *oui* ou *non*, *vrai* ou *faux*, 0 ou 1, etc. La première valeur rencontrée dans l'ensemble de données définit le premier groupe, et l'autre valeur définit le deuxième groupe. Si les variables ne sont pas dichotomiques, vous devez spécifier une césure. La césure affecte les observations avec les valeurs qui sont inférieures ou égales au premier groupe et le reste des observations à un deuxième groupe.

Hypothèses : Les tests non paramétriques ne nécessitent pas d'hypothèses sur la forme de la distribution sous-jacente. On part du principe que les données constituent un échantillon aléatoire.

Pour obtenir un test binomial

▶ A partir des menus, sélectionnez :
Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > Binomial

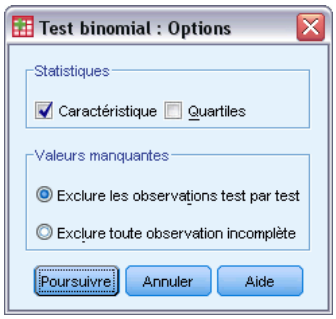
Figure 27-51
Boîte de dialogue Test binomial



- Sélectionnez une ou plusieurs variables numériques à tester.
- Vous pouvez également cliquer sur Options pour les statistiques descriptives, les quartiles et le contrôle du traitement des données manquantes.

Test binomial : Options

Figure 27-52
Boîte de dialogue Test binomial



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :** Affiche la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations non manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour toutes les variables testées sont exclues de toutes les analyses.

Fonctionnalités supplémentaires de la commande NPAR TESTS (Test binomial)

Le langage de syntaxe de commande vous permet aussi de :

- Sélectionner des groupes spécifiques (et en exclure d'autres) lorsqu'une variable comporte plus de deux modalités (avec la sous-commande `BINOMIAL`).
- Spécifier différentes césures ou probabilités pour différentes variables (avec la sous-commande `BINOMIAL`).
- Tester la même variable avec différentes césures ou probabilités (avec la sous-commande `EXPECTED`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Suites en séquences

La procédure Suites en séquences teste si l'ordre d'occurrence de deux valeurs d'une variable est aléatoire. Une séquence est une suite d'observations semblables. Un échantillon comportant trop ou trop peu de séquences suggère que l'échantillon n'est pas aléatoire.

Exemples : Supposons que 20 personnes soient sondées pour déterminer si elles achèteraient un produit donné. On peut douter que l'échantillon soit aléatoire si toutes les personnes sont du même sexe. Les suites en séquences peuvent être utilisées pour déterminer si l'échantillon a été tiré au hasard.

Statistiques : Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles.

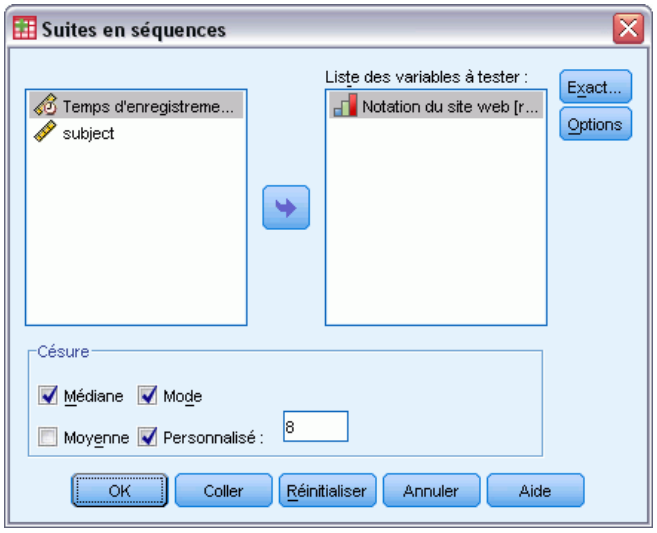
Données. Les variables doivent être numériques. Pour convertir des variables chaînes en variables numériques, utilisez la procédure de recodage automatique, disponible dans le menu Transformer.

Hypothèses : Les tests non paramétriques ne nécessitent pas d'hypothèses sur la forme de la distribution sous-jacente. Utilisez des échantillons à distribution de probabilité continue.

Pour obtenir un test de suites

▶ A partir des menus, sélectionnez :
Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > Séquences

Figure 27-53
Ajout de césures personnalisées



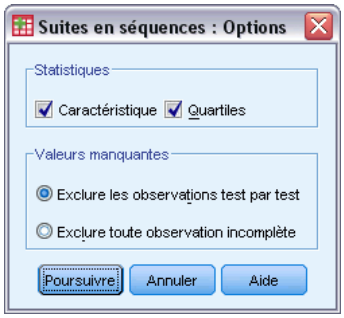
- Sélectionnez une ou plusieurs variables numériques à tester.
- Vous pouvez également cliquer sur Options pour les statistiques descriptives, les quartiles et le contrôle du traitement des données manquantes.

Césure des Suites en séquences

Césure : Spécifie une césure pour dichotomiser les variables que vous avez choisies. Vous pouvez utiliser soit la moyenne, la médiane ou le mode observés, soit une valeur spécifiée comme césure. Les observations dont les valeurs sont inférieures à la césure sont assignées à un groupe et les observations dont les valeurs sont supérieures à la césure sont assignées à l'autre groupe. Un test est réalisé pour chaque césure choisie.

Options des Suites en séquences

Figure 27-54
Boîte de dialogue Suites en séquences : Options



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :** Affiche la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations non manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour l'une ou l'autre variable sont exclues de toutes les analyses.

Fonctionnalités supplémentaires de la commande NPAR TESTS (Suites en séquences)

Le langage de syntaxe de commande vous permet aussi de :

- Spécifier différentes césures pour différentes variables (à l'aide de la sous-commande `RUNS`).
- Tester la même variable par rapport à différentes césures personnalisées (à l'aide de la sous-commande `RUNS`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Test Kolmogorov-Smirnov pour un échantillon

Le test de Kolmogorov-Smirnov pour un échantillon compare la fonction de distribution cumulée observée d'une variable avec une distribution théorique spécifiée, qui peut être normale, uniforme, de Poisson ou exponentielle. Le *Z* de Kolmogorov-Smirnov est calculé à partir de la plus grande différence (en valeur absolue) entre les fonctions de distribution cumulées observées et théoriques. Le test de qualité de l'ajustement contrôle si les observations peuvent avoir été raisonnablement déduites de la distribution spécifiée.

Exemple : La plupart des tests paramétriques nécessitent des variables distribuées normalement. Le test de Kolmogorov-Smirnov pour un échantillon permet de vérifier qu'une variable (par exemple *Revenu*) est distribuée normalement.

Statistiques : Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles.

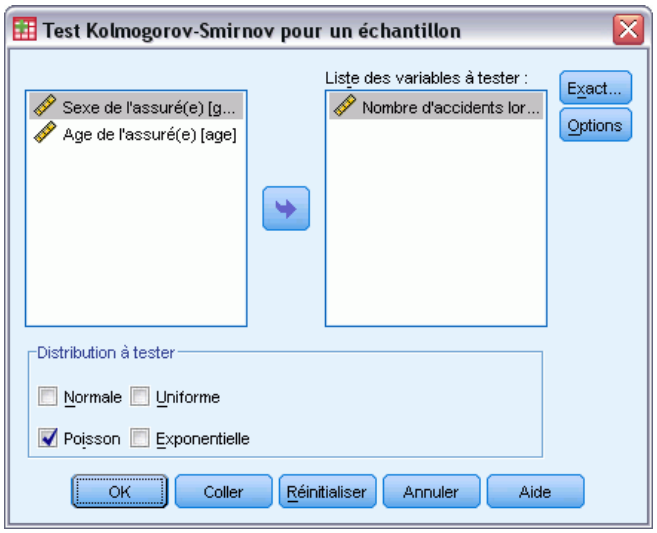
Données. Utilisez des variables quantitatives (mesure d'intervalle ou de rapport).

Hypothèses : Le test de Kolmogorov-Smirnov part du principe que les paramètres de la distribution à tester sont précisés a priori. Cette procédure estime les paramètres à partir d'un échantillon. L'échantillon de moyenne et l'échantillon d'écart type sont les paramètres pour une distribution normale. Les valeurs minimum et maximum de l'échantillon définissent l'intervalle de la distribution uniforme, l'échantillon de moyenne est le paramètre pour la distribution de Poisson et l'échantillon de l'écart-type est le paramètre pour la distribution exponentielle. La puissance du test à détecter les abandons de la distribution hypothétique peut être sérieusement diminuée. Pour le test d'une distribution normale avec des paramètres estimés, considérez le test K-S Lilliefors ajusté (disponible dans la procédure d'exploration).

Pour obtenir un test de Kolmogorov-Smirnov pour un échantillon

- A partir des menus, sélectionnez :
  Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > K-S à 1 échantillon

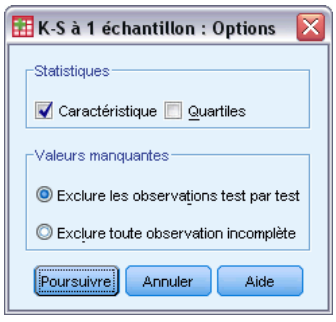
Figure 27-55
Boîte de dialogue Test de Kolmogorov-Smirnov pour un échantillon



- Sélectionnez une ou plusieurs variables numériques à tester. Chaque variable produit un test distinct.
- Vous pouvez également cliquer sur Options pour les statistiques descriptives, les quartiles et le contrôle du traitement des données manquantes.

Options du test de Kolmogorov-Smirnov pour un échantillon

Figure 27-56
Boîte de dialogue K-S pour un échantillon : Options



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :** Affiche la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations non manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour l'une ou l'autre variable sont exclues de toutes les analyses.

Fonctions supplémentaires de commandes NPAR TESTS (test de Kolmogorov-Smirnov pour un échantillon)

Le langage de syntaxe de commande vous permet également de spécifier des paramètres pour la distribution du test (avec la sous-commande `K-S`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Tests pour deux échantillons indépendants

La procédure des tests pour deux échantillons indépendants compare deux groupes d'observations en fonction d'une variable.

Exemple : De nouveaux appareils dentaires qui sont censés être plus confortables, avoir une apparence plus agréable et provoquer des progrès plus rapides pour le redressage des dents ont été développés. Pour savoir si les nouveaux appareils doivent être portés aussi longtemps que les anciens, 10 enfants sont choisis de façon aléatoire pour porter les anciens appareils et 10 autres pour porter les nouveaux appareils. Le test *U* de Mann-Whitney peut par exemple vous montrer que les sujets portant les nouveaux appareils ne doivent pas les porter aussi longtemps que les sujets utilisant les anciens appareils.

Statistiques : Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles. Tests : *U* de Mann-Whitney, réactions extrêmes de Moses, *Z* de Kolmogorov-Smirnov, suites de Wald-Wolfowitz.

Données. Utilisez des variables numériques qui peuvent être ordonnées.

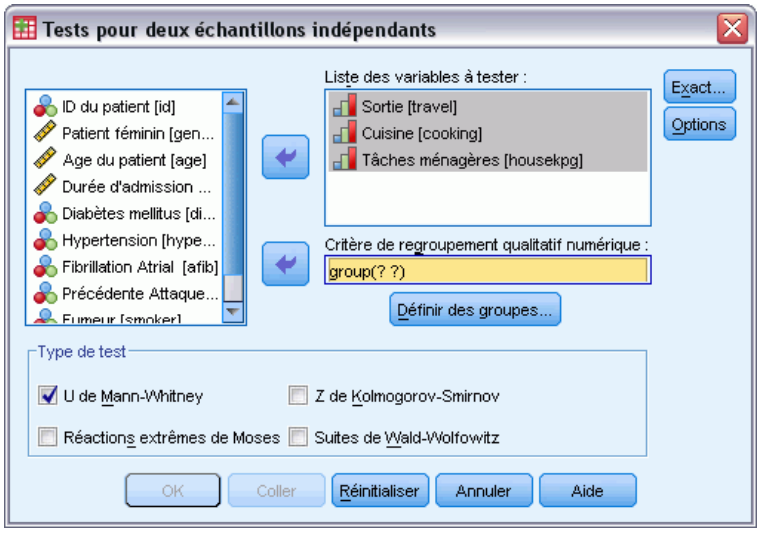
Hypothèses : Utilisez des échantillons indépendants, aléatoires. Le test *U* de Mann-Whitney teste l'égalité de deux distributions. Afin de l'utiliser pour tester les différences entre deux distributions, vous devez supposer que les distributions sont de la même forme.

Pour effectuer les tests pour deux échantillons indépendants

▶ A partir des menus, sélectionnez :

Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > 2 échantillons indépendants

Figure 27-57
Boîte de dialogue Tests pour deux échantillons indépendants



- Sélectionnez une ou plusieurs variables numériques.
- Sélectionnez une variable de regroupement et cliquez sur Définir groupes... pour scinder le fichier en deux groupes ou échantillons.

Types de tests pour deux échantillons indépendants

Type de test : Quatre tests sont disponibles pour tester si deux échantillons (groupes) proviennent de la même population.

Le **test U de Mann-Whitney** est le plus populaire des tests pour deux échantillons indépendants. Il équivaut au test de Wilcoxon et au test de Kruskal-Wallis pour deux groupes. Les tests de Mann-Whitney servent à vérifier que deux échantillons d'une population ont une position équivalente. Les observations des deux groupes sont combinées et ordonnées, et il leur est attribué un rang moyen en cas d'ex aequo. Le nombre d'ex aequo doit être petit par rapport au nombre total d'observations. Si les populations ont une position identique, les rangs doivent être attribués de façon aléatoire entre les deux échantillons. Le test calcule le nombre de fois qu'un résultat du groupe 1 précède un résultat du groupe 2, ainsi que le nombre de fois qu'un résultat du groupe 2 précède un résultat du groupe 1. La statistique du *U* de Mann-Whitney est la plus petite de ces deux nombres. La statistique de la somme des rangs de Wilcoxon *W* est également affichée. *W* est la somme des rangs pour le groupe avec le plus petit rang moyen, sauf si les groupes ont le même rang moyen, auquel cas il s'agit de la somme des rangs du groupe qui a été nommé en dernier dans la boîte de dialogue Définition des deux groupes d'échantillons indépendants..

Le **test Z de Kolmogorov-Smirnov** et les **suites en séquences de Wald-Wolfowitz** sont des tests plus généraux qui détectent les différences de position et la forme des distributions. Le test Z de Kolmogorov-Smirnov est basé sur la différence absolue maximum entre les fonctions de distribution cumulées observées pour les deux échantillons. Lorsque cette différence est significative, on considère que les deux distributions sont différentes. Le test des suites en séquences de Wald-Wolfowitz combine et ordonne les observations des deux groupes. Si les deux

échantillons proviennent de la même population, les deux groupes doivent être dispersés de façon aléatoire dans tout le classement.

Le **test des réactions extrêmes de Moses** part du principe que la variable expérimentale influence certains sujets dans une direction et d'autres sujets dans la direction opposée. Le test vérifie les réponses extrêmes par rapport à un groupe de contrôle. Ce test permet d'étudier l'intervalle du groupe de contrôle et de mesurer à quel point les valeurs extrêmes du groupe expérimental influencent l'amplitude lorsque ce test est associé au groupe de contrôle. Le groupe de contrôle est défini par la valeur du groupe 1 dans la boîte de dialogue Définition des deux groupes d'échantillons indépendants. Les observations des deux groupes sont combinées et ordonnées. L'intervalle du groupe de contrôle se calcule en effectuant la différence entre les rangs des valeurs les plus grandes et les plus petites du groupe de contrôle plus 1. Puisque des valeurs éloignées peuvent occasionnellement et facilement fausser l'intervalle d'amplitude, 5 % des observations de contrôle sont filtrées automatiquement à chaque extrémité.

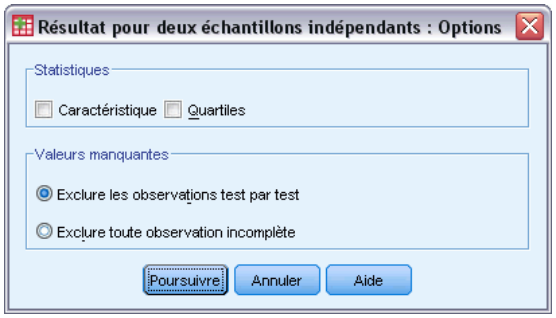
Définition de deux groupes d'échantillons indépendants

Figure 27-58
Boîte de dialogue Tests pour deux échantillons indépendants : Définir groupes

Pour scinder le fichier en deux groupes ou échantillons, indiquez un nombre entier pour le groupe 1 et une autre valeur pour le groupe 2. Les observations avec d'autres valeurs sont exclues de l'analyse.

Tests pour deux échantillons indépendants : Options

Figure 27-59
Boîte de dialogue Deux échantillons indépendants : Options



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :**Indique la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations sans valeurs manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour l'une ou l'autre variable sont exclues de toutes les analyses.

Fonctionnalités supplémentaires de la commande NPAR TESTS (tests pour deux échantillons indépendants)

Le langage de syntaxe de commande vous permet également de spécifier le nombre d'observations devant être filtrées pour le test de Moses (avec la sous-commande `MOSES`).

Reportez-vous à la *Référence de syntaxe de commande* pour une information complète concernant la syntaxe.

Tests pour deux échantillons liés

La procédure des tests pour deux échantillons liés compare les distributions pour deux variables.

Exemple : En général, une famille qui vend sa maison perçoit-elle le prix demandé ? En appliquant le test de Wilcoxon aux données de 10 foyers, vous apprendrez que sept familles perçoivent moins que le prix demandé, qu'une famille perçoit plus que le prix demandé et que deux familles perçoivent le prix demandé.

Statistiques : Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles. Tests : Classement Wilcoxon, Signe, McNemar. Si l'option Tests exacts est installée (disponible uniquement sous les systèmes d'exploitation Windows), le test d'Homogénéité marginale est alors disponible.

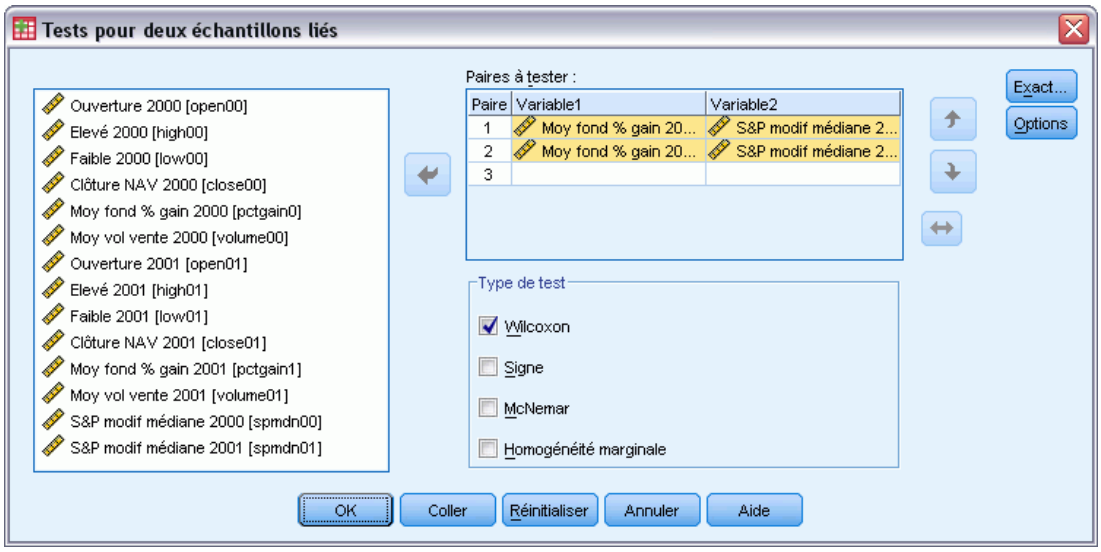
Données. Utilisez des variables numériques qui peuvent être ordonnées.

Hypothèses : Bien qu'aucune distribution particulière ne soit supposée pour les deux variables, on part du principe que la distribution de la population des différences liées est symétrique.

Pour obtenir des tests pour deux échantillons liés

▶ A partir des menus, sélectionnez :
Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > 2 échantillons liés

Figure 27-60
Boîte de dialogue Tests pour deux échantillons liés



▶ Sélectionnez une ou plusieurs paires de variables.

Types de tests pour deux échantillons liés

les tests de cette section comparent les distributions pour deux variables liées. Le test qu'il convient d'utiliser dépend du type de données.

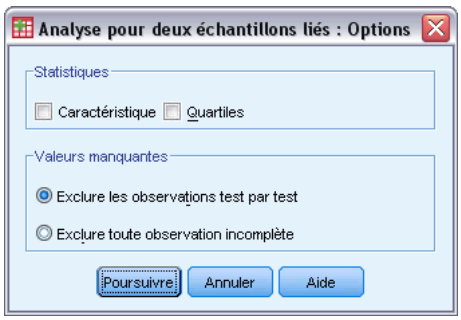
Si vos données sont continues, utilisez le test de Signe ou le test de Wilcoxon. Le **test de signe** calcule les différences entre les deux variables pour toutes les observations, et classe les différences comme étant positives, négatives ou liées. Si les deux variables sont réparties de la même manière, le nombre de différences positives et le nombre de différences négatives ne diffèrent pas de façon significative. Le **test de Wilcoxon** prend en compte les informations relatives au signe des différences, ainsi qu'à l'amplitude des différences entre paires. Comme le test de Wilcoxon intègre plus de renseignements sur les données, il est plus puissant que le test des signes.

Si vos données sont binaires, utilisez le **test de McNemar**. Ce test s'utilise fréquemment lors de situations de mesures répétées, au cours desquelles la réponse du sujet est provoquée deux fois, une fois avant qu'un événement spécifié se produise et une fois après qu'un événement spécifié s'est produit. Le test de McNemar détermine si le taux de réponses initial (avant l'événement) est égal au taux de réponse final (après l'événement). Ce test est utile pour détecter les changements dans les réponses dues à une intervention expérimentale dans les plans avant et après.

Si vos données sont qualitatives, utilisez le **test d'Homogénéité marginale**. Ce test est un développement du test de McNemar d'une réponse binaire à une réponse multinomiale. Il recherche les changements de réponse en utilisant la distribution Khi-deux et permet de détecter les changements de réponse dus à une intervention expérimentale dans les plans avant et après. Le test d'homogénéité marginale n'est disponible que si vous avez installé Exact Tests.

Tests pour deux échantillons liés : Options

Figure 27-61
Boîte de dialogue Tests pour deux échantillons liés : Options



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :**Indique la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations sans valeurs manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour l'une ou l'autre variable sont exclues de toutes les analyses.

Fonctionnalités supplémentaires de la commande NPAR (Deux échantillons liés)

Le langage de syntaxe de commande vous permet également de tester une variable avec chaque variable d'une liste.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Tests pour plusieurs échantillons indépendants

La procédure de Tests pour Plusieurs Echantillons Indépendants compare deux groupes d'observations ou plus sur une variable.

Exemple : Trois marques d'ampoules 100 watts diffèrent-elles par leur durée moyenne de fonctionnement ? A partir de l'analyse de variance d'ordre 1 de Kruskal-Wallis, vous apprendrez peut-être que les trois marques diffèrent par leur durée de vie moyenne.

Statistiques : Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles. Tests : *H* de Kruskal-Wallis, médiane.

Données. Utilisez des variables numériques qui peuvent être ordonnées.

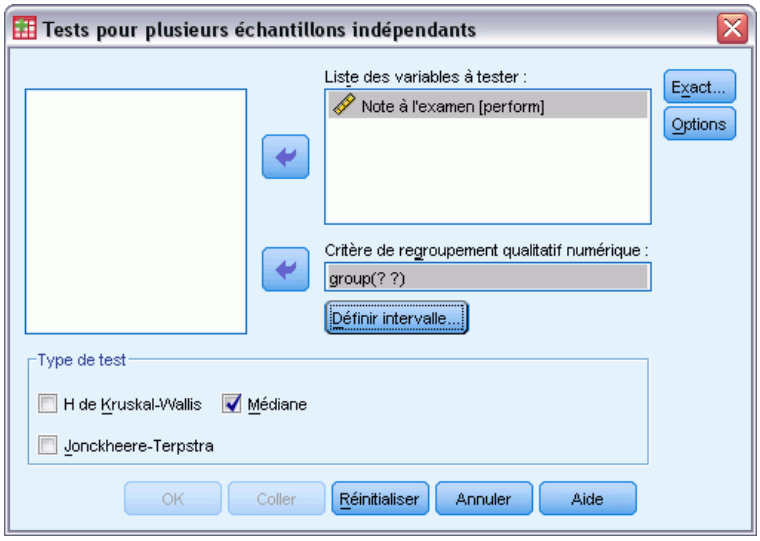
Hypothèses : Utilisez des échantillons indépendants, aléatoires. Le test du *H* de Kruskal-Wallis nécessite que les échantillons testés soient de forme similaire.

Pour obtenir des tests pour plusieurs échantillons indépendants

- A partir des menus, sélectionnez :

  Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > K échantillons indépendants

Figure 27-62
Définition du test de la médiane



- Sélectionnez une ou plusieurs variables numériques.
- Sélectionnez une variable de regroupement et cliquez sur Définir intervalle pour spécifier les valeurs entières minimale et maximale pour la variable de regroupement.

Tests pour Plusieurs Echantillons Indépendants : Types de tests

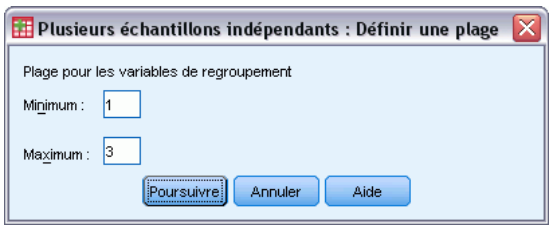
Trois tests permettent de déterminer si plusieurs échantillons indépendants proviennent de la même population. Le test du *H* de Kruskal-Wallis, le test de la Médiane et le test de Jonckheere-Terpstra testent tous si plusieurs échantillons indépendants proviennent de la même population.

Le test **H de Kruskal-Wallis**, extension du test du *U* de Mann-Whitney, est l'équivalent non paramétrique de l'analyse de variance d'ordre 1 et détecte les différences dans la position de la distribution. Le **test de la médiane**, test plus général mais moins puissant, détecte les différences de position et de forme des distributions. Le test du *H* de Kruskal-Wallis et le test de la médiane supposent qu'il n'existe aucun classement *a priori* des *k* populations à partir desquelles les échantillons sont tirés.

Lorsqu'il *existe* un classement naturel *a priori* (ascendant ou descendant) des *k* populations, le test de **Jonckheere-Terpstra** est plus puissant. Par exemple, les *k* populations peuvent représenter *k* températures croissantes. L'hypothèse selon laquelle différentes températures produisent la même distribution des réponses est testée contre l'hypothèse alternative selon laquelle l'accroissement de température fait augmenter la magnitude de la réponse. Ici, l'hypothèse alternative est ordonnée ; le test de Jonckheere-Terpstra est donc le plus approprié. Le test de Jonckheere-Terpstra n'est disponible que si vous avez installé le module complémentaire Tests Exacts.

Tests pour Plusieurs Echantillons Indépendants : Définir l'Intervalle

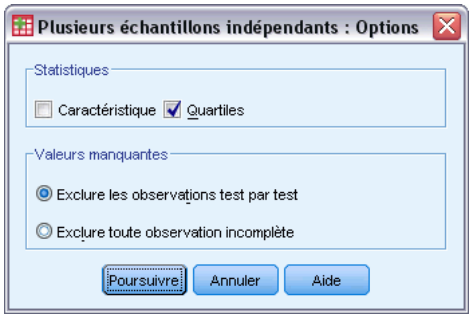
Figure 27-63
Boîte de dialogue Plusieurs échantillons indépendants : Définir l'intervalle



Pour définir l'intervalle, entrez des valeurs entières pour Minimum et Maximum qui correspondent à la modalité la plus basse et à la plus haute du critère de regroupement. Les observations dont les valeurs se trouvent à l'extérieur des limites sont exclues. Par exemple, si vous spécifiez une valeur minimale de 1 et une valeur maximale de 3, seules les valeurs entières comprises entre 1 et 3 seront utilisées. La valeur minimale doit être inférieure à la valeur maximale, et les deux valeurs doivent être spécifiées.

Options des Tests pour Plusieurs Echantillons Indépendants

Figure 27-64
Boîte de dialogue Plusieurs échantillons indépendants : Options



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :**Indique la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations sans valeurs manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour l'une ou l'autre variable sont exclues de toutes les analyses.

Fonctionnalités supplémentaires de la commande NPAR TESTS (K échantillons indépendants)

Le langage de syntaxe de commande vous permet également de spécifier une valeur différente de la médiane observée pour le test de la médiane (avec la sous-commande `MEDIAN`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Tests pour plusieurs échantillons liés

La procédure de Tests pour Plusieurs Echantillons Liés compare les distributions de deux variables ou plus.

Exemple : Le public associe-t-il différents niveaux de prestige à un docteur, un avocat, un officier de police et un enseignant ? On demande à dix personnes de classer ces quatre métiers par ordre de prestige. Le test de Friedman indique que le public associe effectivement différents niveaux de prestige à ces quatre professions.

Statistiques : Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles. Tests : Friedman, *W* de Kendall et *Q* de Cochran.

Données. Utilisez des variables numériques qui peuvent être ordonnées.

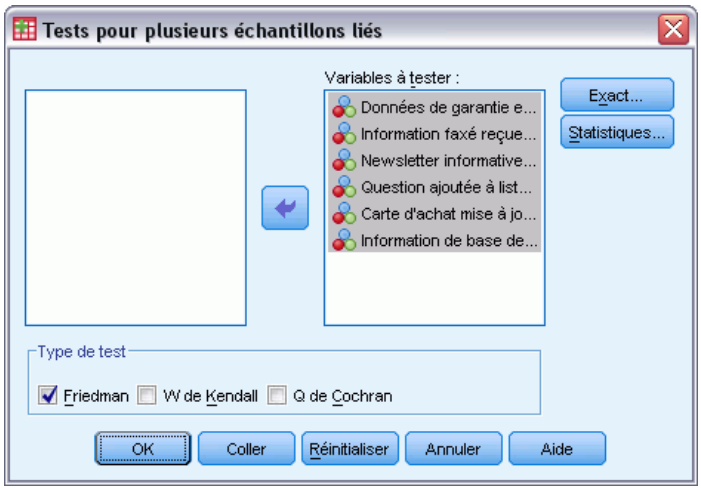
Hypothèses : Les tests non paramétriques ne nécessitent pas d'hypothèses sur la forme de la distribution sous-jacente. Utilisez des échantillons dépendants, aléatoires.

Pour obtenir des tests pour plusieurs échantillons liés

▶ A partir des menus, sélectionnez :

Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > K échantillons liés

Figure 27-65
Sélection de Cochran comme type de test



▶ Sélectionnez deux variables numériques ou plus à tester.

Tests pour plusieurs échantillons liés de types de tests

Trois tests sont disponibles pour comparer les distributions de plusieurs variables liées.

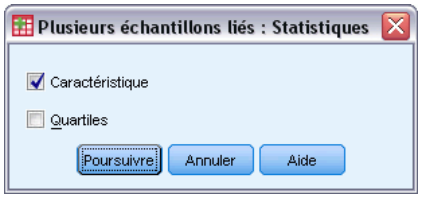
Le **test** de **Friedman** est l'équivalent non paramétrique d'un plan de mesures répétées sur un échantillon ou d'une analyse de variance d'ordre 2 avec une observation par cellule. Le test de Friedman teste l'hypothèse nulle selon laquelle *k* variables liées proviennent de la même population. Pour chaque observation, les variables *k* sont classées de 1 à *k*. La statistique de test est basée sur ces classements.

Le test **W de Kendall** est une standardisation de la statistique de Friedman. Le test *W* de Kendall peut être interprété comme le coefficient de concordance, qui est une mesure de l'accord entre les évaluateurs. Chaque observation est un juge ou un indicateur, et chaque variable est une personne ou un élément jugé. Pour chaque variable, la somme des rangs est calculée. Le *W* de Kendall se situe entre 0 (pas d'accord) et 1 (accord total).

Le **Q de Cochran** est identique au test de Friedman mais s'applique lorsque toutes les réponses sont binaires. Ce test est une extension du test de McNemar à *K* échantillons. Le *Q* de Cochran teste l'hypothèse nulle selon laquelle plusieurs variables dichotomiques liées ont la même moyenne. Les variables sont mesurées sur le même individu ou sur des individus comparables.

Statistiques des tests pour plusieurs échantillons liés

Figure 27-66
Boîte de dialogue Statistiques pour Plusieurs Echantillons Liés



Vous pouvez choisir les statitistiques.

- **Caractéristique :**Indique la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations sans valeurs manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Fonctionnalités supplémentaires de la commande NPAR TESTS (K échantillons liés)

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Test binomial

La procédure de test binomial compare les fréquences observées des deux modalités d'une variable dichotomique avec les fréquences que l'on peut attendre d'une distribution binomiale avec un paramètre de probabilité spécifié. Par défaut, le paramètre de probabilité pour les deux groupes est de 0,5. Pour modifier les probabilités, vous pouvez entrer un test de proportion pour le premier groupe. La probabilité pour le second groupe sera de 1 moins la probabilité spécifiée pour le premier groupe.

Exemple : Quand vous lancez une pièce, la probabilité de tomber sur le côté face est de 1/2. Sur la base de cette hypothèse, une pièce est lancée 40 fois, et les résultats sont enregistrés (pile ou face). Du test binomial, il se peut que vous observiez que les 3/4 des lancements sont tombés sur le côté face et que le seuil de signification observé est bas (0,0027). Ces résultats indiquent qu'il est peu probable que la probabilité pour que la pièce tombe sur le côté face soit égale à 1/2. La pièce est probablement truquée.

