IBM SPSS Statistics Base 23

Important
Avant d'utiliser le présent document et le produit associé, prenez connaissance des informations générales figurant à la section «Remarques», à la page 211.

Notice d'édition

La présente édition s'applique à la version 23.0.0 d'IBM SPSS Statistics et à toutes les éditions et modifications ultérieures sauf mention contraire dans les nouvelles éditions.

Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. Les informations qui y sont fournies sont susceptibles d'être modifiées avant que les produits décrits ne deviennent eux-mêmes disponibles. En outre, il peut contenir des informations ou des références concernant certains produits, logiciels ou services non annoncés dans ce pays. Cela ne signifie cependant pas qu'ils y seront annoncés.

Pour plus de détails, pour toute demande d'ordre technique, ou pour obtenir des exemplaires de documents IBM, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial.

Vous pouvez également consulter les serveurs Internet suivants :

- http://www.fr.ibm.com (serveur IBM en France)
- http://www.ibm.com/ca/fr (serveur IBM au Canada)
- http://www.ibm.com (serveur IBM aux Etats-Unis)

Compagnie IBM France
Direction Qualité
17, avenue de l'Europe
92275 Bois-Colombes Cedex

Table des matières

Avis aux lecteurs canadiens

Le présent document a été traduit en France. Voici les principales différences et particularités dont vous devez tenir compte.

Illustrations

Les illustrations sont fournies à titre d'exemple. Certaines peuvent contenir des données propres à la France.

Terminologie

La terminologie des titres IBM peut différer d'un pays à l'autre. Reportez-vous au tableau ci-dessous, au besoin.

IBM France	**IBM Canada**
ingénieur commercial	représentant
agence commerciale	succursale
ingénieur technico-commercial	informaticien
inspecteur	technicien du matériel

Claviers

Les lettres sont disposées différemment : le clavier français est de type AZERTY, et le clavier français-canadien de type QWERTY.

OS/2 et Windows - Paramètres canadiens

Au Canada, on utilise :

- les pages de codes 850 (multilingue) et 863 (français-canadien),
- le code pays 002,
- le code clavier CF.

Nomenclature

Les touches présentées dans le tableau d'équivalence suivant sont libellées différemment selon qu'il s'agit du clavier de la France, du clavier du Canada ou du clavier des États-Unis. Reportez-vous à ce tableau pour faire correspondre les touches françaises figurant dans le présent document aux touches de votre clavier.

France	Canada	Etats-Unis
(Pos1)		Home
Fin	Fin	End
(PgAr)		PgUp
(PgAv)		PgDn
Inser	Inser	Ins
Suppr	Suppr	Del
Echap	Echap	Esc
Attn	Intrp	Break
Impr écran	ImpEc	PrtSc
Verr num	Num	Num Lock
Arrêt défil	Défil	Scroll Lock
(Verr maj)	FixMaj	Caps Lock
AltGr	AltCar	Alt (à droite)

Brevets

Il est possible qu'IBM détienne des brevets ou qu'elle ait déposé des demandes de brevets portant sur certains sujets abordés dans ce document. Le fait qu'IBM vous fournisse le présent document ne signifie pas qu'elle vous accorde un permis d'utilisation de ces brevets. Vous pouvez envoyer, par écrit, vos demandes de renseignements relatives aux permis d'utilisation au directeur général des relations commerciales d'IBM, 3600 Steeles Avenue East, Markham, Ontario, L3R 9Z7.

Assistance téléphonique

Si vous avez besoin d'assistance ou si vous voulez commander du matériel, des logiciels et des publications IBM, contactez IBM direct au 1 800 465-1234.

Chapitre 1. Livre de codes

Le livre de codes indique les informations du dictionnaire, telles que les noms de variables, les libellés de variables, les libellés de valeurs, les valeurs manquantes, ainsi que les statistiques récapitulatives de toutes les variables (ou celles spécifiées) et les jeux de réponses multiples dans le jeu de données actif. Pour les variables ordinales et nominales ainsi que pour les jeux de réponses multiples, 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 de codes ignore le statut du 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

1. A partir des menus, sélectionnez :
   Analyse > **Rapports** > **Livre de codes**
2. Cliquez sur l'onglet Variables.
3. Sélectionnez une ou plusieurs variables et/ou des jeux de réponses multiples.

Sinon, vous pouvez :

- Contrôler les informations de variables affichées.
- Contrôler les statistiques affichées (ou exclure toutes les statistiques récapitulatives).
- Contrôler l'ordre d'affichage des variables et des jeux de réponses multiples.
- Modifier le niveau de mesure de toute variable dans la liste source afin de modifier les statistiques récapitulatives affichées. Pour plus d'informations, voir «Onglet Statistiques du livre de codes», à la page 4.

Modification des niveaux de mesure

Vous pouvez modifier temporairement le niveau de mesure des variables. (Vous ne pouvez pas modifier celui des jeux de réponses multiples. Ils sont toujours traités comme nominaux.)

1. Cliquez avec le bouton droit de la souris sur une variable de la liste source.
2. Sélectionnez un niveau de mesure dans le menu contextuel.

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 de codes.

Onglet Sortie du livre de codes

L'onglet Sortie contrôle les informations de variables disponibles pour chaque variable et jeu de réponses multiples, leur ordre d'affichage et le contenu de la table d'informations des fichiers en option.

Informations sur les variables

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

Position : Nombre entier qui représente la position de la variable dans l'ordre des fichiers. Non disponible pour les jeux de réponses multiples.

Libellé : Libellé descriptif associé à la variable ou au jeu de réponses multiples.

Type : Type de données fondamental. Les valeurs admises sont *Numérique*, *Chaîne* ou *Jeu de réponses multiples*.

Format : Format d'affichage de la variable, tel que *A4*, *F8.2* ou *DATE11*. Non disponible pour les jeux de réponses multiples.

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 jeux de réponses multiples.

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. Pour plus d'informations, voir .

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

Libellés de valeurs : Libellés descriptifs associés à des valeurs de données spécifiques.

- Si l'option Effectif ou Pourcentage est sélectionnée dans l'onglet Statistiques, les libellés de valeurs définis sont compris dans les sorties même si vous ne sélectionnez pas Libellés de valeur ici.
- Pour les jeux de dichotomies multiples, les "libellés de valeur" sont les libellés des variables élémentaires du jeu ou les libellés des valeurs comptées, selon la définition du jeu. Pour plus d'informations, voir .

Valeurs manquantes : Valeurs manquantes définies par l'utilisateur. Si l'option Effectif ou Pourcentage est sélectionnée dans l'onglet Statistiques, les libellés de valeurs définis sont compris dans les sorties même si vous ne sélectionnez pas Valeurs manquantes ici. Non disponible pour les jeux de réponses multiples.

Attributs personnalisés : Attributs de variable personnalisés. Les sorties comprennent à la fois les noms et les valeurs pour tout attribut de variables personnalisé associé à chaque variable. Pour plus d'informations, voir . Non disponible pour les jeux de réponses multiples.

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 sorties comprennent à la fois les noms et les valeurs pour tout attribut système associé à chaque variable. Non disponible pour les jeux de réponses multiples.

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 le jeu de données n'a jamais été enregistré au format IBM 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, le jeu de données actif ne comporte pas de nom de fichier.)

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

Nombre d'observations : Nombre d'observations dans le jeu 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.

Libellé : Fichier de libellé (si disponible) défini par la commande `FILE LABEL`.

Documents : Texte de document de fichier de données.

Statut de la pondération : Si la pondération est activée, le nom de la variable de pondération est affiché. Pour plus d'informations, voir .

Attributs personnalisés : Attributs de fichiers de données personnalisés définis par l'utilisateur. Les attributs de fichiers de données sont 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 sorties 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 jeux de réponses multiples :

Alphabétique : Ordre alphabétique par nom de variable.

Fichier : Ordre d'affichage des variables dans le jeu de données (leur ordre d'affichage dans l'Editeur de données). Dans l'ordre croissant, les jeux de réponses multiples sont affichés en dernier, après toutes les variables sélectionnées.

Niveau de mesure : Tri par niveau de mesure. Ceci crée quatre groupes de tri (nominal, ordinal, échelle et inconnu). Les jeux de réponses multiples 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 : Ordre d'affichage des variables et des jeux de réponses multiples 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 sorties comprennent les libellés 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 de codes

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

Nombres et pourcentages

Pour les variables nominales et ordinales, les jeux de réponses multiples et les valeurs de libellé 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 la dispersion des valeurs autour de 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. Indique les valeurs correspondant au 25e, 50e et 75e percentiles.

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

Pour obtenir un rapport de fréquences et un graphique à barres, vous pouvez trier les différentes valeurs par ordre croissant ou décroissant, ou bien classer les catégories en fonction de leurs fréquences. Le rapport de fréquences peut être supprimé lorsqu'une variable a plusieurs valeurs distinctes. Vous pouvez libeller les graphiques 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 ? La sortie 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 €.

Tracés et statistiques : Effectifs de fréquences, pourcentages, pourcentages cumulés, moyenne, médiane, mode, somme, écart type, variance, plage, valeurs minimale et maximale, erreur standard de la moyenne, asymétrie et kurtosis (avec leurs erreurs standard), quartiles, percentiles choisis par l'utilisateur, graphiques à barres, graphiques circulaires et histogrammes.

Remarques sur les données de fréquences

Données : Utilisez des codes numériques ou alphanumériques pour coder les variables catégorielles (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 catégories 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 percentiles, sont appropriées pour les variables quantitatives, qu'elles répondent ou non au critère de normalité.

Pour obtenir des tables de fréquences

1. A partir des menus, sélectionnez :
   Analyse > **Statistiques descriptives** > **Effectifs...**
2. Sélectionnez une ou plusieurs variables catégorielles ou quantitatives.

Sinon, vous pouvez :

- Cliquer sur **Statistiques** pour obtenir des statistiques descriptives pour des variables quantitatives.
- Cliquer sur **Graphiques** pour obtenir des graphiques à barres, des graphiques circulaires ou des histogrammes.
- Cliquer sur **Format** pour définir l'ordre de présentation des résultats.

Statistiques des effectifs

Percentiles : Valeurs d'une variable quantitative qui divisent les données triées en groupes par centième. Les quartiles (25e, 50e et 75e percentiles) divisent les observations en quatre groupes de taille égale. Si vous souhaitez un nombre égal de groupes différent de quatre, sélectionnez **Partition en n groupes égaux**. Vous pouvez également spécifier des percentiles particuliers (par exemple, le 95e percentile, 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 et au-dessous de laquelle se trouve la moitié des observations (50e percentile). 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, la plage, le minimum, le maximum et l'erreur standard (ES) de la moyenne.

- *Ecart type*. Mesure de la dispersion des valeurs autour de 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 la dispersion des valeurs 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.
- *Plage*. Différence entre la valeur maximale et la valeur minimale d'une variable numérique (maximum - minimum).
- *Minimum*. Plus petite valeur d'une variable numérique.
- *Maximum*. Plus grande valeur d'une variable numérique.
- *Erreur standard de la moyenne*. Mesure du taux de variation de la valeur de la moyenne sur des échantillons provenant de la 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 kurtosis 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 a une valeur d'asymétrie égale à 0. Une distribution caractérisée par une importante asymétrie positive présente une partie droite plus 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.
- *Kurtosis*. 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 kurtosis est égale à zéro. Un kurtosis 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 kurtosis 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 groupes : Si les valeurs dans vos données représentent des centres de groupes (par exemple, les âges des individus trentenaires sont représentés par le code 35), sélectionnez cette option pour estimer la médiane et les percentiles des données originales, non regroupées.

Graphiques des effectifs

Type de graphique : Un graphique circulaire montre la participation de chaque partie à l'ensemble. Chaque tranche du graphique circulaire correspond à un groupe défini par une simple variable de regroupement. Un graphique à barres montre l'effectif de chaque valeur ou de chaque catégorie sous la forme d'une barre distincte, ce qui vous permet de comparer les catégories visuellement. Un histogramme est également constitué de barres mais ils sont répartis à intervalles égaux. La hauteur de chaque barre 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 graphique : Dans les graphiques à barres, l'axe peut être libellé par effectifs ou pourcentages de fréquences.

Format des effectifs

Ordre d'affichage : La table de fréquences peut être affichée en fonction des valeurs réelles des données ou de l'effectif (fréquence d'occurrence) de ces valeurs, et organisée par valeurs croissantes ou décroissantes. Cependant, si vous demandez un histogramme ou des percentiles, le format des 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 la sortie par variables**).

Supprimer les tableaux comportant plusieurs catégories : Cette option évite l'affichage des tableaux ayant plus que le nombre spécifié de valeurs.

Chapitre 3. Descriptives

La procédure Descriptives affiche les statistiques récapitulatives 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 graphiques, 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, plage, somme, erreur standard de la moyenne, et kurtosis et asymétrie avec leurs erreurs standard (ES).

Remarques sur les données de Descriptives

Données : Utilisez des variables numériques après les avoir visualisées graphiquement en cherchant des erreurs d'enregistrement, les valeurs extrêmes 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 catégories 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

1. A partir des menus, sélectionnez :

   Analyse > **Statistiques descriptives** > **Descriptives...**
2. Sélectionnez une ou plusieurs variables.

Sinon, vous pouvez :

- Cliquer 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

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, la plage, le minimum, le maximum, et l'erreur standard (ES) de la moyenne.

- *Ecart type*. Mesure de la dispersion des valeurs autour de 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 la dispersion des valeurs 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.
- *Plage*. Différence entre la valeur maximale et la valeur minimale d'une variable numérique (maximum - minimum).
- *Minimum*. Plus petite valeur d'une variable numérique.
- *Maximum*. Plus grande valeur d'une variable numérique.
- *Erreur standard de la moyenne*. Mesure du taux de variation de la valeur de la moyenne sur des échantillons provenant de la 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 : Kurtosis et l'asymétrie sont des statistiques qui caractérisent la forme et la symétrie de la distribution. Ces statistiques sont présentées avec leurs erreurs standard.

- *Kurtosis*. 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 kurtosis est égale à zéro. Un kurtosis 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 kurtosis 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 a une valeur d'asymétrie égale à 0. Une distribution caractérisée par une importante asymétrie positive présente une partie droite plus 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.

Fonctions supplémentaires de la commande DESCRIPTIVES

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Enregistrer les scores standardisés (écarts *z*) pour certaines variables uniquement (à l'aide de la sous-commande `VARIABLES`).
- Spécifier le nom des nouvelles variables contenant des scores standardisés (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 au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 4. Explorer

La procédure Explorer produit des statistiques récapitulatives 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 balayage de données, l'identification des valeurs extrêmes, la description, la vérification d'hypothèses et la caractérisation des différences parmi les sous populations (groupes d'observations). Le balayage 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 : Examinez 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 tracés tige et feuille résument graphiquement la répartition des temps d'apprentissage pour chacun des groupes.

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

Remarques sur les Données d'Explorer

Données : La procédure d'Explorer peut être utilisée pour les variables quantitatives (Mesures de niveaux d'intervalle ou de rapport). Un facteur (utilisé pour répartir les données en groupes d'observations) doit avoir un nombre raisonnable de valeurs distinctes (catégories). Ces valeurs peuvent être des chaînes de caractères courtes ou numériques. La variable de libellé par observation, utilisée pour libeller les valeurs extrêmes dans les boîtes à moustaches, 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

1. A partir des menus, sélectionnez :
   Analyse > **Statistiques descriptives** > **Explorer...**
2. Sélectionnez au moins une variable dépendante.

Sinon, vous pouvez :

- Sélectionner un ou plusieurs facteurs, dont les valeurs définiront les groupes d'observations.
- Sélectionner une variable d'identification pour libeller les observations.
- Cliquer sur **Statistiques** pour les M-estimateurs, les Valeurs extrêmes, les Percentiles et les tables de fréquences.
- Cliquer sur **Tracés** pour les histogrammes, les tracés de probabilités gaussiens avec tests et la dispersion/niveau avec test de Levene.

- Cliquer sur **Options** pour le traitement des valeurs manquantes.

Statistiques d'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, la plage, et la plage interquartile. Les statistiques descriptives comprennent aussi les mesures de la forme des répartitions. L'asymétrie et kurtosis 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 extrêmes : Affiche les cinq plus grandes et cinq plus petites valeurs avec les libellés d'observations.

Percentiles : Affiche les valeurs pour le 5e, 10e, 25e, 50e, 75e, 90e et 95e percentiles.

Tracés d'Explorer

Boîtes à moustaches : Ces alternatives contrôlent l'affichage de boîtes à moustaches quand vous avez plus d'une variable dépendante. L'option **Niveaux de facteur** génère un affichage séparé pour chaque variable dépendante. Dans un affichage, les boîtes à moustaches sont données pour chacun des groupes définis par un facteur. L'option **Dépendantes** génère un affichage séparé pour chaque groupe défini par un facteur. 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 tracés tige et feuille et les histogrammes.

Tracés de répartition gaussiens avec tests : Affiche les tracés de probabilités gaussiens et les résidus. La statistique de Kolmogorov-Smirnov avec un niveau 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 tracés de dispersion par niveau. Pour tous les tracés 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 aucun facteur n'est sélectionné, les tracés de dispersion par niveau ne sont pas produits. L'option **Estimation d'exposants** produit un tracé des logs naturels des plages 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 tracé de dispersion par niveau aide à déterminer l'exposant pour qu'une transformation stabilise (rende plus égales) les variances entre groupes. L'option **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 tracés des données transformées. La plage interquartile et la médiane des données transformées sont dessinés. L'option **Sans transformation** produit des tracés 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 tracés 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 :** L'inverse de la racine carrée est calculée pour chaque valeur des données.
- **Inverse :** L'inverse 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

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 facteurs 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 facteurs sont traitées comme une catégorie séparée. Toute sortie est produite pour cette catégorie supplémentaire. Les tables de fréquences contiennent les catégories pour les valeurs manquantes. Les valeurs manquantes pour un facteur sont inclues, mais libellées comme manquantes.

Fonctions supplémentaires de la commande EXAMINE

La procédure Explorer utilise la syntaxe de la commande `EXAMINE`. Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- demander le total des sorties et des tracés en complément des sorties et tracés relatifs aux groupes définis par les facteurs (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 facteurs (au moyen de la sous-commande `VARIABLES`) ;
- spécifier les percentiles autres que ceux par défaut (au moyen de la sous-commande `PERCENTILES`) ;
- calculer les percentiles 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 tracés 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`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 5. Tableaux croisés

La procédure 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 catégories déterminent les tests ou mesures à 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 couche de facteur (variable de contrôle), la procédure Tableaux croisés forme un tableau de statistiques et de mesures pour chaque valeur de la couche de facteur (ou une combinaison de valeurs pour deux variables de contrôle ou plus). Par exemple, si le *sexe* est un facteur de couche 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 Cramér, lambdas symétriques et asymétriques, tau de Goodman et Kruskal, coefficient d'incertitude, gamma, *d* de Somers, tau-*b* de Kendall, tau-*c* de Kendall, coefficient êta, kappa de Cohen, estimations de risque relatif, rapport de cotes, test de McNemar, statistiques de Cochran et de Mantel-Haenszel et statistiques des proportions de colonne.

Remarques sur les Données pour tableau croisé

Données : Pour définir les catégories 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 catégories ordonnées (données ordinales) ou de valeurs quantitatives (données d'intervalle ou données de type rapport), tel que décrit dans la section sur les statistiques. D'autres sont valides lorsque les variables du tableau ont des catégories désordonnées (données nominales). Pour les statistiques basées sur le khi-deux (phi, *V* de Cramér et coefficient de contingence), les données doivent provenir d'un échantillon aléatoire avec une distribution multinomiale.

Remarque : Les variables ordinales peuvent être des codes numériques représentant des catégories (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 catégories. Par exemple, pour une variable de chaîne comportant des valeurs *Faible*, *Moyen*, *Elevé*, l'ordre des catégories 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

1. A partir des menus, sélectionnez :
   Analyse > **Statistiques descriptives** > **Tableaux croisés...**
2. 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-tables.
- Cliquer sur **Cellules** pour les valeurs observées et théoriques, les pourcentages et les résidus.
- Cliquer sur **Format** pour contrôler l'ordre des catégories.

Couches de tableaux croisés

Si vous sélectionnez des variables de couche, un tableau croisé séparé est produit pour chacune des catégories de variable de couche (variable de contrôle). Par exemple, si vous avez une variable de ligne, une variable de colonne, et une variable de couche avec deux catégories, vous obtenez un tableau à deux entrées pour chacune des catégories de la variable de couche. Pour créer une autre couche de variables de contrôle, cliquez sur **Suivant**. Les sous-tables sont produites 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-tables à deux entrées.

Graphiques à barres en cluster de tableaux croisés

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

Tableaux croisés affichant les variables de couche dans des couches de tableau

Afficher les variables de couche dans les couches du tableau : Vous pouvez choisir d'afficher les variables de couche (variables de contrôle) sous forme de couches 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 catégories des variables de couche.

Vous trouverez ci-dessous un exemple utilisant le fichier de données *demo.sav* (disponible dans le dossier Echantillons du répertoire d'installation) ayant été obtenu comme suit :

1. Sélectionnez *Catégorie 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 couche.
2. Sélectionnez l'option **Afficher les variables de couche dans des couches de tableau**.
3. Sélectionnez **Colonne** dans la sous-boîte de dialogue Contenu des cases.
4. Exécutez la procédure Tableaux croisés, double-cliquez sur le tableau croisé et sélectionnez **Diplôme universitaire** dans la liste déroulante Niveau d'éducation.

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

Statistiques de tableaux croisés

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 et Juif, vous pouvez sélectionner le **Coefficient de contingence**, le coefficient **Phi** et **V de Cramér**, **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*. Mesure d'association calculée à partir du khi-deux et 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 catégories de colonnes à partir des catégories 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 Somers*. 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 / intervalle : Lorsqu'une variable est catégorielle et l'autre quantitative, sélectionnez **Eta**. La variable catégorielle 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 catégories (par exemple, le sexe). Deux valeurs êta 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 de 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. Le rapport des cotes 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 un facteur dichotomique et une variable de réponse dichotomique, selon les modèles de covariable définis par des variables (contrôles) de couche. Remarque : alors que les autres statistiques sont calculées couche par couche, les statistiques de Cochran et de Mantel-Haenszel sont calculées une seule fois pour toutes les couches.

Affichage de cellules (cases) de tableaux croisés

Pour vous aider à découvrir des motifs 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 appariées 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 appariées 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 couche) 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 fournissent la différence entre les valeurs observées et les valeurs théoriques. Les résidus standardisés et standardisés ajustés sont également disponibles.

- *Non standardisés*. Différence entre une valeur observée et la valeur attendue. 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 écart type. Egalement appelés résiduels de Pearson, les résiduels 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.

- *Effectifs de cellules arrondis*. Les pondérations d'observation sont utilisées telles quelles, mais les pondérations cumulées dans les cellules sont arrondies avant le calcul de toute statistique.
- *Effectifs de cellules tronqués*. Les pondérations d'observations sont utilisées en l'état, mais les pondérations cumulées dans les cellules sont arrondies avant le calcul de toute statistique.
- *Pondérations des observations arrondies*. Les pondérations des observations sont arrondies avant leur utilisation.
- *Pondérations des observations tronquées*. Les pondérations d'observations sont tronquées 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

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

Chapitre 6. Récapitulatif

La procédure Récapituler calcule les statistiques de sous-groupes pour les variables à l'intérieur des catégories 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 catégories. Les valeurs des données dans chaque catégorie 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 catégorie de la variable de regroupement, valeur de la variable pour la dernière catégorie de la variable de regroupement, écart type, variance, kurtosis, erreur standard de kurtosis, 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.

Récapituler les commentaires de données

Données : Les variables de regroupement sont des variables catégorielles dont les valeurs peuvent être de type numérique ou chaîne. Le nombre de catégories 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 la plage, conviennent aux variables quantitatives qui confirment ou infirment l'hypothèse de normalité.

Obtenir des récapitulatifs des observations

1. A partir des menus, sélectionnez :
   Analyse > **Rapports** > **Récapitulatif des observations**
2. 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 de la sortie, ajouter une légende au-dessous de la sortie, 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 premières** *n* ou désélectionner cet élément pour répertorier toutes les observations.

Options de Récapituler

Récapituler vous permet de modifier le titre de votre sortie 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 retour à la 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 la sortie par un point ou un astérisque. Saisissez 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 la sortie.

Statistiques de Récapituler

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

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

Moyenne géométrique. Nième racine du produit des valeurs de données, no 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 inverses des tailles de l'échantillon.

Kurtosis. 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 kurtosis est égale à zéro. Un kurtosis 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 kurtosis 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 et au-dessous de laquelle se trouve la moitié des observations (50e percentile). 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. Plus petite valeur d'une variable numérique.

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

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

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

Plage. 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 a une valeur d'asymétrie égale à 0. Une distribution caractérisée par une importante asymétrie positive présente une partie droite plus 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.

Ecart type. Mesure de la dispersion des valeurs autour de 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.

Erreur standard de Kurtosis. Rapport de kurtosis avec son erreur standard pouvant être utilisé comme test de normalité (autrement dit, vous pouvez conclure à une anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur kurtosis positive importante indique que les extrémités de la distribution sont plus allongées que celles d'une distribution normale ; une valeur kurtosis négative présente des extrémités plus courtes (semblables à celles d'une distribution uniforme sous forme de boîtes).

Erreur standard de la moyenne. Mesure du taux de variation de la valeur de la moyenne sur des échantillons provenant de la 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).

Erreur standard d'asymétrie. Rapport de l'asymétrie avec son erreur standard pouvant être utilisé comme test de normalité (autrement dit, vous pouvez conclure à une 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 la dispersion des valeurs 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 Moyennes calcule les moyennes de sous-groupes et les statistiques univariées correspondantes pour des variables dépendantes au sein des catégories d'une ou de plusieurs variables indépendantes. Vous pouvez également obtenir une analyse de variance à un facteur, 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 de variance à un facteur 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 catégorie de la variable de regroupement, valeur de la variable pour la dernière catégorie de la variable de regroupement, écart type, variance, kurtosis, erreur standard de kurtosis, 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 variance, un coefficient êta, un coefficient êta-carré, ainsi que des tests de linéarité *R* et R^2.

Remarques sur les données des moyennes

Données : Les variables dépendantes sont quantitatives et les variables indépendantes sont catégorielles. Les valeurs des variables catégorielles 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 variance résiste aux écarts par rapport à la normalité, à condition que les données de chaque cellule soient symétriques. L'analyse de 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

1. A partir des menus, sélectionnez :

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

Options de Moyennes

Vous pouvez choisir l'une des statistiques de sous-groupe suivantes pour les variables à l'intérieur de chaque catégorie de chacune des variables de regroupement : somme, nombre d'observations, moyenne, médiane, médiane groupée, erreur standard pour la moyenne, minimum, maximum, plage, valeur de la variable pour la première catégorie de la variable de regroupement, valeur de la variable pour la dernière catégorie de la variable de regroupement, écart type, variance, kurtosis, erreur standard de kurtosis, asymétrie, erreur standard d'asymétrie, pourcentage de la somme totale, pourcentage de *N* total, pourcentage de la somme dans, pourcentage de *N* dans, 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 la sortie. Les statistiques récapitulatives sont aussi affichées pour chaque variable à travers toutes les catégories.

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

Moyenne géométrique. Nième racine du produit des valeurs de données, no 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 inverses des tailles de l'échantillon.

Kurtosis. 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 kurtosis est égale à zéro. Un kurtosis 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 kurtosis 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 et au-dessous de laquelle se trouve la moitié des observations (50e percentile). 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. Plus petite valeur d'une variable numérique.

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

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

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

Plage. 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 a une valeur d'asymétrie égale à 0. Une distribution caractérisée par une importante asymétrie positive présente une partie droite plus 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.

Ecart type. Mesure de la dispersion des valeurs autour de 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.

Erreur standard de Kurtosis. Rapport de kurtosis avec son erreur standard pouvant être utilisé comme test de normalité (autrement dit, vous pouvez conclure à une anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur kurtosis positive importante indique que les extrémités de la distribution sont plus allongées que celles d'une distribution normale ; une valeur kurtosis négative présente des extrémités plus courtes (semblables à celles d'une distribution uniforme sous forme de boîtes).