Statistiques. Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles.

Données. Les variables testées doivent être numériques et dichotomiques. Pour convertir des variables chaînes en variables numériques, utilisez la procédure de recodage automatique, disponible dans le menu Transformer. Une **variable dichotomique** est une variable qui ne peut prendre que deux valeurs possibles : *oui* ou *non*, *vrai* ou *faux*, 0 ou 1, etc. La première valeur rencontrée dans l'ensemble de données définit le premier groupe, et l'autre valeur définit le deuxième groupe. Si les variables ne sont pas dichotomiques, vous devez spécifier une césure. La césure affecte les observations avec les valeurs qui sont inférieures ou égales au premier groupe et le reste des observations à un deuxième groupe.

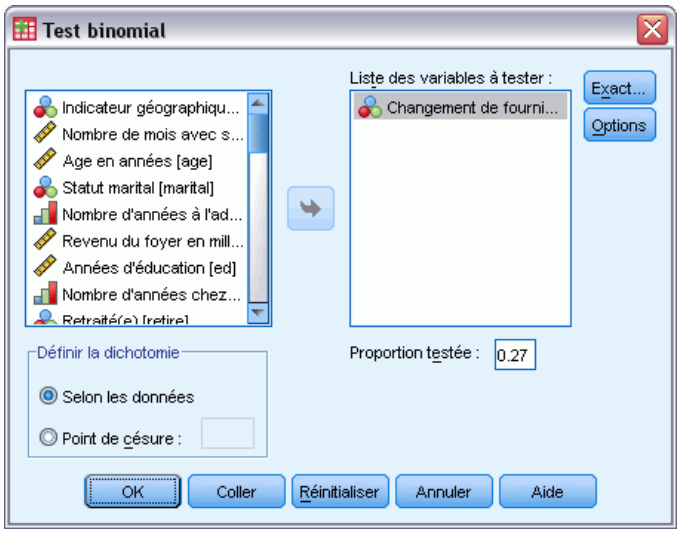
Hypothèses : Les tests non paramétriques ne nécessitent pas d'hypothèses sur la forme de la distribution sous-jacente. On part du principe que les données constituent un échantillon aléatoire.

Pour obtenir un test binomial

- A partir des menus, sélectionnez :

  Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > Binomial

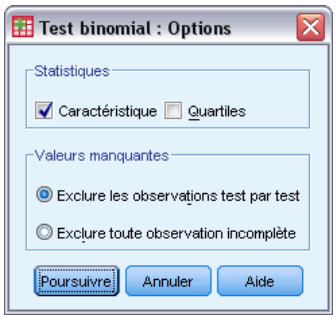
Figure 27-67
Boîte de dialogue Test binomial



- Sélectionnez une ou plusieurs variables numériques à tester.
- Vous pouvez également cliquer sur Options pour les statistiques descriptives, les quartiles et le contrôle du traitement des données manquantes.

Test binomial : Options

Figure 27-68
Boîte de dialogue Test binomial



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :** Affiche la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations non manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour toutes les variables testées sont exclues de toutes les analyses.

Fonctionnalités supplémentaires de la commande NPAR TESTS (Test binomial)

Le langage de syntaxe de commande vous permet aussi de :

- Sélectionner des groupes spécifiques (et en exclure d'autres) lorsqu'une variable comporte plus de deux modalités (avec la sous-commande `BINOMIAL`).
- Spécifier différentes césures ou probabilités pour différentes variables (avec la sous-commande `BINOMIAL`).
- Tester la même variable avec différentes césures ou probabilités (avec la sous-commande `EXPECTED`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Suites en séquences

La procédure Suites en séquences teste si l'ordre d'occurrence de deux valeurs d'une variable est aléatoire. Une séquence est une suite d'observations semblables. Un échantillon comportant trop ou trop peu de séquences suggère que l'échantillon n'est pas aléatoire.

Exemples : Supposons que 20 personnes soient sondées pour déterminer si elles achèteraient un produit donné. On peut douter que l'échantillon soit aléatoire si toutes les personnes sont du même sexe. Les suites en séquences peuvent être utilisées pour déterminer si l'échantillon a été tiré au hasard.

Statistiques : Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles.

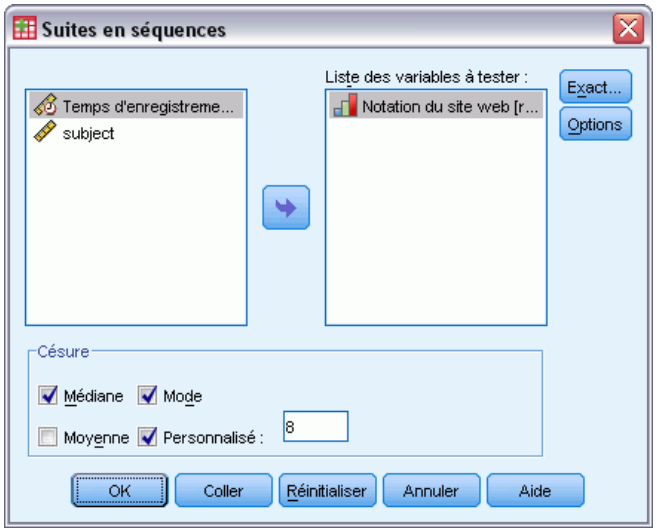
Données. Les variables doivent être numériques. Pour convertir des variables chaînes en variables numériques, utilisez la procédure de recodage automatique, disponible dans le menu Transformer.

Hypothèses : Les tests non paramétriques ne nécessitent pas d'hypothèses sur la forme de la distribution sous-jacente. Utilisez des échantillons à distribution de probabilité continue.

Pour obtenir un test de suites

▶ A partir des menus, sélectionnez :
Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > Séquences

Figure 27-69
Ajout de césures personnalisées



▶ Sélectionnez une ou plusieurs variables numériques à tester.

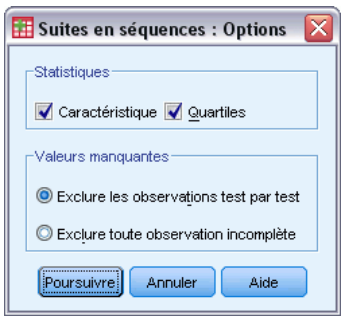
▶ Vous pouvez également cliquer sur Options pour les statistiques descriptives, les quartiles et le contrôle du traitement des données manquantes.

Césure des Suites en séquences

Césure : Spécifie une césure pour dichotomiser les variables que vous avez choisies. Vous pouvez utiliser soit la moyenne, la médiane ou le mode observés, soit une valeur spécifiée comme césure. Les observations dont les valeurs sont inférieures à la césure sont assignées à un groupe et les observations dont les valeurs sont supérieures à la césure sont assignées à l'autre groupe. Un test est réalisé pour chaque césure choisie.

Options des Suites en séquences

Figure 27-70
Boîte de dialogue Suites en séquences : Options



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :** Affiche la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations non manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour l'une ou l'autre variable sont exclues de toutes les analyses.

Fonctionnalités supplémentaires de la commande NPAR TESTS (Suites en séquences)

Le langage de syntaxe de commande vous permet aussi de :

- Spécifier différentes césures pour différentes variables (à l'aide de la sous-commande `RUNS`).
- Tester la même variable par rapport à différentes césures personnalisées (à l'aide de la sous-commande `RUNS`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Test Kolmogorov-Smirnov pour un échantillon

Le test de Kolmogorov-Smirnov pour un échantillon compare la fonction de distribution cumulée observée d'une variable avec une distribution théorique spécifiée, qui peut être normale, uniforme, de Poisson ou exponentielle. Le *Z* de Kolmogorov-Smirnov est calculé à partir de la plus grande différence (en valeur absolue) entre les fonctions de distribution cumulées observées et théoriques. Le test de qualité de l'ajustement contrôle si les observations peuvent avoir été raisonnablement déduites de la distribution spécifiée.

Exemple : La plupart des tests paramétriques nécessitent des variables distribuées normalement. Le test de Kolmogorov-Smirnov pour un échantillon permet de vérifier qu'une variable (par exemple *Revenu*) est distribuée normalement.

Statistiques : Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles.

Données. Utilisez des variables quantitatives (mesure d'intervalle ou de rapport).

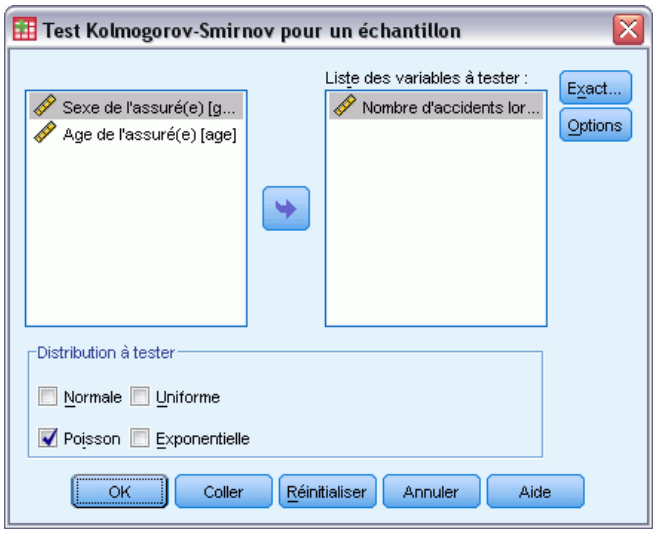
Hypothèses : Le test de Kolmogorov-Smirnov part du principe que les paramètres de la distribution à tester sont précisés a priori. Cette procédure estime les paramètres à partir d'un échantillon. L'échantillon de moyenne et l'échantillon d'écart type sont les paramètres pour une distribution normale. Les valeurs minimum et maximum de l'échantillon définissent l'intervalle de la distribution uniforme, l'échantillon de moyenne est le paramètre pour la distribution de Poisson et l'échantillon de l'écart-type est le paramètre pour la distribution exponentielle. La puissance du test à détecter les abandons de la distribution hypothétique peut être sérieusement diminuée. Pour le test d'une distribution normale avec des paramètres estimés, considérez le test K-S Lilliefors ajusté (disponible dans la procédure d'exploration).

Pour obtenir un test de Kolmogorov-Smirnov pour un échantillon

- A partir des menus, sélectionnez :

  Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > K-S à 1 échantillon

Figure 27-71
Boîte de dialogue Test de Kolmogorov-Smirnov pour un échantillon

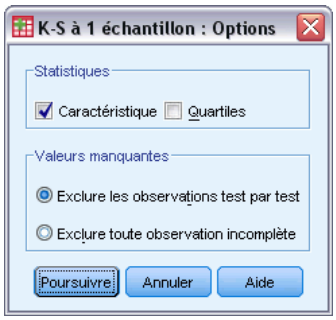


- Sélectionnez une ou plusieurs variables numériques à tester. Chaque variable produit un test distinct.

- Vous pouvez également cliquer sur Options pour les statistiques descriptives, les quartiles et le contrôle du traitement des données manquantes.

Options du test de Kolmogorov-Smirnov pour un échantillon

Figure 27-72
Boîte de dialogue K-S pour un échantillon : Options



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :** Affiche la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations non manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour l'une ou l'autre variable sont exclues de toutes les analyses.

Fonctions supplémentaires de commandes NPAR TESTS (test de Kolmogorov-Smirnov pour un échantillon)

Le langage de syntaxe de commande vous permet également de spécifier des paramètres pour la distribution du test (avec la sous-commande `K-S`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Tests pour deux échantillons indépendants

La procédure des tests pour deux échantillons indépendants compare deux groupes d'observations en fonction d'une variable.

Exemple : De nouveaux appareils dentaires qui sont censés être plus confortables, avoir une apparence plus agréable et provoquer des progrès plus rapides pour le redressage des dents ont été développés. Pour savoir si les nouveaux appareils doivent être portés aussi longtemps que les anciens, 10 enfants sont choisis de façon aléatoire pour porter les anciens appareils et 10 autres pour porter les nouveaux appareils. Le test *U* de Mann-Whitney peut par exemple vous montrer que les sujets portant les nouveaux appareils ne doivent pas les porter aussi longtemps que les sujets utilisant les anciens appareils.

Statistiques : Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles. Tests : *U* de Mann-Whitney, réactions extrêmes de Moses, *Z* de Kolmogorov-Smirnov, suites de Wald-Wolfowitz.

Données. Utilisez des variables numériques qui peuvent être ordonnées.

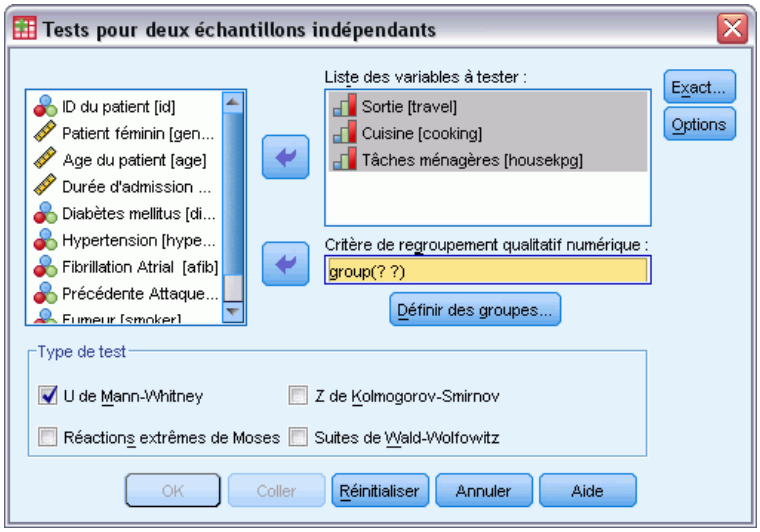
Hypothèses : Utilisez des échantillons indépendants, aléatoires. Le test *U* de Mann-Whitney teste l'égalité de deux distributions. Afin de l'utiliser pour tester les différences entre deux distributions, vous devez supposer que les distributions sont de la même forme.

Pour effectuer les tests pour deux échantillons indépendants

- A partir des menus, sélectionnez :

  Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > 2 échantillons indépendants

Figure 27-73
Boîte de dialogue Tests pour deux échantillons indépendants



- Sélectionnez une ou plusieurs variables numériques.
- Sélectionnez une variable de regroupement et cliquez sur Définir groupes... pour scinder le fichier en deux groupes ou échantillons.

Types de tests pour deux échantillons indépendants

Type de test : Quatre tests sont disponibles pour tester si deux échantillons (groupes) proviennent de la même population.

Le **test U de Mann-Whitney** est le plus populaire des tests pour deux échantillons indépendants. Il équivaut au test de Wilcoxon et au test de Kruskal-Wallis pour deux groupes. Les tests de Mann-Whitney servent à vérifier que deux échantillons d'une population ont une position équivalente. Les observations des deux groupes sont combinées et ordonnées, et il leur est attribué un rang moyen en cas d'ex aequo. Le nombre d'ex aequo doit être petit par rapport au nombre total d'observations. Si les populations ont une position identique, les rangs doivent être attribués

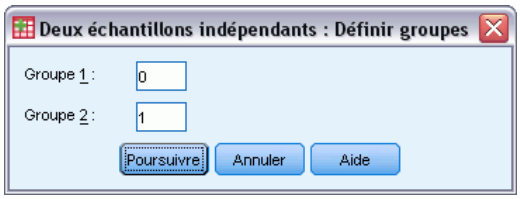
de façon aléatoire entre les deux échantillons. Le test calcule le nombre de fois qu'un résultat du groupe 1 précède un résultat du groupe 2, ainsi que le nombre de fois qu'un résultat du groupe 2 précède un résultat du groupe 1. La statistique du *U* de Mann-Whitney est la plus petite de ces deux nombres. La statistique de la somme des rangs de Wilcoxon *W* est également affichée. *W* est la somme des rangs pour le groupe avec le plus petit rang moyen, sauf si les groupes ont le même rang moyen, auquel cas il s'agit de la somme des rangs du groupe qui a été nommé en dernier dans la boîte de dialogue Définition des deux groupes d'échantillons indépendants..

Le **test Z de Kolmogorov-Smirnov** et les **suites en séquences de Wald-Wolfowitz** sont des tests plus généraux qui détectent les différences de position et la forme des distributions. Le test Z de Kolmogorov-Smirnov est basé sur la différence absolue maximum entre les fonctions de distribution cumulées observées pour les deux échantillons. Lorsque cette différence est significative, on considère que les deux distributions sont différentes. Le test des suites en séquences de Wald-Wolfowitz combine et ordonne les observations des deux groupes. Si les deux échantillons proviennent de la même population, les deux groupes doivent être dispersés de façon aléatoire dans tout le classement.

Le **test des réactions extrêmes de Moses** part du principe que la variable expérimentale influence certains sujets dans une direction et d'autres sujets dans la direction opposée. Le test vérifie les réponses extrêmes par rapport à un groupe de contrôle. Ce test permet d'étudier l'intervalle du groupe de contrôle et de mesurer à quel point les valeurs extrêmes du groupe expérimental influencent l'amplitude lorsque ce test est associé au groupe de contrôle. Le groupe de contrôle est défini par la valeur du groupe 1 dans la boîte de dialogue Définition des deux groupes d'échantillons indépendants. Les observations des deux groupes sont combinées et ordonnées. L'intervalle du groupe de contrôle se calcule en effectuant la différence entre les rangs des valeurs les plus grandes et les plus petites du groupe de contrôle plus 1. Puisque des valeurs éloignées peuvent occasionnellement et facilement fausser l'intervalle d'amplitude, 5 % des observations de contrôle sont filtrées automatiquement à chaque extrémité.

Définition de deux groupes d'échantillons indépendants

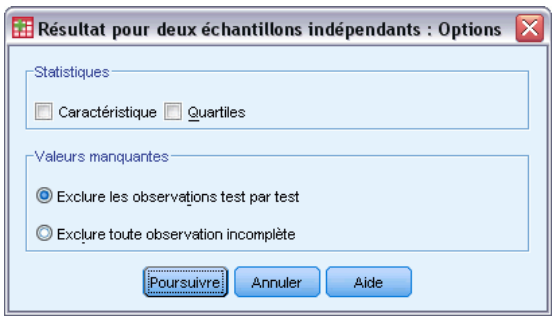
Figure 27-74
Boîte de dialogue Tests pour deux échantillons indépendants : Définir groupes



Pour scinder le fichier en deux groupes ou échantillons, indiquez un nombre entier pour le groupe 1 et une autre valeur pour le groupe 2. Les observations avec d'autres valeurs sont exclues de l'analyse.

Tests pour deux échantillons indépendants : Options

Figure 27-75
Boîte de dialogue Deux échantillons indépendants : Options



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :**Indique la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations sans valeurs manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour l'une ou l'autre variable sont exclues de toutes les analyses.

Fonctionnalités supplémentaires de la commande NPAR TESTS (tests pour deux échantillons indépendants)

Le langage de syntaxe de commande vous permet également de spécifier le nombre d'observations devant être filtrées pour le test de Moses (avec la sous-commande `MOSES`).

Reportez-vous à la *Référence de syntaxe de commande* pour une information complète concernant la syntaxe.

Tests pour deux échantillons liés

La procédure des tests pour deux échantillons liés compare les distributions pour deux variables.

Exemple : En général, une famille qui vend sa maison perçoit-elle le prix demandé ? En appliquant le test de Wilcoxon aux données de 10 foyers, vous apprendrez que sept familles perçoivent moins que le prix demandé, qu'une famille perçoit plus que le prix demandé et que deux familles perçoivent le prix demandé.

Statistiques : Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles. Tests : Classement Wilcoxon, Signe, McNemar. Si l'option Tests exacts est installée (disponible uniquement sous les systèmes d'exploitation Windows), le test d'Homogénéité marginale est alors disponible.

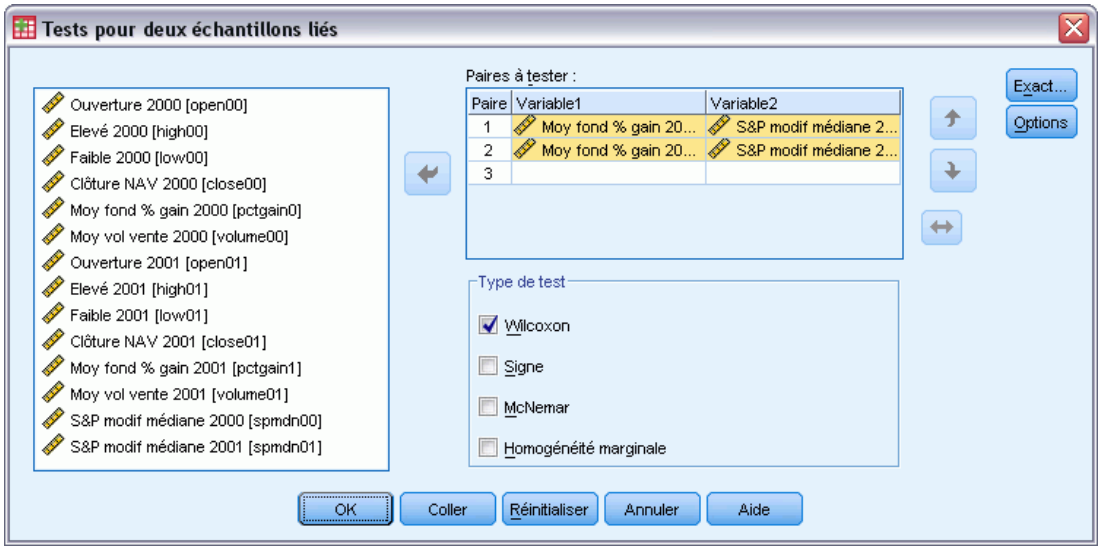
Données. Utilisez des variables numériques qui peuvent être ordonnées.

Hypothèses : Bien qu'aucune distribution particulière ne soit supposée pour les deux variables, on part du principe que la distribution de la population des différences liées est symétrique.

Pour obtenir des tests pour deux échantillons liés

- A partir des menus, sélectionnez :
  Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > 2 échantillons liés

Figure 27-76
Boîte de dialogue Tests pour deux échantillons liés



- Sélectionnez une ou plusieurs paires de variables.

Types de tests pour deux échantillons liés

les tests de cette section comparent les distributions pour deux variables liées. Le test qu'il convient d'utiliser dépend du type de données.

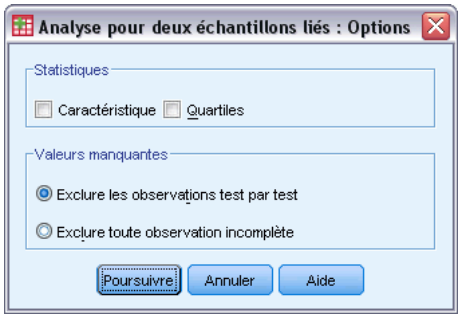
Si vos données sont continues, utilisez le test de Signe ou le test de Wilcoxon. Le **test de signe** calcule les différences entre les deux variables pour toutes les observations, et classe les différences comme étant positives, négatives ou liées. Si les deux variables sont réparties de la même manière, le nombre de différences positives et le nombre de différences négatives ne diffèrent pas de façon significative. Le **test de Wilcoxon** prend en compte les informations relatives au signe des différences, ainsi qu'à l'amplitude des différences entre paires. Comme le test de Wilcoxon intègre plus de renseignements sur les données, il est plus puissant que le test des signes.

Si vos données sont binaires, utilisez le **test de McNemar**. Ce test s'utilise fréquemment lors de situations de mesures répétées, au cours desquelles la réponse du sujet est provoquée deux fois, une fois avant qu'un événement spécifié se produise et une fois après qu'un événement spécifié s'est produit. Le test de McNemar détermine si le taux de réponses initial (avant l'événement) est égal au taux de réponse final (après l'événement). Ce test est utile pour détecter les changements dans les réponses dues à une intervention expérimentale dans les plans avant et après.

Si vos données sont qualitatives, utilisez le **test d'Homogénéité marginale**. Ce test est un développement du test de McNemar d'une réponse binaire à une réponse multinomiale. Il recherche les changements de réponse en utilisant la distribution Khi-deux et permet de détecter les changements de réponse dus à une intervention expérimentale dans les plans avant et après. Le test d'homogénéité marginale n'est disponible que si vous avez installé Exact Tests.

Tests pour deux échantillons liés : Options

Figure 27-77
Boîte de dialogue Tests pour deux échantillons liés : Options



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :**Indique la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations sans valeurs manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour l'une ou l'autre variable sont exclues de toutes les analyses.

Fonctionnalités supplémentaires de la commande NPAR (Deux échantillons liés)

Le langage de syntaxe de commande vous permet également de tester une variable avec chaque variable d'une liste.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Tests pour plusieurs échantillons indépendants

La procédure de Tests pour Plusieurs Echantillons Indépendants compare deux groupes d'observations ou plus sur une variable.

Exemple : Trois marques d'ampoules 100 watts diffèrent-elles par leur durée moyenne de fonctionnement ? A partir de l'analyse de variance d'ordre 1 de Kruskal-Wallis, vous apprendrez peut-être que les trois marques diffèrent par leur durée de vie moyenne.

Statistiques : Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles. Tests : *H* de Kruskal-Wallis, médiane.

Données. Utilisez des variables numériques qui peuvent être ordonnées.

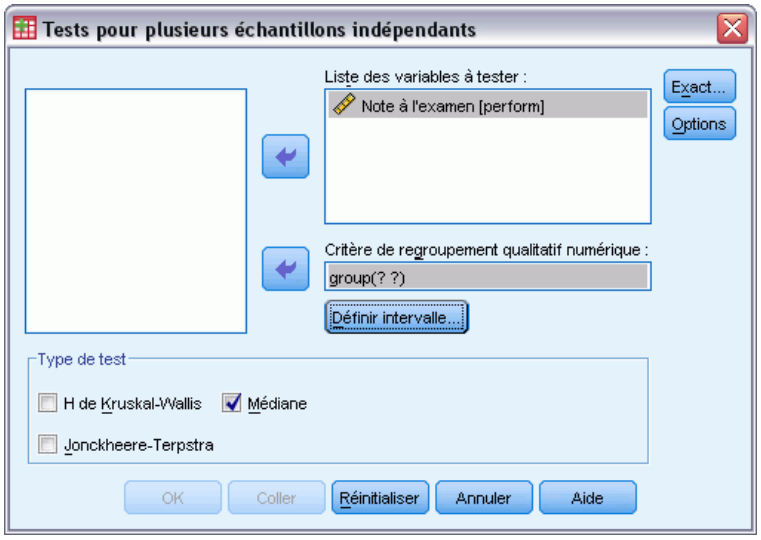
Hypothèses : Utilisez des échantillons indépendants, aléatoires. Le test du *H* de Kruskal-Wallis nécessite que les échantillons testés soient de forme similaire.

Pour obtenir des tests pour plusieurs échantillons indépendants

- A partir des menus, sélectionnez :

  Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > K échantillons indépendants

Figure 27-78
Définition du test de la médiane



- Sélectionnez une ou plusieurs variables numériques.
- Sélectionnez une variable de regroupement et cliquez sur Définir intervalle pour spécifier les valeurs entières minimale et maximale pour la variable de regroupement.

Tests pour Plusieurs Echantillons Indépendants : Types de tests

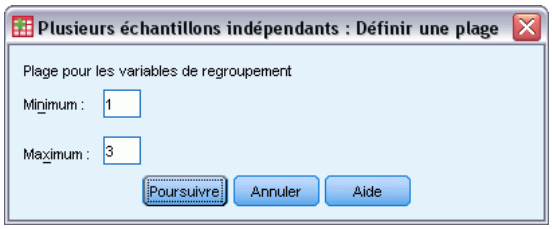
Trois tests permettent de déterminer si plusieurs échantillons indépendants proviennent de la même population. Le test du *H* de Kruskal-Wallis, le test de la Médiane et le test de Jonckheere-Terpstra testent tous si plusieurs échantillons indépendants proviennent de la même population.

Le test **H de Kruskal-Wallis**, extension du test du *U* de Mann-Whitney, est l'équivalent non paramétrique de l'analyse de variance d'ordre 1 et détecte les différences dans la position de la distribution. Le **test de la médiane**, test plus général mais moins puissant, détecte les différences de position et de forme des distributions. Le test du *H* de Kruskal-Wallis et le test de la médiane supposent qu'il n'existe aucun classement *a priori* des *k* populations à partir desquelles les échantillons sont tirés.

Lorsqu'il *existe* un classement naturel *a priori* (ascendant ou descendant) des *k* populations, le test de **Jonckheere-Terpstra** est plus puissant. Par exemple, les *k* populations peuvent représenter *k* températures croissantes. L'hypothèse selon laquelle différentes températures produisent la même distribution des réponses est testée contre l'hypothèse alternative selon laquelle l'accroissement de température fait augmenter la magnitude de la réponse. Ici, l'hypothèse alternative est ordonnée ; le test de Jonckheere-Terpstra est donc le plus approprié. Le test de Jonckheere-Terpstra n'est disponible que si vous avez installé le module complémentaire Tests Exacts.

Tests pour Plusieurs Echantillons Indépendants : Définir l'Intervalle

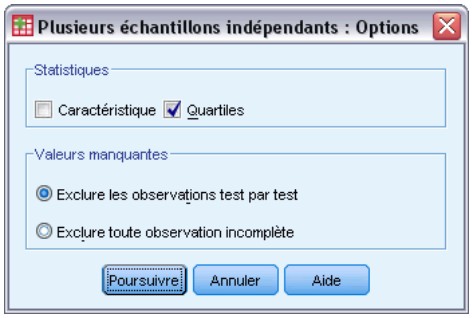
Figure 27-79
Boîte de dialogue Plusieurs échantillons indépendants : Définir l'intervalle



Pour définir l'intervalle, entrez des valeurs entières pour Minimum et Maximum qui correspondent à la modalité la plus basse et à la plus haute du critère de regroupement. Les observations dont les valeurs se trouvent à l'extérieur des limites sont exclues. Par exemple, si vous spécifiez une valeur minimale de 1 et une valeur maximale de 3, seules les valeurs entières comprises entre 1 et 3 seront utilisées. La valeur minimale doit être inférieure à la valeur maximale, et les deux valeurs doivent être spécifiées.

Options des Tests pour Plusieurs Echantillons Indépendants

Figure 27-80
Boîte de dialogue Plusieurs échantillons indépendants : Options



Statistiques : Vous pouvez choisir l'une des statistiques récapitulatives, ou bien les deux.

- **Caractéristique :**Indique la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations sans valeurs manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Valeurs manquantes : Contrôle le traitement des valeurs manquantes.

- **Exclure les observations test par test :** Lorsque plusieurs tests sont indiqués, chaque test est effectué séparément selon le nombre des valeurs manquantes.
- **Exclure toute observation incomplète :** Les observations avec des valeurs manquantes pour l'une ou l'autre variable sont exclues de toutes les analyses.

Fonctionnalités supplémentaires de la commande NPAR TESTS (K échantillons indépendants)

Le langage de syntaxe de commande vous permet également de spécifier une valeur différente de la médiane observée pour le test de la médiane (avec la sous-commande `MEDIAN`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Tests pour plusieurs échantillons liés

La procédure de Tests pour Plusieurs Echantillons Liés compare les distributions de deux variables ou plus.

Exemple : Le public associe-t-il différents niveaux de prestige à un docteur, un avocat, un officier de police et un enseignant ? On demande à dix personnes de classer ces quatre métiers par ordre de prestige. Le test de Friedman indique que le public associe effectivement différents niveaux de prestige à ces quatre professions.

Statistiques : Moyenne, écart-type, minimum, maximum, nombre d'observations avec des valeurs non manquantes et quartiles. Tests : Friedman, *W* de Kendall et *Q* de Cochran.

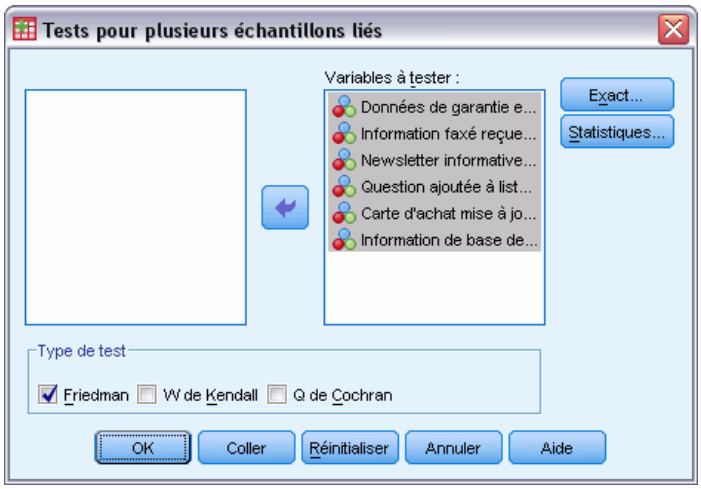
Données. Utilisez des variables numériques qui peuvent être ordonnées.

Hypothèses : Les tests non paramétriques ne nécessitent pas d'hypothèses sur la forme de la distribution sous-jacente. Utilisez des échantillons dépendants, aléatoires.

Pour obtenir des tests pour plusieurs échantillons liés

▶ A partir des menus, sélectionnez :
Analyse > Tests non paramétriques > Boîtes de dialogue ancienne version > K échantillons liés

Figure 27-81
Sélection de Cochran comme type de test



► Sélectionnez deux variables numériques ou plus à tester.

Tests pour plusieurs échantillons liés de types de tests

Trois tests sont disponibles pour comparer les distributions de plusieurs variables liées.

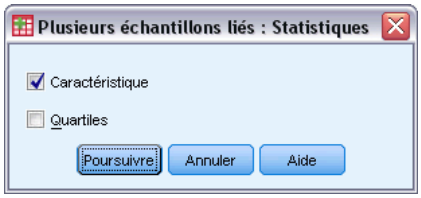
Le **test** de **Friedman** est l'équivalent non paramétrique d'un plan de mesures répétées sur un échantillon ou d'une analyse de variance d'ordre 2 avec une observation par cellule. Le test de Friedman teste l'hypothèse nulle selon laquelle *k* variables liées proviennent de la même population. Pour chaque observation, les variables *k* sont classées de 1 à *k*. La statistique de test est basée sur ces classements.