Erreur standard de la moyenne. Mesure du taux de variation de la valeur de la moyenne sur des échantillons provenant de la 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).

Erreur standard d'asymétrie. Rapport de l'asymétrie avec son erreur standard pouvant être utilisé comme test de normalité (autrement dit, vous pouvez conclure à une 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 la dispersion des valeurs 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 couche

Tableau Anova et êta. Affiche un tableau d'analyse de variance unifactorielle 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 Cubes OLAP (Online Analytical Processing) calcule les totaux, les moyennes et autres statistiques univariées pour des variables récapitulatives continues à l'intérieur de catégories d'une ou de plusieurs variables de regroupement catégorielles. Une couche séparée dans le tableau est créée pour chaque catégorie 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, plage, valeur de la variable pour la première catégorie de la variable de regroupement, valeur de la variable pour la dernière catégorie de la variable de regroupement, écart type, variance, kurtosis, erreur standard de kurtosis, 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.

Remarques sur les données des cubes OLAP

Données : Les variables récapitulatives sont quantitatives (variables continues mesurées sur une échelle d'intervalle ou de rapport) et les variables de regroupement sont catégorielles. Les valeurs des variables catégorielles 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 la plage, conviennent aux variables quantitatives qui confirment ou infirment l'hypothèse de normalité.

Pour obtenir des cubes OLAP

1. A partir des menus, sélectionnez :
   Analyse > **Rapports** > **Cubes OLAP...**
2. Sélectionnez une ou plusieurs variables récapitulatives continues.
3. Sélectionnez une ou plusieurs variables de regroupement catégorielles.

Eventuellement :

- Sélectionnez d'autres statistiques récapitulatives (cliquez sur **Statistiques...**). Vous devez sélectionner un ou plusieurs critères de regroupement pour pouvoir sélectionner les statistiques récapitulatives.
- Calculez les différences entre des paires de variables et des paires de groupes définies par un critère de regroupement (cliquez sur **Différences**).
- Créez des titres de tableaux personnalisés (cliquez sur **Titre**).
- Masquez 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.

Statistiques des Cubes OLAP

Vous pouvez choisir l'une des statistiques de sous-groupe suivantes pour les variables récapitulatives à l'intérieur de chaque catégorie de chacune des variables de regroupement : somme, nombre d'observations, moyenne, médiane, médiane groupée, erreur standard pour la moyenne, minimum, maximum, plage, valeur de la variable pour la première catégorie de la variable de regroupement, valeur de la variable pour la dernière catégorie de la variable de regroupement, écart type, variance, kurtosis,

erreur standard de kurtosis, asymétrie, erreur standard d'asymétrie, pourcentage total des observations, pourcentage de la somme totale, pourcentage total des observations 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 la sortie. Les statistiques récapitulatives sont aussi affichées pour chaque variable à travers toutes les catégories.

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

Moyenne géométrique. Nième racine du produit des valeurs de données, no 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 inverses des tailles de l'échantillon.

Kurtosis. 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 kurtosis est égale à zéro. Un kurtosis 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 kurtosis 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 et au-dessous de laquelle se trouve la moitié des observations (50e percentile). 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. Plus petite valeur d'une variable numérique.

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

Pourcentage de N dans. Pourcentage du nombre d'observations pour la variable de regroupement spécifiée dans les catégories des autres variables 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 pour la variable de regroupement spécifiée dans les catégories 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 catégorie.

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

Plage. 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 a une valeur d'asymétrie égale à 0. Une distribution caractérisée par une importante asymétrie positive présente une partie droite plus 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.

Ecart type. Mesure de la dispersion des valeurs autour de 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.

Erreur standard de Kurtosis. Rapport de kurtosis avec son erreur standard pouvant être utilisé comme test de normalité (autrement dit, vous pouvez conclure à une anormalité si ce rapport est inférieur à -2 ou supérieur à +2). Une valeur kurtosis positive importante indique que les extrémités de la distribution sont plus allongées que celles d'une distribution normale ; une valeur kurtosis négative présente des extrémités plus courtes (semblables à celles d'une distribution uniforme sous forme de boîtes).

Erreur standard de la moyenne. Mesure du taux de variation de la valeur de la moyenne sur des échantillons provenant de la 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).

Erreur standard d'asymétrie. Rapport de l'asymétrie avec son erreur standard pouvant être utilisé comme test de normalité (autrement dit, vous pouvez conclure à une 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 la dispersion des valeurs 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.

Différences des Cubes OLAP

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 catégorie (catégorie moins) dans chaque paire sont soustraites des valeurs des statistiques récapitulatives de la première catégorie de la paire. Les différences de pourcentage utilisent la valeur de la statistique récapitulative pour la catégorie 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.

Titre des Cubes OLAP

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 retour à la ligne dans le texte.

Chapitre 9. Tests T

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 sorties 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 de la sortie par défaut.

Test T pour échantillons indépendants

La procédure 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 : taille d'échantillon, moyenne, écart type et erreur standard de la moyenne. Pour la différence de moyennes : moyenne, erreur standard et niveau de confiance (vous pouvez préciser le niveau de confiance). Tests : test de Levene sur l'égalité des variances et tests *t* des variances regroupées en pool et séparées pour l'égalité des moyennes.

Remarques sur les Données du Test T pour Echantillons Indépendants

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 extrêmes.

Obtenir un test t pour échantillons indépendants

1. A partir des menus, sélectionnez :

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

Définition de Groupes Test T pour échantillons indépendants

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.

Pour les variables de regroupement alphanumériques, entrez une chaîne pour le Groupe 1 et une autre pour le Groupe 2, par exemple *oui* et *non*. Les observations qui ont une autre valeur sont exclues de l'analyse.

Options de Test T pour échantillons indépendants

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 Test T pour échantillons 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 : moyenne, taille d'échantillon, écart type et erreur standard de la moyenne. Pour chaque paire de variables : 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.

Commentaires relatifs au test T pour échantillons appariés

Données : Pour chaque test apparié, précisez deux variables continues (niveau d'intervalle de mesure ou niveau de rapport 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

1. A partir des menus, sélectionnez :

   Analyse > **Comparer les moyennes** > **Test T pour échantillons appariés...**
2. Sélectionnez une ou plusieurs paires de variables
3. Vous pouvez également cliquer sur **Options** pour contrôler le traitement des données manquantes et le niveau de l'intervalle de confiance.

Options de 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.

Fonctions supplémentaires de la commande T-TEST

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Produire à la fois des tests t pour un échantillon et pour des échantillons 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`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Test T pour échantillon unique

La procédure 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 la pondération moyenne des échantillons diffère de 1,3 livre à l'intervalle de confiance 95 %.

Statistiques : Pour chaque variable de 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).

Commentaires sur les Données du Test T pour échantillon unique

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

1. A partir des menus, sélectionnez :

   Analyse > **Comparer les moyennes** > **Test T pour échantillon unique...**
2. Sélectionnez au moins une variable à tester par rapport à la même valeur supposée.
3. Entrez une valeur test numérique à laquelle vous souhaitez comparer chaque moyenne d'échantillon.
4. Vous pouvez également cliquer sur **Options** pour contrôler le traitement des données manquantes et le niveau de l'intervalle de confiance.

Options de 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.

Fonctions supplémentaires de la commande T-TEST

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Produire à la fois des tests t pour un échantillon et pour des échantillons 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`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Fonctions supplémentaires de la commande T-TEST

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Produire à la fois des tests t pour un échantillon et pour des échantillons 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`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

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 catégories.

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 variance et tests d'égalité des moyennes pour chaque variable dépendante, contrastes a priori spécifiés par l'utilisateur et tests de plage et comparaisons multiples post hoc : Bonferroni, Sidak, test de Tukey, GT2 de Hochberg, Gabriel, Dunnett, test *F* de Ryan-Einot-Gabriel-Welsch (*F* de R-E-G-W), test de plage de Ryan-Einot-Gabriel-Welsch (*Q* de R-E-G-W), T2 de Tamhane, T3 de Dunnett, Games-Howell, *C* de Dunnett, test de plage multiple de Duncan, Student-Newman-Keuls (S-N-K), *b* de Tukey, Waller-Duncan, Scheffé et différence la moins significative.

Remarques sur les Données ANOVA à 1 facteur

Données : Les valeurs du facteur 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 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

1. A partir des menus, sélectionnez :
   Analyse > **Comparer les moyennes** > **ANOVA à 1 facteur...**
2. Sélectionnez au moins une variable dépendante.
3. Sélectionnez un facteur indépendant simple.

Contrastes ANOVA à 1 facteur

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

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 du facteur. 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 (catégorie) du facteur 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 catégorie du facteur. Le premier coefficient de la liste correspond à la valeur la plus petite du facteur, et le dernier coefficient correspond à la valeur la plus élevée. Par exemple, s'il y a six catégories de facteurs, 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

Lorsque vous avez déterminé qu'il existe des différences parmi les moyennes, les tests de plages post hoc et de comparaisons multiples appariées peuvent déterminer les moyennes qui diffèrent. Les tests de plage 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 de plage. Il existe d'autres tests de plage, 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 de plage de Ryan-Einot-Gabriel-Welsch) et le Waller-Duncan. Les tests de comparaison multiple disponibles sont les suivants : Bonferroni, Tukey, Sidak, Gabriel, Hochberg, Dunnett, Scheffé et LSD (Différence la moins significative).

- *LSD*. Utilisation de tests t pour effectuer toutes les comparaisons appariées 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 appariées 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 niveau de signification observé est ainsi ajusté en raison des comparaisons multiples réalisées.
- *Sidak*. Test de comparaisons multiples appariées reposant sur la statistique t. Le test de Sidak ajuste le niveau de signification en fonction des comparaisons multiples et fournit des bornes plus étroites que le test de Bonferroni.
- *Scheffé*. Exécute des comparaisons appariées 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 appariées.
- *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 une plage de Student.
- *S-N-K*. Ce test effectue toutes les comparaisons appariées de moyennes, à l'aide de la distribution des plages 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 étape par étape. Les moyennes sont triées dans l'ordre décroissant et les différences extrêmes sont testées en premier.

- *Tukey*. Utilise les statistiques de plages de Student pour effectuer toutes les comparaisons appariées de groupes. Fixe le taux d'erreur expérimental au niveau du taux d'erreur de l'ensemble pour toutes des comparaisons appariées.
- *B de Tukey*. Utilise la distribution des plages de Student pour effectuer des comparaisons de classes deux à deux. 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 appariées en suivant un ordre étape par étape 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 de plage de Student.
- *GT2 de Hochberg*. Test de multiples comparaisons et plages appariées utilisant le modulo maximum de Student. Similaire au test de Tukey.
- *Gabriel*. Test de comparaison appariée qui utilise le modulo 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 appariées comparant un ensemble de traitements à une moyenne de contrôle unique.La dernière catégorie est la catégorie de contrôle par défaut. Vous pouvez également choisir la première catégorie. L'option **Bilatéral** teste que la moyenne à un certain niveau (hormis la catégorie de contrôle) du facteur n'est pas égale à celle de la catégorie de contrôle. L'option **<Contrôle** permet de tester si la moyenne est inférieure, à un certain niveau du facteur, à celle de la catégorie de contrôle. L'option **> Contrôle** permet de tester si la moyenne est supérieure, à un certain niveau du facteur, à celle de la catégorie 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 des comparaisons appariées basé sur le test T. Ce test est opportun lorsque les variances sont inégales.
- *T3 de Dunnett*. Test des comparaisons appariées basé sur le modulo maximal de Student. Ce test est opportun lorsque les variances sont inégales.
- *Games-Howell*. Test de comparaison appariée qui peut parfois être souple. Ce test est opportun lorsque les variances sont inégales.
- *C de Dunnett*. Test des comparaisons appariées basé sur le modulo maximal de Student. Ce test est opportun lorsque les variances sont inégales.

Remarque : Il peut vous paraître plus facile d'interpréter la sortie 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 croisé dynamique activé, choisissez **Propriétés du tableau** dans le menu Format).

Options ANOVA à 1 facteur

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.

Tracé des moyennes : Affiche un graphique qui trace les moyennes de sous-groupes (les moyennes de chaque groupe définies par les valeurs du facteur).

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 le facteur d'une analyse donnée. De même, on n'utilise pas d'observation en dehors de la plage spécifiée pour le facteur.
- **Exclure toute observation incomplète :** Les observations ayant des valeurs manquantes pour le facteur 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.

Fonctions supplémentaires de la commande ONEWAY

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Obtenir des statistiques à effets fixes et aléatoires. Ecart 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, les tests de comparaison multiple Bonferroni, Duncan et Scheffé (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 regroupées en pool et des degrés de liberté des variances regroupées en pool. Ces matrices peuvent être utilisées à la place des données brutes pour obtenir une analyse de variance à un facteur (avec la sous-commande `MATRIX`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 11. Analyse GLM – Univarié

La procédure GLM – Univarié fournit un modèle de régression et une analyse de variance pour plusieurs variables dépendantes par un ou plusieurs facteurs ou variables. Les facteurs 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 (prédicteur) sont spécifiées comme covariables.

Vous pouvez tester les modèles équilibrés comme déséquilibrés. Un plan 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 tracés de profil (tracés 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.

Pondération 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 de plage post hoc et comparaisons multiples : différence la moins significative, Bonferroni, Sidak, Scheffé, *test de F* multiple de Ryan-Einot-Gabriel-Welsch, plage multiple de Ryan-Einot-Gabriel-Welsch, Student-Newman-Keuls, test de Tukey, *b* de Tukey, Duncan, GT2 de Hochberg, test *t* de Gabriel, Waller-Duncan, Dunnett (unilatéral, bilatéral), T2 de Tamhane, T3 de Dunnett, Games-Howell et *C* de Dunnett. Statistiques descriptives : moyennes observées, écarts types et effectifs pour toutes les variables dépendantes de toutes les cellules. Le test de Levene pour l'homogénéité de la variance.

Tracés : Dispersion par niveau, résiduels et profils (interaction).

Remarques sur les données GLM - Univarié

Données : La variable dépendante est quantitative. Les facteurs sont catégoriels. 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 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 tracés de dispersion par niveau. Vous pouvez également étudier les résidus et les tracés résiduels.