Le test **W de Kendall** est une standardisation de la statistique de Friedman. Le test *W* de Kendall peut être interprété comme le coefficient de concordance, qui est une mesure de l'accord entre les évaluateurs. Chaque observation est un juge ou un indicateur, et chaque variable est une personne ou un élément jugé. Pour chaque variable, la somme des rangs est calculée. Le *W* de Kendall se situe entre 0 (pas d'accord) et 1 (accord total).

Le **Q de Cochran** est identique au test de Friedman mais s'applique lorsque toutes les réponses sont binaires. Ce test est une extension du test de McNemar à *K* échantillons. Le *Q* de Cochran teste l'hypothèse nulle selon laquelle plusieurs variables dichotomiques liées ont la même moyenne. Les variables sont mesurées sur le même individu ou sur des individus comparables.

Statistiques des tests pour plusieurs échantillons liés

Figure 27-82
Boîte de dialogue Statistiques pour Plusieurs Echantillons Liés



Vous pouvez choisir les statitistiques.

- **Caractéristique :**Indique la moyenne, l'écart-type, le minimum, le maximum, et le nombre d'observations sans valeurs manquantes.
- **Quartiles :** Indique les valeurs correspondant au 25ème, 50ème et 75ème centiles.

Fonctionnalités supplémentaires de la commande NPAR TESTS (K échantillons liés)

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
28

Analyse des réponses multiples

Deux procédures sont proposées pour l'analyse des vecteurs comportant plusieurs variables dichotomiques ou plusieurs modalités. La procédure Fréquences multiréponses affiche les tableaux de fréquences. La procédure des Tableaux croisés multiréponses affiche des tableaux croisés à deux ou trois dimensions. Avant d'utiliser l'une de ces procédures, vous devez définir des vecteurs multiréponses.

Exemple : Cet exemple illustre l'utilisation des éléments multiréponses dans une étude de marché. Ces données sont fictives et ne doivent pas être considérées comme réelles. Une compagnie aérienne est parfois amenée à interroger les passagers d'un trajet donné pour évaluer la concurrence. Dans cet exemple, American Airlines veut savoir si ses passagers utilisent d'autres compagnies aériennes pour couvrir le trajet Chicago-New York et connaître l'importance relative des horaires et du service dans le choix d'une compagnie aérienne. L'hôtesse distribue à chaque passager un court questionnaire lors de l'embarquement. La première question est la suivante : Entourez toutes les compagnies aériennes par lesquelles vous avez effectué au moins un vol dans les six derniers mois parmi American, United, TWA, USAir et d'autres. Il s'agit d'une question à réponses multiples, car le passager peut entourer plus d'une réponse. Cependant, cette question ne peut pas être codée directement, parce qu'une variable ne peut avoir qu'une valeur pour chaque cas. Vous devez utiliser plusieurs variables pour mapper les réponses à chaque question. Ceci peut être fait de deux manières. L'une consiste à définir une variable correspondant à chaque choix possible (par exemple, American, United, TWA, USAir et d'autres). Si le passager entoure United, le numéro de code 1 est affecté à la variable *united*, sinon c'est le code 0 qui lui est affecté. Il s'agit de la méthode de codage de variables à **dichotomie multiple**. L'autre méthode permettant de mapper les réponses est la **méthode des modalités multiples**, où vous devez estimer le nombre maximal de réponses possibles à la question et définir le même nombre de variables, avec des codes correspondant à la compagnie aérienne empruntée. En utilisant un échantillon de questionnaires, vous vous apercevrez peut-être que personne n'a emprunté plus de trois compagnies différentes pour ce trajet. Qui plus est, vous vous rendrez compte que, du fait de la déréglementation des compagnies aériennes, 10 autres compagnies figurent dans la modalité Autre. A l'aide de la méthode multiréponses, vous pouvez définir trois variables, codées comme suit : 1 = *american*, 2 = *united*, 3 = *twa*, 4 = *usair*, 5 = *delta*, etc. Si un passager entoure American et TWA, la première variable porte le code 1, la seconde le code 3, et la troisième un code sans valeur. Un autre passager a peut-être entouré American et ajouté Delta. Ainsi, la première variable porte le code 1, la seconde le code 5, et la troisième un code sans valeur. Si vous utilisez la méthode des dichotomies multiples, d'un autre côté, vous finissez par vous retrouver avec 14 variables différentes. Les deux méthodes de codage sont possibles dans le cadre de cette enquête. Cependant, votre choix dépendra de la répartition des réponses.

Définition de vecteurs multiréponses

La procédure de définition de vecteurs multiréponses regroupe des variables itemaires dans des vecteurs de dichotomies ou de modalités, pour lesquels vous pouvez obtenir des tableaux de fréquences et des tableaux croisés. Vous pouvez définir jusqu'à 20 vecteurs multiréponses. Chaque vecteur doit avoir un nom unique. Pour éliminer un vecteur, sélectionnez-le dans la liste des vecteurs multiréponses et cliquez sur Eliminer. Pour modifier un vecteur, sélectionnez-le dans la liste, modifiez-en les caractéristiques et cliquez sur Changer.

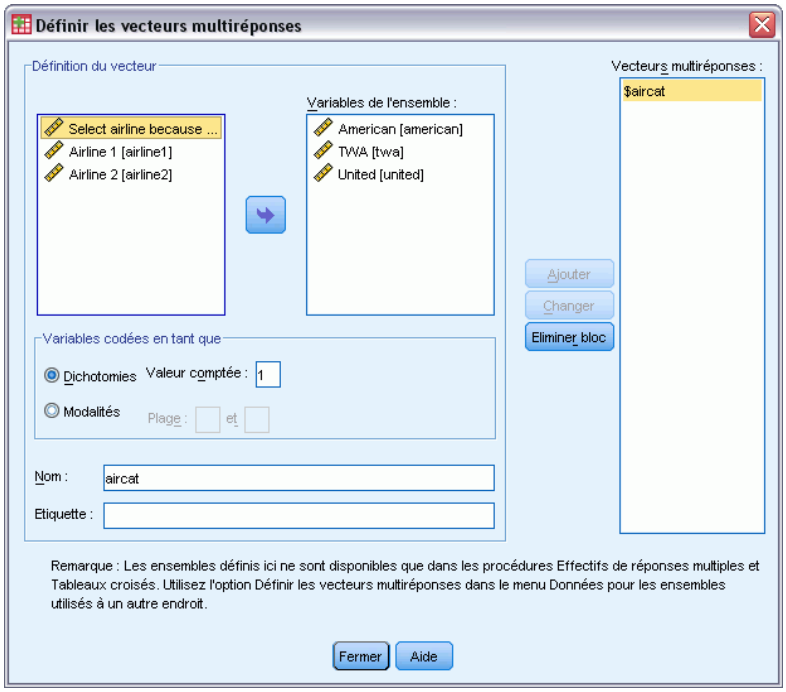
Vous pouvez coder vos variables itemaires sous forme de dichotomies ou de modalités. Pour utiliser des variables dichotomiques, sélectionnez Variables dichotomiques afin de créer un vecteur de dichotomies multiples. Entrez une valeur entière dans Valeur comptée. Chaque variable ayant au moins une occurrence de la valeur comptée devient une modalité du vecteur de dichotomies multiples. Sélectionnez Modalités pour créer un vecteur de modalités multiples ayant le même intervalle de valeurs que les variables qui le composent. Entrez des nombres entiers pour le minimum et le maximum de l'intervalle des modalités du vecteur de modalités multiples. La procédure totalise chaque valeur entière contenue dans l'intervalle pour toutes les variables qui le composent. Les modalités vides ne sont pas tabulées.

A chaque vecteur multiréponses doit être attribué un nom unique de 7 caractères maximum. La procédure ajoute un signe dollar ($) devant le nom que vous avez attribué. Les noms réservés suivants ne doivent pas être utilisés : *casenum*, *sysmis*, *jdate*, *date*, *time*, *length* et *width*. Le nom du vecteur multiréponses doit uniquement être utilisé dans les procédures multiréponses. Vous ne pouvez pas faire référence aux noms des vecteurs multiréponses dans les autres procédures. A titre facultatif, vous pouvez entrer une étiquette de variable décrivant le vecteur multiréponses. Cette étiquette peut comporter jusqu'à 40 caractères.

Définir des vecteurs de réponses multiples

▶ A partir des menus, sélectionnez :
Analyse > Réponses multiples > Définir des groupes de variables…

Figure 28-1
Boîte de dialogue Définition des vecteurs multiréponses



- Sélectionnez deux ou plusieurs variables.
- Si vos variables sont codées comme dichotomies, indiquez la valeur que vous souhaitez calculer. Si elles sont codées comme modalités, définissez leur intervalle.
- Entrez un nom unique pour chaque vecteur multiréponses.
- Cliquez sur Ajouter pour ajouter les vecteurs multiréponses à la liste des vecteurs définis.

Tableaux de fréquences des réponses multiples

La procédure Fréquences multiréponses produit des tableaux de fréquences pour les vecteurs multiréponses. Vous devez d'abord définir un ou plusieurs vecteurs multiréponses (voir " Définir les vecteurs multiréponses ").

Pour les vecteurs de dichotomies multiples, les noms de modalité apparaissant dans le résultat proviennent d'étiquettes de variable définies pour les variables itemaires du groupe. Si les étiquettes de variable ne sont pas définies, les noms de variables servent d'étiquettes. Pour les vecteurs de modalités multiples, les étiquettes des modalités proviennent des étiquettes de valeurs de la première variable du groupe. Si les modalités manquantes de la première variable sont présentes pour d'autres variables du groupe, définissez une étiquette de valeurs pour les modalités manquantes.

Valeurs manquantes : Les cas de valeurs manquantes sont exclus tableau par tableau. Vous pouvez donc choisir l'une ou les deux solutions suivantes :

- **Exclure les observations ayant une information incomplète à l'intérieur des dichotomies :** Ceci permet d'exclure les observations ayant des valeurs manquantes pour toute variable issue du tableau croisé du vecteur de dichotomies multiples. Ceci s'applique seulement aux vecteurs multiréponses définis comme vecteurs de dichotomies. Par défaut, une observation est considérée manquante pour un vecteur de dichotomies multiples si aucune de ses variables composantes ne contient de valeur comptée. Les cas de valeurs manquantes pour certaines variables, mais pas toutes, sont inclus dans les tabulations du groupe si au moins une variable contient la valeur comptée.
- **Exclure toute observation ayant une information incomplète à l'intérieur des modalités :** Cela permet d'exclure les observations ayant des valeurs manquantes pour toute variable provenant du tableau croisé du vecteur des modalités multiples. Ceci s'applique seulement aux vecteurs multiréponses définis comme des vecteurs de modalités. Par défaut, une observation est considérée manquante pour un vecteur de modalités multiples si aucune de ses composantes n'a de valeurs valides à l'intérieur de l'intervalle défini.

Exemple : Chaque variable créée à partir d'une question de l'enquête est une variable élémentaire. Pour analyser un élément multiréponses, vous devez combiner les variables dans l'un des deux types de vecteurs multiréponses : vecteur de modalités multiples ou vecteur de dichotomies multiples. Par exemple, si dans une enquête, une compagnie aérienne vous demande la compagnie (American Airlines, United Airlines ou TWA) que vous avez empruntée au cours des six derniers mois, si vous utilisez des variables dichotomiques et avez défini un **vecteur de dichotomies multiples**, chacune des trois variables du vecteur devient une modalité de la variable de regroupement. Les effectifs et les pourcentages correspondant aux trois compagnies aériennes s'affichent dans un tableau de fréquences. Si vous découvrez qu'aucun des répondants n'a mentionné plus de deux compagnies, vous pouvez créer deux variables, chacune ayant trois codes, un par compagnie aérienne. Si vous définissez un **vecteur de modalités multiples**, les valeurs sont tabulées et les mêmes codes sont ajoutés dans toutes les variables élémentaires. Le vecteur de valeurs résultant est le même que pour chacune des variables itemaires. Par exemple, 30 réponses pour United représentent la somme des cinq réponses United pour la compagnie aérienne 1 et des 25 réponses United pour la compagnie 2. Les effectifs et les pourcentages correspondant aux trois compagnies aériennes s'affichent dans un tableau de fréquences.

Statistiques : Tableaux de fréquences contenant des effectifs, des pourcentages de réponses, des pourcentages de cas, le nombre de cas valables, et le nombre de cas manquants.

Données : Utilisez des vecteurs multiréponses.

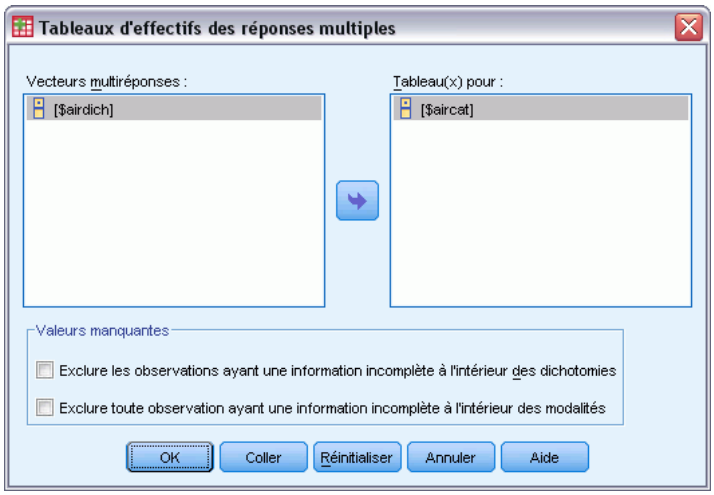
Hypothèses : Les effectifs et pourcentages représentent une description utile des données de n'importe quelle distribution.

Procédures apparentées : La procédure Définir Vecteurs multiréponses vous permet de définir des vecteurs multiréponses.

Pour obtenir des tableaux de fréquences de réponses multiples

▶ A partir des menus, sélectionnez :
Analyse > Réponses multiples > Fréquences

Figure 28-2
Boîte de dialogue Fréquences de réponses multiples



▶ Sélectionnez un ou plusieurs vecteurs multiréponses.

Tableaux croisés des réponses multiples

La procédure Tableaux croisés de réponses multiples classe, par tableaux croisés, des vecteurs multiréponses définis, des variables itemaires ou une combinaison. Vous pouvez également obtenir des pourcentages de cellules basés sur des observations ou des réponses, modifier la gestion des valeurs manquantes ou obtenir des tableaux croisés appariés. Vous devez d'abord définir un ou plusieurs vecteurs multiréponses (veuillez consulter " Pour Définir des vecteurs multiréponses ").

Pour les vecteurs de dichotomies multiples, les noms de modalité apparaissant dans le résultat proviennent d'étiquettes de variable définies pour les variables itemaires du groupe. Si les étiquettes de variable ne sont pas définies, les noms de variables servent d'étiquettes. Pour les vecteurs de modalités multiples, les étiquettes des modalités proviennent des étiquettes de valeurs de la première variable du groupe. Si les modalités manquantes de la première variable sont présentes pour d'autres variables du groupe, définissez une étiquette de valeurs pour les modalités manquantes. La procédure affiche les étiquettes de modalité des colonnes sur trois lignes, avec jusqu'à huit caractères par ligne. Pour éviter de scinder les mots, vous pouvez inverser les éléments lignes et les éléments colonnes ou redéfinir les étiquettes.

Exemple : Les vecteurs de dichotomies multiples et les vecteurs de modalités multiples peuvent être croisés avec d'autres variables dans cette procédure. Dans le cadre d'une enquête menée auprès de passagers de compagnies aériennes, voici ce qui leur a été demandé : Parmi les compagnies aériennes suivantes, entourez toutes celles avec lesquelles vous avez voyagé au moins une fois durant les six derniers mois (American, United, TWA). Est-il plus important de privilégier l'horaire ou le service ? Choisissez une seule réponse. Après avoir saisi les données en tant que dichotomies ou modalités multiples, et après les avoir combinées dans un vecteur, vous pouvez croiser les choix de compagnie aérienne déclarés avec la question relative au service ou aux horaires.

Statistiques : Tableau croisé avec cellule, ligne, colonne, et effectif total, et avec les pourcentages ligne, colonne, et effectif total. Les pourcentages cellule peuvent être basés sur les observations ou les réponses.

Données : Utilisez des vecteurs multiréponses ou des variables qualitatives numériques.

Hypothèses : Les effectifs et pourcentages offrent une description utile des données qui suivent tout type de distribution.

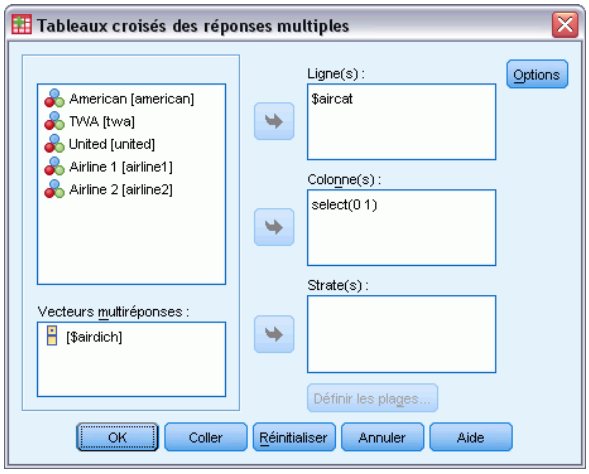
Procédures apparentées : La procédure Définir Vecteurs multiréponses vous permet de définir des vecteurs multiréponses.

Pour obtenir des tableaux croisés des réponses multiples

- A partir des menus, sélectionnez :

  Analyse > Réponses multiples > Tableaux croisés

Figure 28-3
Boîte de dialogue Tableaux Croisés de réponses multiples

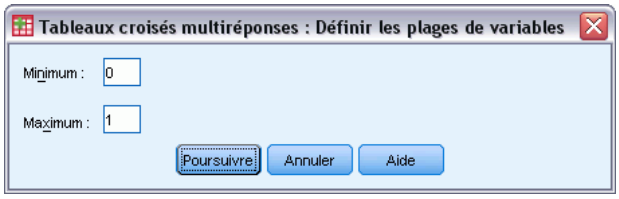


- Sélectionnez une ou plusieurs variables numériques ou vecteurs multiréponses pour chaque dimension de tableau croisé.
- Définissez l'intervalle de chaque variable itemaire.

  Sinon, vous pouvez obtenir un tableau croisé bilatéral pour chaque modalité de variable de contrôle ou chaque vecteur multiréponses. Sélectionnez un ou plusieurs éléments pour la liste de strate(s).

Définir Intervalles Tableaux croisés De réponses multiples

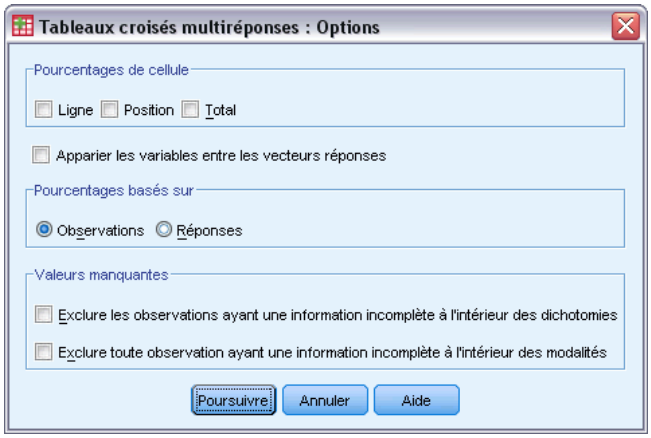
Figure 28-4
Définir boîte de dialogue Intervalle Variables Tableaux croisés de réponses multiples



Les intervalles des valeurs doivent être définis pour toute variable itemaire de tableaux croisés. Entrez les valeurs entières de modalités minimum et maximum que vous souhaitez tabuler. Les modalités se situant en dehors de l'intervalle sont exclues de l'analyse. Les valeurs se situant à l'intérieur de l'intervalle inclusif sont supposées être des nombres entiers (les nombres non entiers sont tronqués).

Options Tableaux croisés de réponses multiples

Figure 28-5
Boîte de dialogue Options Tableaux croisés de réponses multiples



Pourcentages de cellule : Les effectifs des cellules sont toujours affichés. Vous pouvez choisir d'afficher les pourcentages lignes, les pourcentages colonnes, et les pourcentages tableau bilatéral (total).

Pourcentages basés sur : Vous pouvez baser les pourcentages cellules sur les observations (ou répondants). Ceci n'est pas possible si vous sélectionnez la fonction qui permet d'apparier les variables entre les vecteurs de modalités multiples. Vous pouvez aussi baser les pourcentages cellules sur les réponses. Pour les vecteurs de dichotomies multiples, le nombre de réponses est égal au nombre de valeurs comptées à travers les observations. Pour les vecteurs de modalités multiples, le nombre de réponses correspond au nombre de valeurs comprises dans l'intervalle défini.

Valeurs manquantes : Vous avez le choix entre les deux options suivantes :

- **Exclure les observations ayant une information incomplète à l'intérieur des dichotomies :** Ceci permet d'exclure les observations ayant des valeurs manquantes pour toute variable issue du tableau croisé du vecteur de dichotomies multiples. Ceci s'applique seulement aux vecteurs multiréponses définis comme vecteurs de dichotomies. Par défaut, une observation est considérée manquante pour un vecteur de dichotomies multiples si aucune de ses variables composantes ne contient de valeur comptée. Les observations ayant des valeurs manquantes pour certaines, mais pas toutes, les variables sont incluses dans les tableaux croisés du groupe si au moins une variable contient la valeur comptée.
- **Exclure toute observation ayant une information incomplète à l'intérieur des modalités :** Cela permet d'exclure les observations ayant des valeurs manquantes pour toute variable provenant du tableau croisé du vecteur des modalités multiples. Ceci s'applique seulement aux vecteurs

multiréponses définis comme des vecteurs de modalités. Par défaut, une observation est considérée manquante pour un vecteur de modalités multiples si aucune de ses composantes n'a de valeurs valides à l'intérieur de l'intervalle défini.

Par défaut, lorsque vous croisez deux vecteurs de modalités multiples, la procédure tabule chaque variable du premier groupe avec chaque variable du second groupe et additionne les effectifs de chaque cellule. Par conséquent, certaines réponses peuvent apparaître plus d'une fois dans un tableau. Vous pouvez choisir l'option suivante :

Apparier les variables entre les vecteurs réponses : Cela permet d'apparier la première variable du premier groupe avec la première variable du second groupe, etc. Si vous sélectionnez cette option, la procédure basera les pourcentages cellules sur les réponses plutôt que sur les répondants. On ne peut apparier les vecteurs de dichotomies multiples ou les variables itemaires.

Fonctionnalités supplémentaires de la commande MULT RESPONSE

Le langage de syntaxe de commande vous permet aussi de :

- Obtenir des tableaux croisés ayant jusqu'à cinq dimensions (avec la sous-commande `BY`).
- Modifier les options de formatage du résultat, y compris la suppression des étiquettes de valeurs (avec la sous-commande `FORMAT`).

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
29

Tableaux de Résultats

Les listes d'observations et les statistiques descriptives sont des outils de base permettant d'étudier et de présenter des données. Vous pouvez obtenir les listes d'observations à l'aide de l'éditeur de Données ou de la procédure Récapituler, les effectifs de fréquence et les statistiques descriptives à l'aide de la procédure Fréquences, et les statistiques de sous-population à l'aide de la procédure Moyennes. Chacune de ces procédures utilise un format destiné à rendre les informations claires. Si vous souhaitez afficher les informations dans un format différent, les procédures Tableaux de bord en lignes et Tableaux de bord en colonnes vous permettent de contrôler la présentation des données.

Tableaux de bord en lignes

Tableaux de bord en lignes produit des tableaux de bord dans lesquels différentes statistiques récapitulatives sont disposées en lignes. Les listes d'observations sont également disponibles, avec ou sans statistiques récapitulatives.

Exemple : Une société possédant une chaîne de magasins conserve des dossiers sur les employés comprenant le salaire, l'ancienneté, le magasin et le service où chaque employé travaille. Vous pourriez générer un tableau de bord fournissant les informations individuelles sur les employés (liste) divisées par magasin et par division (critères d'agrégation), avec les statistiques récapitulatives (par exemple, salaire moyen) pour chaque magasin, division et par division dans chaque magasin.

Variables en colonnes : Donne la liste des variables de tableau pour lesquelles vous voulez obtenir des listes d'observations ou des statistiques récapitulatives, et contrôle le format d'affichage des Variables en colonnes.

Variables de ventilation : Donne la liste des critères d'agrégation optionnels qui divisent le tableau de bord en groupes et contrôle les statistiques récapitulatives et les formats d'affichage des colonnes de ventilation. Pour les critères d'agrégation multiples, il y aura un groupe séparé pour chaque modalité de chaque critère d'agrégation à l'intérieur des modalités du critère d'agrégation précédent dans la liste. Les critères d'agrégation doivent être des variables qualitatives discrètes qui divisent les observations en un nombre limité de modalités significatives. Les valeurs individuelles de chaque critère d'agrégation apparaissent, triées, dans une colonne séparée à gauche des Variables en colonnes.

Tableau de bord : Contrôle les caractéristiques globales du tableau de bord, y compris les statistiques récapitulatives globales, l'affichage des valeurs manquantes, la numérotation des pages et les titres.

Afficher les observations : Affiche les valeurs réelles (ou les étiquettes de valeurs) des variables de Variables en colonnes pour chaque observation. Cela produit une liste , qui peut être nettement plus longue qu'un tableau de bord.

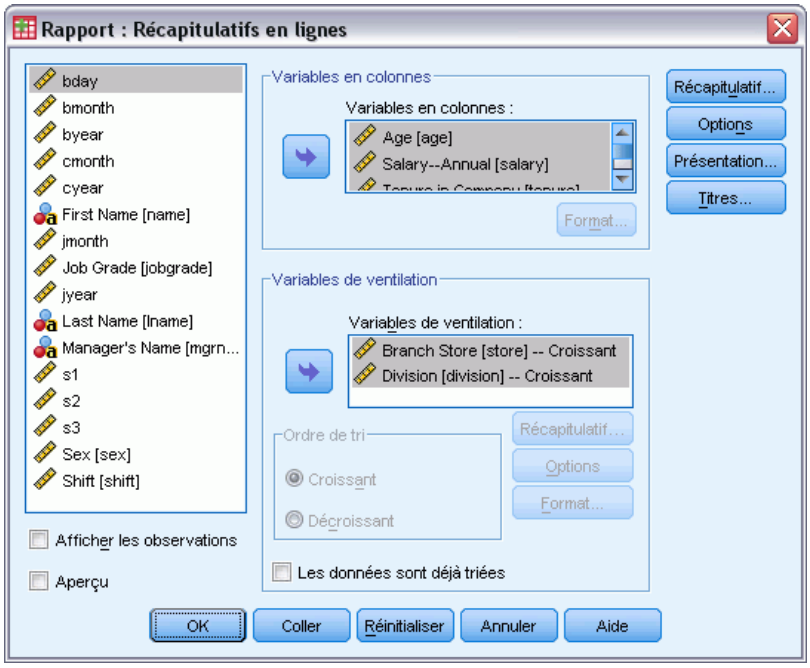
Aperçu. N'affiche que la première page du tableau de bord. Cette option est utile pour avoir un aperçu du format de votre tableau sans traiter le tableau entier.

Les données sont déjà triées : Pour les rapports avec critères d'agrégation, le fichier de données doit être trié par valeur des critères d'agrégation avant de générer le tableau de bord. Si votre fichier de données est déjà trié par valeur des critères d'agrégation, vous pouvez gagner du temps de traitement en sélectionnant cette option. Cette option est particulièrement utile après avoir vu un aperçu du tableau.

Pour obtenir un rapport récapitulatif : Récapitulatifs en lignes

- A partir des menus, sélectionnez :
  Analyse > Rapports > Tableaux de bord en lignes
- Sélectionnez une ou plusieurs variables pour les variables en colonnes. Une colonne est générée dans le tableau de bord pour chaque variable sélectionnée.
- Pour les tableaux triés et affichés par sous-groupe, sélectionnez une ou plusieurs variables pour les critères d'agrégation.
- Pour les tableaux avec statistiques récapitulatives de sous-groupe définies par des critères d'agrégation, sélectionnez le critère d'agrégation dans la liste Variables de ventilation et cliquez sur Tableau récapitulatif dans le groupe Variables de ventilation pour spécifier les mesures récapitulatives.
- Pour les tableaux avec statistiques récapitulatives globales, cliquez sur Tableau récapitulatif pour spécifier les mesures récapitulatives.

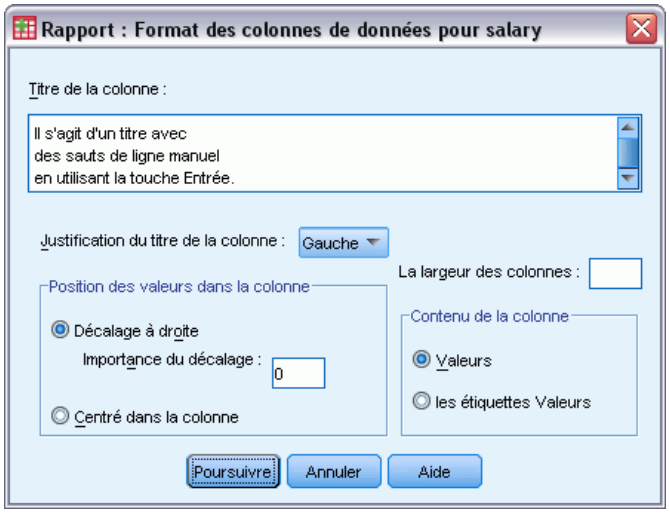
Figure 29-1
Boîte de dialogue Tableaux de bord en lignes



Format des Colonnes de données/Ventilations des Tableaux de bord

Les boîtes de dialogue Format contrôlent les titres et largeurs des colonnes, l'alignement du texte et l'affichage des valeurs de données ou des étiquettes de valeurs. Format colonne de données contrôle le format des Variables en colonnes du côté droit de la page du tableau de bord. Format colonne de ventilation contrôle le format des Colonnes de ventilation du côté gauche.

Figure 29-2
Boîte de dialogue Tableau de bord : Format colonne de données



Titre de la colonne : Pour la variable sélectionnée, contrôle le titre de la colonne. Les titres longs sont automatiquement ajustés dans la colonne. Utilisez la touche Entrée pour insérer manuellement des sauts de lignes aux endroits où vous voulez ajuster les titres.

Position des valeurs dans la colonne : Pour la variable sélectionnée, contrôle l'alignement des valeurs de données ou des étiquettes de données dans la colonne. L'alignement des valeurs ou des étiquettes n'affecte pas l'alignement des titres de colonnes. Vous pouvez soit indenter le contenu de la colonne d'un nombre de caractères donné, soit centrer le contenu de la colonne.

Contenu de la colonne : Pour la variable sélectionnée, contrôle l'affichage soit des valeurs de données, soit des étiquettes de valeurs définies. Les valeurs de données sont affichées pour toutes les valeurs qui ne possèdent pas d'étiquette de valeur définie. (Non disponible pour les Variables en colonnes dans les Tableaux de bord en colonnes)

Fonctions récapitulatives des Tableaux pour/Fonctions récapitulatives Finales

Les deux boîtes de dialogue Fonctions récapitulatives contrôlent l'affichage des statistiques récapitulatives pour les agrégats et pour l'ensemble du tableau de bord. Fonctions récapitulatives contrôle les statistiques de sous-groupe pour chaque catégorie définie par la ou les variables de ventilation. Fonctions récapitulatives Finales contrôle les statistiques globales affichées à la fin du tableau de bord.

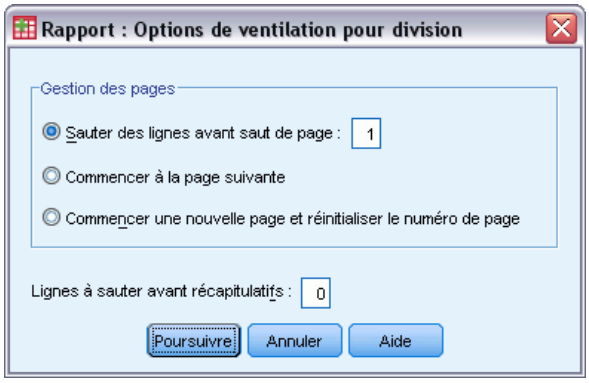
Figure 29-3
Boîte de dialogue Tableau de bord : Fonction récapitulative

Les statistiques récapitulatives disponibles sont la somme des valeurs, la moyenne des valeurs, la valeur minimale, la valeur maximale, le nombre d'observations, le pourcentage d'observations situées au-dessus ou en dessous d'une valeur spécifiée, le pourcentage d'observations comprises à l'intérieur d'un intervalle donné de valeurs, l'écart-type, l'aplatissement, la variance et l'asymétrie.

Options de Ventilation de Tableau de Bord

Options de Ventilation contrôle l'espacement et la pagination des informations de modalité de ventilation.

Figure 29-4
Boîte de dialogue Options de Ventilation des Tableaux de bord



Gestion des pages : Contrôle l'espacement et la pagination des modalités du critère d'agrégation sélectionné. Vous pouvez spécifier un nombre de lignes vides entre les modalités de ventilation ou commencer chaque modalité de ventilation sur une nouvelle page.

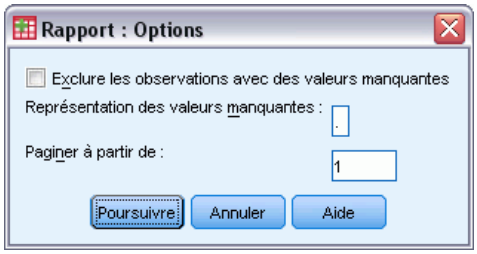
Lignes à sauter avant fonctions élémentaires : Contrôle le nombre de lignes vides entre les étiquettes ou les données des modalités de ventilation et les statistiques récapitulatives. Cette option est particulièrement utile pour les tableaux de bords combinés incluant des listes d'observations

individuelles et des statistiques récapitulatives pour les modalités de ventilation ; dans ces tableaux, vous pouvez insérer des espaces entre les listes d'observations et les statistiques récapitulatives.

Options du Tableau de bord

Options du Tableau de bord contrôle le traitement et l'affichage des valeurs manquantes et la numérotation des pages du tableau de bord.

Figure 29-5
Boîte de dialogue Tableau de bord : Options colonne de ventilation ...



Exclure les observations avec des valeurs manquantes : Elimine (du tableau de bord) toute observation avec des valeurs manquantes pour l'une des variables du tableau de bord.

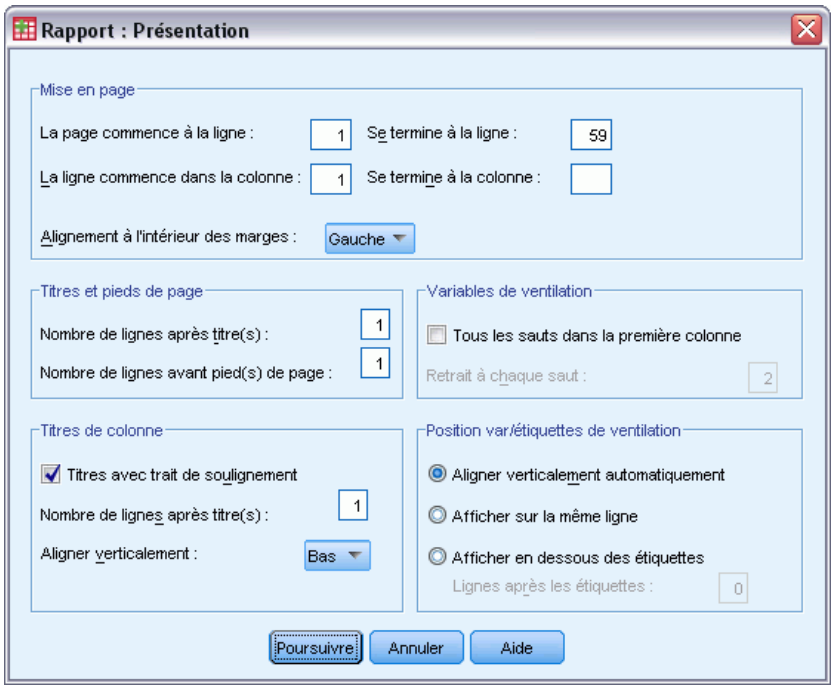
Représentation des valeurs manquantes : Vous permet de spécifier le symbole représentant les valeurs manquantes dans le fichier de données. Ce symbole ne peut comporter qu'un seul caractère et sert à représenter les **valeurs manquantes par défaut** et les **valeurs manquantes utilisateur**.

Paginer à partir de : Vous permet de spécifier un numéro pour la première page du tableau de bord.

Présentation du Tableau de bord

Présentation du Tableau de bord contrôle la largeur et la longueur de chaque page du tableau de bord, l'emplacement du tableau sur la page et l'insertion de lignes vides et d'étiquettes.

Figure 29-6
Boîte de dialogue Tableau : Présentation



Mise en page : Contrôle les marges de page exprimées en lignes (haut et bas) et en caractères (gauche et droite), et reporte l'alignement à l'intérieur des marges.

Titres et bas de page : Contrôle le nombre de lignes séparant les titres et les pieds de page du corps du tableau de bord.

Variables de ventilation : Contrôle l'affichage des colonnes de ventilation. Si des critères d'agrégation multiples sont spécifiés, ils peuvent être affichés en colonnes séparées ou dans la première colonne. Placer tous les critères d'agrégation dans la première colonne produit un tableau de bord plus étroit.

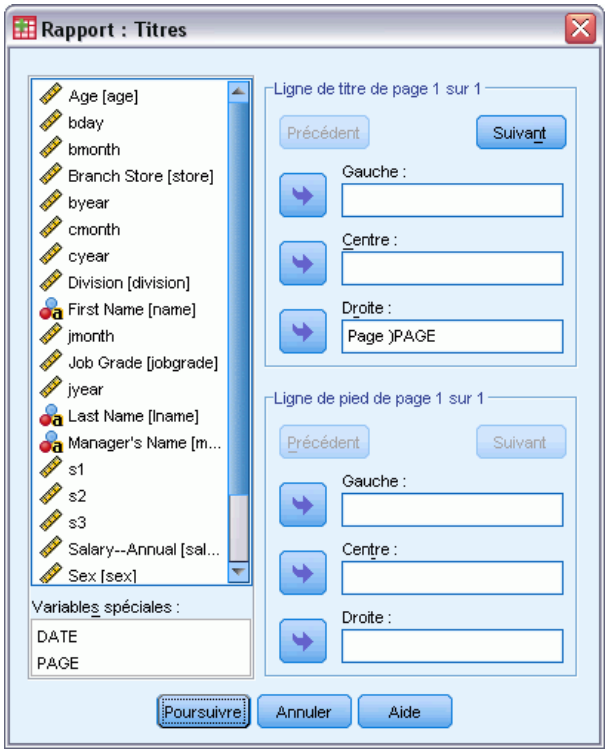
Titres des colonnes : Contrôle l'affichage des titres de colonnes, y compris le soulignement des titres, l'espacement entre les titres et le corps du tableau, et l'alignement vertical des titres de colonnes.

Lignes de variables en colonnes et étiquettes de ventilation. Contrôle l'emplacement des informations de Variables en colonnes (valeurs de données et/ou statistiques récapitulatives) par rapport aux étiquettes de ventilation au début de chaque modalité de ventilation. La première ligne des informations de Variables en colonnes peut commencer soit sur la même ligne que l'étiquette de modalité de ventilation, soit un nombre donné de lignes après cette étiquette. (Non disponible pour les Tableaux de bord en colonnes)

Titres du Tableau de bord

Titres du Tableau de bord contrôle le contenu et l'emplacement des titres et pieds de page du tableau de bord. Vous pouvez spécifier jusqu'à dix lignes de titre et jusqu'à dix lignes de pieds de page, avec des composants justifiés à gauche, centrés et justifiés à droite sur chaque ligne.

Figure 29-7
Boîte de dialogue Tableau : Titres



Si vous insérez des variables dans les titres ou les pieds de page, l'étiquette de valeur actuelle ou la valeur de la variable est affichée dans le titre ou le pied de page. Dans les titres, l'étiquette de valeur correspondant à la valeur de la variable au début de la page est affichée. Dans les pieds de page, l'étiquette de valeur correspondant à la valeur de la variable à la fin de la page est affichée. S'il n'y a aucune étiquette de valeur, la valeur réelle est affichée.

Variables spéciales : Les variables spéciales *DATE* et *PAGE* vous permettent d'insérer la date actuelle ou le numéro de page dans l'une des lignes d'un en-tête ou d'un pied de page. Si votre fichier de données contient des variables nommées *DATE* ou *PAGE*, vous ne pouvez pas utiliser ces variables dans les titres ou les pieds de page des tableaux.

Tableaux de bord en colonnes

Tableaux de bord en colonnes produit des tableaux de bord dans lesquels différentes statistiques récapitulatives apparaissent en colonnes séparées.

Exemple : Une société possédant une chaîne de magasins conserve des dossiers sur les employés comprenant le salaire, l'ancienneté et le service où chaque employé travaille. Vous pourriez générer un tableau de bord fournissant des statistiques récapitulatives sur les salaires (par exemple moyenne, minimum, maximum) pour chaque division.

Variables en colonnes : Fournit la liste des variables du tableau de bord pour lesquelles vous voulez des statistiques récapitulatives et contrôle le format d'affichage et les statistiques récapitulatives affichées pour chaque variable.

Variables de ventilation : Fournit la liste des critères d'agrégation optionnels qui divisent le tableau de bord en groupes et contrôle les formats d'affichage des colonnes de ventilation. Pour les critères d'agrégation multiples, il y aura un groupe séparé pour chaque modalité de chaque critère d'agrégation à l'intérieur des modalités du critère d'agrégation précédent dans la liste. Les critères d'agrégation doivent être des variables qualitatives discrètes qui divisent les observations en un nombre limité de modalités significatives.

Tableau de bord : Contrôle les caractéristiques globales du tableau de bord, y compris l'affichage des valeurs manquantes, la numérotation des pages et les titres.

Aperçu. N'affiche que la première page du tableau de bord. Cette option est utile pour avoir un aperçu du format de votre tableau sans traiter le tableau entier.

Les données sont déjà triées : Pour les rapports avec critères d'agrégation, le fichier de données doit être trié par valeur des critères d'agrégation avant de générer le tableau de bord. Si votre fichier de données est déjà trié par valeur des critères d'agrégation, vous pouvez gagner du temps de traitement en sélectionnant cette option. Cette option est particulièrement utile après avoir vu un aperçu du tableau.

Pour obtenir un rapport récapitulatif : Récapitulatifs en colonnes

- A partir des menus, sélectionnez :
  Analyse > Rapports > Tableaux de bord en colonnes

- Sélectionnez une ou plusieurs variables pour les variables en colonnes. Une colonne est générée dans le tableau de bord pour chaque variable sélectionnée.

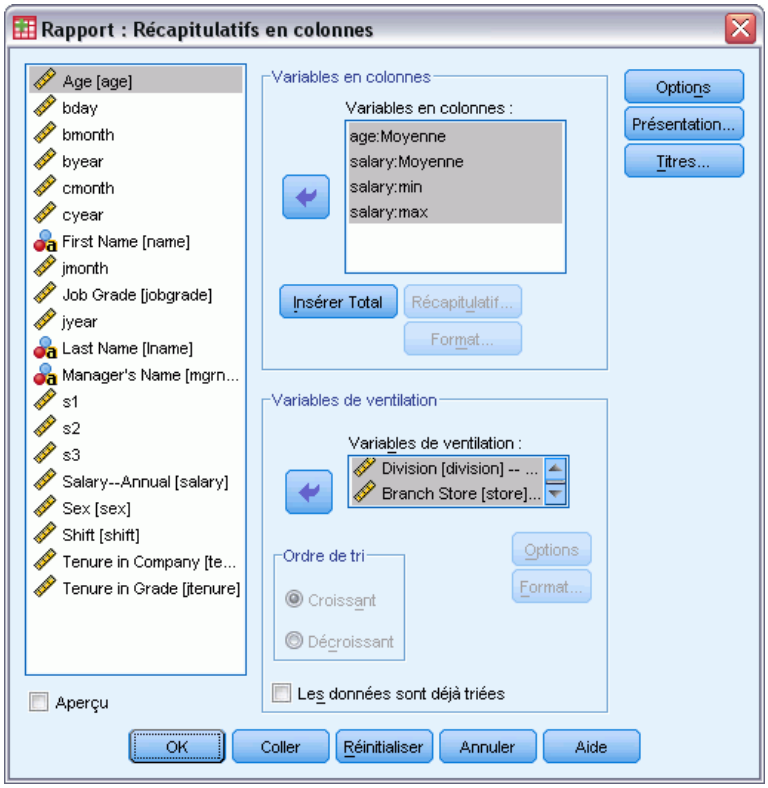
- Pour modifier la mesure récapitulative d'une variable, sélectionnez la variable dans la liste Variables en colonnes et cliquez sur Tableau récapitulatif.

- Pour obtenir plus d'une mesure récapitulative pour une variable, sélectionnez la variable dans la liste source et déplacez-la dans la liste Variables en colonnes plusieurs fois, une fois pour chaque mesure récapitulative que vous souhaitez.

- Pour afficher une colonne contenant la somme, la moyenne, le rapport ou une autre fonction de colonnes existantes, cliquez sur Insérer le total. Une variable appelée *total* est alors placée dans la liste Variables en colonnes.

- Pour les tableaux triés et affichés par sous-groupe, sélectionnez une ou plusieurs variables pour les critères d'agrégation.

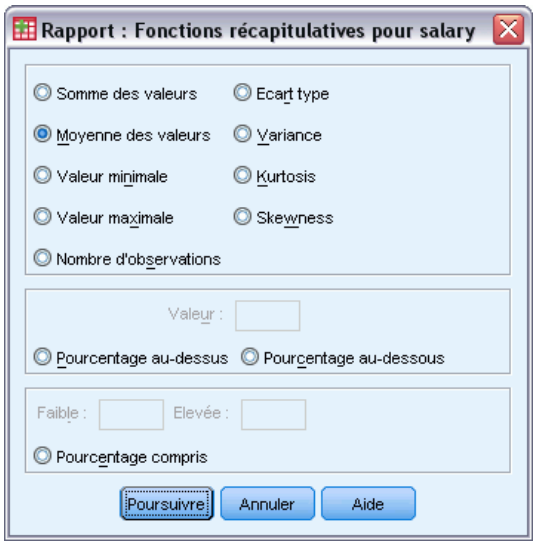
Figure 29-8
Boîte de dialogue Tableaux de bord en colonnes



Fonction récapitulative des Colonnes de données

Fonctions récapitulatives contrôle les statistiques récapitulatives affichées pour la variable de colonne de données sélectionnée.

Figure 29-9
Boîte de dialogue Tableau de bord : Fonction récapitulative



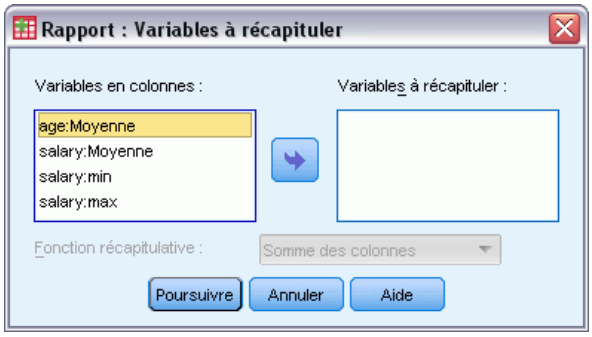
Les statistiques récapitulatives disponibles sont la somme des valeurs, la moyenne des valeurs, la valeur minimale, la valeur maximale, le nombre d'observations, le pourcentage d'observations situées au-dessus ou en dessous d'une valeur spécifiée, le pourcentage d'observations comprises à l'intérieur d'un intervalle donné de valeurs, l'écart-type, l'aplatissement, la variance et l'asymétrie.

Fonction élémentaire des Colonnes de Données pour colonne de total

Variables à récapituler contrôle les statistiques récapitulatives totales qui récapitulent deux ou plusieurs Variables en colonnes.

Les statistiques récapitulatives totales sont la somme des colonnes, la moyenne des colonnes, le minimum, le maximum, la différence entre les valeurs de deux colonnes, le quotient des valeurs d'une colonne divisées par les valeurs d'une autre colonne et le produit des valeurs de colonnes multipliées.

Figure 29-10
Boîte de dialogue Tableau de bord : Colonnes



Somme des colonnes : La colonne *total* représente la somme des colonnes de la liste Variables à récapituler.

Moyenne des colonnes : La colonne *total* représente la moyenne des colonnes de la liste Variables à récapituler.

Minimum des colonnes : La colonne *total* représente la somme minimale des colonnes de la liste Variables à récapituler.

Maximum des colonnes : La colonne *total* représente la somme maximale des colonnes de la liste Variables à récapituler.

1ère colonne – 2ème colonne : La colonne *total* représente la différence des colonnes de la liste Variables à récapituler. La liste Variables à récapituler doit contenir exactement deux colonnes.

1ère colonne / 2ème colonne : La colonne *total* représente le quotient des colonnes de la liste Variables à récapituler. La liste Variables à récapituler doit contenir exactement deux colonnes.

% 1ère colonne / 2ème colonne : La colonne *total* représente le pourcentage de la première colonne par rapport à la seconde colonne de la liste Variables à récapituler. La liste Variables à récapituler doit contenir exactement deux colonnes.

Produit des colonnes : La colonne *total* représente le produit des colonnes de la liste Variables à récapituler.

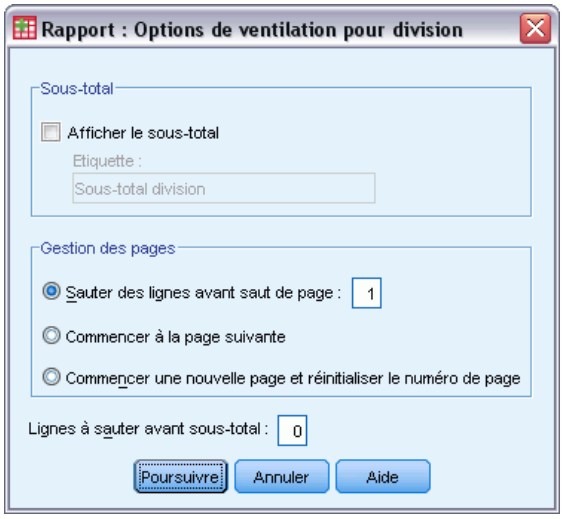
Format des Colonnes du Tableau de bord

Les options de format des Variables en colonnes et de ventilation pour les Tableaux de bord en colonnes sont identiques à celles décrites pour les Tableaux de bord en lignes.

Tableaux de bord en Colonnes : Options de Ventilation

Options de Ventilation contrôle l'affichage des sous-totaux, l'espacement et la pagination des modalités de ventilation.

Figure 29-11
Boîte de dialogue Options de Ventilation des Tableaux de bord



Sous-total : Contrôle l'affichage des sous-totaux pour les modalités de ventilation.

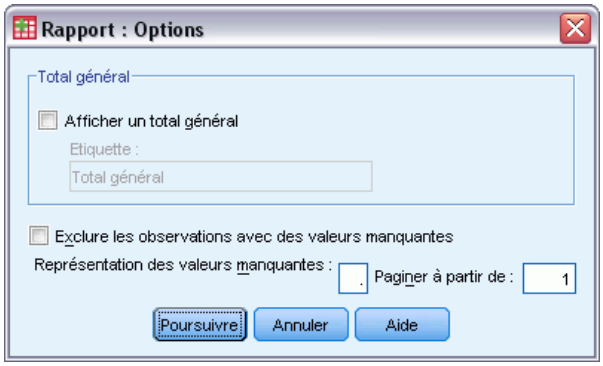
Gestion des pages : Contrôle l'espacement et la pagination des modalités du critère d'agrégation sélectionné. Vous pouvez spécifier un nombre de lignes vides entre les modalités de ventilation ou commencer chaque modalité de ventilation sur une nouvelle page.

Lignes à sauter avant sous-total : Contrôle le nombre de lignes vides entre les données des modalités de ventilation et les sous-totaux.

Options des Tableaux de bord en Colonnes

Options contrôle l'affichage des totaux généraux, l'affichage des valeurs manquantes et la pagination dans les Tableaux de bord en colonnes.

Figure 29-12
Boîte de dialogue Tableau de bord : Options colonne de ventilation ...



Total général : Affiche et étiquette un total général pour chaque colonne ; affiché au bas de la colonne.

Valeurs manquantes : Vous pouvez exclure les valeurs manquantes du tableau ou sélectionner un caractère unique indiquant les valeurs manquantes dans le tableau de bord.

Présentation du Tableau de bord en Colonnes

Les options de présentation pour les Tableaux de bord en colonnes sont identiques à celles présentées pour les Tableaux de bord en lignes.

Fonctionnalités supplémentaires de la commande REPORT

Le langage de syntaxe de commande vous permet aussi de :

- Afficher différentes fonctions récapitulatives dans les colonnes d'une ligne de fonction unique.
- Insérer des fonctions récapitulatives dans les Variables en colonnes pour des variables autres que la variable de la colonne de données, ou pour diverses combinaisons (fonctions composites) de fonctions récapitulatives.
- Utiliser la Médiane, le Mode, la Fréquence et le Pourcentage comme des fonctions récapitulatives.

- Contrôler plus précisément le format d'affichage des statistiques récapitulatives.
- Insérer des lignes vides à divers emplacements du tableau de bord.
- Insérer des lignes vides toutes les *n* observations dans les listes.

Du fait de la complexité de la syntaxe de la commande `REPORT`, vous trouverez peut-être utile, lorsque vous construirez un nouveau tableau de bord avec syntaxe, d'approcher le tableau généré à partir des boîtes de dialogue, de copier et coller la syntaxe correspondante, puis de préciser cette syntaxe afin d'obtenir le tableau de bord exact que vous souhaitez.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
30

Analyse de fiabilité

L'analyse de fiabilité vous permet d'étudier les propriétés des échelles de mesure et des éléments qui les constituent. La procédure d'analyse de fiabilité calcule plusieurs mesures fréquemment utilisées de la fiabilité de l'échelle et propose également des informations sur les relations entre les différents éléments de l'échelle. Les coefficients de corrélation intra-classe peuvent être utilisés pour calculer les estimations de fiabilité inter-coefficients.

Exemple : Mon questionnaire mesure-t-il de façon fidèle la satisfaction de la clientèle ? L'analyse de la fiabilité vous permet de déterminer dans quelle mesure les éléments de votre questionnaire sont liés les uns aux autres et vous procure un indice général de la consistance ou de la cohérence interne de l'échelle dans son ensemble. Elle vous permet enfin d'identifier les éléments qui posent problème et qu'il faudrait exclure de l'échelle.

Statistiques : Descriptions de chaque variable et pour l'échelle, statistiques récapitulatives sur les éléments, corrélations et covariances entre éléments, prévisions de fiabilité, tableau d'ANOVA, coefficients de corrélation intra-classe, T^2 d'Hotelling et test d'additivité de Tukey.

Modèles : Les modèles suivants de fiabilité sont disponibles :

- **Alpha (Cronbach) :** Il s'agit d'un modèle de cohérence interne, fondé sur la corrélation moyenne entre éléments.
- **Split-half :** Ce modèle fractionne l'échelle en deux et examine la corrélation entre les deux parties.
- **Guttman :** Ce modèle calcule les limites minimales de Guttman pour une fiabilité vraie.
- **Parallèle :** Ce modèle part de l'hypothèse que tous les éléments ont des variances égales et des variances d'erreur égales en cas de réplication.
- **Parallèle strict :** Ce modèle se fonde sur les mêmes hypothèses que le modèle parallèle mais envisage également que tous les éléments ont la même moyenne.

Données. Les données peuvent être dichotomiques, ordinales ou constituer des intervalles, mais elles doivent être codées en numérique.

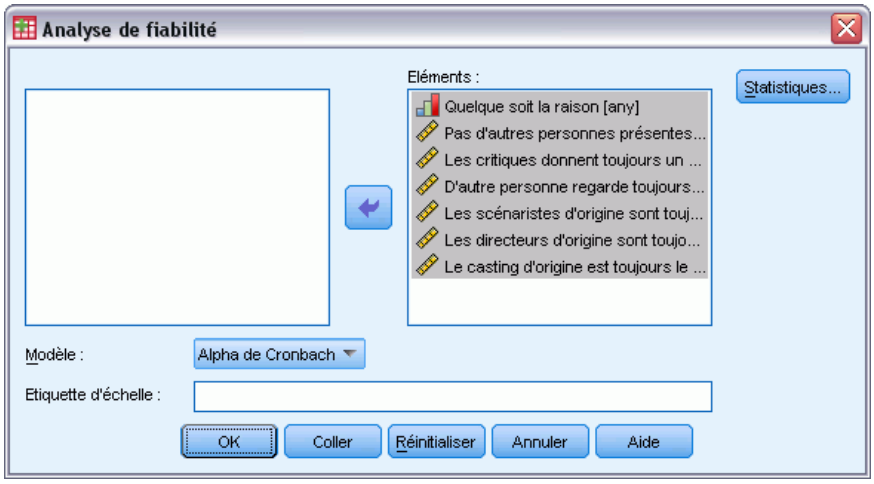
Hypothèses : Les observations doivent être indépendantes, les erreurs ne doivent pas être corrélées entre éléments. Chaque paire d'éléments doit avoir une distribution gaussienne bivariée. Les échelles doivent être additives, de sorte que chaque élément est linéairement relié au total.

Procédures apparentées : Si vous souhaitez explorer la dimensionnalité des éléments de votre échelle (pour voir si plusieurs éléments de base sont nécessaires au modèle des calculs), utilisez Analyse factorielle ou Positionnement multidimensionnel. Pour identifier des groupes homogènes de variables, utilisez la classification hiérarchique pour classer les variables.

Obtenir une analyse de fiabilité

▶ A partir des menus, sélectionnez :
Analyse > Echelle > Analyse de la fiabilité...

Figure 30-1
Boîte de dialogue Analyse d'items

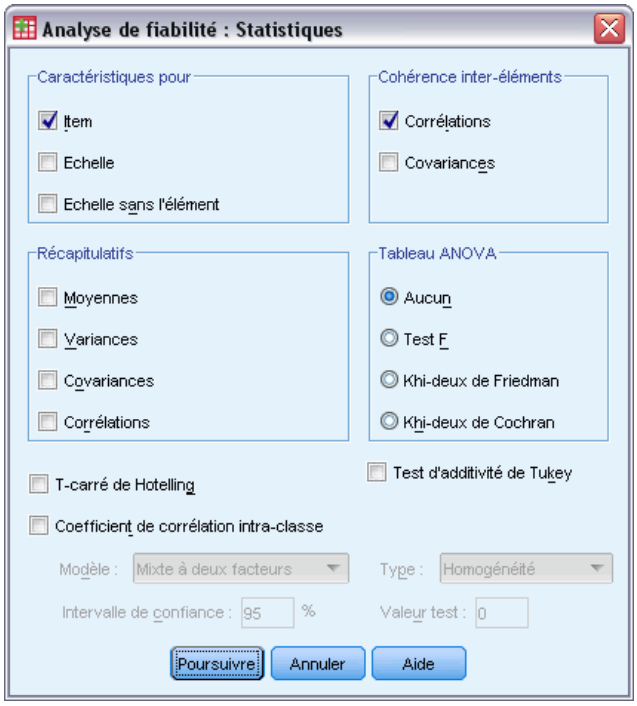


- Sélectionnez deux variables (éléments) au moins en tant que composants potentiels d'une échelle additive.

- Sélectionnez un modèle dans la liste déroulante Modèle.

Statistiques de l'analyse de fiabilité

Figure 30-2
Boîte de dialogue Analyse d'items : Statistiques



Vous pouvez sélectionner différentes statistiques décrivant votre échelle et vos éléments. Les statistiques émises par défaut regroupent le nombre d'observations, le nombre d'éléments et les prévisions de fiabilité de la façon suivante :

- **Modèles alpha.** Coefficient alpha ; pour les données dichotomiques, il s'agit d'un équivalent du coefficient Kuder-Richardson 20 (KR20).
- **Modèles Split-half.** Corrélation entre les sous-échelles, fiabilité Split-half de Guttman, fiabilité de Spearman-Brown (longueur égale ou inégale) et coefficient alpha pour chaque moitié.
- **Modèles de Guttman.** Coefficients de fiabilité lambda 1 à lambda 6.
- **Modèles parallèle et parallèle strict.** Test de qualité de l'ajustement du modèle, estimation de la variance de l'erreur, variance commune et réelle, estimation de la corrélation commune entre éléments, estimation de la fiabilité et estimation de la fiabilité non biaisée.

Caractéristiques pour : Produit des statistiques descriptives pour les échelles ou les éléments sur les observations.

- **Item.** Produit des statistiques descriptives pour les items sur les observations.
- **Echelle.** Produit des statistiques descriptives pour les échelles.
- **Echelle sans l'item :** Affiche les statistiques récapitulatives en comparant chaque item à l'échelle composée des autres items. Les statistiques incluent la moyenne et la variance de l'échelle si l'item a été supprimé de l'échelle, la corrélation entre l'item et l'échelle composée des autres items, et l'alpha de Cronbach si l'item a été supprimé de l'échelle.

Principales statistiques : Fournit des statistiques descriptives de la distribution des éléments sur l'ensemble des éléments dans l'échelle.

- **Moyennes.** Statistiques récapitulatives pour les moyennes d'élément. Les moyennes d'élément minimale, maximale et intermédiaire sont affichées, ainsi que le rapport de la moyenne maximale à la moyenne minimale.
- **Variances.** Statistique récapitulative des variances d'élément. Les valeurs de variance maximale, minimale et moyenne sont affichées, ainsi que la plage et la variance des variances d'élément, et le rapport entre les variances d'élément maximale et minimale.
- **Covariances.** Statistiques récapitulatives pour les covariances inter élément. Les covariances entre éléments minimale, maximale et intermédiaire sont affichées, ainsi que l'intervalle et la variance des covariances entre éléments, et le rapport de la covariance entre éléments maximale à la covariance minimale.
- **Corrélations.** Statistiques récapitulatives pour les corrélations inter élément. Les corrélations entre éléments minimale, maximale et intermédiaire sont affichées, ainsi que l'intervalle et la variance des corrélations entre éléments, et le rapport de la corrélation entre éléments maximale à la corrélation minimale.

Cohérence inter-items : Produit des matrices de corrélations et de covariances entre éléments.

Tableau ANOVA : Produit des tests de moyennes égales.

- **Test F.** Affiche un tableau d'analyse de la variance des mesures répétées.

- **Khi-deux de Friedman.** Affiche le test de Friedman (khi-deux) et le coefficient de concordance de Kendall. Cette option convient aux données organisées sous forme de rangs. Le test du khi-deux remplace le test F habituel dans le tableau ANOVA.
- **Khi-deux de Cochran.** Affiche la valeur Q de Cochran. Cette option est appropriée pour les données dichotomiques. Le Q de Cochran remplace le test F habituel dans le tableau ANOVA.

T-carré de Hotelling : Produit un test multivarié basé sur l'hypothèse nulle que tous les éléments sur l'échelle ont la même moyenne.

Test d'additivité de Tukey : Produit un test basé sur l'hypothèse qu'il n'y a pas d'interaction multiplicative entre les éléments.

Coefficient de corrélation intra-classe : Produit des mesures d'homogénéité ou de cohérence des valeurs par observation.

- **Modèle :** Sélectionnez le modèle de calcul du coefficient de corrélation intra-classe. Les modèles disponibles sont Mixte à deux facteurs, Aléatoire à deux facteurs et Aléatoire à un facteur. Sélectionnez Mixte à deux facteurs lorsque les effets de population sont aléatoires et les effets d'item sont fixes, sélectionnez Aléatoire à deux facteurs lorsque les effets de population et les effets d'items sont aléatoires, ou sélectionnez Aléatoire à un facteur lorsque les gens effectuent un facteur.
- **Type.** Sélectionnez le type d'index. Les types disponibles sont Homogénéité et Cohérence absolue.
- **Intervalle de confiance :** Spécifiez le niveau de l'intervalle de confiance. La valeur par défaut est 95 %.
- **Valeur test :** Spécifiez la valeur hypothétique du coefficient pour le test d'hypothèse. Il s'agit de la valeur par rapport à laquelle la valeur observée est comparée. La valeur par défaut est 0.

Fonctionnalités supplémentaires de la commande RELIABILITY

Le langage de syntaxe de commande vous permet aussi de :

- Lire et analyser une matrice de corrélation.
- Enregistrer une matrice de corrélation à analyser ultérieurement.
- Spécifier un fractionnement autre qu'en deux moitiés égales quant au nombre d'éléments pour la méthode Split-half.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
31
Positionnement multidimensionnel

Le positionnement multidimensionnel tente de déterminer une structure dans un ensemble de mesures de distance entre objets ou observations. Pour cela, il affecte les observations à des positions particulières dans un espace conceptuel (à deux ou trois dimensions généralement) de telle sorte que les distances entre les points dans l'espace correspondent le mieux possible aux dissimilarités données. Dans la plupart des cas, les dimensions de cet espace conceptuel peuvent être interprétées et utilisées pour mieux comprendre les données.

Si vous avez mesuré objectivement les variables, vous pouvez utiliser le positionnement multidimensionnel comme technique de factorisation (le Positionnement multidimensionnel calcule pour vous les distances à partir des données multivariées, le cas échéant). Le positionnement multidimensionnel peut également s'appliquer à des estimations subjectives de dissimilarité entre objets ou concepts. D'autre part, le positionnement multidimensionnel peut gérer les informations de dissimilarité provenant de plusieurs sources, comme c'est le cas lorsqu'il y a plusieurs indicateurs ou plusieurs répondants au questionnaire.

Exemple : Comment les gens perçoivent-ils les relations entre différentes voitures ? Si les données que vous obtenez de vos répondants indiquent des évaluations de similarité entre différentes modèles, le positionnement multidimensionnel peut servir à identifier les dimensions qui décrivent les perceptions des consommateurs. Vous pouvez trouver, par exemple, que le prix et la taille du véhicule définissent un espace à deux dimensions qui tient compte des similarités reportées par les répondants.

Statistiques : Pour chaque modèle : matrice des données, positionnement optimisé des données de la matrice, stress S (de Young), stress S (de Kruskal), RSQ, coordonnées des stimuli, stress moyen et RSQ pour chaque stimulus (modèles RMDS). Pour les modèles des différences individuelles : pondérations des sujets et indice de singularité pour chaque sujet. Pour chaque matrice dans les modèles de positionnement multidimensionnel répliqués : stress et RSQ pour chaque stimulus. Diagrammes : coordonnées des stimulus (à deux ou trois dimensions), diagramme de dispersion des disparités par rapport aux distances.

Données. Si vos données sont dissemblables, toutes les dissemblances doivent être quantitatives et mesurées avec les mêmes unités et échelles. Si vos données sont multivariées, les variables peuvent être quantitatives, binaires mais peuvent aussi être des données d'effectif. Le positionnement des variables est un enjeu de taille : les différences de positionnement peuvent affecter votre solution. Si vos variables présentent de grandes différences de positionnement (par exemple, si une variable est mesurée en dollar et l'autre en année), vous devez envisager de les standardiser (et cela automatiquement par la procédure de positionnement multidimensionnel).