Pour obtenir des tables GLM - Univarié

1. A partir des menus, sélectionnez :
   Analyse > **Modèle linéaire général** > **Univarié...**
2. Sélectionnez une variable dépendante.
3. Sélectionnez des variables pour Facteurs fixés, Facteurs aléatoires et Covariables en fonction de vos données.
4. 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

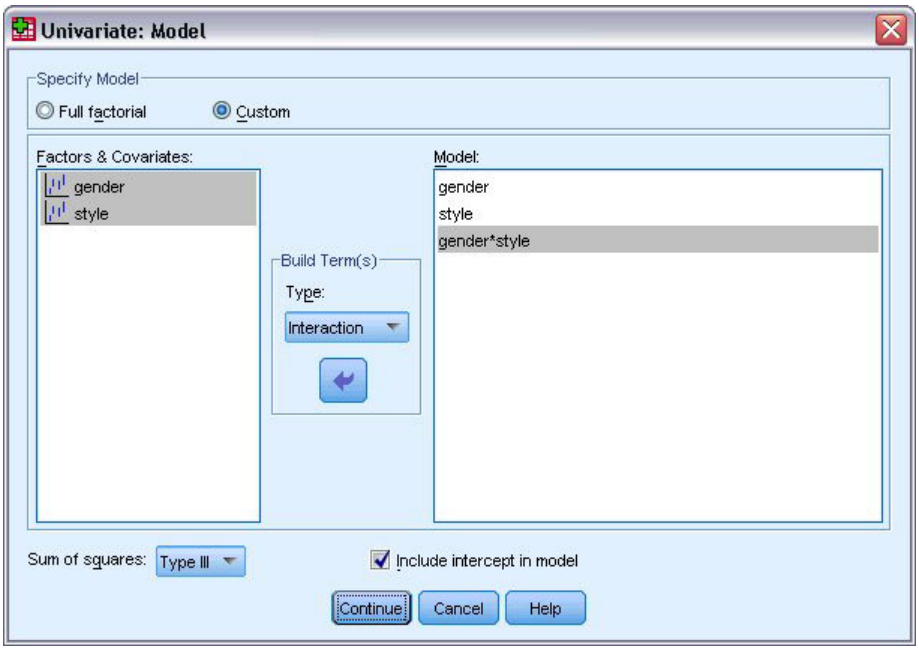

Figure 1. Boîte de dialogue Modèle univarié

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 facteur/covariable. Vous devez indiquer tous les termes à inclure dans le modèle.

Facteurs 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.

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 : La constante 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 des effets factoriels principaux uniquement.
- Tout modèle de régression.
- Un plan par imbrication pur. (Cette forme d'imbrication peut être spécifiée par la syntaxe.)

Type III : Valeur par défaut. Cette méthode calcule les sommes des carrés d'un effet dans le plan comme les sommes des carrés, ajustées pour tout autre effet qui ne le contient pas et orthogonales pour tous les effets qui le contiennent. 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 plan 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 plan, 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

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.

GLM - Univarié : 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.

La sortie 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 déviation, simple, différence, Helmert, répétée et modèle polynomial. Pour les contrastes d'écart et simple, vous pouvez choisir si la catégorie de référence est la première ou la dernière.

Types de contraste

Déviation : Compare la moyenne de chaque niveau (hormis une catégorie 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 catégorie 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.

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

Tracés de profil GLM

Les tracés de profil (tracés d'interaction) sont utiles pour comparer les moyennes marginales dans votre modèle. Un tracé de profil est un tracé en ligne 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 tracé distinct. Tous les facteurs fixés et aléatoire sont disponibles pour les tracés. Pour les analyses multivariées, les tracés de profil 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 tracés de profil. GLM - Multivarié et GLM - Mesures Répétées ne sont disponibles que si vous avez installé l'option Statistiques avancées.

Un tracé de profil 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.

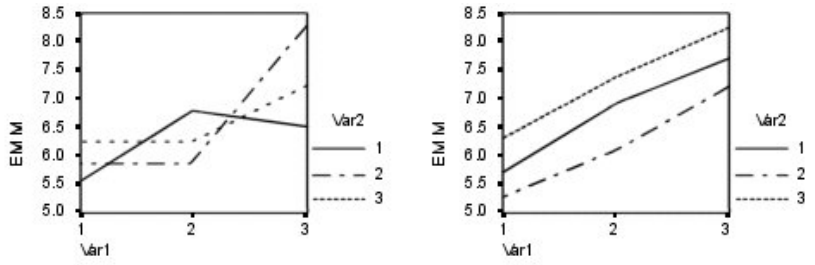

Figure 2. Tracé non parallèle (gauche) et tracé parallèle (droite)

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

Options GLM

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 appariées non corrigées des moyennes marginales estimées pour tout effet principal dans le modèle, à la fois pour les facteurs intersujets et intrasujets. 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 de confiance 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.

Affichage : 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 tracés de dispersion par niveau et résiduels sont utiles pour vérifier les hypothèses sur les données. Ceci n'est pas valable s'il n'y a pas de facteurs. Sélectionnez **Tracés résiduels** pour produire un tracé résiduel observé/estimé/standardisé pour chaque variable dépendante. Ces tracés 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.

Fonctions supplémentaires de la commande UNIANOVA

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Spécifier les effets imbriqués dans un plan (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 de 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 catégorie 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 entre 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`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Comparaisons post hoc GLM

Test de comparaison multiple post hoc : Lorsque vous avez déterminé qu'il existe des différences parmi les moyennes, les tests de plages post hoc et de comparaisons multiples appariées 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 de plage de Student pour effectuer des comparaisons appariées entre les groupes et fixe le taux d'erreur empirique au taux d'erreur du regroupement de toutes les comparaisons appariées. 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 modulo maximum selon Student. Le test de Tukey est généralement plus efficace. Le **test de comparaison appariée de Gabriel** utilise également le modulo 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 appariée de Dunnett** compare un ensemble de traitements à une simple moyenne de contrôle. La dernière catégorie est la catégorie de contrôle par défaut. Vous pouvez également choisir la première catégorie. Vous pouvez également choisir un test unilatéral ou bilatéral. Pour tester que la moyenne à un certain niveau (hormis la catégorie de contrôle) du facteur n'est pas égale à celle de la catégorie de contrôle, utilisez le test double-face. Pour tester si la moyenne est inférieure, à un certain niveau du facteur, à celle de la catégorie 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 de plages 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 la plage de Student. Ces tests sont plus efficaces que le test de plages 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 appariées conservatif fondé sur un test *t*, le **T3 de Dunnett** (test de comparaison appariée fondé sur le modulo maximal de Student), le **test de comparaison appariée** de **Games-Howell** (parfois flexible) ou le **C de Dunnetts** (test de comparaison appariée fondé sur la plage de 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 de plages multiples de Duncan**, Student-Newman-Keuls (**S-N-K**) et le **b de Tukey** sont des tests de plage qui classifient les moyennes de groupe et calculent une valeur de plage. 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 de plage 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 appariée, 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 appariée 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 appariées 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 de

plage 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 de plage.

Options GLM

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 appariées non corrigées des moyennes marginales estimées pour tout effet principal dans le modèle, à la fois pour les facteurs intersujets et intrasujets. 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 de confiance 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.

Affichage : 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 tracés de dispersion par niveau et résiduels sont utiles pour vérifier les hypothèses sur les données. Ceci n'est pas valable s'il n'y a pas de facteurs. Sélectionnez **Tracés résiduels** pour produire un tracé résiduel observé/estimé/standardisé pour chaque variable dépendante. Ces tracés 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.

Fonctions supplémentaires de la commande UNIANOVA

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Spécifier les effets imbriqués dans un plan (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 de 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 catégorie 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 entre 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).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Enregistrement GLM

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.
- *Pondérés*. 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 expliquée pour les unités statistiques qui ont les mêmes valeurs pour les valeurs explicatives.

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 permettant de savoir de combien les résidus de toutes les observations seraient modifiés si une observation donnée était exclue du calcul 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, de 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.
- *Pondérés*. Résidus estimés non standardisés pondérés. Disponibles uniquement lorsqu'une variable WLS a été préalablement sélectionnée.
- *Standardisés*. Résidu, divisé par une estimation de son écart type. Egalement appelés résiduels de Pearson, les résiduels standardisés ont une moyenne de 0 et un écart type de 1.
- *De Student*. 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 nouveau jeu de données de la session en cours ou dans un fichier de données externe au format IBM 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

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 appariées non corrigées des moyennes marginales estimées pour tout effet principal dans le modèle, à la fois pour les facteurs intersujets et intrasujets. 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 de confiance 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.

Affichage : 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 tracés de dispersion par niveau et résiduels sont utiles pour vérifier les hypothèses sur les données. Ceci n'est pas valable s'il n'y a pas de facteurs. Sélectionnez **Tracés résiduels** pour produire un tracé résiduel observé/estimé/standardisé pour chaque variable dépendante. Ces tracés 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.

Fonctions supplémentaires de la commande UNIANOVA

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Spécifier les effets imbriqués dans un plan (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 de 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 catégorie 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 entre 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).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 12. Corrélations bivariées

La procédure Corrélations bivariées calcule le coefficient de corrélation de Pearson, le rho de Spearman et le tau-*b* de Kendall avec leurs niveaux 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 extrêmes (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 nuage de points 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 : nombre d'observations avec des valeurs non manquantes, de moyenne et d'écart type. Pour chaque paire de : coefficient de corrélation de Pearson, rho de Spearman, tau-*b* de Kendall, déviation des produits en croix et covariance.

Remarques sur les données des corrélations bivariées

Données : Utilisez des variables quantitatives symétriques pour le coefficient de corrélation de Pearson, et des variables quantitatives ou des variables avec des catégories 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...**

1. 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 catégories 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

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.
- **Déviation des produits en croix et covariances :** Indiqués pour chaque paire de variables. Le produit des déviations 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 à la déviation des produits en croix divisée 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.

Fonctions supplémentaires des commandes CORRELATIONS et NONPAR CORR

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- 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`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 13. Corrélations partielles

La procédure Corrélations partielles calcule les coefficients de corrélation partielle qui décrivent le rapport linéaire entre deux variables tout en contrôlant les effets d'une ou de 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 : nombre d'observations avec des valeurs non manquantes, de moyenne et d'écart type. Matrices de corrélation partielle et simple, avec degrés de liberté et niveaux de signification.

Remarques sur les données des corrélations partielles

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

1. A partir des menus, sélectionnez :
   Analyse > **Corrélation** > **Partielle...**
2. Sélectionnez au moins deux variables numériques pour lesquelles vous voulez calculer des corrélations partielles.
3. 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 niveau 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 niveau 0,05 sont identifiés par une astérisque, les coefficients significatifs au niveau 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

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.

Fonctions supplémentaires de la commande PARTIAL CORR

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- 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`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 14. Distances

La procédure Distances 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, l'analyse de cluster 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 la taille du moteur, 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 analyse de cluster 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, la différence 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

1. A partir des menus, sélectionnez :

   Analyse > **Corrélation** > **Distances...**
2. Sélectionnez au minimum une ou deux variables numériques pour calculer respectivement les distances existant entre des observations ou des variables.
3. Sélectionnez une possibilité dans le groupe Calculer les distances pour calculer les proximités existant entre des observations ou des variables.

Distances : 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, différence 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 les scores *z*, la plage –1 à 1, la plage 0 à 1, l'amplitude maximale de 1, la moyenne de 1 ou l'écart type de 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 disponibles sont Valeurs absolues, Inverser le signe et Rééchelonner entre 0 et 1.

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 les scores *z*, la plage –1 à 1, la plage 0 à 1, l'amplitude maximale de 1, la moyenne de 1 et l'écart type de 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 disponibles sont Valeurs absolues, Inverser le signe et Rééchelonner entre 0 et 1.

Fonctions supplémentaires de la commande PROXIMITIES

La procédure Distances utilise la syntaxe de la commande `PROXIMITIES`. Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

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

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 15. Modèles linéaires

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

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 décisions) sur le même jeu 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. Pour plus d'informations, voir .

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 restrictions sur les niveaux de mesure des prédicteurs (variables d'entrée) ; les champs catégoriels (nominaux et ordinaux) sont utilisés comme facteurs dans le modèle et les champs continus sont utilisés comme covariables.

Remarque : Si un champ catégoriel comprend plus de 1 000 catégories, la procédure ne s'exécute pas et aucun modèle n'est créé.

Obtention d'un modèle linéaire

Cette fonction nécessite l'option Statistiques de base.

A partir des menus, sélectionnez :

Analyse > **Régression** > **Modèles linéaires automatiques...**

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

Objectifs

Quel est votre objectif principal ? Sélectionnez l'objectif de votre choix.

- **Créer un modèle standard** : Cette méthode permet de créer un modèle unique afin de prédire la cible à l'aide de prédicteurs. En général, les modèles standard sont plus faciles à interpréter et peuvent être plus rapidement évalués que des jeux 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 scores 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 du jeu 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 scores peut être plus longue qu'avec un modèle standard.

  L'agrégation de bootstrap (bagging) produit des réplicats du jeu de données d'apprentissage en effectuant un échantillonnage avec remplacement à partir du jeu de données d'origine. Ceci crée des échantillons de bootstrap de taille identique à celle du jeu de données d'origine. Ensuite, un modèle de composant est créé 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.
- **Crée un modèle pour des jeux de données très volumineux (nécessite IBM SPSS Statistics Server)** : Cette méthode permet de créer un modèle d'ensemble en scindant le jeu de données en blocs de données distincts. Choisissez cette option si votre jeu de données est trop important pour que vous puissiez créer l'un des modèles ci-dessus, ou pour la génération d'un modèle incrémental. La construction de cette option peut être moins longue, mais l'obtention des scores peut être plus longue qu'avec un modèle standard. Cette option requiert une connectivité à IBM SPSS Statistics Server.

Pour plus d'informations sur les paramètres relatifs au boosting, au bagging et aux jeux de données très volumineux, voir «Ensembles», à la page 64.

Bases

Préparer automatiquement les données : Cette option permet à la procédure de transformer la cible et les prédicteurs en interne afin de maximiser le pouvoir prédictif 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 prédicteur de date est transformé en une nouveau prédicteur continu qui contient la durée écoulée depuis une date de référence (01/01/1970). Chaque prédicteur d'heure est transformé en une nouveau prédicteur continu qui contient la durée écoulée depuis une heure de référence (00:00:00).
- **Régler le niveau de mesure :** Les prédicteurs continus ayant moins de 5 valeurs distinctes sont reconvertis en prédicteurs ordinales. Les prédicteurs ordinaux ayant plus de 10 valeurs distinctes sont reconvertis en prédicteurs continus.
- **Traitement des valeurs extrêmes :** Les valeurs de prédicteurs continus 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 prédicteurs nominaux sont remplacées par le mode de la partition d'apprentissage. Les valeurs manquantes de prédicteurs ordinaux sont remplacées par la médiane de la partition d'apprentissage. Les valeurs manquantes de prédicteurs continus 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 catégories similaires sont identifiées 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 0,1), 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 prédicteur.

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

Méthodes de choix du modèle : Choisissez l'une des méthodes de sélection du modèle (détails ci-dessous) ou **Inclure tous les prédicteurs**, qui entre simplement tous les prédicteurs disponibles en tant que termes du modèle des effets principaux. Le modèle **Etape par étape 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 étape par étape.

- **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 du surajustement. L'ensemble de prévention du surajustement est un sous-échantillon aléatoire d'environ 30 % du jeu 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 étape par étape 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 de pas :** L'algorithme étape par étape 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 étape par étape 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 du surajustement. L'ensemble de prévention du surajustement est un sous-échantillon aléatoire d'environ 30 % du jeu 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 calcule 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

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 jeux 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

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 du 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

Enregistrer les prévisions dans le jeu de données actif : 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

La vue Récapitulatif des modèles est un instantané permettant de consulter en un coup d'oeil 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.

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

Préparation automatique des données

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 :

- **Dériver les durées de date** calcule le temps écoulé, en mois, à partir des valeurs d'un champ contenant des dates jusqu'à la date système en cours.
- **Dériver les durées temporelles** calcule le temps écoulé, en heures, à partir des valeurs d'un champ contenant des heures jusqu'à l'heure système en cours.
- **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.
- **Tronquer les valeurs extrêmes** définit les valeurs des prédicteurs continus 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 cible**identifie les catégories de prédicteurs 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 prédicteurs constants / après le traitement des valeurs extrêmes / après la fusion des catégories** supprime les prédicteurs qui possèdent une valeur unique, éventuellement une fois les autres actions ADP effectuées.

Importance des prédicteurs

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 graphique d'importance des prédicteurs peut vous y aider en indiquant l'importance relative de chaque prédicteur en estimant le modèle. Etant donné que les valeurs sont relatives, la somme des valeurs pour l'ensemble des prédicteurs affichés est 1,0. L'importance des prédicteurs n'a aucun rapport avec la précision du modèle. Elle est juste liée à l'importance de chaque prédicteur pour la réalisation d'une prévision, peu importe si cette dernière est précise ou non.

Valeurs prévues en fonction des valeurs observées

Ceci affiche un nuage de points 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

Ceci affiche un graphique de diagnostique des résidus du modèle.

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

- **Histogramme :** Il s'agit d'un histogramme à intervalles des résidus de Student 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.
- **Tracé P-P :** Il s'agit d'un tracé probabilité-probabilité mis en intervalles qui compare des résidus de Student à une distribution normale. Si la pente des points tracés 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 tracés ont une courbe en S, la distribution des résidus est asymétrique.

Valeurs extrêmes

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

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

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

- **Graphique :** Il s'agit d'un graphique dans lequel les effets sont triés de haut en bas en diminuant l'importance du prédicteur. 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 infobulle 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 du prédicteur. 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 prédicteurs : Il existe un curseur de l'importance du prédicteur qui contrôle ceux qui sont affichés dans la vue. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les prédicteurs les plus importants. Par défaut, les 10 premiers effets sont affichés.

Signification : Il existe un curseur de signification qui offre des contrôles plus avancées sur les effets affichés dans la vue, en plus de celles affichées en fonction de l'importance du prédicteur 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

Cette vue affiche la valeur de chaque coefficient du modèle. Veuillez noter que les facteurs (prédicteurs catégoriels) 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** .

- **Graphique :** Il s'agit d'un graphique qui affiche d'abord la constante, puis trie les effets de haut en bas en diminuant l'importance du prédicteur. 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 infobulle 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 du prédicteur. 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 infobulle affiche le nom du paramètre, l'effet auquel il est associé, et (pour les prédicteurs catégoriels) les libellés 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'un prédicteur catégoriel.

Importance des prédicteurs : Il existe un curseur de l'importance du prédicteur qui contrôle ceux qui sont affichés dans la vue. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les prédicteurs les plus importants. Par défaut, les 10 premiers effets sont affichés.

Signification : Il existe un curseur de signification qui offre des contrôles plus avancées sur les coefficients affichés dans la vue, en plus de celle affichée en fonction de l'importance du prédicteur. 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

Il s'agit de graphiques affichés pour des prédicteurs significatifs. Le graphique affiche la valeur estimée par le modèle de la cible sur l'axe vertical pour chaque valeur du prédicteur de l'axe horizontal en conservant tous les autres prédicteurs. Il offre une visualisation pratique des effets des coefficients de chaque prédicteur sur la cible.

Remarque : Si aucun prédicteur n'est significatif, aucune moyenne estimée n'est générée.

Récapitulatif de génération de modèle

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 génération de modèle.

Pas à pas ascendante : Lorsque l'algorithme de sélection est pas à pas ascendante, le tableau affiche les 10 dernières étapes de l'algorithme pas à pas. Pour chaque pas, 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 procédure 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 nuage de points 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 : nombre d'observations valides, moyenne et écart type. Pour chaque modèle : coefficient de régression, matrice de corrélation, mesure et corrélations partielles, *R* multiple, R^2, R^2 ajusté, modification dans R^2, erreur standard de l'estimation, tableau d'analyse de 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 informations de diagnostic des observations. Tracés : nuages de points, tracés partiels, histogrammes et tracés de probabilités gaussiens.

Régression linéaire : Remarques sur les données

Données : Les variables dépendantes et indépendantes doivent être quantitatives. Les variables catégorielles, 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

1. A partir des menus, sélectionnez :

   Analyse > **Régression** > **Linéaire...**
2. Dans la boîte de dialogue Régression linéaire, sélectionnez une variable numérique dépendante.
3. 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 tracés.
- Sélectionnez une variable de pondération WLS numérique pour une analyse des moindres carrés pondérés.

WLS. Permet d'obtenir un modèle pondéré des moindres carrés. 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 dans laquelle toutes les variables d'un bloc sont introduites en une seule étape.
- *Etape par étape*. 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 supprimée.
- *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 introduites dans l'équation, puis éliminées une à une. La variable ayant la plus petite corrélation partielle avec la variable dépendante est la variable dont la suppression 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 de suppression.
- *Sélection ascendante*. Procédure de sélection étape par étape 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 signification dans vos sorties sont basées sur l'adéquation à un modèle unique. Par conséquent, les valeurs de signification ne sont généralement pas valables lorsqu'on utilise une méthode détaillée (étape par étape, 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 étape par étape, 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

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.

Tracés de régression linéaire

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

Nuages de points : Vous pouvez tracer deux des éléments suivants : la variable dépendante, les prévisions standardisées, les résidus standardisés, les résidus supprimés, les prévisions ajustées, les résidus de Student et les résidus supprimés de Student. Tracer 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 tracés partiels : Affiche des nuages de points 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 tracé partiel.

Tracés résiduels standardisés : Vous pouvez obtenir des histogrammes des résidus standardisés et des tracés de probabilités gaussiens en comparant la répartition des résidus standardisés à une répartition gaussienne.

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

Régression linéaire : Enregistrement des nouvelles variables

Vous pouvez enregistrer les prévisions, les résidus et les autres statistiques utiles pour les informations de diagnostic. 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*. Valeur prédite pour une observation lorsque celle-ci 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 permettant de savoir de combien les résidus de toutes les observations seraient modifiés si une observation donnée était exclue du calcul 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*. Bornes supérieure et inférieure (deux variables) de l'intervalle de prévision de la variable dépendante pour une observation particulière.
- *Intervalle de confiance*. Saisissez une valeur comprise entre 1 et 99,99 pour spécifier le niveau de confiance des deux intervalles de prévision. Vous devez sélectionner Moyenne ou Individuelle avant d'entrer cette valeur. Les niveaux 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 écart type. Egalement appelés résiduels de Pearson, les résiduels standardisés ont une moyenne de 0 et un écart type de 1.
- *De Student*. 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.
- *De Student supprimés*. 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.

- *DfBêta(s)*. 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 de prévision*. 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 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 jeu de données ou dans un fichier de données. Les jeux 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 jeux 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

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*. L'option **Intervalles de confiance** affiche les intervalles de confiance avec le niveau spécifié de confiance pour chaque coefficient de régression 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 introduites et éliminées du modèle sont répertoriées et les statistiques de qualité d'ajustement suivantes sont affichées : *R* multiple, R^2 et R^2 ajusté, erreur standard de l'estimation et 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 un bon prédicteur 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 niveau 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.

Mesure. 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 condition 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 résultats et les informations de diagnostic des observations correspondant au critère de sélection (valeurs extrêmes de *n* écarts types).

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.

- *Choisir la probabilité de F*. Une variable est entrée dans le modèle si le niveau de signification de la valeur F est inférieur à la valeur Entrée ; la variable est éliminée si ce niveau est supérieur à la valeur Suppression. La valeur Entrée doit être inférieure à la valeur Suppression 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 Suppression.
- *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 Suppression. La valeur Entrée doit être supérieure à la valeur Suppression 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 Suppression.

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.

Fonctions supplémentaires de la commande REGRESSION

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- 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 au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 17. Régression ordinale

La procédure 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 plan 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 plages 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 catégories 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.

Tracés et statistiques : Fréquences observées et théoriques, et fréquences cumulées, résiduels de Pearson pour les fréquences cumulées, probabilités observées et théoriques, probabilités observées et cumulées théoriques de chaque catégorie de réponse par motif de covariable, corrélation asymptotique et matrices de covariance des estimations des paramètres, khi-deux de Pearson et khi-deux du rapport de vraisemblance, statistiques de 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.

Régression ordinale : Remarques sur les données

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 catégorie. Les facteurs sont supposés être catégoriels. 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 motif 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

1. A partir des menus, sélectionnez :
   Analyse > **Régression** > **Ordinale...**
2. Sélectionnez une variable dépendante.
3. 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.

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 de vraisemblance :** L'algorithme s'interrompt si la modification absolue ou relative apportée au log de 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 de 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 suivantes sont disponibles.

- **Logit :** $f(x)=\log(x/(1-x))$. Cette option est généralement utilisée pour les catégories distribuées de façon égale.
- **Log-log complémentaire :** $f(x)=\log(-\log(1-x))$. Cette option est généralement utilisée lorsque les catégories supérieures sont les plus probables.
- **Log-log négatif :** $f(x)=-\log(-\log(x))$. Cette option est généralement utilisée lorsque les catégories inférieures sont les plus probables.
- **Probit :** $f(x)=\Phi^{-1}(x)$. Cette option est généralement utilisée lorsque la variable de latence est normalement distribuée.
- **Cauchit (Cauchy inverse) :** $f(x)=\tan(\pi(x-0.5))$. Cette option est généralement utilisée lorsque la variable de latence possède un grand nombre de valeurs extrêmes.

Régression ordinale : Sortie

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

Affichage : Génère des tableaux pour :

- **Historique des itérations d'impression :** Le log de 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.
- **Statistiques de qualité d'ajustement :** Statistiques du khi-deux du rapport de vraisemblance et de Pearson. 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, résiduels de Pearson pour les fréquences cumulées, probabilités observées et théoriques, probabilités observées et cumulées de chaque catégorie de réponse par motif de covariable. Notez que pour les modèles comportant plusieurs motifs 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 motif de classement d'un motif de facteur/covariable dans les catégories de réponse. Il existe autant de probabilités que de nombre de catégories de réponse.
- **Catégorie estimée :** Catégorie de réponse contenant la probabilité estimée maximale pour un motif de facteur/covariable.
- **Probabilité de catégorie estimée :** Probabilité estimée de classement d'un motif de facteur/covariable au sein d'une catégorie prévue. Cette probabilité représente également le maximum de probabilités estimées du motif de facteur/covariable.
- **Probabilité de catégorie actuelle :** Probabilité estimée de classement d'un motif de facteur/covariable au sein de la catégorie actuelle.

Imprimer un log de vraisemblance : Contrôle l'affichage du log de 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.

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 entre facteurs ou covariables.

Facteurs/covariables : 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.

Facteurs/covariables : 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.

Fonctions 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 également d'effectuer les actions suivantes :

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

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 18. Estimation de courbe

La procédure Estimation de courbe produit des statistiques de régression d'estimation de courbe et les tracés relatifs pour 11 modèles différents de régression d'estimation de courbe. 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 nuage de points 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* multiple, R^2, R^2 ajusté, erreur standard de l'estimation, tableau d'analyse de variance, prévisions, résidus et intervalles de la prévision. Modèles : linéaire, logarithmique, inverse, quadratique, cubique, de puissance, composé, courbe en S, logistique, de croissance et exponentiel.

Remarques sur les données de l'estimation de courbe

Données : Les variables dépendantes et indépendantes doivent être quantitatives. Si vous sélectionnez **Temps** à partir du jeu de données actif comme variable indépendante (au lieu de sélectionner une variable), la procédure Estimation de courbe 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 temporelles. L'analyse des séries temporelles nécessite une structure de fichier de données dans lequel chaque observation (ligne) 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 un modèle linéaire est utilisé, les hypothèses suivantes doivent être satisfaites : 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 une estimation de courbe

1. A partir des menus, sélectionnez :

   Analyse > **Régression** > **Estimation de courbe...**
2. Sélectionnez au moins une variable dépendante. Un modèle différent est produit pour chaque variable dépendante.
3. Sélectionnez une variable indépendante (une variable dans le jeu de données actif ou dans le jeu de données actif ou **Temps**).
4. Eventuellement :

- Sélectionner une variable pour libeller des observations dans les nuages de points. Pour chaque point du nuage de points, utilisez l'outil de sélection de points pour afficher la valeur de la variable avec Libellé 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.
- **Tracer sous forme graphique :** Trace graphiquement les valeurs de la variable dépendante et chaque modèle sélectionné face à la variable indépendante. Un graphique séparé est produit pour chaque variable dépendante.
- **Afficher le tableau ANOVA :** Affiche une table récapitulative de l'analyse de variance pour chaque modèle sélectionné.