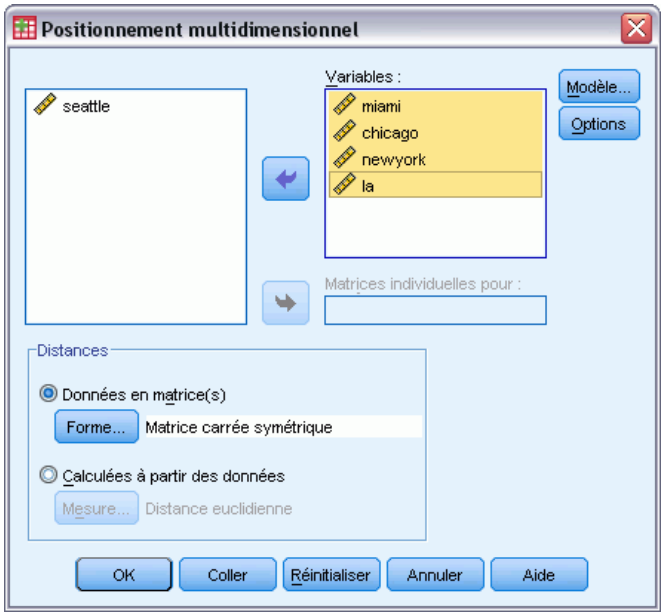
Hypothèses : La procédure de positionnement multidimensionnel est relativement indépendante de toute hypothèse de distribution. Assurez-vous que vous avez sélectionné le niveau de mesure approprié (ordinal, intervalle ou rapport) dans la boîte de dialogue Positionnement multidimensionnel : Options afin de garantir la justesse des résultats.

Procédures apparentées : Si votre but est la factorisation, vous pouvez également envisager l'analyse factorielle, plus particulièrement si vos données sont quantitatives. Si vous souhaitez identifier des groupes d'observations similaires, envisagez de compléter votre analyse par positionnement multidimensionnel avec une analyse des *nuées dynamiques* ou une analyse de la classification hiérarchique.

Obtenir une analyse par positionnement multidimensionnel

- A partir des menus, sélectionnez :
  Analyse > Echelle > Positionnement multidimensionnel

Figure 31-1
Boîte de dialogue Positionnement multidimensionnel



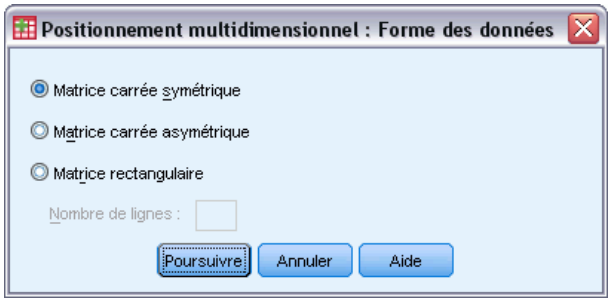
- Sélectionnez au moins quatre variables numériques pour l'analyse.
- Dans le groupe Distances, sélectionnez Données en matrice(s) ou Calculées à partir des données.
- Si vous avez sélectionné Calculées à partir des données, vous pouvez également sélectionner une variable de regroupement pour les matrices individuelles. La variable de regroupement peut être numérique ou être une variable chaîne.

Eventuellement, vous pouvez aussi :

- Indiquez la forme de la matrice lorsque les données sont des distances.
- Spécifiez la mesure de la distance à utiliser lors de la création de distances à partir des données.

Forme des données du positionnement multidimensionnel

Figure 31-2
Boîte de dialogue Positionnement multidimensionnel : Forme

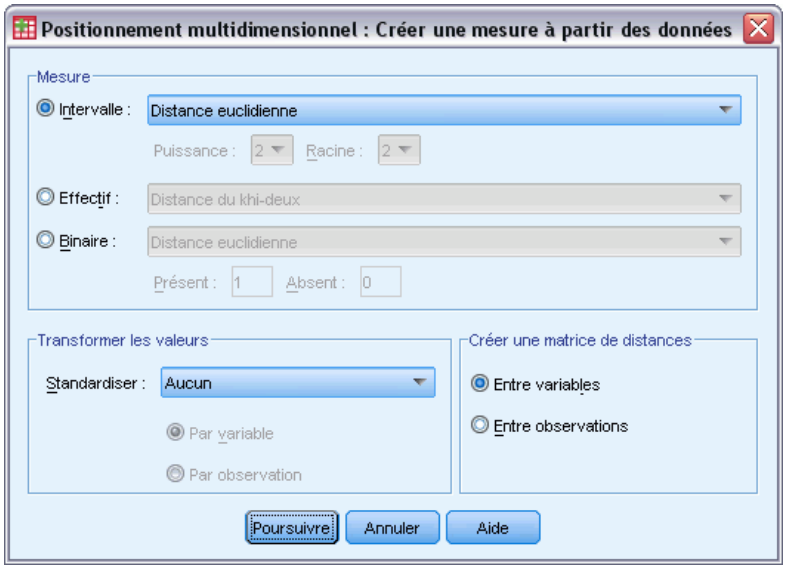


Si votre ensemble de données actif représente les distances au sein d'un ensemble d'objets ou entre deux ensembles d'objets, vous devez indiquer la forme de votre matrice de données afin d'obtenir des résultats corrects.

Remarque : Vous pouvez sélectionner Carré symétrique si la boîte de dialogue Modèle indique une conditionnalité de ligne.

Positionnement multidimensionnel : créer une mesure

Figure 31-3
Boîte de dialogue Positionnement multidimensionnel : Calcul de l'indice de dissimilarité



Le positionnement multidimensionnel utilise les données de dissimilarité pour créer une solution de codage. Si vos données sont multivariées (valeurs des variables mesurées), vous devez créer des données de dissimilarité afin de calculer une solution de positionnement multidimensionnel. Vous pouvez spécifier les détails de création de mesures de dissimilarité à partir de vos données.

Mesure : Vous permet de spécifier la mesure de dissimilarité adaptée à votre analyse. Sélectionnez une possibilité dans le groupe Mesure correspondant à votre type de données, puis sélectionnez l'une des mesures dans la liste déroulante correspondant à ce type de mesure. Les possibilités sont :

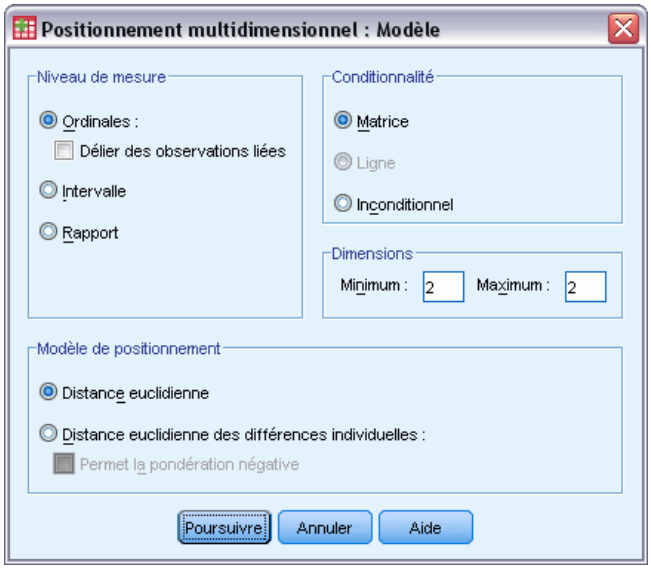
- **Intervalle :** Distance Euclidienne, Carré de la distance Euclidienne, Distance de Tchebycheff, Distance de Manhattan, Distance de Minkowski ou Autre.
- **Effectif** : Distance du Khi-deux ou Distance du phi-deux.
- **Binaire :** Distance Euclidienne, Carré de la distance Euclidienne, Ecart de taille, Différence de motif, Variance ou Lance et Williams.

Créer une matrice de distances : Vous permet de choisir l'unité d'analyse. Les possibilités sont Par variables ou Par observations.

Transformer les valeurs : Dans certains cas, comme lorsque les variables sont mesurées selon des échelles très différentes, vous voudrez peut-être standardiser des valeurs avant de calculer les proximités (ne s'applique pas aux données binaires). Sélectionnez une méthode de standardisation dans la liste déroulante. Si aucune standardisation n'est requise, choisissez Aucune.

Modèle de positionnement multidimensionnel

Figure 31-4
Boîte de dialogue Positionnement multidimensionnel : Modèle



Une estimation correcte d'un modèle de positionnement multidimensionnel dépend des aspects des données et du modèle lui-même.

Niveau de mesure. Vous permet de spécifier le niveau de vos données. Les possibilités sont Ordinales, Intervalle ou Rapport. Si vos variables sont ordinales, la sélection de l'option Délier des observations liées demande qu'elles soient traitées en tant que variables continues, de telle sorte que les liens (mêmes valeurs pour des observations différentes) soient résolus de manière optimale.

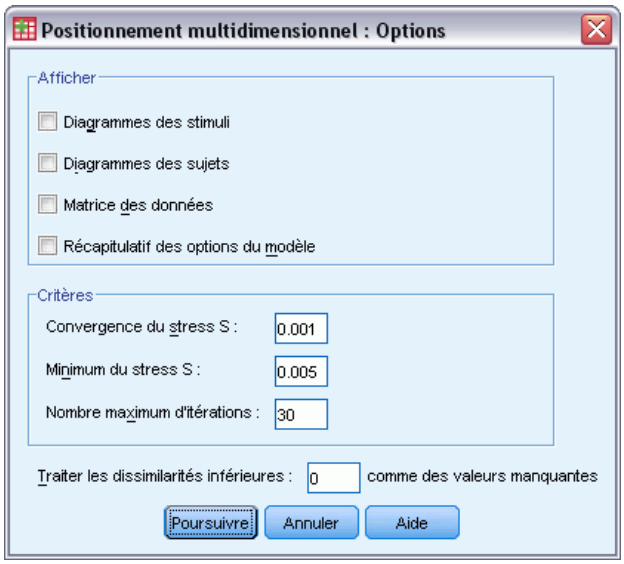
Conditionnalité : Vous permet de spécifier les comparaisons pertinentes. Les possibilités sont Matrice, Ligne et Inconditionnel.

Dimensions : Vous permet de spécifier la dimensionnalité de la ou des solutions de positionnement. Une seule solution est calculée pour chaque nombre de l'intervalle. Indiquez des nombres entiers entre 1 et 6. La valeur minimale 1 n'est autorisée que si vous sélectionnez l'option Distance Euclidienne comme modèle de positionnement. Pour n'obtenir qu'une seule solution, indiquez le même nombre en tant que minimum et maximum.

Modèle de positionnement. Vous permet de spécifier les hypothèses sous lesquelles le positionnement est effectué. Les possibilités existantes sont Distance Euclidienne ou Distance Euclidienne des différences individuelles (connue également en tant que INDSCAL). Pour le modèle Distance Euclidienne des différences individuelles, vous pouvez sélectionner l'option Permet la pondération négative, si cela convient à vos données.

Positionnement multidimensionnel : Options

Figure 31-5
Boîte de dialogue Positionnement multidimensionnel : Options



Vous pouvez spécifier les options de votre analyse par positionnement multidimensionnel.

Afficher : Vous permet d'afficher les différents types d'affichage. Les options possibles sont Diagrammes des stimuli, Diagrammes des sujets, Matrice des données et Récapitulatif des options du modèle.

Critères : Vous permet de déterminer quand l'itération doit s'interrompre. Pour modifier les valeurs par défaut, entrez des valeurs pour la Convergence du stress S, le Minimum du stress S et le Maximum des itérations.

Traiter les dissimilarités inférieures à n comme des valeurs manquantes : Ces distances sont exclues de l'analyse.

Fonctionnalités supplémentaires de la commande ALSCAL

Le langage de syntaxe de commande vous permet aussi de :

- Utiliser trois types de modèle supplémentaires, ASCAL, AINDS et GEMSCAL dans la documentation relative au Positionnement multidimensionnel.
- Effectuer des transformations polynomiales sur l'intervalle et les données de type ratio.
- Analyser les similarités (plutôt que les distances) avec des données ordinales.
- Analyser les données nominales.
- Enregistrer diverses matrices de coordonnées et de pondération dans des fichiers et les relire pour l'analyse.
- Contraindre le dépliage multidimensionnel.

Pour obtenir des renseignements complets sur la syntaxe, reportez-vous au manuel *Command Syntax Reference*.

Chapitre
32

Statistiques de ratio

La procédure Statistiques de ratio permet d'obtenir la liste exhaustive des statistiques récapitulatives qui servent à décrire le rapport entre deux variables d'échelle.

Vous pouvez trier le résultat sur la base des valeurs d'une variable de regroupement, dans l'ordre croissant ou décroissant. Vous pouvez supprimer le rapport des statistiques de ratio dans le document de sortie et enregistrer les résultats dans un fichier externe.

Exemple : Le rapport existant entre le prix estimatif et le prix de vente des maisons est-il uniforme dans chacun de ces cinq comtés ? D'après les résultats, vous pouvez conclure que la distribution des rapports varie considérablement d'un comté à l'autre.

Statistiques. Médiane, moyenne, moyenne pondérée, intervalles de confiance, coefficient de dispersion (COD), coefficient de variation avec médiane centrée, coefficient de variation avec moyenne centrée, différentiel lié au prix (PRD), écart-type, écart absolu moyen (AAD), intervalle, valeurs minimale et maximale, et index de concentration calculés pour un intervalle ou un pourcentage défini par l'utilisateur dans le rapport médian.

Données. Utilisez des codes numériques ou alphanumériques pour coder les variables de regroupement (mesures de niveau nominal ou ordinal).

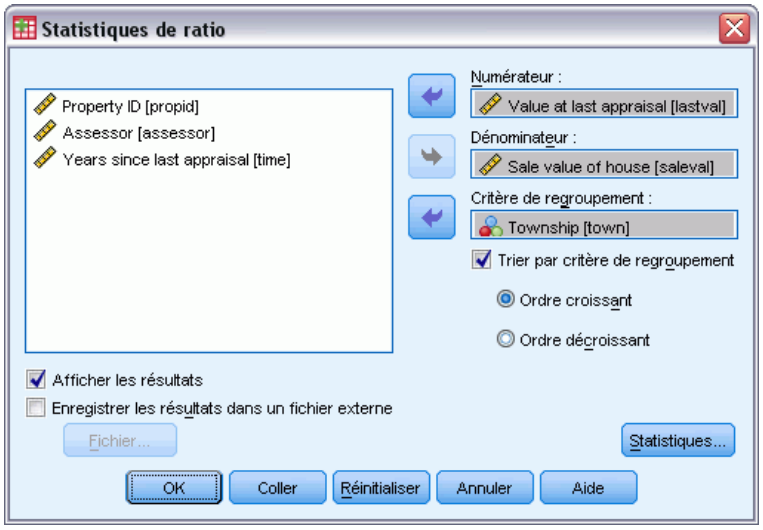
Hypothèses : Vous devez utiliser des variables d'échelle acceptant les valeurs positives pour les variables qui définissent le numérateur et le dénominateur du rapport.

Pour obtenir des statistiques de ratio

▶ A partir des menus, sélectionnez :

Analyse > Statistiques descriptives > Ratio

Figure 32-1
Boîte de dialogue Statistiques de ratio



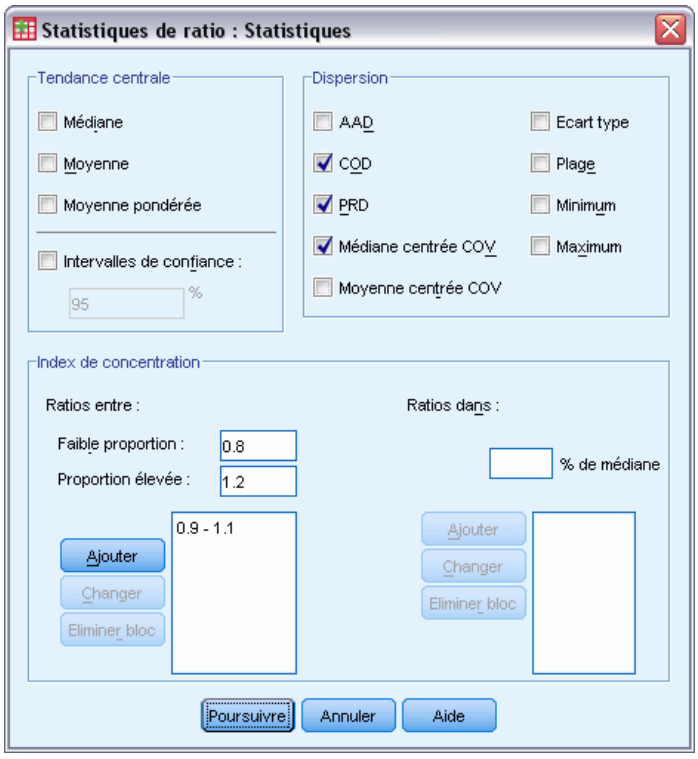
- Sélectionnez la variable du numérateur.
- Sélectionnez la variable du dénominateur.

Éventuellement :

- Sélectionner une variable de regroupement et préciser l'ordre de présentation des groupes dans le résultat.
- Choisissez si vous souhaitez afficher les résultats dans l'Editeur de résultats.
- Choisissez ou non d'enregistrer les résultats dans un fichier externe en vue d'une utilisation ultérieure, et précisez le nom du fichier dans lequel les résultats sont enregistrés.

Statistiques de ratio

Figure 32-2
Boîte de dialogue Statistiques de ratio



Tendance centrale : Les mesures de tendance centrale sont des statistiques qui décrivent la distribution des rapports.

- **Médiane.** La valeur telle que le nombre de ratios inférieurs à cette valeur est identique au nombre de ratios supérieurs à cette valeur.
- **Moyenne :** Résultat de la somme des ratios divisée par le nombre total de ratios.

- **Moyenne pondérée.** Résultat de la division de la moyenne du numérateur par la moyenne du dénominateur. La moyenne pondérée correspond également à la moyenne des ratios pondérée par le dénominateur.
- **Intervalles de confiance.** Affiche les intervalles de confiance de la moyenne, de la médiane et de la moyenne pondérée (si demandée). Affectez une valeur supérieure ou égale à 0 et inférieure à 100 au niveau de confiance.

Dispersion : Ces statistiques permettent de mesurer le degré de variation, ou de répartition, au niveau des valeurs observées.

- **AAD :** L'écart absolu moyen est égal à la somme des écarts absolus des ratios relatifs à la médiane, divisée par le nombre total de ratios.
- **COD :** Le coefficient de dispersion résulte de l'expression de l'écart moyen absolu en pourcentage de la médiane.
- **PRD :** Le différentiel lié au prix, ou index de régressivité, résulte de la division de la moyenne par la moyenne pondérée.
- **Médiane centrée COV.** Le coefficient de variation avec médiane centrée résulte de l'expression de la racine de la moyenne des carrés de l'écart par rapport à la médiane en pourcentage de la médiane.
- **Moyenne centrée COV.** Le coefficient de variation avec moyenne centrée résulte de l'expression de l'écart-type en tant que pourcentage de la moyenne.
- **Ecart type :** L'écart-type est la racine carrée positive de la somme des carrés des écarts des ratios relatifs à la moyenne divisée par le nombre total des ratios moins un.
- **Intervalle :** Résultat de la soustraction du ratio minimal au ratio maximal.
- **Minimum :** Le minimum est le plus petit ratio.
- **Maximum.** Le maximum est le plus grand ratio.

Index de concentration. Le coefficient de concentration mesure le pourcentage des ratios compris dans un intervalle. Vous pouvez le calculer de deux manières :

- **Ratios entre :** Dans ce cas, vous définissez l'intervalle de manière explicite en précisant les valeurs minimale et maximale. Entrez les valeurs des proportions inférieure et supérieure, puis cliquez sur Ajouter pour obtenir un intervalle.
- **Ratios dans :** Dans ce cas, vous définissez l'intervalle de manière implicite en indiquant le pourcentage de la médiane. Entrez une valeur comprise entre 0 et 100, et cliquez sur Ajouter. La limite inférieure de l'intervalle est égale à (1 – 0,01×valeur)×médiane et la limite supérieure est égale à (1 + 0,01×valeur)×médiane.

Chapitre
33

Courbes ROC

Cette procédure constitue un moyen efficace d'évaluer les performances des méthodes de classement ne mettant en œuvre qu'une seule variable à deux modalités et utilisées pour la classification des sujets.

Exemple : Une banque envisage de classer correctement ses clients en modalités, à savoir ceux qui assumeront ou non le remboursement de leur prêt. Des méthodes particulières sont développées afin de supporter la prise de décision. Les courbes ROC peuvent être utilisées pour évaluer le mode de fonctionnement optimal de ces méthodes.

Statistiques. La zone inférieure à la courbe ROC comporte un intervalle de confiance ainsi que les coordonnées de cette courbe. Diagrammes : Courbe ROC

Méthodes. L'estimation de la zone située sous la courbe ROC peut être calculée de façon paramétrique ou non à l'aide du modèle exponentiel binégatif.

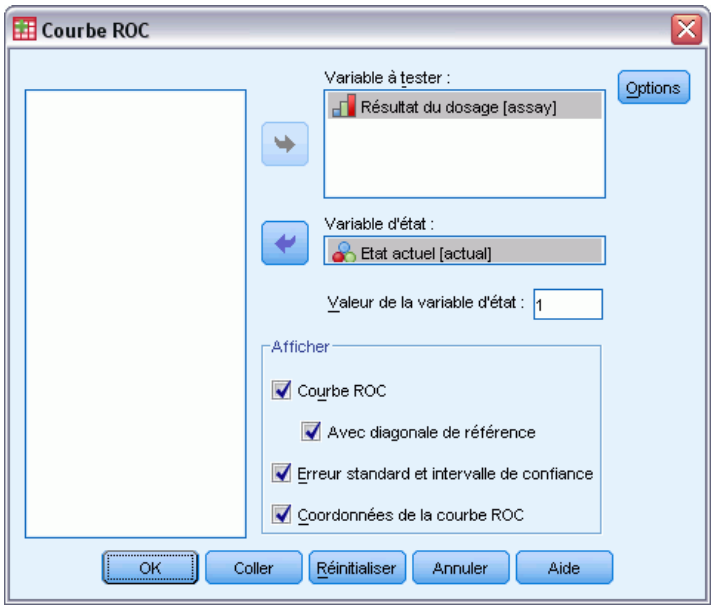
Données. Les variables de test sont quantitatives. Les variables de test sont souvent composées des probabilités issues d'une analyse discriminante ou d'une régression logistique, ou bien des scores indiqués sur une échelle arbitraire et spécifiant la « force de conviction » d'un indicateur lorsqu'un sujet se rapporte à l'une ou l'autre des modalités. La variable d'état peut être d'un type quelconque et indique la véritable modalité à laquelle un sujet appartient. La valeur de la variable d'état indique la modalité à considérer comme *positive*.

Hypothèses : Les nombres croissants d'une échelle d'indicateurs confirment que le sujet appartient à une modalité, tandis que les nombres décroissants d'une échelle confirment qu'il appartient à une autre modalité. L'utilisateur doit choisir la direction *positive*. On suppose également que la *véritable* modalité à laquelle chaque sujet appartient est connue.

Pour obtenir une courbe ROC

- A partir des menus, sélectionnez :
  Analyse > Courbe ROC...

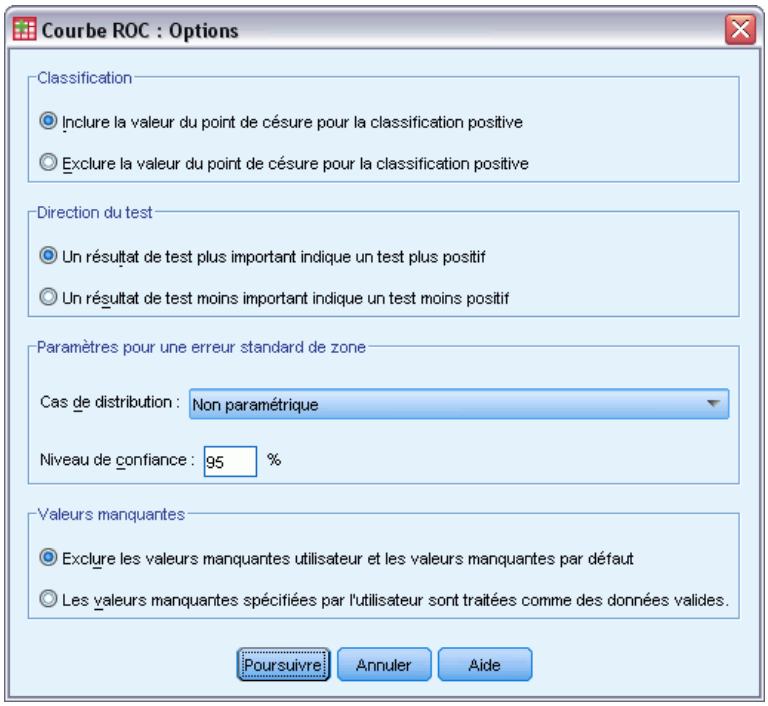
Figure 33-1
Boîte de dialogue Courbe ROC



- Sélectionnez une ou plusieurs variables de probabilité de test.
- Sélectionnez une variable d'état.
- Identifiez la valeur *positive* de la variable d'état.

Courbe ROC : Options

Figure 33-2
Boîte de dialogue Courbe ROC : Options



Vous pouvez indiquer les options suivantes pour votre analyse ROC :

Classification Permet de spécifier si la valeur du point de césure doit être incluse ou exclue lors d'une classification *positive*. Ce paramètre n'a pas de conséquence sur le résultat.

Direction du test Permet de spécifier la direction de l'échelle en fonction de la modalité *positive*.

Paramètres pour une erreur standard de zone Vous permet de spécifier la méthode utilisée pour estimer l'erreur standard de la zone située sous la courbe. Les méthodes disponibles sont des valeurs exponentielles non paramétriques et bi-négatives. Vous permet également de définir le niveau de l'intervalle de confiance. Les valeurs de l'intervalle se situent entre 50,1 % et 99,9 %.

Valeurs manquantes : Vous permet de spécifier comment traiter les valeurs manquantes.

Chapitre
34

Simulation

Les modèles prédictifs, tels que les modèles à régression linéaire, nécessitent un ensemble d'entrées connues afin de prédire un résultat ou une valeur cible. Toutefois, dans de nombreuses applications de la vie réelle, les valeurs des entrées sont incertaines. La simulation vous permet de prendre en compte l'incertitude relative aux entrées des modèles prédictifs et d'évaluer la probabilité de divers résultats du modèle en présence de cette incertitude. Par exemple, vous avez un modèle de profit qui comprend les coûts des matériaux en entrée, mais il existe une incertitude quant à ces coûts en raison de l'instabilité du marché. Dans ce cas, vous pouvez utiliser la simulation pour modéliser cette incertitude et déterminer les effets qu'elle a sur les profits.

La simulation offerte dans IBM® SPSS® Statistics utilise la méthode de Monte Carlo. Les entrées incertaines sont modélisées à l'aide de distributions de probabilité (telle que la distribution triangulaire), et des valeurs simulées sont générées pour ces entrées d'après ces distributions. Les entrées dont les valeurs sont connues sont fixées selon les valeurs connues. Le modèle prédictif est évalué à l'aide d'une valeur simulée pour chaque entrée incertaine et des valeurs fixes des entrées connues, afin de calculer la cible (ou les cibles) du modèle. Ce processus est répété plusieurs fois (en général des dizaines ou des centaines de milliers de fois) et résulte en une distribution des valeurs cible qui peut être utilisée pour répondre à des questions de nature probabiliste. Dans le contexte de SPSS Statistics, chaque répétition du processus génère une observation distincte (enregistrement) de données qui consiste en l'ensemble des valeurs simulées des entrées incertaines, des valeurs des entrées fixes et de la ou des cible(s) prédite(s) du modèle.

Pour lancer une simulation, vous devez spécifier des informations telles que le modèle prédictif, les distributions de probabilité des entrées incertaines et les corrélations entre ces entrées et les valeurs des entrées fixes. Une fois toutes les informations requises pour la simulation spécifiées, vous pouvez l'exécuter et éventuellement enregistrer les spécifications dans un fichier de **plan de simulation**. Il est possible de partager le plan de simulation avec d'autres utilisateurs, qui peuvent à leur tour exécuter la simulation sans avoir besoin de savoir exactement comment elle a été créée.

Deux interfaces sont disponibles pour utiliser les simulations. Le Générateur de simulation est une interface avancée destinée aux utilisateurs qui conçoivent et exécutent des simulations. Il offre l'ensemble complet de fonctionnalités permettant la conception de simulations, l'enregistrement des spécifications dans un fichier de plan de simulation, la spécification du résultat et l'exécution de la simulation. Il est possible de construire une simulation basée sur un fichier de modèle IBM SPSS ou sur un ensemble d'équations personnalisées que vous définissez dans le Générateur de simulation. Il est également possible de charger un plan de simulation existant dans le Générateur de simulation, de modifier ses paramètres et d'exécuter la simulation, ainsi que d'enregistrer le plan mis à jour au besoin. Pour les utilisateurs disposant d'un plan de simulation et qui souhaitent principalement exécuter une simulation, une interface simplifiée est disponible. Elle vous permet de modifier des paramètres afin d'exécuter la simulation dans différentes conditions, mais n'offre pas l'ensemble des fonctionnalités du Générateur de simulation qui permettent de concevoir des simulations.

Pour concevoir une simulation pour un modèle prédictif défini dans un fichier de modèle

Pour concevoir une simulation pour un modèle prédictif défini par des équations personnalisées

Pour exécuter une simulation à partir d'un plan de simulation

Pour concevoir une simulation basée sur un fichier de modèle

- A partir des menus, sélectionnez :
  Analyse > Simulation...
- Cliquez sur Sélectionner un fichier de modèle SPSS, puis cliquez sur Poursuivre.
- Ouvrez le fichier de modèle souhaité.

  Le fichier de modèle peut être un fichier XML ou une archive ZIP qui contient le modèle au format PMML créé à partir de IBM® SPSS® Statistics ou de IBM® SPSS® Modeler. Pour plus d'informations, reportez-vous à la section Onglet Modèle sur p. 319.
- Sur l'onglet Simulation (dans le Générateur de simulation), spécifiez les distributions de probabilité des entrées simulées et les valeurs des entrées fixes. Si l'ensemble de données actif contient des données historiques relatives aux entrées simulées, cliquez sur Tout ajuster pour déterminer automatiquement la distribution la mieux adaptée à chaque entrée, ainsi que les corrélations entre celles-ci.
- Cliquez sur Exécuter pour exécuter la simulation. Par défaut, le plan de simulation qui spécifie les détails de la simulation, est enregistré à l'emplacement indiqué dans les paramètres d'enregistrement.

  Vous pouvez au besoin procéder aux actions suivantes :
  - Modifier l'emplacement du plan de simulation enregistré.
  - Spécifier les corrélations connues existant entre les entrées simulées.
  - Spécifier l'analyse de sensibilité pour vérifier l'effet de la variation de la valeur d'une entrée fixe ou d'un paramètre de distribution d'une entrée simulée.
  - Spécifiez des options avancées telles que le nombre maximum d'observations à générer ou un échantillonage des extrémités.
  - Personnaliser le résultat.
  - Enregistrer les données simulées dans un fichier de données.

Pour concevoir une simulation basée sur des équations personnalisées

- A partir des menus, sélectionnez :
  Analyse > Simulation...
- Cliquez sur Entrer les équations puis cliquez sur Poursuivre.
- Cliquez sur Nouvelle équation sur l'onglet Modèle (du Générateur de simulation) afin de définir chaque équation dans votre modèle prédictif.
- Cliquez sur l'onglet Simulation et spécifiez les distributions de probabilité des entrées simulées et les valeurs des entrées fixes. Si l'ensemble de données actif contient des données historiques

relatives aux entrées simulées, cliquez sur Tout ajuster pour déterminer automatiquement la distribution la mieux adaptée à chaque entrée, ainsi que les corrélations entre celles-ci.

▶ Cliquez sur Exécuter pour exécuter la simulation. Par défaut, le plan de simulation qui spécifie les détails de la simulation, est enregistré à l'emplacement indiqué dans les paramètres d'enregistrement.

Vous pouvez au besoin procéder aux actions suivantes :

- Modifier l'emplacement du plan de simulation enregistré.
- Spécifier les corrélations connues existant entre les entrées simulées.
- Spécifier l'analyse de sensibilité pour vérifier l'effet de la variation de la valeur d'une entrée fixe ou d'un paramètre de distribution d'une entrée simulée.
- Spécifiez des options avancées telles que le nombre maximum d'observations à générer ou un échantillonage des extrémités.
- Personnaliser le résultat.
- Enregistrer les données simulées dans un fichier de données.

Pour exécuter une simulation à partir d'un plan de simulation

Deux options sont disponibles pour exécuter une simulation à partir d'un plan de simulation. Vous pouvez utiliser soit la boîte de dialogue Exécuter la simulation, conçue principalement pour exécuter une simulation à partir d'un plan de simulation, soit le Générateur de simulation.

Pour utiliser la boîte de dialogue Exécuter la simulation :

▶ A partir des menus, sélectionnez :
Analyse > Simulation...

▶ Cliquez sur Ouvrir un plan de simulation existant.

▶ Vérifiez que la case Ouvrir dans le Générateur de simulation n'est pas cochée et cliquez sur Poursuivre.

▶ Ouvrez le plan de simulation souhaité.

▶ Cliquez sur Exécuter dans la boîte de dialogue Exécuter la simulation.