Modèles d'estimation de courbe

Vous pouvez choisir des modèles de régression d'estimation de courbe. Pour déterminer quel modèle utiliser, tracez 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 nuage de points de vos données. Si le tracé 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).

De puissance. 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).

Courbe 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 limite supérieure. Après avoir sélectionné la logistique, précisez la valeur limite 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.

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

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

Enregistrement de l'estimation de courbe

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 libellés descriptifs s'affichent dans un tableau dans la fenêtre de sortie.

Calculer une prévision : Si, dans le jeu 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 temporelle. 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 Plage 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 fonction peut être utilisée pour prévoir les valeurs au-delà de la dernière observation de la série temporelle. 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 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 de prédicteur présentent une forte corrélation ou lorsque le nombre de prédicteurs dépasse le nombre d'observations.

PLS combine des fonctions 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.

Tableaux : Proportion de variance expliquée (par facteur latent), pondérations de facteurs latents, chargements 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.

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

Remarques sur les données de régression des moindres carrés partiels

Niveau de mesure : Les variables dépendantes et indépendantes (prédicteur) 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 catégorielles (nominales ou ordinales) sont traitées de manière équivalente par la procédure.

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

Pondérations de fréquence : 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 indicateur représentant les variables catégorielles.

Pour obtenir la régression des moindres carrés partiels

A partir des menus, sélectionnez :

Analyse > **Régression** > **Moindres carrés partiels...**

1. Sélectionnez au moins une variable dépendante.

2. Sélectionnez au moins une variable indépendante.

Sinon, vous pouvez :

- Indiquer une catégorie de référence pour les variables catégorielles dépendantes (nominale ou ordinale).
- Indiquer une variable à utiliser comme identificateur unique pour les jeux de données enregistrés et les sorties par observation.
- Indiquer une limite supérieure sur le nombre de facteurs latents à extraire.

Prérequis

La procédure Régression des moindres carrés partiels représente une commande d'extension Python et requiert IBM SPSS Statistics - Essentials for Python, qui est installé par défaut avec votre produit IBM SPSS Statistics. Cette procédure nécessite également la présence des bibliothèques Python NumPy et SciPy, qui sont toutes les deux disponibles gratuitement.

Remarque : Pour les utilisateurs travaillant en mode d'analyse distribuée (IBM SPSS Statistics Server requis), les bibliothèques NumPy et SciPy doivent être installées sur le serveur. Contactez l'administrateur système pour obtenir de l'aide.

Utilisateurs Windows et Mac

Sous Windows et Mac, les bibliothèques NumPy et SciPy doivent être installées sur une version de Python 2.7 différente de celle installée avec IBM SPSS Statistics. Si vous ne disposez pas d'une version autre que Python 2.7, vous pouvez la télécharger sur http://www.python.org. Vous pouvez ensuite installer les bibliothèques NumPy et SciPy pour Python version 2.7. Les programmes d'installation sont disponibles à l'adresse http://www.scipy.org/Download.

Pour activer l'utilisation de NumPy et SciPy, définissez l'emplacement de la version de Python 2.7 sur le répertoire d'installation des bibliothèques NumPy et SciPy. Vous pouvez configurer l'emplacement de Python à partir de l'onglet Emplacements des fichiers de la boîte de dialogue Options (Edition > Options).

Utilisateurs Linux

Nous vous suggérons de télécharger vous-même les versions source et compilée des bibliothèques NumPy et SciPy. La source est disponible à l'adresse http://www.scipy.org/Download. Vous pouvez installer NumPy et SciPy dans la version de Python 2.7 qui est installée avec IBM SPSS Statistics. Elle se trouve dans le répertoire `Python` sous le répertoire d'installation d'IBM SPSS Statistics.

Si vous choisissez d'installer NumPy et SciPy dans une version de Python 2.7 autre que celle installée avec IBM SPSS Statistics, vous devez définir l'emplacement de Python sur cette version. Vous pouvez configurer l'emplacement de Python à partir de l'onglet Emplacements des fichiers de la boîte de dialogue Options (Edition > Options).

Serveur Windows et Unix

Les bibliothèques NumPy et SciPy doivent être installées, sur le serveur, sur une version de Python 2.7 différente de celle installée avec IBM SPSS Statistics. S'il n'existe pas de version distincte de Python 2.7 sur le serveur, elle peut être téléchargée depuis http://www.python.org. Les bibliothèques NumPy et SciPy pour Python 2.7 sont disponibles à partir du site http://www.scipy.org/Download. Pour activer l'utilisation de NumPy et SciPy, définissez sur le serveur l'emplacement de la version de Python 2.7 sur le répertoire d'installation des bibliothèques NumPy et SciPy. L'emplacement de Python est défini à partir de la IBM SPSS Statistics Administration Console.

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.

Facteurs 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

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

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

- **Enregistrer les estimations concernant les observations individuelles :** Enregistre les estimations de modèle par observation suivantes : prévisions, résidus, distance au modèle de facteur latent et scores factoriels latents. Elle permet également de tracer les scores factoriels latents.
- **Enregistrer les estimations concernant les facteurs latents :** Enregistre les chargements de facteurs latents et les pondérations de facteurs latents. Elle permet également de tracer les pondérations de facteurs latents.
- **Enregistrer 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 appelée *k*.

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.

Remarques sur l'analyse du voisin le plus proche

Cible et fonctions : La cible et les fonctions peuvent être :

- *Nominal*. Une variable peut être traitée comme étant nominale si ses valeurs représentent des catégories 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 catégories 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 catégories 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 catégorielles 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 :

Tableau 1. Icônes de niveau de mesure

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

Tableau 1. Icônes de niveau de mesure (suite)

	Numérique	Chaîne	Date	Heure
Nominal				

Codification des variables indicatrices : La procédure recode provisoirement les variables de prédicteur catégorielles et les variables dépendantes via la codification un-de-*c* pour la durée de la procédure. S'il existe des catégories *c* d'une variable, la variable est stockée comme vecteurs c, la première catégorie étant identifiée par (1,0,...,0), la suivante par (0,1,0,...,0), ... et la dernière par (0,0,...,0,1).

Ce schéma de codification augmente la dimensionnalité de l'espace des fonctions. Plus particulièrement, le nombre total de dimensions correspond au nombre de prédicteurs d'échelle plus le nombre de catégories sur l'ensemble des prédicteurs catégoriels. En conséquence, ce système de codification 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 catégories dans vos prédicteurs catégoriels en combinant des catégories similaires ou en supprimant les observations comportant des catégories extrêmement rares avant de lancer la procédure.

Toute codification un-de-*c* repose sur les données d'apprentissage, même si un échantillon restant est défini (voir «Partitions», à la page 90). Ainsi, si l'échantillon restant contient des observations avec des catégories de prédicteurs absentes des données de formation, ces observations ne seront pas évaluées. Si l'échantillon restant contient des observations avec des catégories de variable dépendantes absentes des données de formation, ces observations seront évaluées.

Rééchelonnement : Les fonctions d'échelle sont normalisées par défaut. Le rééchelonnement repose entièrement sur les données d'apprentissage, même si un échantillon restant est défini (voir «Partitions», à la page 90). Si vous spécifiez une variable pour définir des partitions, il est important que ces fonctions présentent des distributions similaires à travers les échantillons de formation et les échantillons restants. Par exemple, utilisez la procédure Explorer pour examiner les distributions à travers les partitions.

Pondérations de fréquence : Cette procédure ignore les pondérations de fréquence.

Réplication de résultats : La procédure utilise la génération de nombres aléatoires pendant l'affectation 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 une valeur de départ pour le Mersenne Twister (voir «Partitions», à la page 90) 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...**

1. Spécifiez une ou plusieurs fonctions, qui peuvent être considérées comme des prédicteurs si une cible existe.

   Cible (facultative) : Si aucune cible (variable dépendante ou réponse) n'est spécifiée, la procédure cherche uniquement les voisins les plus proches de *k* : aucune classification ni prévision n'est effectuée.

   Normaliser les fonctions d'échelle : Les fonctions normalisées possèdent la même plage 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 scolaires

d'un certain quartier (l'observation focale) sont comparables à ceux de quartiers similaires. Il utilise l'analyse du voisin le plus proche pour connaître les districts scolaires les plus identiques selon un ensemble de fonctions donné. Il compare ensuite les scores 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 fonctions, dans le graphique des homologues et sur la carte des quadrants. Les informations sur les observations locales sont enregistrées dans les fichiers spécifiés sur l'onglet Sortie.

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.

Libellé d'observation (facultatif) : Les observations sont libellées à l'aide de ces valeurs sur le graphique de l'espace des fonctions, dans le graphique des homologues 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 de plusieurs variables (champs) du jeu 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.

Analyser les données : Lit les données dans le jeu de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si le jeu de données est important, cette action peut prendre un certain temps.

Affecter 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 la vue de variable 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

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 une plage de valeurs et permettre à la procédure de choisir le nombre "optimal" de voisins au sein de cette plage. La méthode pour déterminer le nombre de voisins les plus proches dépend si la sélection des fonctions est requise par l'onglet Fonctions ou non.

- Si oui, la sélection des fonctions sera alors exécutée pour chaque valeur de *k* dans la plage requise, et le *k* ainsi que l'ensemble des fonctions l'accompagnant, 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 sélection des fonctions n'est pas activée, la validation croisée de niveau *V* est utilisée pour sélectionner le nombre "optimal" de voisins. Reportez-vous à l'onglet Partitions pour contrôler l'affectation 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 fonctions selon leur importance normalisée lors du calcul des distances. L'importance des fonctions pour un prédicteur est calculée par le rapport du taux d'erreur ou l'erreur de la somme des carrés du modèle avec le prédicteur supprimé 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 fonctions 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.

Fonctions

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

Critère d'arrêt : A chaque étape, la fonction dont l'addition au modèle entraîne l'erreur la plus faible (calculée comme le taux d'erreur pour une cible catégorielle 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 fonctions spécifié :** L'algorithme ajoute un nombre fixe de fonctions 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 fonctions importantes. L'augmentation des valeurs du nombre à sélectionner capturera toutes les fonctions importantes, au risque d'ajouter des fonctions qui en réalité alimentent l'erreur du modèle.
- **Changement minimal dans le rapport d'erreur absolue :** L'algorithme prend fin lorsque le changement dans le rapport d'erreur absolue indique que le modèle ne peut pas être davantage amélioré par l'ajout de nouvelles fonctions. Indiquez un nombre positif. La diminution des valeurs pour le changement minimal aura tendance à inclure davantage de fonctions, 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 fonctions, au risque de perdre des fonctions 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 fonctions pour pouvoir déterminer quelles sont les fonctions les plus importantes. Pour plus d'informations, voir «Journal d'erreur de sélection des fonctions», à la page 95.

Partitions

L'onglet Partitions vous permet de diviser le jeu 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 du jeu de données actif en échantillons d'apprentissage et restants. 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 le jeu de données doit être affecté à l'échantillon d'apprentissage pour l'obtention d'un modèle. L'**échantillon restant** est un ensemble indépendant d'enregistrements de données

utilisé pour évaluer le modèle final ; l'erreur pour l'échantillon restant donne une estimation "honnête" de la capacité de prévision du modèle parce que les observations restantes n'ont pas été utilisées pour construire le modèle.

- **Affecter aléatoirement des observations aux partitions :** Spécifiez le pourcentage d'observations à affecter à l'échantillon d'apprentissage. Le reste est affecté à l'échantillon restant.
- **Utiliser une variable pour affecter des observations :** Indiquez une variable numérique qui affecte chaque observation du jeu de données actif à l'échantillon d'apprentissage et restant. 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 restant. Les observations contenant des valeurs système manquantes sont exclues de l'analyse. Les valeurs manquantes de 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 des fonctions 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 :** Spécifiez 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 :** Indiquez une variable numérique qui affecte chaque observation du jeu de données actif à un niveau. La variable doit être numérique et accepter des valeurs comprises entre 1 et *V*. Si une valeur manque dans cette plage, et que sur toutes les scissions les fichiers scindés sont activés, cela provoquera une erreur.

Définir la valeur de départ pour Mersenne Twister : Définir une valeur de départ vous permet de reproduire les analyses. L'utilisation de ce contrôle 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 de la valeur de départ dans cette boîte de dialogue conserve l'état actuel du générateur de nombres aléatoires et restaure cet état une fois l'analyse terminée.

Enregistrement

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 catégorie prévue :** Cette option enregistre la valeur prévue pour une cible d'échelle ou la catégorie prévue pour une cible catégorielle.
- **Probabilité prédite :** Enregistre les probabilités prévues pour une cible catégorielle. Une variable distincte est enregistrée pour chacune des *n* premières catégories, *n* étant spécifié dans le contrôle **Catégories maximales à enregistrer pour la cible catégorielle**.
- **Variable de partition de formation/traitement :** Si des observations sont affectées aléatoirement aux échantillons d'apprentissage et aux échantillons restants 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.

Sortie

Sortie du visualiseur

- **Récapitulatif du traitement des observations :** Affiche la table récapitulative de traitement des observations, qui récapitule le nombre d'observations incluses et exclues dans l'analyse, au total et par échantillon de formation et par échantillon restant.
- **Graphiques et tableaux :** Affiche les sorties liées au modèle, y compris les tableaux et les graphiques. Les tables du modèle incluent les *k* voisins les plus proches et les distances pour observations focales, classement des variables de réponse catégorielle, ainsi qu'un récapitulatif d'erreur. Les sorties graphiques dans l'affichage du modèle incluent un journal d'erreur de sélection, un graphique d'importance des fonctions, un graphique d'espace des fonctions, un graphique des homologues et une carte des quadrants. Pour plus d'informations, voir «Vue du modèle».