Vous pouvez également procéder aux actions suivantes :

- Configurer ou modifier l'analyse de sensibilité pour vérifier l'effet de la variation de la valeur d'une entrée fixe ou d'un paramètre de distribution d'une entrée simulée.
- Réajuster les distributions et les corrélations des entrées simulées aux nouvelles données.
- Modifier la distribution d'une entrée simulée.
- Personnaliser le résultat.
- Enregistrer les données simulées dans un fichier de données.

Pour exécuter la simulation à partir du Générateur de simulation :

- A partir des menus, sélectionnez :
  Analyse > Simulation...
- Cliquez sur Ouvrir un plan de simulation existant.
- Sélectionnez la case Ouvrir dans le Générateur de simulation et cliquez sur Poursuivre.
- Ouvrez le plan de simulation souhaité.
- Modifiez les paramètres requis sur l'onglet Simulation.
- Cliquez sur Exécuter pour exécuter la simulation.

Vous pouvez également procéder aux actions suivantes :

- Enregistrer les paramètres modifiés dans un nouveau plan de simulation ou remplacer le plan de simulation existant.
- Réajuster les distributions et les corrélations des entrées simulées aux nouvelles données.
- Personnaliser le résultat.
- Enregistrer les données simulées dans un fichier de données.

Générateur de simulation

Le Générateur de simulation fournit l'ensemble complet des fonctionnalités permettant de concevoir et d'exécuter des simulations. Il vous permet de réaliser les tâches générales suivantes :

- Concevoir et exécuter une simulation pour un modèle IBM SPSS défini dans un fichier de modèle PMML.
- Concevoir et exécuter une simulation pour un modèle prédictif défini par un ensemble d'équations personnalisées.
- Exécuter une simulation basée sur un plan de simulation existant, et éventuellement modifier des paramètres du plan.

Onglet Modèle

L'onglet Modèle spécifie la source du modèle prédictif utilisée pour la simulation.

Sélectionner un fichier de modèle SPSS. Cette option spécifie que le modèle prédictif est défini dans un fichier de modèle IBM SPSS. Un fichier de modèle IBM SPSS est un fichier XML ou une archive ZIP qui contient le modèle au format PMML créé à partir de IBM® SPSS® Statistics ou de IBM® SPSS® Modeler. Les modèles prédictifs sont créés par des procédures telles que la régression linéaire et les arbres de décision dans SPSS Statistics, et ils peuvent être exportés vers un fichier de modèle. Vous pouvez sélectionner un fichier de modèle différent en cliquant sur Parcourir et en accédant à son emplacement.

Modèles PMML pris en charge par les simulations

Régression linéaire
Modèle linéaire généralisé
Régression logistique binaire
Régression logistique multinomiale
Régression multinomiale ordinale
Régression de Cox
TREE
Arbre boosté (C5)
Discriminant
Classification en deux étapes
Nuées dynamiques
Réseau de neurones
Ensemble de règles (Liste de décisions)

Remarque : les modèles PMML qui possèdent plusieurs champs cible (variables) ou scissions ne sont pas pris en charge par les simulations.

Entrez les équations pour le modèle. Cette option spécifie que le modèle prédictif consiste en une ou plusieurs équations personnalisées que vous devez créer. Créez les équations en cliquant sur Nouvelle équation. Cette action ouvre l'Editeur d'équation. Vous pouvez modifier des équations existantes, les copier et les utiliser en tant que modèles pour de nouvelles équations, les réorganiser et les supprimer.

- Le Générateur de simulation ne prend pas en charge les systèmes d'équations simultanées ou les équations non linéaires dans la variable cible.
- Les équations personnalisées sont évaluées en fonction de leur ordre de spécification. Si l'équation d'une cible donnée dépend d'une autre cible, alors la seconde cible doit être définie par une équation précédente.

  Par exemple, considérons l'ensemble des trois équations suivantes, l'équation du *profit* dépend des valeurs des *revenus* et des *dépenses*, par conséquent les équations des *revenus* et des *dépenses* doivent précéder celle du *profit*.

```
revenus = prix*volume

dépenses= fixes + volume*(coûts_matériaux_unité +
coûts_main-d'oeuvre_unité)

profit = revenus - dépenses
```

Editeur d'équation

L'Editeur d'équation vous permet de créer ou de modifier une équation personnalisée de votre modèle prédictif.

- L'expression de l'équation peut contenir des champs provenant de l'ensemble de données actif ou de nouveaux champs d'entrée que vous définissez dans l'Editeur d'équation.

- Vous pouvez spécifier les propriétés de la cible, par exemple son niveau de mesure et ses étiquettes de valeur, et choisir si un résultat est généré pour la cible.
- Vous pouvez utiliser des cibles provenant d'équations précédemment définies en tant qu'entrées de l'équation en cours, ce qui vous permet de créer des équations couplées.
- Vous pouvez aussi ajouter un commentaire descriptif à l'équation. Les commentaires s'afichent sur l'onglet Modèle en même temps que l'équation.

▶ Saisissez le nom de la cible. Si vous le souhaitez, cliquez sur Modifier en dessous de la zone de texte Cible pour ouvrir la boîte de dialogue Entrées définies, qui vous permet de modifier les propriétés par défaut de la cible.

▶ Pour construire une expression, vous pouvez soit coller les composants dans le champ Expression numérique, soit les saisir au clavier directement dans ce même champ.

- Il est possible de construire une expression en utilisant des champs provenant de l'ensemble de données actif ou de définir de nouvelles entrées en cliquant sur le bouton Nouveau. Cette action ouvre la boîte de dialogue Entrées définies.
- Vous pouvez coller des fonctions en sélectionnant un groupe dans la liste Groupe de fonctions, puis en double-cliquant sur la fonction désirée dans la liste Fonctions (ou sélectionnez la fonction, puis cliquez sur la flèche adjacente à la liste Groupe de fonctions). Définissez tous les paramètres indiqués par un point d'interrogation. Le groupe de fonctions étiqueté Tous répertorie toutes les fonctions disponibles. Une brève description de la fonction sélectionnée apparaît dans une zone particulière de la boîte de dialogue.
- Les constantes chaîne doivent être présentées entre guillemets.
- Si des valeurs contiennent des chiffres décimaux, utilisez la virgule comme indicateur décimal.

Remarque : les simulations ne prennent pas en charge les équations personnalisées contenant des cibles chaîne.

Entrées définies

La boîte de dialogue Entrées définies vous permet de définir de nouvelles entrées et de définir les propriétés des cibles.

Nom. Spécifiez le nom de la cible ou de l'entrée.

Cible. Vous pouvez spécifier le niveau de mesure de la cible. Le niveau de mesure par défaut est continu. Vous pouvez également indiquer si un résultat sera créé pour la cible. Par exemple, pour un ensemble d'équations couplées, il se peut que vous soyez intéressé uniquement par la génération du résultat de la cible de l'équation finale, par conséquent dans un tel cas vous supprimerez le résultat des autres cibles.

Entrée à simuler. Cette option indique que les valeurs de l'entrée seront simulées en fonction d'une distribution de probabilité spécifique (cette distribution de probabilité est spécifiée sur l'onglet Simulation). Le niveau de mesure détermine l'ensemble par défaut des distributions considérées lors de la recherche de la distribution la mieux adaptée à l'entrée (cliquez sur Ajuster ou Tout ajuster sur l'onglet Simulation). Par exemple, si le niveau de mesure est ordinal, la distribution binomiale (appropriée aux données ordinales) sera considérée mais la distribution normale ne le sera pas.

Entrée de valeur fixe. Cette option spécifie que la valeur de l'entrée est connue et sera fixée à la valeur connue. Les entrées fixes peuvent être des valeurs numériques ou des chaînes. Spécifiez la valeur de l'entrée fixe. Les valeurs de chaîne ne doivent pas être entre guillemets.

Etiquettes de valeurs. Vous pouvez spécifier des étiquettes de valeurs pour les cibles, les entrées simulées et les entrées fixes. Les étiquettes de valeur sont utilisées dans les diagrammes et les tableaux de résultats.

Onglet Simulation.

L'onglet Simulation spécifie toutes les propriétés de la simulation, autres que le modèle prédictif associé. Vous pouvez réaliser les tâches générales suivantes sur l'onglet Simulation :

- Spécifier les distributions de probabilité des entrées simulées et les valeurs des entrées fixes.
- Spécifier les corrélations existant entre les entrées simulées.
- Spécifier les options avancées telles que l'échantillonage des extrémités et les critères d'ajustement des distributions aux données historiques.
- Personnaliser le résultat.
- Spécifier l'emplacement d'enregistrement du plan de simulation et éventuellement enregistrer les données simulées.

Champs simulés

Pour exécuter une simulation, chaque entrée du modèle prédictif doit être spécifiée en tant qu'entrée fixe ou entrée simulée. Les entrées simulées sont celles dont les valeurs sont incertaines et seront générées à partir d'une distribution de probabilité spécifique. Les distributions les mieux adaptées aux entrées simulées, ainsi que les corrélations entre ces entrées, peuvent être déterminées automatiquement à partir des données historiques des entrées concernées. Il est également possible de spécifier manuellement des distributions ou des corrélations si aucune donnée historique n'est disponible, ou si vous avez besoin d'utiliser des distributions ou des corrélations spécifiques.

Les entrées fixes sont celles dont les valeurs sont connues. Elles restent constantes pour chaque observation générée dans la simulation. Par exemple, vous avez un modèle de régression linéaire concernant les ventes qui est fonction d'un certain nombre d'entrées, dont le prix, et vous souhaitez que le prix soit fixe et corresponde au prix du marché actuel. Vous spécifierez donc le prix en tant qu'entrée fixe.

Ajustement automatique des distributions et calcul automatique des corrélations pour les entrées simulées. Si l'ensemble de données actif contient des données historiques relatives aux entrées à simuler, vous pouvez déterminer automatiquement la distribution la mieux adaptée à ces entrées, ainsi que les corrélations entre celles-ci. Les étapes à suivre sont les suivantes :

▶ Vérifiez que chaque entrée du modèle que vous souhaitez simuler est mise en correspondance avec le champ approprié dans l'ensemble de données actif. Les entrées du modèle sont répertoriées dans la colonne Entrée et la colonne Ajuster à affiche le champ correspondant dans l'ensemble

de données actif. Vous pouvez mettre les entrées en correspondance avec un champ différent de l'ensemble de données actif en sélectionnant un élément différent dans la liste déroulante Ajuster à.

La valeur *-Aucun-* dans la colonne Ajuster à indique que l'entrée ne peut être automatiquement mise en correspondance avec un champ de l'ensemble de données actif. Par défaut, les entrées de modèle sont mises en correspondance avec des champs de l'ensemble de données dont le nom et le niveau de mesure correspondent. Si l'ensemble de données actif ne contient pas de données historiques pour une entrée particulière, déterminez manuellement la distribution à utiliser pour cette entrée, ou spécifiez que l'entrée est une entrée fixe. *Remarque* : l'ajustement de la distribution ne prend pas en charge l'ajustement à des champs chaîne. Si votre modèle prédictif contient une entrée chaîne, vous devez la spécifier en tant qu'entrée fixe.

▶ Cliquez sur Tout ajuster.

La distribution la mieux adaptée et ses paramètres associés s'affichent alors dans la colonne Distribution en même temps qu'un diagramme de la distribution superposé à un histogramme (diagramme en bâtons) représentant les données historiques. Les corrélations entre les entrées simulées s'affichent dans les paramètres Corrélations. Vous pouvez examiner les résultats de l'ajustement et personnaliser l'ajustement automatique de la distribution d'une entrée particulière en sélectionnant la ligne de cette entrée et en cliquant sur Détails de l'ajustement. Pour plus d'informations, reportez-vous à la section Détails de l'ajustement sur p. 325.

Vous pouvez exécuter l'ajustement automatique de la distribution d'une entrée particulière en sélectionnant la ligne de cette entrée et en cliquant sur Ajuster. Les corrélations de toutes les entrées simulées mises en correspondance à des champs de l'ensemble de données actif sont calculées automatiquement.

Remarque : Pour les entrées continues et ordinales, s'il est impossible de trouver un ajustement acceptable pour une des distributions testées, la distribution empirique est alors utilisée. Pour les entrées continues, la distribution empirique est la fonction de distribution cumulée des données historiques. Pour les entrées ordinales, la distribution empirique est la distribution qualitative des données historiques.

Spécification manuelle des distributions. Vous pouvez spécifier manuellement la distribution de probabilité de n'importe quelle entrée simulée en sélectionnat la distribution souhaitée dans la liste déroulante Type et en saisissant les paramètres de la distribution dans la grille Paramètres. Une fois les paramètres d'une distribution définis, un diagramme de la distribution s'appuyant sur les paramètres spécifiés s'affiche à côté de la grille Paramètres. Voici quelques remarques sur des distributions particulières :

- **Qualitative :** La distribution Qualitative décrit un champ d'entrée comportant un nombre fixe de valeurs numériques, appelées modalités. Chaque modalité possède une probabilité associée, de sorte que la somme des probabilités de toutes les modalités est égale à 1. Pour entrer une modalité, cliquez sur la colonne de gauche dans la grille Paramètres et remplacez la valeur textuelle par la valeur numérique représentant la modalité. Entrez la probabilité associée à la modalité dans la colonne de droite.
- **Binomiale négative - Echecs.** Décrit la distribution du nombre d'échecs dans une séquence de tentatives avant qu'un nombre spécifique de succès ne soit observé. Le paramètre *seuil* est le nombre spécifique de succès et le paramètre *prob* est la probabilité de succès pour chaque tentative donnée.

- **Binomiale négative - Tentatives.** Décrit la distribution du nombre de tentatives requis avant qu'un nombre spécifique de succès ne soit observé. Le paramètre *seuil* est le nombre spécifique de succès et le paramètre *prob* est la probabilité de succès pour chaque tentative donnée.
- **Intervalle :** Cette distribution consiste en un ensemble d'intervalles, chaque intervalle ayant une probabilité qui lui a été affecté, de sorte que la somme de toutes les probabilités des intervalles est égale à 1. Les valeurs à l'intérieur d'un intervalle donné sont tirées d'une distribution uniforme définie sur cet intervalle. Les intervalles sont spécifiés en entrant une valeur minimale, une valeur maximale et une probabilité.

  Par exemple, vous pensez que le coût des matières premières a 40 % de chance de chuter d'environ 10 - 15 $ par unité et une chance de 60 % de chuter d'environ 15 - 20 $ par unité. Vous procédez à la modélisation du coût à l'aide d'une distribution Intervalle consistant en deux intervalles [10 - 15] et [15 - 20], définissez la probabilité associée au premier intervalle à 0,4 et la probabilité associée au second intervalle à 0,6. Les intervalles n'ont pas besoin d'être contigus et ils peuvent même se chevaucher. Par exemple, vous pouvez spécifier les intervalles 10 - 15 $ et 20-25 $ ou 10-15 $ et 13-16 $.
- **Weibull.** Le paramètre *c* est un paramètre d'emplacement facultatif qui spécifie l'emplacement de l'origine de la distribution.

Les paramètres des distributions suivantes ont la même signification que dans les fonctions de variable aléatoire associées disponibles dans la boîte de dialogue Calculer la variable : Bernoulli, Bêta, Binomiale, Exponentielle, Gamma, Lognormale, Binomiale négative (Echecs et Tentatives), Normale, Poisson et Uniforme.

Spécification des entrées fixes. Spécifiez une entrée fixe en sélectionnant Fixe dans la liste déroulante Type située dans la colonne Distribution et en saisissant la valeur de l'entrée fixe. Cette valeur peut être numérique ou alphanumérique selon que l'entrée est une entrée numérique ou chaîne. Les valeurs de chaîne ne doivent pas être entre guillemets.

Spécification des bornes sur les valeurs simulées. La plupart des distributions prennent en charge les bornes supérieure et inférieure sur les valeurs simulées. La borne inférieure peut être spécifiée en saisissant sa valeur dans la zone de texte Min et la borne supérieure peut être spécifiée en saisissant sa valeur dans la zone de texte Max.

Verrouillage des entrées simulées. Le verrouillage d'une entrée simulée, en cochant la case correspondante dans la colonne indiquée par une icône en forme de verrou, exclut cette entrée de l'ajustement automatique de la distribution. Cette option est très utile lorsque vous spécifiez manuellement une distribution et souhaitez vous assurer qu'elle ne sera pas affectée par un ajustement automatique. Le verrouillage est également très utile si vous avez l'intention de partager votre plan de simulation avec des utilisateurs qui l'exécuteront dans la boîte de dialogue Exécuter la simulation, et que vous souhaitez empêcher ces utilisateurs de modifier certaines distributions. A cette fin, les distributions d'entrées verrouillées ne peuvent être modifiées dans la boîte de dialogue Exécuter la simulation.

Analyse de sensibilité. L'analyse de sensibilité vous permet de vérifier l'effet de modifications systématiques apportées à une entrée fixe ou à un paramètre de distribution associé à une entrée simulée, en générant un ensemble indépendant d'observations simulées, —soit une simulation

distincte— pour chaque valeur spécifiée. Pour spécifier l'analyse de sensiblité, sélectionnez une entrée simulée ou une entrée fixe et cliquez sur Analyse de sensibilité. L'analyse de sensibilité est limitée à une seule entrée fixe ou un seul paramètre de distribution associé à une entrée simulée. Pour plus d'informations, reportez-vous à la section Analyse de sensibilité sur p. 326.

Icônes d'état d'ajustement

Les icônes de la colonne Ajuster à indiquent l'état d'ajustement de chaque champ d'entrée.

Aucune distribution n'a été spécifiée pour l'entrée et l'entrée n'a pas été spécifiée en tant qu'entrée fixe. Pour exécuter la simulation, vous devez soit spécifier une distribution pour cette entrée, soit la définir en tant qu'entrée fixe et lui affecter une valeur.

L'entrée a été précédemment ajustée à un champ qui n'existe pas dans l'ensemble de données actif. Aucune action n'est nécessaire sauf si vous souhaitez réajuster la distribution de l'entrée à l'ensemble de données actif.

La distribution la mieux adaptée a été remplacée par une autre distribution dans la boîte de dialogue Détails de l'ajustement.

L'entrée est définie sur la distribution la mieux adaptée.

La distribution a été spécifiée manuellement ou des itérations de l'analyse de sensibilité ont été spécifiées pour cette entrée.

Détails de l'ajustement

La boîte de dialogue Détails de l'ajustement affiche les résultats de l'ajustement automatique de la distribution d'une entrée spécifique. Les distributions sont ordonnées en fonction de la qualité de l'ajustement, la distribution la mieux adaptée étant répertoriée en tête de liste. Vous pouvez remplacer la distribution la mieux adaptée en sélectionnant le bouton radio correspondant à la distribution souhaitée dans la colonne Utilisation. En sélectionnant un bouton radio dans la colonne Utilisation, vous affichez également un diagramme de la distribution superposé à un histogramme (diagramme en bâtons) représentant les données historiques de l'entrée.

Statistiques de l'ajustement. Par défaut, pour les champs continus, le test d'Anderson-Darling est utilisé pour déterminer la qualité de l'ajustement. Pour les champs continus uniquement, vous pouvez aussi choisir le test de Kolmogorov-Smirnoff pour déterminer la qualité de l'ajustement. Ce choix peut être fait dans les paramètres des Options avancées. Pour les entrées continues, les résultats des deux tests s'affichent dans la colonne Statistiques de l'ajustement avec une indication du test choisi (A pour Anderson-Darling et K pour Kolmogorov-Smirnoff) pour l'ordonnancement des distributions. Pour les entrées ordinales et nominales, le test du Khi-deux est utilisé. Les valeurs p associées aux tests sont également affichées.

Paramètres. Les paramètres de distribution associés à chaque distribution ajustée sont affichés dans la colonne Paramètres. Les paramètres des distributions suivantes ont la même signification que dans les fonctions de variable aléatoire associées disponibles dans la boîte de dialogue Calculer la

variable : Bernoulli, Bêta, Binomiale, Exponentielle, Gamma, Lognormale, Binomiale négative (Echecs et Tentatives), Normale, Poisson et Uniforme. Pour la distribution Qualitative, les noms des paramètres sont les modalités et les valeurs des paramètres sont les probabilités associées.

Réajustement avec un ensemble de distribution personnalisé. Par défaut, le niveau de mesure de l'entrée est utilisé pour déterminer l'ensemble des distributions considéré pour l'ajustement automatique de la distribution. Par exemple, des distributions continues telles que les distributions Lognormale et Gamma sont prises en considération lors de l'ajustement d'une entrée continue mais les distributions discrètes telles que les distributions Binomiale et de Poisson ne le sont pas. Vous pouvez choisir un sous-ensemble des distributions par défaut en sélectionnant les distributions souhaitées dans la colonne Réajuster. Vous pouvez également remplacer l'ensemble des distributions par défaut en sélectionnant un niveau de mesure différent dans la liste déroulante Traiter en tant que (Mesure) puis en sélectionnant les distributions souhaitées dans la colonne Réajuster. Cliquez sur Exécuter le réajustement pour réajuster l'ensemble de distributions personnalisé.

Analyse de sensibilité

L'analyse de sensibilité vous permet de vérifier l'effet de la variation d'une entrée fixe ou d'un paramètre de distribution associé à une entrée simulée, pour un ensemble spécifique de valeurs. Un ensemble indépendant d'observations simulées - en fait, une simulation distincte - est généré pour chaque valeur spécifiée, ce qui vous permet de vérifier les effets de la variation de l'entrée. Chaque ensemble d'observations simulées est appelé **itération**.

Itérer. Cette option vous permet d'indiquer l'ensemble des valeurs appliquées à l'entrée et en fonction desquelles l'entrée va varier.

- Si vous préférez faire varier la valeur d'un paramètre de distribution, sélectionnez ce paramètre dans la liste déroulante. Entrez l'ensemble des valeurs dans la grille représentant les valeurs de paramètres par itération. Cliquez sur Poursuivre pour ajouter les valeurs spécifiées à la grille Paramètres de l'entrée associée, avec un indice indiquant le nombre d'itérations de la valeur.
- Pour les distributions Qualitative et Intervalle, les probabilités des modalités ou des intervalles, respectivement, peuvent être modifiées mais les valeurs des modalités et les extrema des intervalles ne peuvent l'être. Sélectionnez une modalité ou un intervalle dont vous souhaitez faire varier la probabilité et spécifiez l'ensemble des probabilités dans la grille des valeurs de paramètres par itération. Les probabilités des autres modalités ou intervalles seront automatiquement ajustées en conséquence.

Aucune itération. Utilisez cette option pour annuler les itérations d'une entrée. Cliquez sur Poursuivre pour supprimer les itérations.

Corrélations

Les entrées des modèles prédictifs sont souvent corrélées, comme par exemple, la taille et le poids. Les corrélations existant entre les entrées qui seront simulées doivent être prises en compte afin d'assurer que les valeurs simulées conservent ces corrélations.

Recalculer les corrélations lors de l'ajustement. Cette option indique que les corrélations entre les entrées simulées sont automatiquement calculées lors de l'ajustement des distributions à l'ensemble de données actif à l'aide de l'action Tout ajuster ou Ajuster située dans les paramètres Champs simulés.

Ne pas recalculer les corrélations lors de l'ajustement. Choisissez cette option si vous souhaitez spécifier manuellement les corrélations et éviter qu'elles ne soient remplacées lors de l'ajustement automatique des distributions à l'ensemble de données actif. Les valeurs saisies dans la grille Corrélations doivent être comprises entre -1 et 1. Une valeur égale à 0 indique qu'il n'y a pas de corrélation entre les entrées appariées.

Restaurer. Cette option rétablit toutes les corrélations sur 0.

Options avancées

Nombre maximum d'observations. Cette option indique le nombre maximum d'observations de données simulées et les valeurs cible associées à générer. Lorsque l'analyse de sensibilité est sélectionnée, cette option indique le nombre maximal d'observations pour chaque itération.

Cible des critères d'arrêt. Si votre modèle prédictif contient plusieurs cibles, vous pouvez choisir la cible à laquelle s'appliquent les critères d'arrêt.

Critères d'arrêt. Ces options permettent de spécifier les critères d'arrêt de la simulation, en général avant que le nombre maximal d'observations autorisé ait été généré.

- **Poursuivre jusqu'à atteindre le maximum.** Cette option indique que des observations simulées seront générées tant que le nombre maximal d'observations n'est pas atteint.
- **Arrêter une fois les extrémités échantillonnées.** Utilisez cette option si vous souhaitez vous assurer que l'une des extrémités d'une distribution cible spécifique a bien été échantillonnée. Les observations simulées seront générées tant que l'échantillonnage des extrémités spécifié n'est pas terminé ou que le nombre maximal d'observations n'est pas atteint. Si votre modèle prédictif contient plusieurs cibles, choisissez celle à laquelle s'appliquent ces critères d'arrêt dans la liste déroulante Cible des critères d'arrêt.

  Type. Vous pouvez définir la limite de la zone d'extrémité en spécifiant une valeur de cible, par ex. 10 000 000 ou un centile, par ex. le 99e centile. Si vous choisissez Valeur dans la liste déroulante Type, saisissez la valeur de la limite dans la zone de texte Valeur et choisissez dans la liste déroulante Côté le côté de la zone d'extrémité (Gauche ou Droite). Si vous choisissez Centile dans la liste déroulante Type, saisissez une valeur dans la zone de texte Centile.

  Effectif. Spécifiez ici le nombre de valeurs de la cible qui doivent se situer dans la zone d'extrémité afin de vous assurer que l'extrémité a été correctement échantillonnée. La génération d'observations va cesser une fois ce nombre atteint.
- **Arrêter lorsque l'intervalle de confiance de la moyenne atteint le seuil spécifié.** Utilisez cette option si vous souhaitez vous assurer que la moyenne d'une cible donnée est connue avec un degré de précision spécifique. Les observations simulées seront générées tant que le degré de précision spécifique ou le nombre maximal d'observations n'est pas atteint. Pour utiliser cette option, vous devez spécifier un niveau de confiance et un seuil. Les observations simulées seront générées tant que l'intervalle de confiance associé au niveau spécifique

n'atteint pas le seuil. Par exemple, vous pouvez utiliser cette option pour spécifier que les observations doivent être générées tant que l'intervalle de confiance de la moyenne à un niveau de confiance de 95 % n'atteint pas 5 % de la valeur moyenne. Si votre modèle prédictif contient plusieurs cibles, choisissez celle à laquelle s'appliquent ces critères d'arrêt dans la liste déroulante Cible des critères d'arrêt.

Type de seuil. Le seuil peut être spécifié en tant que valeur numérique ou en tant que pourcentage de la moyenne. Si vous choisissez Valeur dans la liste déroulante Type de seuil, saisissez le seuil dans la zone de texte Seuil sous forme de valeur. Si vous choisissez Pourcentage dans la liste déroulante Type de seuil, saisissez une valeur dans la zone de texte Seuil sous forme de pourcentage.

Nombre d'observations à échantillonner. Cette option permet de spécifier le nombre d'observations à utiliser lors de l'ajustement automatique des distributions des entrées simulées à l'ensemble de données actif. Si votre ensemble de données est très grand, vous voudrez peut-être limiter le nombre d'observations à utiliser pour l'ajustement de la distribution. En sélectionnant Limiter à N observations, seules les N premières observations sont utilisées.

Critères de qualité de l'ajustement (continus). Pour les entrées continues, vous pouvez utiliser le test d'Anderson-Darling ou le test de Kolmogorov-Smirnoff relatif à la qualité de l'ajustement afin de classer les distributions des entrées simulées lors de leur ajustement à l'ensemble de données actif. Le test d'Anderson-Darling est sélectionné par défaut et est particulièrement recommandé si vous souhaitez assurer le meilleur ajustement possible dans les zones d'extrémité.

Distribution empirique. Applicable aux entrées continues, la distribution empirique est la fonction de distribution cumulée des données historiques. Cette option vous permet de spécifier le nombre d'intervalles utilisés pour calculer la distribution empirique des entrées continues. La valeur par défaut est 100 et la valeur maximale est 1000.

Dupliquer les résultats. Définir un générateur aléatoire vous permet de dupliquer votre simulation. Spécifiez un entier ou cliquez sur Générer, ce qui crée un entier pseudo-aléatoire compris entre 1 et 2147483647, inclus. La valeur par défaut est 498654860.

Fonctions de densité

Ces paramètres vous permettent de personnaliser le résultat des fonctions de densité de probabilité et des fonctions de distribution cumulée pour les cibles continues, ainsi que les diagrammes en bâtons des valeurs prédites pour les cibles qualitatives.

Fonction de densité de probabilité (FDP). La fonction de densité de probabilité affiche la distribution des valeurs cible. Pour les cibles continues, elle vous permet de déterminer la probabilité que la cible se trouve dans une zone donnée. Pour les cibles qualitatives (cibles dont le niveau de mesure est nominal ou ordinal), un diagramme en bâtons est généré, qui affiche le pourcentage d'observations situées dans chaque modalité de la cible. Des options supplémentaires sont disponibles pour les cibles qualitatives des modèles PMML, dont le paramètre Valeurs de modalité à rapporter, décrit ultérieurement.

Pour les modèles de classification en deux étapes et les modèles de classification en nuées dynamiques, un diagramme en bâtons représentant l'appartenance aux classes est généré.

Fonction de distribution cumulée (FDC). La fonction de distribution cumulée affiche la probabilité qu'une valeur de la cible soit inférieure ou égale à une valeur donnée. Elle est disponible uniquement pour les variables continues.

Lignes de référence (continues). Vous pouvez demander l'ajout de plusieurs lignes de référence verticales aux fonctions de densité de probabilité et aux fonctions de distribution cumulée pour les cibles continues.

- **Sigmas.** Des lignes de références peuvent être ajoutées à plus et moins un nombre spécifique d'écarts-type de la moyenne de la cible.
- **Centiles :** Des lignes de références peuvent être ajoutées à un ou deux centiles de la distribution pour chaque cible, en entrant des valeurs dans les zones de texte Inférieure et Supérieure. Par exemple, une valeur de 95 dans la zone de texte Supérieure représente le 95e centile, qui est la valeur en dessous de laquelle se trouvent 95 % des observations. De même, une valeur de 5 dans la zone de texte Inférieure représente le 5e centile, qui est la valeur en dessous de laquelle se trouvent 5% des observations.
- **Lignes de référence personnalisées.** Des lignes de références peuvent être ajoutées à des valeurs spécifiques de la cible.

Remarque : lorsque plusieurs fonctions de densité ou de distribution sont affichées sur un diagramme unique (en raison de cibles ou résultats multiples provenant des itérations de l'analyse de sensibilité), des lignes de référence (autres que les lignes personnalisées) sont appliquées respectivement à chaque fonction.

Superposer les résultats provenant de cibles continues distinctes. Dans le cas de cibles continues multiples, cette option spécifie si les fonctions de distribution de ces cibles s'affichent sur une seule représentation graphique, superposant le diagramme des fonctions de densité de probabilité et celui des fonctions de distribution cumulée. Si cette option n'est pas sélectionnée, les résultats de chaque cible s'affichent sur un diagramme distinct.

Valeurs de modalité à rapporter. Pour les modèles PMML comportant des cibles qualitatives, le résultat du modèle est un ensemble de probabilités prédites, une pour chaque modalité, que la valeur de la cible tombe dans chaque modalité. La modalité présentant la probabilité la plus élevée est considérée comme la modalité prédite et est utilisée pour générer le diagramme en bâtons décrit dans le paramètre Fonction de densité de probabilité sus-mentionné. Sélectionnez Modalité prédite pour générer le diagramme en bâtons. Sélectionnez Probabilités prédites pour générer les histogrammes de la distribution des probabilités prédites de chaque modalité de la cible.

Regroupement pour l'analyse de sensibilité. Les simulations qui comprennent une analyse de sensibilité génèrent un ensemble indépendant de valeurs cible prédites pour chaque itération définie par l'analyse (une itération pour chaque valeur d'entrée soumise à une variation). Lorsque des itérations sont présentes, le diagramme en bâtons de la modalité prédite d'une cible qualitative est représenté sous forme de diagramme en bâtons regroupés qui comprend les résultats de toutes les itérations. Vous pouvez choisir de regrouper les modalités ou les itérations.

Résultat

Diagrammes tornado. Les diagrammes Tornado sont des diagrammes en bâtons qui représentent les relations entre des cibles et des entrées simulées à l'aide de plusieurs mesures.

- **Corrélation de cible avec entrée.** Cette option crée un diagramme tornado des coefficients de corrélation existant entre une cible donnée et chacune de ses entrées simulées. Ce type de diagramme tornado ne prend pas en charge les cibles ou les entrées simulées dont le niveau de mesure est nominal.
- **Contribution à la variance.** Cette option crée un diagramme tornado qui affiche la contribution à la variance d'une cible provenant de chacune de ses entrées simulées. Cela vous permet d'évaluer le degré de contribution de chaque entrée à l'incertitude globale de la cible. Ce type de diagramme tornado ne prend pas en charge les cibles ou les entrées simulées dont le niveau de mesure est nominal ou ordinal.
- **Sensibilité de la cible à modifier.** Cette option crée un diagramme tornado qui représente les effets sur la cible de la modification de chaque entrée simulée, en ajoutant ou en retirant un nombre spécifique d'écarts-type de la distribution associée à l'entrée. Ce type de diagramme tornado ne prend pas en charge les cibles ou les entrées simulées dont le niveau de mesure est nominal ou ordinal.

Boîtes à moustaches des distributions cible. Les boîtes à moustaches sont disponibles pour les cibles continues. Sélectionnez Superposer les résultats provenant de cibles distinctes si votre modèle prédictif contient plusieurs cibles continues et que vous souhaitez afficher les boîtes à moustaches de toutes les cibles sur un même diagramme.

Diagrammes de dispersion comparant les cibles et les entrées. Les diagrammes de dispersion comparant les cibles et les entrées simulées sont disponibles à la fois pour les cibles continues et les cibles qualitatives, et comprennent les dispersions des cibles ayant à la fois des entrées continues et qualitatives associées. Les appariements impliquant une cible qualitative ou une entrée qualitative sont affichés sous la forme d'une carte de chaleur.

Créer un tableau des valeurs de centiles. Pour les cibles continues, vous pouvez obtenir un tableau des centiles des distributions cible spécifiés. Les quartiles (25e, 50e et 75e centiles) divisent les observations en quatre classes de taille égale. Si vous souhaitez un nombre égal de groupes différent de quatre, sélectionnez Intervalles et indiquez le nombre souhaité. Sélectionnez Centiles personnalisés pour indiquer des centiles personnalisés, par ex. le 99e centile.

Statistiques descriptives des distributions cible. Cette option crée des tableaux de statistiques descriptives pour les cibles continues et qualitatives, ainsi que pour les entrées continues. Pour les cibles continues, le tableau comprend la moyenne, l'écart-type, la médiane, les valeurs minimale et maximale, l'intervalle de confiance de la moyenne au niveau spécifié et les 5e et 95e centiles de la distribution cible. Pour les cibles qualitatives, le tableau comprend le pourcentage d'observations qui se situent dans chaque modalité de la cible. Pour les cibles qualitatives des modèles PMML, le tableau comprend également la probabilité moyenne de chaque modalité de la cible. Pour les entrées continues, le tableau comprend la moyenne, l'écart-type et les valeurs minimale et maximale.

Entrées simulées à inclure dans le résultat. Par défaut, toutes les entrées simulées sont incluses dans les résultats. Vous pouvez exclure certaines entrées simulées des résultats. Dans ce cas, elles seront exclues des diagrammes tornado, des diagrammes de dispersion et des résultats sous forme de tableau.

Formats d'affichage. Vous pouvez définir le format utilisé pour l'affichage des valeurs de cibles et d'entrées (aussi bien les entrées simulées que les entrées fixes).

Enregistrer

Enregistrer le plan de cette simulation. Vous pouvez enregistrer les spécifications de votre simulation dans un fichier de plan de simulation. Les fichiers de plan de simulation possèdent l'extension *.splan*. Vous pouvez rouvrir le plan dans le Générateur de simulation, éventuellement lui apporter des modifications et y exécuter la simulation. Il est possible de partager le plan de simulation avec d'autres utilisateurs, qui peuvent à leur tour exécuter la simulation dans la boîte de dialogue Exécuter la simulation. Les plans de simulation contiennent toutes les spécifications sauf les suivantes : paramètres des fonctions de densité, paramètres de résultats des diagrammes et tableaux, paramètres d'options avancées pour l'ajustement, la distribution empirique et le générateur aléatoire.

Enregistrer les données simulées dans un nouveau fichier de données. Vous pouvez enregistrer les entrées simulées, les entrées fixes et les valeurs cible prédites dans un fichier de données SPSS Statistics, un nouvel ensemble de données de la session en cours ou un fichier de données Excel. Chaque observation (ou ligne) du fichier de données contient les valeurs prédites des cibles ainsi que les entrées simulées et les entrées fixes qui ont généré les valeurs cible. Lorsque l'analyse de sensibilité est spécifiée, chaque itération génère un ensemble d'observations continues étiqueté par le numéro d'itération.

Boîte de dialogue Exécuter la simulation

La boîte de dialogue Exécuter la simulation est destinée aux utilisateurs disposant d'un plan de simulation et qui souhaitent principalement l'exécuter. Elle offre les fonctions nécessaires à l'exécution de la simulation dans différentes conditions. Elle vous permet de réaliser les tâches générales suivantes :

- Configurer ou modifier l'analyse de sensibilité pour vérifier l'effet de la variation de la valeur d'une entrée fixe ou d'un paramètre de distribution d'une entrée simulée.
- Réajuster les distributions de probabilité des entrées incertaines (et les corrélations existant entre ces entrées) aux nouvelles données.
- Modifier la distribution d'une entrée simulée.
- Personnaliser le résultat.
- Exécuter la simulation.

Onglet Simulation.

L'onglet Simulation vous permet de spécifier l'analyse de sensibilité, de réajuster les distributions de probabilité des entrées simulées et des corrélations entre elles aux nouvelles données, et de modifier la distribution de probabilité associée à une entrée simulée particulière.

La grille Entrées simulées contient un ensemble de données pour chaque entrée de votre modèle prédictif. Ces données comprennent le nom de l'entrée et le type de distribution de probabilité associée à l'entrée, ainsi qu'un diagramme de la courbe de distribution associée. Chaque entrée se voit également affecter une icône d'état (un cercle de couleur avec une coche), utile lors du réajustement des distributions aux nouvelles données. En outre, les entrées peuvent présenter une icône en forme de verrou qui indique que l'entrée est verrouillée et ne peut être modifiée ou réajustée aux nouvelles données dans la boîte de dialogue Exécuter la simulation. Pour modifier une entrée verrouillée, vous devez ouvrir le plan de simulation dans le Générateur de simulation.

Chaque entrée peut être soit simulée soit fixe. Les entrées simulées sont celles dont les valeurs sont incertaines et seront générées à partir d'une distribution de probabilité spécifique. Les entrées fixes sont celles dont les valeurs sont connues. Elles restent constantes pour chaque observation générée dans la simulation. Pour utiliser une entrée particulière, sélectionnez l'ensemble de données correspondant dans la grille Entrées simulées.

Spécification de l'analyse de sensibilité

L'analyse de sensibilité vous permet de vérifier l'effet de modifications systématiques apportées à une entrée fixe ou à un paramètre de distribution associé à une entrée simulée, en générant un ensemble indépendant d'observations simulées, —soit une simulation distincte— pour chaque valeur spécifiée. Pour spécifier l'analyse de sensiblité, sélectionnez une entrée simulée ou une entrée fixe et cliquez sur Analyse de sensibilité. L'analyse de sensibilité est limitée à une seule entrée fixe ou un seul paramètre de distribution associé à une entrée simulée. Pour plus d'informations, reportez-vous à la section Analyse de sensibilité sur p. 326.

Réajustement des distributions aux nouvelles données

Pour réajuster automatiquement les distributions de probabilité des entrées simulées (et les corrélations existant entre ces entrées) aux données de l'ensemble de données actif :

- Vérifiez que chaque entrée du modèle est mise en correspondance avec le champ approprié dans l'ensemble de données actif. Chaque entrée simulée est ajustée au champ de l'ensemble de données actif spécifié dans la liste déroulante Champ associée à cette entrée. Il est facile d'identifier les entrées non mises en correspondance, il suffit de rechercher les entrées dont l'icône d'état représente une coche et un point d'interrogation, tel qu'illustré ci-dessous.

- Au besoin, modifiez les champs de mise en correspondance en sélectionnant l'option Ajuster à un champ de l'ensemble de données puis le champ souhaité dans la liste.

▶ Cliquez sur Tout ajuster.

Pour chaque entrée ajustée, la distribution la mieux adaptée s'affiche en même temps qu'un diagramme de la distribution superposé à un histogramme (diagramme en bâtons) représentant les données historiques de l'entrée. S'il est impossible de trouver un ajustement acceptable, la distribution empirique est alors utilisée. Pour les entrées ajustées par distribution empirique, seul un histogramme des données historiques s'affiche car la distribution empirique est en fait représentée par cet histogramme.

Remarque : pour obtenir la liste complète des icônes d'état, reportez-vous à la rubrique Champs simulés sur p. 322.

Modification des distributions de probabilité

Vous pouvez modifier la distribution de probabilité d'une entrée simulée et éventullement changer une entrée simulée en entrée fixe, et vice-versa.

▶ Sélectionnez l'entrée souhaitée et sélectionnez Définir manuellement la distribution.

▶ Sélectionnez le type de distribution souhaité et spécifiez les paramètres de la distribution. Pour changer une entrée simulée en entrée fixe, sélectionnez Fixe dans la liste déroulante Type.

Une fois les paramètres d'une distribution définis, le diagramme représentant la distribution (affiché dans l'ensemble de données de l'entrée dans la grille) sera mis à jour afin de refléter vos modifications. Pour plus d'informations sur la spécification manuelle des distributions de probabilité, consultez la rubrique Champs simulés sur p. 322.

Onglet Résultat

L'onglet Résultat vous permet de personnaliser les résultats générés par la simulation.

Fonctions de densité. Les fonctions de densité sont les principaux moyens permettant de sonder l'ensemble de résultats généré par votre simulation.

- **Fonction de densité de probabilité.** La fonction de densité de probabilité affiche la distribution des valeurs cible, vous permettant de déterminer la probabilité que la cible se situe dans une zone donnée. Pour les cibles dont le résultat est fixe, par ex. "niveau de service faible", "niveau de service correct", "bon niveau de service" et "excellent niveau de service", un diagramme en bâtons est généré, qui affiche le pourcentage d'observations situées dans chaque modalité de la cible.
- **Fonction de distribution cumulée.** La fonction de distribution cumulée affiche la probabilité qu'une valeur de la cible soit inférieure ou égale à une valeur donnée.

Diagrammes tornado. Les diagrammes Tornado sont des diagrammes en bâtons qui représentent les relations entre des cibles et des entrées simulées à l'aide de plusieurs mesures.

- **Corrélation de cible avec entrée.** Cette option crée un diagramme tornado des coefficients de corrélation existant entre une cible donnée et chacune de ses entrées simulées.

- **Contribution à la variance.** Cette option crée un diagramme tornado qui affiche la contribution à la variance d'une cible provenant de chacune de ses entrées simulées. Cela vous permet d'évaluer le degré de contribution de chaque entrée à l'incertitude globale de la cible.
- **Sensibilité de la cible à modifier.** Cette option crée un diagramme tornado qui représente les effets sur la cible de la modification de chaque entrée simulée, en ajoutant ou en retirant un écart-type de la distribution associée à l'entrée.

Diagrammes de dispersion comparant les cibles et les entrées. Cette option génère des diagrammes de dispersion des cibles par rapport aux entrées simulées. Les appariements impliquant une cible présentant un ensemble fixe de résultats (ou une entrée simulée présentant un ensemble fixe de valeurs) sont affichés sous la forme d'une carte de chaleur.

Boîtes à moustaches des distributions cible. Cette option génère des boîtes à moustaches des distributions cible.

Tableau de quartiles. Cette option génère un tableau des quartiles des distributions cible. Les quartiles d'une distribution sont les 25e, 50e et 75e centiles de cette distribution et ils divisent les observations en quatre classes de taille égale.

Superposer les résultats provenant de cibles distinctes. Si le modèle prédictif que vous simulez contient plusieurs cibles, vous pouvez choisir d'afficher ou non les résultats des différentes cibles sur un même diagramme. Ce paramètre s'applique aux diagrammes des fonctions de densité de probabilité, des fonctions de distribution cumulée et des boîtes à moustaches. Par exemple, si vous sélectionnez cette option, les fonctions de densité de probabilité de toutes les cibles s'afficheront sur un même diagramme.

Enregistrer le plan de cette simulation. Vous pouvez enregistrer les modifications apportées à votre simulation dans un fichier de plan de simulation. Les fichiers de plan de simulation possèdent l'extension *.splan*. Vous pouvez rouvrir le plan dans le Générateur de simulation ou dans la boîte de dialogue Exécuter la simulation. Les plans de simulation contiennent toutes les spécifications sauf les paramètres de résultats.

Enregistrer les données simulées dans un nouveau fichier de données. Vous pouvez enregistrer les entrées simulées, les entrées fixes et les valeurs cible prédites dans un fichier de données SPSS Statistics, un nouvel ensemble de données de la session en cours ou un fichier de données Excel. Chaque observation (ou ligne) du fichier de données contient les valeurs prédites des cibles ainsi que les entrées simulées et les entrées fixes qui ont généré les valeurs cible. Lorsque l'analyse de sensibilité est spécifiée, chaque itération génère un ensemble d'observations continues étiqueté par le numéro d'itération.

Si vous avez besoin de personnaliser davantage les résultats, exécutez plutôt votre simulation dans le Générateur de simulation. Pour plus d'informations, reportez-vous à la section Pour exécuter une simulation à partir d'un plan de simulation sur p. 318.

Utilisation du résultat graphique créé par la simulation

Plusieurs diagrammes générés par la simulation offrent des fonctionnalités interactives vous permettant de personnaliser leur affichage. Ces fonctionnalités interactives sont disponibles en activant (par double-clic) l'objet de diagramme dans le Viewer de résultats. Tous les diagrammes de simulation sont des visualisations graphiques.

Diagrammes de fonction de densité de probabilité pour les cibles continues. Ce type de diagramme présente deux lignes de référence verticales pouvant être déplacées par glissement, qui divisent le diagramme en zones distinctes. Le tableau en dessous du diagramme affiche la probabilité que la cible se trouve dans chacune des zones. Si plusieurs fonctions de densité sont affichées sur le même diagramme, le tableau comporte une ligne distincte pour les probabilités associées à chaque fonction de densité. Chacune des lignes de référence présente un curseur (triangle inversé) qui vous permet de déplacer la ligne. D'autres fonctionnalités supplémentaires sont disponibles en cliquant sur le bouton Options de diagramme situé sur le diagramme. Vous pouvez notamment définir explicitement les positions des curseurs, ajouter des lignes de référence fixes et modifier l'affichage du diagramme, pour passer par exemple d'une courbe continue à un histogramme ou vice-versa. Pour plus d'informations, reportez-vous à la section Options de diagramme sur p. 335.

Diagrammes de fonction de distribution cumulée pour les cibles continues. Ce type de diagramme présente les mêmes deux lignes de référence verticales et le même tableau associé décrits ci-dessus pour les diagrammes de fonction de densité de probabilité. Il offre également un accès à la boîte de dialogue Options de diagramme qui vous permet de définir explicitement les positions des curseurs, d'ajouter des lignes de référence fixes et de spécifier si oui ou non la fonction de distribution cumulée est affichée en tant que fonction croissante (par défaut) ou fonction décroissante. Pour plus d'informations, reportez-vous à la section Options de diagramme sur p. 335.

Diagrammes en bâtons pour les cibles qualitatives avec itérations d'analyse de sensibilité. Pour les cibles qualitatives avec itérations d'analyse de sensibilité, les résultats de la modalité cible prédite sont affichés sous la forme d'un diagramme en bâtons regroupés qui comprend les résultats de toutes les itérations. Le diagramme comprend une liste déroulante qui vous permet de regrouper par modalité ou par itération. Pour les modèles de classification en deux étapes et les modèles de classification en nuées dynamiques, il est possible de regrouper par classe ou par itération.

Boîtes à moustaches pour les cibles multiples avec itérations d'analyse de sensibilité. S'applique aux modèles prédictifs avec plusieurs cibles continues et des itérations d'analyse de sensibilité. Permet d'afficher des boîtes à moustaches pour toutes les cibles sur un même diagramme unique, sous la forme d'une boîte à moustache juxtaposée. Le diagramme comprend une liste déroulante qui vous permet de regrouper par cible ou par itération.

Options de diagramme

La boîte de dialogue Options de diagramme vous permet de personnaliser l'affichage des diagrammes de fonctions de densité de probabilité et de fonctions de distribution cumulée actifs, générés par une simulation.

Affichage. La liste déroulante Affichage s'applique uniquement au diagramme de fonction de densité de probabilité. Elle vous permet de faire basculer l'affichage d'une courbe continue à un histogramme. Cette fonctionnalité n'est pas disponible lorsque plusieurs fonctions de densité sont affichées sur le même diagramme. Dans un tel cas, les fonctions de densité peuvent être affichées uniquement sous la forme de courbes continues.

Ordre. La liste déroulante Ordre s'applique uniquement au diagramme de fonction de distribution cumulée. Elle indique si la fonction de distribution cumulée s'affiche sous la forme d'une fonction croissante (par défaut) ou d'une fonction décroissante. Si elle s'affiche sous la forme d'une fonction décroissante, la valeur de la fonction à un point donné sur l'axe horizontal est la probabilité que la cible se situe à la droite de ce point.

Positions des curseurs. Vous pouvez définir explicitement les positions des lignes de référence ajustables en entrant des valeurs dans les zones de texte Supérieur et Inférieur. Vous pouvez supprimer la ligne de gauche en sélectionnant Infini -, ce qui a pour effet de définir la position du curseur sur l'infini négatif, et vous pouvez supprimer la ligne de droite en sélectionnant Infini, ce qui défini sa position sur l'infini.

Lignes de référence. Vous pouvez ajouter différentes lignes de référence verticales fixes aux fonctions de densité de probabilité et aux fonctions de distribution cumulée.

- **Sigmas.** Des lignes de références peuvent être ajoutées à plus et moins un nombre spécifique d'écarts-type de la moyenne de la cible.
- **Centiles :** Des lignes de références peuvent être ajoutées à un ou deux centiles de la distribution pour chaque cible, en entrant des valeurs dans les zones de texte Inférieure et Supérieure. Par exemple, une valeur de 95 dans la zone de texte Supérieure représente le 95e centile, qui est la valeur en dessous de laquelle se trouvent 95 % des observations. De même, une valeur de 5 dans la zone de texte Inférieure représente le 5e centile, qui est la valeur en dessous de laquelle se trouvent 5% des observations.
- **Positions personnalisées.** Des lignes de références peuvent être ajoutées à des valeurs spécifiques le long de l'axe horizontal.

Les lignes de référence peuvent être supprimées en décochant l'option associée dans la boîte de dialogue Options de diagramme avant de cliquer sur Poursuivre.

Remarque : lorsque plusieurs fonctions de densité ou de distribution sont affichées sur un diagramme unique (en raison de cibles ou résultats multiples provenant des itérations de l'analyse de sensibilité), des lignes de référence (autres que les lignes personnalisées) sont appliquées respectivement à chaque fonction.

Annexe

A

Remarques

Ces informations ont été développées pour les produits et services offerts dans le monde.

Il est possible qu'IBM n'offre pas dans les autres pays les produits, services et fonctionnalités décrits dans ce document. Contactez votre représentant local IBM pour obtenir des informations sur les produits et services actuellement disponibles dans votre région. Toute référence à un produit, programme ou service IBM n'implique pas que les seuls les produits, programmes ou services IBM peuvent être utilisés. Tout produit, programme ou service de fonctionnalité équivalente qui ne viole pas la propriété intellectuelle IBM peut être utilisé à la place. Cependant l'utilisateur doit évaluer et vérifier l'utilisation d'un produit, programme ou service non IBM.

IBM peut posséder des brevets ou des applications de brevet en attente qui couvrent les sujets décrits dans ce document. L'octroi de ce document n'équivaut aucunement à celui d'une licence pour ces brevets. Vous pouvez envoyer par écrit des questions concernant la licence à :

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, États-Unis

Pour obtenir des informations de licence concernant la configuration de caractères codés sur deux octets (DBCS), veuillez contacter dans votre pays le département chargé de la propriété intellectuelle chez IBM ou envoyez vos commentaires par écrit à :

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japon.

Le paragraphe suivant ne s'applique pas au Royaume-Uni ni à aucun pays dans lequel ces dispositions sont contraires au droit local : INTERNATIONAL BUSINESS MACHINES FOURNIT CETTE PUBLICATION « EN L'ETAT » SANS GARANTIE D'AUCUNE SORTE, IMPLICITE OU EXPLICITE, Y COMPRIS, MAIS SANS ETRE LIMITE AUX GARANTIES IMPLICITES DE NON VIOLATION, DE QUALITE MARCHANDE OU D'ADAPTATION POUR UN USAGE PARTICULIER. Certains états n'autorisent pas l'exclusion de garanties explicites ou implicites lors de certaines transactions, par conséquent, il est possible que cet énoncé ne vous concerne pas.

Ces informations peuvent contenir des erreurs techniques ou des erreurs typographiques. Ces informations sont modifiées de temps en temps ; ces modifications seront intégrées aux nouvelles versions de la publication. IBM peut apporter des améliorations et/ou modifications des produits et/ou des programmes décrits dans cette publications à tout moment sans avertissement préalable.

Toute référence dans ces informations à des sites Web autres qu'IBM est fournie dans un but pratique uniquement et ne sert en aucun cas de recommandation pour ces sites Web. Le matériel contenu sur ces sites Web ne fait pas partie du matériel de ce produit IBM et l'utilisation de ces sites Web se fait à vos propres risques.

IBM peut utiliser ou distribuer les informations que vous lui fournissez, de la façon dont il le souhaite, sans encourir aucune obligation envers vous.

Les personnes disposant d'une licence pour ce programme et qui souhaitent obtenir des informations sur celui-ci pour activer : (i) l'échange d'informations entre des programmes créés de manière indépendante et d'autres programmes (notamment celui-ci) et (ii) l'utilisation mutuelle des informations qui ont été échangées, doivent contacter :

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, États-Unis.

Ces informations peuvent être disponibles, soumises à des conditions générales, et dans certains cas payantes.

Le programme sous licence décrit dans ce document et toute la documentation sous licence disponible pour ce programme sont fournis par IBM en conformité avec les conditions de l'accord du client IBM, avec l'accord de licence du programme international IBM et avec tout accord équivalent entre nous.

les informations concernant les produits autres qu'IBM ont été obtenues auprès des fabricants de ces produits, leurs annonces publiques ou d'autres sources publiques disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances, leur compatibilité ou toute autre fonctionnalité associée à des produits autres qu'IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits.

Ces informations contiennent des exemples de données et de rapports utilisés au cours d'opérations quotidiennes standard. Pour les illustrer le mieux possible, ces exemples contiennent des noms d'individus, d'entreprises, de marques et de produits. Tous ces noms sont fictifs et toute ressemblance avec des noms et des adresses utilisés par une entreprise réelle ne serait que pure coïncidence.

Si vous consultez la version papier de ces informations, il est possible que certaines photographies et illustrations en couleurs n'apparaissent pas.

Marques commerciales

IBM, le logo IBM, ibm.com et SPSS sont des marques commerciales d'IBM Corporation, déposées dans de nombreuses juridictions du monde entier. Une liste à jour des marques IBM est disponible sur Internet à l'adresse *http://www.ibm.com/legal/copytrade.shtml*.

Adobe, le logo Adobe, PostScript et le logo PostScript sont des marques déposées ou des marques commerciales de Adobe Systems Incorporated aux Etats-Unis et/ou dans d'autres pays.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, et Pentium sont des marques commerciales ou des marques déposées de Intel Corporation ou de ses filiales aux Etats-Unis et dans d'autres pays.

Java et toutes les marques et logos Java sont des marques commerciales de Sun Microsystems, Inc. aux Etats-Unis et/ou dans d'autres pays.

Linux est une marque déposée de Linus Torvalds aux Etats-Unis et/ou dans d'autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques commerciales de Microsoft Corporation aux Etats-Unis et/ou dans d'autres pays.

UNIX est une marque déposée de The Open Group aux Etats-Unis et dans d'autres pays.

Ce produit utilise WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com/*.

Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés.

Les captures d'écran des produits Adobe sont reproduites avec l'autorisation de Adobe Systems Incorporated.

Les captures d'écran des produits Microsoft sont reproduites avec l'autorisation de Microsoft Corporation.

Index

§2S R^2 de Cox et Snell
  Régression ordinale, 115

§AB R^2 de McFadden
  Régression ordinale, 115
§AB R^2 de Nagelkerke
  Régression ordinale, 115
Affectation de mémoire
  Dans l'analyse TwoStep Cluster, 173
ajustement automatique de la distribution
  dans la simulation, 322
Ajustement de fonctions, 120
  Analyse de la variance, 120
  Enregistrement de prévisions, 122
  Enregistrement de résidus, 122
  Enregistrement d'intervalles de prévision, 122
  Inclusion de la constante, 120
  Modèles, 122
  Prévision, 122
ajustement de la distribution
  dans la simulation, 322
Alpha de Cronbach
  Dans l'analyse de fiabilité, 300–301
Alpha-maximisation, 164
Analyse de fiabilité, 300
  Coefficient de corrélation intra-classe, 301
  Corrélations et covariances entre éléments, 301
  descriptives, 301
  exemple, 300
  Fonctionnalités supplémentaires, 303
  Kuder-Richardson 20, 301
  statistiques, 300–301
  T^2 de Hotelling , 301
  Tableau ANOVA, 301
  Test d'additivité de Tukey, 301
Analyse de la variance
  Ajustement de fonctions, 120
  Dans ANOVA à 1 facteur, 54
  Dans la régression linéaire, 109
  Dans Moyennes, 38
analyse de sensibilité
  dans la simulation, 326
Analyse de séries chronologiques
  Prévision, 122
  Prévision d'observations, 122
analyse d'hypothèses
  dans la simulation, 326
Analyse discriminante, 153
  Coefficients de la fonction, 156
  Critères, 157
  Définition d'intervalles, 155
  Diagrammes, 158
  Distance de Mahalanobis, 157
  Enregistrement des variables du classement, 160
  exemple, 153
  Exportation des informations du modèle, 160
  Fonctionnalités supplémentaires, 160
  Lambda de Wilks, 157
  Matrice de covariance, 158
  Matrices, 156
  Méthodes de l'analyse discriminante, 157
  Méthodes pas à pas, 153
  Options d'affichage, 157–158
  Probabilités a priori, 158
  Sélection d'observations, 155
  statistiques, 153, 156
  Statistiques descriptives, 156
  V de Rao , 157
  Valeurs manquantes, 158
  Variables de regroupement, 153
  Variables indépendantes, 153
Analyse du voisin le plus proche, 129
  enregistrement de variables, 138
  Options, 140
  partitions, 136
  Résultats, 139
  sélection des descriptives, 135
  voisins, 133
  vue du modèle, 141
Analyse en composantes principale, 161, 164
Analyse factorielle, 161
  Aperçu, 161
  Cartes factorielles, 166
  Convergence, 164, 166
  descriptives, 163
  exemple, 161
  Facteurs, 167
  Fonctionnalités supplémentaires, 168
  Format d'affichage des projections, 168
  Méthodes de rotation, 166
  Méthodes d'extraction, 164
  Sélection d'observations, 162
  statistiques, 161, 163
  Valeurs manquantes, 168
Analyse multiréponses
  Tableau croisé, 283
  Tableaux croisés des réponses multiples, 283
  tableaux de fréquences, 281
  Tableaux de fréquences des réponses multiples, 281
Analyse TwoStep Cluster, 170
  Enregistrer dans le fichier de travail, 175
  Enregistrer dans le fichier externe, 175
  Options, 173
  statistiques, 175
Andrew
  Dans Explorer, 18
ANOVA
  Dans ANOVA à 1 facteur, 54
  dans des modèles linéaires, 97
  dans GLM - Univarié, 60
  Dans Moyennes, 38
  Modèle, 62

ANOVA à 1 facteur, 54
  Comparaisons multiples, 56
  Contrastes, 55
  Contrastes polynomiaux, 55
  Fonctionnalités supplémentaires, 59
  Options, 58
  statistiques, 58
  Tests post hoc, 56
  Valeurs manquantes, 58
  Variables actives, 54
Aplatissement
  Dans Explorer, 18
  Dans les cubes OLAP, 43
  dans les Descriptives, 14
  Dans les effectifs, 9
  Dans les tableaux de bord en colonnes, 295
  Dans les tableaux de bord en lignes, 289
  Dans Moyennes, 38
  Dans Récapituler, 33
Arbres hiérarchiques
  Classification hiérarchique, 192
Association linéaire par linéaire
  Tableaux croisés, 25
Asymétrie
  Dans Explorer, 18
  Dans les cubes OLAP, 43
  dans les Descriptives, 14
  Dans les effectifs, 9
  Dans les tableaux de bord en colonnes, 295
  Dans les tableaux de bord en lignes, 289
  Dans Moyennes, 38
  Dans Récapituler, 33

B de Tukey
  Dans ANOVA à 1 facteur, 56
  GLM, 66
bagging
  dans des modèles linéaires, 85
boîtes à moustaches
  Comparaison des niveaux de facteur, 19
  Comparaison des variables, 19
  Dans Explorer, 19
  dans la simulation, 330
Bonferroni
  Dans ANOVA à 1 facteur, 56
  GLM, 66
boosting
  dans des modèles linéaires, 85

C de Dunnett
  Dans ANOVA à 1 facteur, 56
  GLM, 66
Carré de la distance euclidienne
  Distances, 80
carte des quadrants
  dans l'analyse du voisin le plus proche, 148
Cartes factorielles
  Dans Analyse factorielle, 166
centiles
  dans la simulation, 330
Centiles
  Dans Explorer, 18
  Dans les effectifs, 9
Classification, 176
  affichage de classes, 177
  affichage général, 177
  Choix d'une procédure, 169
  Classification hiérarchique, 189
  Dans Courbe ROC, 313
  Efficacité, 196
  Nuées dynamiques, 194
Classification hiérarchique, 189
  Arbres hiérarchiques, 192
  Chaîne des agrégations, 191
  Classes d'affectation, 191, 193
  Classification de variables, 189
  Classification d'observations, 189
  Diagrammes en stalactite, 192
  Enregistrement de nouvelles variables, 193
  exemple, 189
  Fonctionnalités supplémentaires, 193
  Matrices de distance, 191
  Mesures de distance, 190
  Mesures de similarité, 190
  Méthodes de classification, 190
  Orientation du diagramme, 192
  statistiques, 189, 191
  Transformation de mesures, 190
  Transformation de valeurs, 190
Coefficient alpha
  Dans l'analyse de fiabilité, 300–301
Coefficient de concordance de Kendall (W)
  Tests non paramétriques pour échantillons liés, 214
Coefficient de contingence
  Tableaux croisés, 25
Coefficient de corrélation de Spearman
  Corrélations bivariées, 73
  Tableaux croisés, 25
Coefficient de corrélation intra-classe
  Dans l'analyse de fiabilité, 301
Coefficient de corrélation par rang
  Corrélations bivariées, 73
Coefficient de corrélation *r*
  Corrélations bivariées, 73
  Tableaux croisés, 25
Coefficient de dispersion (COD)
  Dans les statistiques de ratio, 311
Coefficient de variation (COV)
  Dans les statistiques de ratio, 311
Coefficient d'incertitude
  Tableaux croisés, 25
Coefficients bêta
  Dans la régression linéaire, 109

Coefficients de régression
  Dans la régression linéaire, 109
Colonne de total
  Dans les tableaux de bord, 296
Comparaison des groupes
  Dans les cubes OLAP, 45
Comparaison des variables
  Dans les cubes OLAP, 45
Comparaisons multiples
  Dans ANOVA à 1 facteur, 56
Comparaisons multiples post hoc, 56
comparaisons par paire
  tests non paramétriques, 242
Contrastes
  Dans ANOVA à 1 facteur, 55
  GLM, 64
Contrastes à la précédente
  GLM, 64
Contrastes de Helmert
  GLM, 64
Contrastes déviation
  GLM, 64
Contrastes différence
  GLM, 64
Contrastes polynomiaux
  Dans ANOVA à 1 facteur, 55
  GLM, 64
Contrastes simples
  GLM, 64
Convergence
  Analyse de classification des nuées dynamiques, 196
  Dans Analyse factorielle, 164, 166
Correction pour la continuité de Yates
  Tableaux croisés, 25
Corrélation de Pearson
  Corrélations bivariées, 73
  Tableaux croisés, 25
corrélations
  dans la simulation, 326
Corrélations
  Corrélations bivariées, 73
  Dans Corrélations partielles, 76
  Ordre zéro, 77
  Tableaux croisés, 25
Corrélations bivariées
  Coefficients de corrélation, 73
  Fonctionnalités supplémentaires, 75
  Options, 75
  Seuil de signification, 73
  statistiques, 75
  Valeurs manquantes, 75
Corrélations partielles, 76
  Corrélations simples, 77
  Dans la régression linéaire, 109
  Fonctionnalités supplémentaires, 78
  Options, 77
  statistiques, 77
  Valeurs manquantes, 77
Corrélations simples
  Dans Corrélations partielles, 77
Courbe ROC, 313
  Diagrammes et statistiques, 315
critère de prévention du surajustement
  dans des modèles linéaires, 87
Critère d'information d'Akaike
  dans des modèles linéaires, 87
critères d'informations
  dans des modèles linéaires, 87
Cubes OLAP, 41
  statistiques, 43
  titres, 46

D
  Tableaux croisés, 25
D de Somers
  Tableaux croisés, 25
Décomposition hiérarchique, 63
Définir les vecteurs multiréponses, 280
  Définition des étiquettes, 280
  Définition des noms, 280
  Dichotomies, 280
  modalités, 280
Dernier
  Dans les cubes OLAP, 43
  Dans Moyennes, 38
  Dans Récapituler, 33
Descriptives, 13
  Enregistrement des écarts *z*, 13
  Fonctionnalités supplémentaires, 16
  Ordre d'affichage, 14
  statistiques, 14
Diagnostic des observations
  Dans la régression linéaire, 109
diagramme de dispersion
  dans la simulation, 330
diagramme de l'espace des descriptives
  dans l'analyse du voisin le plus proche, 142
Diagramme de répartition gaussien
  Dans Explorer, 19
  Dans la régression linéaire, 105
Diagramme de répartition hors tendance
  Dans Explorer, 19
diagrammes
  Dans Courbe ROC, 313
  Etiquettes d'observations, 120
Diagrammes de dispersion
  Dans la régression linéaire, 105
Diagrammes de profils
  GLM, 65
Diagrammes des résidus
  dans GLM - Univarié, 70
Diagrammes dispersion/niveau
  Dans Explorer, 19
  dans GLM - Univarié, 70