Fichiers

- **Exporter le modèle dans un fichier XML :** Vous pouvez utiliser ce fichier de modèle pour appliquer les informations du modèle aux autres fichiers de données à des fins d'évaluation. Cette option n'est pas disponible si des fichiers scindés ont été définis.
- **Exporter les distances entre les observations focales et les voisins les plus proches de k :** 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

Valeurs manquantes de l'utilisateur : Les variables catégorielles doivent avoir des valeurs valides pour qu'une observation puisse être incluse dans l'analyse. Ces contrôles vous permettent d'indiquer si les valeurs manquantes de l'utilisateur sont considérées comme valides parmi les variables catégorielles.

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

Vue du modèle

Lorsque vous sélectionnez **Graphiques et tableaux** dans l'onglet Sortie, la procédure produit un objet de Voisin le plus proche dans le visualiseur. 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 fonction du modèle lorsque l'utilisateur fait défiler une partie de la vue principale.

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

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

Espace des fonctions

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

Clés : En plus des valeurs de fonctions, les points du tracé 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 catégories d'une cible catégorielle, et les ombres indiquant la plage 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.

Contrôles et intéractivité : Un certain nombre de contrôles dans le graphique vous permettent d'explorer l'espace des fonctions.

- Vous pouvez choisir quel sous-ensemble de fonctions vous souhaitez afficher dans le graphique et modifier les fonctions à représenter dans les dimensions.
- Les "observations focales" sont tout simplement des points sélectionnés dans le graphique de l'espace des fonctions. 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 "contrôles" habituels pour la sélection des points sont appliqués. 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 graphique des homologues, sont automatiquement mis à jour en fonction des observations sélectionnées dans l'espace des fonctions.
- Vous pouvez modifier le nombre de voisins les plus proches (*k*) à afficher pour les observations focales.
- Positionner le curseur sur un point du graphique affiche une infobulle avec la valeur du libellé d'observation, ou le nombre d'observations si les libellés d'observation ne sont pas définis, et les valeurs des cibles observées et prévues.
- Le bouton "Réinitialiser" permet de restaurer l'état d'origine de l'espace des fonctions.

Ajout et suppression des champs et des variables

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

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 visualiseur de modèles doit être en mode Edition et une observation doit être sélectionnée dans l'espace des fonctions.

1. Pour mettre le visualiseur de modèle en mode Edition, sélectionnez à partir des menus :
   Affichage > **Mode Edition**
2. Une fois le mode Edition sélectionné, cliquez sur l'une des observations de l'espace des fonctions.
3. 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 fonctions. L'icône en regard du nom de la variable indique le niveau de mesure de celle-ci.

4. 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 fonctions. Pour afficher les zones, faites glisser une variable depuis la palette des variables ou sélectionnez **Afficher les zones**.

L'espace des fonctions 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.

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 graphique 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

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

Comportement de lien : Le graphique des homologues est relié à l'espace des fonctions de deux manières différentes.

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

Distances du voisin le plus proche

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 de libellé des observations pour l'observation focale. Si les libellés d'observations ne sont pas définis, 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 de libellé d'observation pour le *i*ème voisin le plus proche de l'observation focale. Si les libellés d'observations ne sont pas définis, 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

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

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

Journal d'erreur de sélection des fonctions

Les points du graphique affichent l'erreur (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 fonction sur l'axe *x* (plus toutes les fonctions à gauche sur l'axe *x*). Ce graphique est disponible si une cible existe et si la sélection des fonctions est activée.

Journal d'erreur de la sélection de k

Les points du graphique affichent l'erreur (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 graphique est disponible si une cible existe et si la sélection de *k* est activée.

Journal d'erreur de k et de la fonction de sélection

Il s'agit d'un graphique de sélection de fonction (voir «Journal d'erreur de sélection des fonctions»), affiché par *k*. Ce graphique est disponible si une cible existe et que *k* et la fonction de sélection sont toutes les deux activées.

Table de classification

Cette table affiche par partition la classification croisée des valeurs prévues de la cible observées contre celles prévues. Elle est disponible si une cible existe et si elle est catégorielle.

- La ligne **(Manquante)** de la partition traitée contient des observations restantes contenant des valeurs manquantes sur la cible. Ces observations contribuent à l'échantillon restant : les valeurs de pourcentage global mais pas les valeurs de pourcentage correct.

Récapitulatif d'erreur

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 catégorielle.

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 de prédicteur 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 de prédicteur, 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 étape par étape, vous pouvez découvrir que vous n'avez pas forcément besoin d'inclure les quatre variables dans la fonction.

Statistiques : Pour chaque variable : moyennes, écarts types, ANOVA à 1 facteur. Pour chaque analyse : *Test M* de Box, matrice de corrélations intra-groupe, matrice de covariance intra-groupe, matrice de covariance de chaque groupe, matrice de covariance totale. Pour chaque fonction discrimante canonique : valeur propre, pourcentage de la variance, corrélation canonique, lambda de Wilks, khi-deux. Pour chaque pas : probabilités a priori,coefficients de fonction de Fisher, coefficients de fonction non standardisés, lambda de Wilks pour chaque fonction canonique.

Remarques sur les données de l'analyse discriminante

Données : La variable de regroupement doit avoir un nombre limité de catégories 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 de prédicteur 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 catégorielle. 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

1. A partir des menus, sélectionnez :

Analyse > **Classification** > **Analyse discriminante...**

2. Sélectionnez une variable de regroupement à valeur entière et cliquez sur **Définir plage** pour spécifier les catégories à considérer.
3. Sélectionnez les variables indépendantes, ou de prédicteur. (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.)
4. 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 détaillée étape par étape :** Utilise l'analyse étape par étape pour contrôler l'entrée et la suppression de variables.
5. Vous pouvez également sélectionner les observations avec une variable de sélection.

Définition de plages pour l'analyse discriminante

Spécifiez la valeur minimum et maximum de la variable de regroupement pour l'analyse. Les observations avec des valeurs hors de cette plage 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

Pour sélectionner les observations pour votre analyse :

1. Dans la boîte de dialogue Analyse discriminante, sélectionnez une variable de sélection.
2. 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

Descriptives : Les options disponibles sont moyennes (y compris les écarts types), ANOVA à 1 facteur et Test *M* de Box.

- *Moyennes*. Affiche le total et la moyenne de groupe ainsi que l'écart type des variables indépendantes.
- *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.
- *Test de Box*. Test d'égalité des matrices de covariance des groupes. 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 coefficient de la fonction discriminante non standardisés.

Matrices : Les matrices de coefficients pour variables indépendantes disponibles sont la matrice de corrélation intra-groupe, la matrice de covariance intra-groupe, la matrice de covariance de chaque groupe et la matrice de covariance totale.

- *Corrélation intra-groupe*. Affiche une matrice de corrélations intragroupes regroupée en pool, en calculant la moyenne des matrices de covariance distinctes pour tous les groupes avant de calculer les corrélations.
- *Covariance intra-groupe*. Affiche une matrice de covariances intragroupes regroupée en pool, 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 groupe*. Affiche des matrices de covariance distinctes pour chaque groupe.
- *Covariance totale*. Affiche une matrice de covariance de toutes les observations comme si elles provenaient d'un seul échantillon.

Méthode détaillée étape par étape de l'analyse discriminante

Méthode : Sélectionnez la statistique à utiliser pour introduire ou éliminer 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 introduire une variable.

- *Lambda de Wilks*. Méthode de sélection des variables pour une analyse discriminante étape par étape qui sélectionne les variables à introduire 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 étape par étape, 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 introduire et éliminer 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 Suppression. La valeur Entrée doit être supérieure à la valeur Suppression 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 Suppression.
- *Choisir la probabilité de F*. Une variable est entrée dans le modèle si le niveau de signification de la valeur F est inférieur à la valeur Entrée ; la variable est éliminée si ce niveau est supérieur à la valeur Suppression. La valeur Entrée doit être inférieure à la valeur Suppression 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 Suppression.

Affichage : 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.

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 un groupe.

- **Egales pour tous les groupes :** Des probabilités à priori égales sont supposées pour tous les groupes; ceci n'a aucune incidence sur les coefficients.
- **A calculer selon les effectifs :** Les tailles de groupes observés dans votre échantillon déterminent les probabilités à priori du groupe d'appartenance. Par exemple, si 50% des observations comprises dans l'analyse appartiennent au premier groupe, 25 % au deuxième, et 25 % au troisième, les coefficients de classification sont ajustés pour accroître la probabilité d'affectation du premier groupe par rapport aux deux autres.

Affichage : Les options d'affichage disponibles sont les résultats par observation, la table récapitulative 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 chacun des groupes 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 de covariance : Vous pouvez choisir de classer les observations en utilisant une matrice de covariance intra-groupe ou une matrice de covariance groupe par groupe.

- *Intra-groupe*. La matrice de covariances intragroupes regroupée en pool est utilisée pour classifier les observations.
- *Groupe par groupe*. Les matrices de covariance 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.

Tracés : Les options de tracé disponibles sont groupes combinés, groupe par groupe, et carte territoriale d'affectation.

- *Tous groupes combinés*. Crée un nuage de points 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.
- *Groupe par groupe*. Crée des nuages de points groupe par groupe 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 territoriale*. Tracé 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

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. Vous pouvez utiliser ce fichier de modèle pour appliquer les informations du modèle aux autres fichiers de données à des fins d'évaluation.

Fonctions supplémentaires de la commande DISCRIMINANT

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- 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 au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 22. Analyse factorielle

L'analyse factorielle essaie d'identifier des variables sous-jacentes, ou **facteurs**, qui permettent d'expliquer le motif des corrélations à l'intérieur d'un ensemble de variables observées. L'analyse factorielle est souvent utilisée pour réduire un jeu de données. L'analyse factorielle est souvent utilisée dans la réduction de données, 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 scores factoriels, 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 une enquête politique ? L'examen des corrélations parmi les éléments d'une enquête 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 scores factoriels pour chaque répondant, facteurs que vous pouvez utiliser pour des analyses ultérieures. Par exemple, sur la base des scores factoriels, vous pouvez développer un modèle logistique de régression pour prévoir le comportement de vote.

Statistiques : Pour chaque : nombre d'observations valides, moyenne et écart type. Pour chaque analyse factorielle : matrice de corrélation des variables, incluant des niveaux 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 chargements des 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 les rotations obliques : motif et matrices de structure après rotation ; matrice de coefficients de scores factoriels et matrice de facteur de covariance. Tracés : tracé d'effondrement et carte de chargement du premier, du deuxième et du troisième facteur.

Remarques sur les données d'analyse factorielle

Données : Les variables doivent être quantitatives au niveau de l'*intervalle* ou du *rapport*. Les données catégorielles (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

1. A partir des menus, sélectionnez :

   Analyse > **Réduction des dimensions** > **Analyse factorielle...**
2. Sélectionnez les variables pour l'analyse factorielle.

Sélection des observations pour l'analyse factorielle

Pour sélectionner les observations pour votre analyse :

1. Sélectionnez une variable de sélection.
2. 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.

Caractéristiques d'analyse factorielle

Statistiques : L'option **Caractéristiques univariées** inclut 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 niveaux 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 de corrélation 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 négatifs des coefficients de corrélation partielle ; la matrice de covariance des anti-images contient les négatifs 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

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 des 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 normale 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 de corrélation initiale où les coefficients de corrélation multiple au carré sont placés sur la diagonale comme estimation initiale des qualités. Ces chargements factoriel sont utilisés 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 de Cronbach*. 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.

Affichage : Vous permet de demander la solution factorielle avant rotation et un tracé d'effondrement.

- *Structure factorielle sans rotation*. Affiche les chargements factoriels sans rotation (matrice de la structure factorielle), les qualités de représentation et les valeurs propres de la solution factorielle.
- *Tracé d'effondrement*. Tracé représentant la variance associée à chaque facteur. Permet de déterminer le nombre de facteurs à conserver. Généralement, le tracé montre une rupture 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 d'étapes que l'algorithme peut utiliser pour estimer la solution.

Rotation d'analyse factorielle

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 forts chargements 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.
- *Méthode 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 jeux de données.

Affichage : Vous permet d'inclure la sortie de la structure après rotation, et également d'afficher les tracés de chargement 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.
- *Tracé des chargements des facteurs*. Tracé en trois dimensions des chargements des trois premiers facteurs. Pour une solution à deux facteurs, un tracé en deux dimensions est affiché. Le tracé n'est pas affiché si un seul facteur est extrait. Les tracés affichent des solutions ayant subi une rotation si cette dernière est demandée.

Maximum des itérations pour converger : Vous permet de spécifier le nombre maximum d'étapes que l'algorithme peut utiliser pour réaliser la rotation.

Scores d'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 scores factoriels sont la Régression, Bartlett, et Anderson-Rubin.

- *Méthode de régression*. Méthode d'estimation des coefficients de scores factoriels. Les écarts obtenus ont une moyenne de 0 et une variance égale au carré de la corrélation multiple entre les scores factoriels estimés et les vraies valeurs du facteur. Les écarts peuvent être corrélés même lorsque les facteurs sont orthogonaux.
- *Scores de Bartlett*. Méthode d'estimation des coefficients de scores factoriels. Les scores 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 de scores 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 des scores factoriels : Vous permet de montrer les coefficients par lesquels les variables sont multipliées pour obtenir les scores factoriels. Cela permet également de montrer les corrélations entre les scores factoriels.

Options d'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 sortie. Triez les coefficients par leur taille et supprimez les coefficients dont la valeur absolue est inférieure à la valeur spécifiée.

Fonctions supplémentaires de la commande FACTOR

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- spécifier des critères de convergence pour les itérations lors de l'extraction et de la rotation ;
- définir des tracés factoriels individuels après rotation ;
- indiquer le nombre de scores factoriels à enregistrer ;
- spécifier les valeurs des diagonales pour la méthode de factorisation en axes principaux ;

- créer des matrices de corrélation ou des matrices de chargement factoriel sur un disque pour une analyse ultérieure ;
- lire et analyser des matrices de corrélation ou des matrices de chargement factoriel.

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 23. Choix d'une procédure de classification

Vous pouvez effectuer des analyses de cluster à l'aide de la procédure TwoStep, hiérarchique ou de nuées dynamiques. Chaque procédure utilise un algorithme différent pour la création des clusters, et chacune d'elles comporte des options qui ne sont pas disponibles dans les autres procédures.