Diagrammes en bâtons
  Dans les effectifs, 11
diagrammes en secteurs
  Dans les effectifs, 11
Diagrammes en stalactite
  Classification hiérarchique, 192
Diagrammes partiels
  Dans la régression linéaire, 105
Diagrammes tige et feuille
  Dans Explorer, 19
diagrammes tornado
  dans la simulation, 330
Dictionnaire
  Livre de codes, 1
Différence de bêta
  Dans la régression linéaire, 107
Différence de prévision
  Dans la régression linéaire, 107
Différence la moins significative
  Dans ANOVA à 1 facteur, 56
  GLM, 66
Différence significative de Tukey
  Dans ANOVA à 1 facteur, 56
  GLM, 66
Différences entre les groupes
  Dans les cubes OLAP, 45
Différences entre les variables
  Dans les cubes OLAP, 45
Différentiel lié au prix (PRD)
  Dans les statistiques de ratio, 311
Distance de Cook
  Dans la régression linéaire, 107
  GLM, 68
Distance de Mahalanobis
  Analyse discriminante, 157
  Dans la régression linéaire, 107
Distance de Manhattan
  dans l'analyse du voisin le plus proche, 133
  Distances, 80
Distance de Minkowski
  Distances, 80
Distance de Tchebycheff
  Distances, 80
Distance du Khi-deux
  Distances, 80
Distance euclidienne
  dans l'analyse du voisin le plus proche, 133
  Distances, 80
distance Manhattan
  dans l'analyse du voisin le plus proche, 133
Distances, 79
  Calcul des distances existant entre des observations, 79
  Calcul des distances existant entre des variables, 79
  exemple, 79
  Fonctionnalités supplémentaires, 82
  Mesures de dissimilarité, 80
  Mesures de similarité, 81
  statistiques, 79
  Transformation de mesures, 80–81
  Transformation de valeurs, 80–81
distances du voisin le plus proche
  dans l'analyse du voisin le plus proche, 147
Division
  Division dans les colonnes de tableaux, 296

Ecart absolu moyen (AAD)
  Dans les statistiques de ratio, 311
Ecart *z*
  dans les Descriptives, 13
  Enregistrement sous forme de variables, 13
Ecart-type
  Dans Explorer, 18
  dans GLM - Univarié, 70
  Dans les cubes OLAP, 43
  dans les Descriptives, 14
  Dans les effectifs, 9
  Dans les statistiques de ratio, 311
  Dans les tableaux de bord en colonnes, 295
  Dans les tableaux de bord en lignes, 289
  Dans Moyennes, 38
  Dans Récapituler, 33
échantillon d'apprentissage
  dans l'analyse du voisin le plus proche, 136
échantillon traité
  dans l'analyse du voisin le plus proche, 136
Echantillons liés, 271, 276
Echelle
  Dans l'analyse de fiabilité, 300
  Dans le positionnement multidimensionnel, 304
Effectif observé
  Tableaux croisés, 28
Effectif théorique
  Tableaux croisés, 28
Effectifs, 8
  diagrammes, 11
  Formats, 12
  Ordre d'affichage, 12
  statistiques, 9
  Suppression de tableaux, 12
Elimination descendante
  Dans la régression linéaire, 104
ensembles
  dans des modèles linéaires, 89
erreur standard
  Dans Courbe ROC, 315
Erreur standard
  Dans Explorer, 18
  dans les Descriptives, 14
  Dans les effectifs, 9
  GLM, 68, 70
Erreur standard d'aplatissement
  Dans les cubes OLAP, 43
  Dans Moyennes, 38
  Dans Récapituler, 33

Erreur standard d'asymétrie
Dans les cubes OLAP, 43
Dans Moyennes, 38
Dans Récapituler, 33
Erreur standard de la moyenne
Dans les cubes OLAP, 43
Dans Moyennes, 38
Dans Récapituler, 33
Estimation de l'intensité des effets
dans GLM - Univarié, 70
Estimations de Hodges-Lehman
Tests non paramétriques pour échantillons liés, 214
Estimations de puissance
dans GLM - Univarié, 70
Estimations des paramètres
dans GLM - Univarié, 70
Régression ordinale, 115
êta
Dans Moyennes, 38
Tableaux croisés, 25
Eta carré
dans GLM - Univarié, 70
Dans Moyennes, 38
Etude appariée
Test T pour échantillons appariés, 50
Etude de contrôle d'observation
Test T pour échantillons appariés, 50
Explorer, 17
Diagrammes, 19
Fonctionnalités supplémentaires, 21
Options, 20
statistiques, 18
Transformations de l'exposant, 20
Valeurs manquantes, 20

F de R-E-G-W (Ryan-Einot-Gabriel-Welsch)
Dans ANOVA à 1 facteur, 56
GLM, 66
F multiple de Ryan-Einot-Gabriel-Welsch
Dans ANOVA à 1 facteur, 56
GLM, 66
Facteur d'inflation de la variance
Dans la régression linéaire, 109
Facteurs, 167
Facteurs d'Anderson-Rubin, 167
Facteurs de Bartlett, 167
Factorisation en axes principaux, 164
Factorisation en projections, 164
Fiabilité de Spearman-Brown
Dans l'analyse de fiabilité, 301
Fiabilité Split-half
Dans l'analyse de fiabilité, 300–301
fonctions de densité de probabilité
dans la simulation, 328
fonctions de distribution cumulée
dans la simulation, 328
Formatage
Colonnes dans les tableaux de bord, 289
Fréquences cumulées
Régression ordinale, 115
Fréquences de classe
Dans l'analyse TwoStep Cluster, 175
Fréquences observées
Régression ordinale, 115
Fréquences théoriques
Régression ordinale, 115

gamma
Tableaux croisés, 25
Gamma de Goodman et Kruskal
Tableaux croisés, 25
Générateur de simulation, 319
Gestion des pages
Dans les tableaux de bord en colonnes, 297
Dans les tableaux de bord en lignes, 291
Gestion du bruit
Dans l'analyse TwoStep Cluster, 173
GLM
Diagrammes de profils, 65
Enregistrement de matrices, 68
Enregistrement de variables, 68
Modèle, 62
Somme des carrés, 62
Tests post hoc, 66
GLM - Univarié, 60, 71
Affichage, 70
Contrastes, 64
Diagnostics, 70
Moyennes marginales estimées, 70
Options, 70
GT2 de Hochberg
Dans ANOVA à 1 facteur, 56
GLM, 66

H de Kruskal-Wallis
Tests pour deux échantillons indépendants, 273
Histogrammes
Dans Explorer, 19
Dans la régression linéaire, 105
Dans les effectifs, 11
Historique des itérations
Régression ordinale, 115

ICC. *Voir* Coefficient de corrélation intra-classe, 301
importance des valeurs prédites
modèles linéaires, 93
importance des variables
dans l'analyse du voisin le plus proche, 146
Index de concentration
Dans les statistiques de ratio, 311

informations sur les champs continus
tests non paramétriques, 241
informations sur les champs qualitatifs
tests non paramétriques, 240
Intervalle multiple de Ryan-Einot-Gabriel-Welsch
Dans ANOVA à 1 facteur, 56
GLM, 66
Intervalles de Clopper-Pearson
Tests non paramétriques à un échantillon, 202
intervalles de confiance
Dans Courbe ROC, 315
Intervalles de confiance
Dans ANOVA à 1 facteur, 58
Dans Explorer, 18
Dans la régression linéaire, 109
Enregistrement dans la régression linéaire, 107
GLM, 64, 70
Test T pour échantillon unique, 53
Test T pour échantillons appariés, 51
Test T pour échantillons indépendants, 49
Intervalles de Jeffreys
Tests non paramétriques à un échantillon, 202
Intervalles de prévision
Enregistrement dans la régression linéaire, 107
Enregistrement dans l'ajustement de fonctions, 122
intervalles du rapport de vraisemblance
Tests non paramétriques à un échantillon, 202
Itérations
Analyse de classification des nuées dynamiques, 196
Dans Analyse factorielle, 164, 166

Kappa
Tableaux croisés, 25
Kappa de Cohen
Tableaux croisés, 25
Khi-deux, 244
Association linéaire par linéaire, 25
Correction pour la continuité de Yates, 25
Indépendance, 25
Intervalle théorique, 246
Options, 246
Pearson, 25
Rapport de vraisemblance, 25
statistiques, 246
Tableaux croisés, 25
Test à un échantillon, 244
Test exact de Fisher, 25
Valeurs manquantes, 246
Valeurs théoriques, 246
Khi-deux de Pearson
Régression ordinale, 115
Tableaux croisés, 25
Khi-deux du rapport de vraisemblance
Régression ordinale, 115
Tableaux croisés, 25
KR20
Dans l'analyse de fiabilité, 301
Kuder-Richardson 20 (KR20)
Dans l'analyse de fiabilité, 301

Lambda
Tableaux croisés, 25
Lambda de Goodman et Kruskal
Tableaux croisés, 25
Lambda de Wilks
Analyse discriminante, 157
Lien
Régression ordinale, 114
Liste des observations, 31
Livre de codes, 1
Résultats, 3
statistiques, 5
LSD de Fisher
GLM, 66

M-estimateur de Huber
Dans Explorer, 18
M-estimateur redescendant de Hampel
Dans Explorer, 18
M-estimateurs
Dans Explorer, 18
marques commerciales, 338
Matrice de corrélation
Analyse discriminante, 156
Dans Analyse factorielle, 161, 163
Régression ordinale, 115
Matrice de covariance
Analyse discriminante, 156, 158
Dans la régression linéaire, 109
GLM, 68
Régression ordinale, 115
Matrice de transformation
Dans Analyse factorielle, 161
Matrice des projections factorielles
Dans Analyse factorielle, 161
Maximum
Comparaison des colonnes de tableaux, 296
Dans Explorer, 18
Dans les cubes OLAP, 43
dans les Descriptives, 14
Dans les effectifs, 9
Dans les statistiques de ratio, 311
Dans Moyennes, 38
Dans Récapituler, 33
Maximum de vraisemblance
Dans Analyse factorielle, 164
médiane
Dans Explorer, 18
Dans les cubes OLAP, 43
Dans les effectifs, 9
Dans les statistiques de ratio, 311
Dans Moyennes, 38
Dans Récapituler, 33

Médiane de groupes
  Dans les cubes OLAP, 43
  Dans Moyennes, 38
  Dans Récapituler, 33
meilleurs sous-ensembles
  dans des modèles linéaires, 87
mentions légales, 337
Mesure de dissimilarité de Lance et Williams, 80
  Distances, 80
Mesure de distance du Phi-deux
  Distances, 80
Mesure d'écart de la taille
  Distances, 80
Mesure d'écart de structures
  Distances, 80
Mesures de distance
  Classification hiérarchique, 190
  dans l'analyse du voisin le plus proche, 133
  Distances, 80
Mesures de la dispersion
  Dans Explorer, 18
  dans les Descriptives, 14
  Dans les effectifs, 9
  Dans les statistiques de ratio, 311
Mesures de la distribution
  dans les Descriptives, 14
  Dans les effectifs, 9
Mesures de la tendance centrale
  Dans Explorer, 18
  Dans les effectifs, 9
  Dans les statistiques de ratio, 311
Mesures de similarité
  Classification hiérarchique, 190
  Distances, 81
Minimum
  Comparaison des colonnes de tableaux, 296
  Dans Explorer, 18
  Dans les cubes OLAP, 43
  dans les Descriptives, 14
  Dans les effectifs, 9
  Dans les statistiques de ratio, 311
  Dans Moyennes, 38
  Dans Récapituler, 33
Modalité de référence
  GLM, 64
Mode
  Dans les effectifs, 9
Modèle composé
  Ajustement de fonctions, 122
Modèle cubique
  Ajustement de fonctions, 122
Modèle de croissance
  Ajustement de fonctions, 122
Modèle de Guttman
  Dans l'analyse de fiabilité, 300–301
Modèle de puissance
  Ajustement de fonctions, 122
Modèle d'échelle
  Régression ordinale, 118
Modèle d'emplacement
  Régression ordinale, 117
Modèle en S
  Ajustement de fonctions, 122
Modèle exponentiel
  Ajustement de fonctions, 122
Modèle inverse
  Ajustement de fonctions, 122
Modèle linéaire
  Ajustement de fonctions, 122
Modèle logarithmique
  Ajustement de fonctions, 122
Modèle logistique
  Ajustement de fonctions, 122
Modèle parallèle
  Dans l'analyse de fiabilité, 300–301
Modèle parallèle strict
  Dans l'analyse de fiabilité, 300–301
Modèle quadratique
  Ajustement de fonctions, 122
Modèles factoriels complets
  GLM, 62
modèles linéaires, 83
  choix du modèle, 87
  Coefficients, 98
  critère d'informations, 91
  duplication des résultats, 90
  ensembles, 89
  importance des valeurs prédites, 93
  moyennes estimées, 100
  niveau de confiance, 86
  objectifs, 85
  options de modèle, 90
  préparation automatique des données, 86, 92
  récapitulatif de création de modèle, 101
  récapitulatif du modèle, 91
  règles de combinaison, 89
  Résidus, 95
  Statistique R-deux, 91
  Tableau ANOVA, 97
  valeurs éloignées, 96
  valeurs prédites en fonction des valeurs observées, 94
Modèles personnalisés
  GLM, 62
Moindres carrés généralisés
  Dans Analyse factorielle, 164
Moindres carrés non pondérés
  Dans Analyse factorielle, 164
moindres carrés pondérés
  Dans la régression linéaire, 102
Moyenne
  Dans ANOVA à 1 facteur, 58
  Dans Explorer, 18
  Dans les cubes OLAP, 43
  dans les Descriptives, 14

Dans les effectifs, 9
Dans les statistiques de ratio, 311
Dans les tableaux de bord en colonnes, 295
Dans les tableaux de bord en lignes, 289
Dans Moyennes, 38
Dans Récapituler, 33
Des colonnes de tableaux multiples, 296
Sous-groupe, 36, 41
Moyenne géométrique
Dans les cubes OLAP, 43
Dans Moyennes, 38
Dans Récapituler, 33
Moyenne harmonique
Dans les cubes OLAP, 43
Dans Moyennes, 38
Dans Récapituler, 33
Moyenne pondérée
Dans les statistiques de ratio, 311
Moyenne tronquée
Dans Explorer, 18
Moyennes, 36
Options, 38
statistiques, 38
Moyennes des groupes, 36, 41
Moyennes des sous-groupes, 36, 41
Moyennes marginales estimées
dans GLM - Univarié, 70
Moyennes observées
dans GLM - Univarié, 70
Multiplication
Multiplication dans les colonnes de tableaux, 296

Newman-Keuls
GLM, 66
Nombre d'observations
Dans les cubes OLAP, 43
Dans Moyennes, 38
Dans Récapituler, 33
Nombre maximal de branches
Dans l'analyse TwoStep Cluster, 173
Nuées dynamiques
Aperçu, 194
Classes d'affectation, 197
Critères de convergence, 196
Distances entre les classes, 197
Efficacité, 196
Enregistrement des informations sur les classes, 197
Exemples, 194
Fonctionnalités supplémentaires, 198
Itérations, 196
Méthodes, 194
statistiques, 194, 197
Valeurs manquantes, 197
Numérotation des pages
Dans les tableaux de bord en colonnes, 298
Dans les tableaux de bord en lignes, 291

pairs
dans l'analyse du voisin le plus proche, 147
pas à pas ascendant
dans des modèles linéaires, 87
Phi
Tableaux croisés, 25
plage
Dans les cubes OLAP, 43
dans les Descriptives, 14
Dans les effectifs, 9
Dans les statistiques de ratio, 311
Dans Moyennes, 38
Dans Récapituler, 33
Plum
Régression ordinale, 113
Positionnement multidimensionnel, 304
Conditionnalité, 307
Création de matrices de distance, 306
Critères, 308
Définition de la forme des données, 306
Dimensions, 307
exemple, 304
Fonctionnalités supplémentaires, 309
Mesures de distance, 306
Modèles de positionnement, 307
Niveaux de mesure, 307
Options d'affichage, 308
statistiques, 304
Transformation de valeurs, 306
pourcentages
Tableaux croisés, 28
Pourcentages en colonne
Tableaux croisés, 28
Pourcentages en ligne
Tableaux croisés, 28
Pourcentages totaux
Tableaux croisés, 28
Premier
Dans les cubes OLAP, 43
Dans Moyennes, 38
Dans Récapituler, 33
préparation automatique des données
dans des modèles linéaires, 92
Prévision
Ajustement de fonctions, 122
Prévisions
Enregistrement dans la régression linéaire, 107
Enregistrement dans l'ajustement de fonctions, 122
Prévisions pondérées
GLM, 68
Profondeur d'arborescence
Dans l'analyse TwoStep Cluster, 173
Proximités
Classification hiérarchique, 189

Q de Cochran
Dans les tests pour plusieurs échantillons liés, 277

Q de R-E-G-W (Ryan-Einot-Gabriel-Welsch)
Dans ANOVA à 1 facteur, 56
GLM, 66
Qualité de l'ajustement
Régression ordinale, 115
Quartiles
Dans les effectifs, 9

R multiple
Dans la régression linéaire, 109
R-deux
dans des modèles linéaires, 91
R-deux ajusté
dans des modèles linéaires, 87
R^2
Dans la régression linéaire, 109
Dans Moyennes, 38
modification R^2, 109
R^2 ajusté
Dans la régression linéaire, 109
Rapport de covariance
Dans la régression linéaire, 107
récapitulatif de l'intervalle de confiance
tests non paramétriques, 220–221, 225
récapitulatif d'erreur
dans l'analyse du voisin le plus proche, 152
récapitulatif d'hypothèses
Tests non paramétriques, 219
Récapituler, 31
Options, 33
statistiques, 33
règles de combinaison
dans des modèles linéaires, 89
Régression
Diagrammes, 105
Régression linéaire, 102
Régression multiple, 102
Régression des moindres carrés partiels, 124
exporter des variables, 127
Modèle, 126
Régression linéaire, 102
Blocs, 102
Diagrammes, 105
Enregistrement de nouvelles variables, 107
Exportation des informations du modèle, 107
Fonctionnalités supplémentaires, 112
Méthodes de sélection des variables, 104, 111
Pondérations, 102
Résidus, 107
statistiques, 109
Valeurs manquantes, 111
Variable de sélection, 105
Régression multiple
Dans la régression linéaire, 102
Régression ordinale , 113
Fonctionnalités supplémentaires, 119
Lien, 114
Modèle d'échelle, 118
Modèle d'emplacement, 117
Options, 114
statistiques, 113
Réponses multiples
Fonctionnalités supplémentaires, 286
Résidu non standardisé
GLM, 68
Résidus
Enregistrement dans la régression linéaire, 107
Enregistrement dans l'ajustement de fonctions, 122
Tableaux croisés, 28
Résidus de Pearson
Régression ordinale, 115
Résidus de Student
Dans la régression linéaire, 107
Résidus standardisés
Dans la régression linéaire, 107
GLM, 68
Résidus supprimés
Dans la régression linéaire, 107
GLM, 68
Rho
Corrélations bivariées, 73
Tableaux croisés, 25
Risque
Tableaux croisés, 25
risque relatif
Tableaux croisés, 25
Rotation equamax
Dans Analyse factorielle, 166
Rotation oblimin directe
Dans Analyse factorielle, 166
Rotation quartimax
Dans Analyse factorielle, 166
Rotation Varimax
Dans Analyse factorielle, 166

S
Tableaux croisés, 25
Sélection ascendante
Dans la régression linéaire, 104
dans l'analyse du voisin le plus proche, 135
sélection de k
dans l'analyse du voisin le plus proche, 150
sélection de k et des descriptives
dans l'analyse du voisin le plus proche, 151
sélection des descriptives
dans l'analyse du voisin le plus proche, 149
Sélection progressive
Dans la régression linéaire, 104
Seuil initial
Dans l'analyse TwoStep Cluster, 173
simulation, 316, 325
ajustement de la distribution, 322
analyse de sensibilité, 326
analyse d'hypothèses, 326

boîtes à moustaches, 330
centiles des distributions cible, 330
corrélations entre entrées, 326
création de nouvelles entrées, 321
création d'un plan de simulation, 317
critères d'arrêt, 327
diagrammes de dispersion, 330
diagrammes interactifs, 335
diagrammes tornado, 330
échantillonnage des extrémités, 327
éditeur d'équation, 320
enregistrer les données simulées, 331
enregistrer un plan de simulation, 331
exécution d'un plan de simulation, 318, 331
fonction de densité de probabilité, 328
fonction de distribution cumulée, 328
formats d'affichage des cibles et des entrées, 330
Générateur de simulation, 319
modèles pris en charge, 319
options de diagramme, 335
personnalisation de l'ajustement de la distribution, 325
réajustement des distributions aux nouvelles données, 332
résultat, 328, 330
résultats de l'ajustement de la distribution, 325
spécification du modèle, 319
Simulation de Monte Carlo, 316
Somme
Dans les cubes OLAP, 43
dans les Descriptives, 14
Dans les effectifs, 9
Dans Moyennes, 38
Dans Récapituler, 33
Somme des carrés, 63
GLM, 62
sous-ensembles homogènes
tests non paramétriques, 243
Sous-totaux
Dans les tableaux de bord en colonnes, 297
Standardisation
Dans l'analyse TwoStep Cluster, 173
Statistique de Brown-Forsythe
Dans ANOVA à 1 facteur, 58
Statistique de Cochran
Tableaux croisés, 25
Statistique de Durbin-Watson
Dans la régression linéaire, 109
Statistique de Mantel-Haenszel
Tableaux croisés, 25
Statistique de Welch
Dans ANOVA à 1 facteur, 58
statistique F
dans des modèles linéaires, 87
Statistique *R*
Dans la régression linéaire, 109
Dans Moyennes, 38
Statistiques de ratio, 310
statistiques, 311
Statistiques des proportions de colonne
Tableaux croisés, 28
Statistiques descriptives
Dans Explorer, 18
dans GLM - Univarié, 70
Dans l'analyse TwoStep Cluster, 175
dans les Descriptives, 13
Dans les effectifs, 9
Dans les statistiques de ratio, 311
Dans Récapituler, 33
strates
Tableaux croisés, 23
Stress
Dans le positionnement multidimensionnel, 304
Stress S
Dans le positionnement multidimensionnel, 304
Student-Newman-Keuls
Dans ANOVA à 1 facteur, 56
GLM, 66
Suites de Wald-Wolfowitz
Tests pour deux échantillons indépendants, 269
suites en séquences
Tests non paramétriques à un échantillon, 201, 205
Suites en séquences
Césures, 264–265
Fonctionnalités supplémentaires, 266
Options, 266
statistiques, 266
Valeurs manquantes, 266

T^2 de Hotelling
Dans l'analyse de fiabilité, 300–301
T2 de Tamhane
Dans ANOVA à 1 facteur, 56
GLM, 66
T3 de Dunnett
Dans ANOVA à 1 facteur, 56
GLM, 66
Tableau croisé
Multiréponses, 283
Tableaux croisés, 22
tableau de classification
dans l'analyse du voisin le plus proche, 152
Tableaux croisés, 22
Affichage de cellules, 28
Diagrammes en bâtons juxtaposés, 24
Formats, 30
statistiques, 25
strates, 23
Suppression de tableaux, 22
Variables de contrôle, 23
Tableaux croisés des réponses multiples, 283
Appariement des variables entre les vecteurs, 285
Définition des plages de valeurs, 284
Pourcentages basés sur les observations, 285

Pourcentages basés sur les réponses, 285
Pourcentages dans les cellules, 285
Valeurs manquantes, 285
Tableaux de bord
Colonne de total, 296
Comparaison des colonnes, 296
Division des valeurs de colonnes, 296
Multiplication des valeurs de colonnes, 296
Tableaux de bord en colonnes, 293
Tableaux de bord en lignes, 287
Totaux composites, 296
Tableaux de bord en colonnes, 293
Colonne de total, 296
Fonctionnalités supplémentaires, 298
Format de colonne, 289
Gestion des pages, 297
Mise en page, 291
Numérotation des pages, 298
Sous-totaux, 297
Total général, 298
Valeurs manquantes, 298
Tableaux de bord en lignes, 287
Critères d'agrégation, 287
Espacement de ventilation, 290
Fonctionnalités supplémentaires, 298
Format de colonne, 289
Gestion des pages, 290
Mise en page, 291
Numérotation des pages, 291
Pieds de page, 293
Séquences de tri, 287
titres, 293
Valeurs manquantes, 291
Variables dans les titres, 293
Variables en colonnes, 287
Tableaux de contingence, 22
tableaux de fréquences
Dans Explorer, 18
Dans les effectifs, 8
Tableaux de fréquences des réponses multiples, 281
Valeurs manquantes, 281
Tau de Goodman et Kruskal
Tableaux croisés, 25
Tau de Kruskal
Tableaux croisés, 25
Tau-*b*
Tableaux croisés, 25
Tau-*b* de Kendall
Corrélations bivariées, 73
Tableaux croisés, 25
Tau-*c*
Tableaux croisés, 25
Tau-*c* de Kendall , 25
Tableaux croisés, 25
Termes construits, 62, 118
Termes d'interaction, 62, 118
test binomial
Tests non paramétriques à un échantillon, 201–202
Test binomial, 262
Dichotomies, 262
Fonctionnalités supplémentaires, 264
Options, 264
statistiques, 264
Valeurs manquantes, 264
Test d'additivité de Tukey
Dans l'analyse de fiabilité, 300–301
Test de comparaison par paire de Gabriel
Dans ANOVA à 1 facteur, 56
GLM, 66
Test de comparaison par paire de Games et Howell
Dans ANOVA à 1 facteur, 56
GLM, 66
Test de Friedman
Dans les tests pour plusieurs échantillons liés, 277
Tests non paramétriques pour échantillons liés, 214
Test de Kolmogorov-Smirnov
Tests non paramétriques à un échantillon, 201, 204
Test de la médiane
Tests pour deux échantillons indépendants, 273
Test de Levene
Dans ANOVA à 1 facteur, 58
Dans Explorer, 19
dans GLM - Univarié, 70
Test de Lilliefors
Dans Explorer, 19
Test de McNemar
Tableaux croisés, 25
Tests non paramétriques pour échantillons liés, 214–215
Tests pour deux échantillons liés, 271
test de Scheffé
Dans ANOVA à 1 facteur, 56
GLM, 66
Test de Shapiro-Wilks
Dans Explorer, 19
Test de sphéricité de Bartlett
Dans Analyse factorielle, 163
Test de Wilcoxon
Tests non paramétriques à un échantillon, 201
Tests non paramétriques pour échantillons liés, 214
Tests pour deux échantillons liés, 271
Test des droites parallèles
Régression ordinale, 115
Test des réactions extrêmes de Moses
Tests pour deux échantillons indépendants, 269
Test des signes
Tests non paramétriques pour échantillons liés, 214
Tests pour deux échantillons liés, 271
Test d'Homogénéité marginale
Tests non paramétriques pour échantillons liés, 214
Tests pour deux échantillons liés, 271
Test d'intervalle multiple de Duncan
Dans ANOVA à 1 facteur, 56
GLM, 66

test du Khi-deux
  Tests non paramétriques à un échantillon, 201, 203
Test exact de Fisher
  Tableaux croisés, 25
Test Kolmogorov-Smirnov pour un échantillon, 266
  Distribution à tester, 266
  Fonctionnalités supplémentaires, 268
  Options, 268
  statistiques, 268
  Valeurs manquantes, 268
Test M de Box
  Analyse discriminante, 156
test pour échantillons indépendants
  tests non paramétriques, 232
Test Q de Cochran
  Tests non paramétriques pour échantillons liés, 214, 216
Test *T*
  dans GLM - Univarié, 70
  Test T pour échantillon unique, 52
  Test T pour échantillons appariés, 50
  Test T pour échantillons indépendants, 47
Test *t* de Dunnett
  Dans ANOVA à 1 facteur, 56
  GLM, 66
Test *t* de Sidak
  Dans ANOVA à 1 facteur, 56
  GLM, 66
test *t* de Student, 47
Test *t* de Waller-Duncan
  Dans ANOVA à 1 facteur, 56
  GLM, 66
Test *t* dépendant
  Test T pour échantillons appariés, 50
test *t* pour deux échantillons
  Test T pour échantillons indépendants, 47
Test T pour échantillon unique, 52
  Fonctionnalités supplémentaires, 53
  Intervalles de confiance, 53
  Options, 53
  Valeurs manquantes, 53
Test T pour échantillons appariés, 50
  Options, 51
  Sélection de variables appariées, 50
  Valeurs manquantes, 51
Test T pour échantillons indépendants, 47
  Définition de groupes, 49
  Intervalles de confiance, 49
  Options, 49
  Valeurs manquantes, 49
  Variables chaîne, 49
  Variables de regroupement, 49
Tests de colinéarité
  Dans la régression linéaire, 109
Tests de l'indépendance
  Khi-deux, 25
Tests de linéarité
  Dans Moyennes, 38
Tests de normalité
  Dans Explorer, 19
Tests d'homogénéité de la variance
  Dans ANOVA à 1 facteur, 58
  dans GLM - Univarié, 70
tests non paramétriques
  Tests pour deux échantillons indépendants, 268
Tests non paramétriques
  Khi-deux, 244
  Suites en séquences, 264
  Test Kolmogorov-Smirnov pour un échantillon, 266
  Tests pour deux échantillons liés, 271
  Tests pour plusieurs échantillons indépendants, 273
  Tests pour plusieurs échantillons liés, 276
  vue du modèle, 218
Tests non paramétriques à un échantillon, 199
  champs, 200
  suites en séquences, 205
  test binomial, 202
  Test de Kolmogorov-Smirnov, 204
  test du Khi-deux, 203
Tests non paramétriques pour échantillons indépendants, 206
  Onglet Champs, 208
Tests non paramétriques pour échantillons liés, 211
  champs, 213
  Test de McNemar, 215
  Test Q de Cochran, 216
Tests pour deux échantillons indépendants, 268
  Définition de groupes, 270
  Fonctionnalités supplémentaires, 271
  Options, 271
  statistiques, 271
  Types de test, 269
  Valeurs manquantes, 271
  Variables de regroupement, 270
Tests pour deux échantillons liés, 271
  Fonctionnalités supplémentaires, 273
  Options, 273
  statistiques, 273
  Types de test, 272
  Valeurs manquantes, 273
Tests pour plusieurs échantillons indépendants, 273
  Définition de l'intervalle, 275
  Fonctionnalités supplémentaires, 276
  Options, 275
  statistiques, 275
  Types de test, 274
  Valeurs manquantes, 275
  Variables de regroupement, 275
Tests pour plusieurs échantillons liés, 276
  Fonctionnalités supplémentaires, 278
  statistiques, 277
  Types de test, 277
titres
  Dans les cubes OLAP, 46

Tolérance
  Dans la régression linéaire, 109
Totaux généraux
  Dans les tableaux de bord en colonnes, 298
Tukey
  Dans Explorer, 18

U de Mann-Whitney
  Tests pour deux échantillons indépendants, 269

V de Crameré
  Tableaux croisés, 25
V de Rao
  Analyse discriminante, 157
valeurs éloignées
  Dans la régression linéaire, 105
  Dans l'analyse TwoStep Cluster, 173
Valeurs éloignées
  Dans Explorer, 18
valeurs extrêmes
  Dans Explorer, 18
Valeurs influentes
  Dans la régression linéaire, 107
  GLM, 68
Valeurs manquantes
  Corrélations bivariées, 75
  Dans Analyse factorielle, 168
  Dans ANOVA à 1 facteur, 58
  Dans Corrélations partielles, 77
  Dans Courbe ROC, 315
  Dans Explorer, 20
  Dans la régression linéaire, 111
  dans l'analyse du voisin le plus proche, 140
  Dans les suites en séquences, 266
  Dans les tableaux croisés des réponses multiples, 285
  Dans les tableaux de bord en colonnes, 298
  Dans les tableaux de bord en lignes, 291
  Dans les tableaux de fréquences des réponses multiples, 281
  Dans Test Kolmogorov-Smirnov pour un échantillon, 268
  Test binomial, 264
  Test du Khi-deux, 246
  Test T pour échantillon unique, 53
  Test T pour échantillons appariés, 51
  Test T pour échantillons indépendants, 49
  Tests pour deux échantillons indépendants, 271
  Tests pour deux échantillons liés, 273
  Tests pour plusieurs échantillons indépendants, 275
Valeurs propres
  Dans Analyse factorielle, 163–164
  Dans la régression linéaire, 109
Valeurs standardisées
  dans les Descriptives, 13
Variable de sélection
  Dans la régression linéaire, 105
Variables de contrôle
  Tableaux croisés, 23
Variance
  Dans Explorer, 18
  Dans les cubes OLAP, 43
  dans les Descriptives, 14
  Dans les effectifs, 9
  Dans les tableaux de bord en colonnes, 295
  Dans les tableaux de bord en lignes, 289
  Dans Moyennes, 38
  Dans Récapituler, 33
vecteurs multiréponses
  Livre de codes, 1
viewer de classes
  à propos des modèles de classe, 176
  affichage du contenu des cellules, 181
  Aperçu, 177
  comparaison des classes, 185
  distribution des cellules, 184
  faire basculer les classes et les caractéristiques, 180
  filtrage des enregistrements, 187
  importance des valeurs prédites, 182
  récapitulatif du modèle, 178
  taille des classes, 183
  transposer les classes et les caractéristiques, 180
  trier l'affichage des caractéristiques, 181
  trier l'affichage des classes, 181
  trier le contenu des cellules, 181
  trier les caractéristiques, 181
  trier les classes, 181
  Utilisation, 186
  vue de base, 182
  vue de comparaison des classes, 185
  vue de la distribution des cellules, 184
  vue de la taille des classes, 183
  vue de l'importance des variables prédites de classe, 182
  vue des centres de classes, 179
  vue des classes, 179
  vue récapitulative, 178
visualisation
  modèles de classification, 177
vue du modèle
  dans l'analyse du voisin le plus proche, 141
  Tests non paramétriques, 218

W de Kendall
  Dans les tests pour plusieurs échantillons liés, 277

Z de Kolmogorov-Smirnov
  Dans Test Kolmogorov-Smirnov pour un échantillon, 266
  Tests pour deux échantillons indépendants, 269