Analyse de cluster TwoStep : La procédure Analyse de cluster TwoStep est la méthode privilégiée pour de nombreuses applications. Elle offre les fonctions spécifiques suivantes :

- Sélection automatique du meilleur nombre de clusters, en plus des mesures de sélection parmi des modèles de cluster.
- Possibilité de créer simultanément des modèles de cluster sur la base de variables catégorielles et continues.
- Possibilité d'enregistrer le modèle de cluster dans un fichier XML externe, puis de lire ce fichier et de mettre à jour le modèle de cluster à l'aide des données les plus récentes.

En outre, la procédure Analyse de cluster TwoStep permet d'analyser des fichiers de données volumineux.

Analyse de cluster hiérarchique : La procédure d'analyse de cluster hiérarchique est limitée à des fichiers de données plus petits (centaines d'objets à classer), mais offre les fonctions spécifiques suivantes :

- Possibilité de classer des observations ou des variables.
- Possibilité de calculer plusieurs solutions possibles et d'enregistrer des clusters d'affectation pour chacune de ces solutions.
- Plusieurs méthodes de formation de clusters, de transformation de variables et de mesure de la dissimilarité entre les clusters.

Tant que toutes les variables sont du même type, la procédure d'analyse de cluster hiérarchique peut analyser des variables d'intervalle (continues), d'effectif ou binaires.

Analyse de cluster des nuées dynamiques : La procédure d'analyse de cluster de nuées dynamiques est limitée aux données continues et exige que vous indiquiez au préalable le nombre de clusters. Elle offre néanmoins les fonctions spécifiques suivantes :

- Possibilité d'enregistrer les distances à partir des centres de clusters pour chaque objet.
- Possibilité de lire les centres de clusters initiaux à partir d'un fichier IBM SPSS Statistics externe et d'enregistrer les centres de clusters finaux dans ce même fichier.

En outre, la procédure d'analyse de cluster de nuées dynamiques permet d'analyser des fichiers de données volumineux.

Chapitre 24. analyse de cluster TwoStep

L'analyse de cluster TwoStep est un outil d'exploration conçu pour révéler des groupements naturels (ou clusters) au sein d'un jeu de données. L'algorithme utilisé par cette procédure possède plusieurs fonctions qui le distinguent des techniques de classification standard :

- **Gestion des données catégorielles et continues :** En supposant que les variables soient indépendantes, une distribution jointe multinomiale-normale peut être placée sur des variables catégorielles et continues.
- **Sélection automatique du nombre de clusters :** En comparant les valeurs d'un critère de modèle-choix dans différentes solutions de classification, la procédure peut déterminer automatiquement le nombre optimal de clusters.
- **Evolutivité :** En construisant une arborescence de fonctions de cluster (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.

Mesure de distance : Cette sélection détermine la façon dont la similarité entre deux clusters est calculée.

- **Log de 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 catégorielles 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 clusters. Elle peut être utilisée uniquement lorsque toutes les variables sont continues.

Nombre de clusters : Cette sélection vous permet d'indiquer la façon dont le nombre de clusters doit être déterminé.

- **Déterminer automatiquement :** Cette procédure déterminera automatiquement le « meilleur » nombre de clusters 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 clusters que la procédure doit prendre en compte.
- **Indiquer une valeur fixe :** Vous permet d'indiquer le nombre de clusters (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, voir «Options de la procédure d'analyse de cluster TwoStep», à la page 112.

Critère de classification : Cette sélection détermine la façon dont l'algorithme de classification automatique détermine le nombre de clusters. Vous pouvez spécifier le critère d'information bayésien (BIC) ou le critère d'information d'Akaike (AIC).

Remarques sur les données de la procédure d'analyse de cluster TwoStep

Données : Cette procédure fonctionne avec des variables continues et catégorielles. 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 fonctions de cluster 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 cluster sont indépendantes. De plus, chaque variable continue est considérée comme ayant une distribution normale (gaussienne) et chaque variable catégorielle 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 des deux variables continues. Utilisez la procédure Tableaux croisés pour tester l'indépendance de deux variables catégorielles. Utilisez la procédure Moyennes pour tester l'indépendance entre une variable continue et une variable catégorielle. Utilisez la procédure Explorer pour tester la normalité d'une variable continue. Utilisez la procédure Test du Khi-deux pour tester si une variable catégorielle possède une distribution multinomiale spécifique.

Pour effectuer une procédure d'analyse de cluster TwoStep

1. A partir des menus, sélectionnez :

   Analyse > **Classification** > **Cluster TwoStep...**
2. Sélectionnez une ou plusieurs variables catégorielles ou continues.

Sinon, vous pouvez :

- Ajuster les critères sur lesquels est basée la construction des clusters.
- Sélectionner les paramètres de gestion du bruit, d'affectation de mémoire, de standardisation de variable et d'entrée de modèle de cluster.
- Demander les sorties 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 de cluster TwoStep

Traitement des valeurs extrêmes : Ce groupe permet de traiter les valeurs extrêmes, notamment lors de la classification, si l'arborescence des fonctions de cluster (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 extrêmes sont placées dans l'arborescence CF si cela est possible. Sinon, les valeurs extrêmes 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 un cluster deviennent des valeurs extrêmes libellées. Un numéro d'identification de -1 est affecté au cluster de valeur extrême et cette dernière n'est pas prise en compte dans le nombre de clusters.

Allocation de mémoire : Ce groupe vous permet d'indiquer la quantité de mémoire maximale en mégaoctets (Mo) que l'algorithme de cluster 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 spécifié de clusters 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 :** Indique le nombre maximal de noeuds de l'arborescence CF pouvant être générés par la procédure, d'après la fonction $(b^{d+1} - 1) / (b - 1)$, où b correspond au 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 cluster : Ce groupe vous permet d'importer et de mettre à jour un modèle de cluster 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, voir «Sortie de l'analyse de cluster TwoStep», à la page 114.

Si vous avez indiqué la mise à jour d'un modèle de cluster, 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 cluster, la procédure considère qu'aucune des observations sélectionnées dans le jeu de données actif n'a été utilisée pour créer le modèle de cluster 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 catégorielles 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 de cluster TwoStep sur les groupes d'observations combinés pour obtenir de meilleurs résultats.

Sortie de l'analyse de cluster TwoStep

Sortie : Ce groupe fournit des options d'affichage pour les résultats de la classification.

- **Tableaux croisés dynamiques :** Les résultats sont affichés dans des tableaux croisés dynamiques.
- **Tableaux et graphiques dans le visualiseur de modèle :** Les résultats sont affichés dans le visualiseur de modèle.
- **Champs d'évaluation :** Calcule les données de cluster pour les variables non utilisées dans la création des clusters. Les champs d'évaluation peuvent être affichés en même temps que les fonctions 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 le jeu de données actif.

- **Créer une variable d'appartenance à un cluster :** Cette variable contient un numéro d'identification de cluster 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 du jeu de données actif effectuée par cette procédure au cours d'une session.

Fichiers XML : Le modèle de cluster final et l'arborescence CF représentent deux types de fichiers de sortie pouvant être exportés au format XML.

- **Exporter le modèle final :** Le modèle de cluster 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 cluster et de le mettre à jour ultérieurement en utilisant des données plus récentes.

Visualiseur de clusters

Les modèles de cluster sont généralement utilisés pour trouver des groupes (ou des clusters) 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 l'analyse de cluster 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 cluster :

- Examiner les clusters afin de déterminer les caractéristiques uniques de ce cluster. *Est-ce qu'un cluster contient tous les emprunteurs à revenus élevés ? Est-ce que ce cluster contient davantage d'enregistrements que les autres ?*
- Examiner les champs des clusters afin de déterminer comment les valeurs sont distribuées parmi les clusters. *Est-ce que le niveau d'éducation détermine l'appartenance à un cluster ? Est-ce qu'une cote de solvabilité élevée permet de distinguer l'appartenance à un cluster spécifique ?*

Grâce à l'utilisation des vues principales et des différentes vues liées dans le visualiseur de clusters, vous pouvez avoir un bon aperçu qui vous aidera à répondre à ces questions.

Pour consulter les informations concernant un modèle de cluster, activez (double-cliquez) sur l'objet visualiseur de modèle dans le visualiseur.

Visualiseur de clusters

Le visualiseur de cluster 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, voir «Vue récapitulative du modèle».
- Clusters. Pour plus d'informations, voir «Vue des clusters».

Il existe quatre vues liées/auxiliaires :

- Importance des prédicteurs. Pour plus d'informations, voir «Vue de l'importance des prédicteurs de cluster», à la page 117.
- Taille des clusters (par défaut). Pour plus d'informations, voir «Vue de la taille des clusters», à la page 117.
- Distribution des cellules. Pour plus d'informations, voir «Vue de la distribution des cellules», à la page 117.
- Comparaison des clusters. Pour plus d'informations, voir «Vue de comparaison des clusters», à la page 117.

Vue récapitulative du modèle

La vue récapitulative du modèle affiche un instantané, ou un récapitulatif, du modèle de cluster, y compris une mesure par silhouette de la cohésion et de la séparation des clusters 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 cluster 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 du cluster. Dans la vue récapitulative du modèle, d'après le classement de Kaufman et Rousseeuw, un bon résultat équivaut à des données qui indiquent une preuve raisonnable ou forte de la structure du cluster, 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 son cluster et B correspond à la distance de l'enregistrement au centre du cluster le plus proche auquel il n'appartient pas. Un coefficient de silhouette de 1 signifie que toutes les observations sont situées directement au centre de leurs clusters. Une valeur de −1 signifie que toutes les observations sont situées au centre de cluster d'autres clusters. Une valeur de 0 signifie, en moyenne, que les observations sont équidistantes du centre de leur propre cluster et du centre de l'autre cluster la plus proche.

Le récapitulatif inclut un tableau qui contient les informations suivantes :

- **Algorithme** : L'algorithme de classification utilisé, par exemple "TwoStep".
- **Fonctions d'entrée** : Le nombre de champs, ou d'**entrées** ou de **prédicteurs**.
- **Clusters** : Le nombre de clusters dans la solution.

Vue des clusters

La vue des clusters contient une grille présentant les clusters par fonctions, qui inclut les noms des clusters, leurs tailles et les profils de chaque cluster.

Les colonnes de la grille contiennent les informations suivantes :

- **Cluster** : Les nombres de clusters créées par l'algorithme.
- **Libellé** : Tous les libellés appliqués à chaque cluster (celle-ci est vierge par défaut). Double-cliquez dans la cellule pour saisir un libellé qui décrit les contenus de cluster ; par exemple "Acheteurs de voitures de luxe".
- **Description** : Toutes les descriptions du contenu de cluster (celle-ci est vierge par défaut). Double-cliquez dans la cellule pour saisir une description du cluster ; par exemple "professionnels, plus de 55 ans, gagnant plus de 100 000 $".

- **Taille :** La taille de chaque cluster sous la forme d'un pourcentage de l'échantillon général des clusters. Chaque cellule de taille de la grille affiche une barre verticale qui indique le pourcentage de taille au sein du cluster, un pourcentage de taille au format numérique et les effectifs d'observation du cluster.
- **Fonctions :** Les entrées ou prédicteurs individuels, triés par importance générale par défaut. Si des colonnes ont la même taille, elles sont affichées par ordre croissant des numéros de cluster.

  L'importance des fonctions générales est indiquée par la couleur d'ombrage de l'arrière-plan de la cellule ; la fonction 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 fonction.

Lorsque vous passez la souris sur une cellule, le nom complet/le libellé de la fonction 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 fonction. Dans la vue Centres de clusters, cela inclut les statistiques et la valeur de la cellule ; par exemple : "Moyenne : 4,32". Pour les fonctions catégorielles, la cellule affiche le nom de la catégorie la plus fréquente (modale) et son pourcentage.

Dans la vue des clusters, vous pouvez sélectionner plusieurs manières d'afficher les informations des clusters :

- Transposer les clusters et les fonctions. Pour plus d'informations, voir «Transposition des clusters et des fonctions».
- Trier les fonctions. Pour plus d'informations, voir «Tri des fonctions».
- Trier les clusters. Pour plus d'informations, voir «Tri des clusters».
- Sélectionner le contenu des cellules. Pour plus d'informations, voir «Contenu des cellules».

Transposition des clusters et des fonctions : Par défaut, les clusters sont affichés en tant que colonnes et les fonctions sont affichées en tant que lignes. Pour inverser cet affichage, cliquez sur le bouton **Transposer les clusters et les fonctions** à gauche des boutons **Trier les fonctions par**. Par exemple, vous pouvez réaliser ceci lorsque de nombreux clusters sont affichés, afin de réduire le défilement horizontal nécessaire pour visualiser les données.

Tri des fonctions : Le bouton **Trier les fonctions par** vous permet de sélectionner la façon dont les cellules de fonctions sont affichées :

- **Importance générale** : Il s'agit de l'ordre de tri par défaut. Les fonctions sont triées en ordre décroissant de l'importance générale, et l'ordre de tri est le même pour tous les clusters. Si des fonctions ont des valeurs d'importance liées, les fonctions liées sont répertoriées par ordre croissant des noms de fonctions.
- **Importance intra-cluster** : Les fonctions sont triées par rapport à leur importance pour chaque cluster. Si des fonctions ont des valeurs d'importance liées, les fonctions liées sont répertoriées par ordre croissant des noms de fonctions. Lorsque cette option est sélectionnée, l'ordre de tri varie généralement d'un clusters à l'autre.
- **Nom** : Les fonctions sont triées par nom dans l'ordre alphabétique.
- **Ordre des données** : Les fonctions sont triées selon leur ordre dans le jeu de données.

Tri des clusters : Par défaut, les clusters sont triées en ordre de taille décroissante. Les boutons **Trier les clusters par** vous permettent de les trier par nom dans l'ordre alphabétique, ou, si vous avez créé des libellés uniques, par ordre alphabétique des libellés.

Les fonctions ayant le même libellé sont triées selon le nom de cluster. Si des clusters sont triées par libellé et que vous modifiez le libellé d'un cluster, 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 fonctions et d'évaluation.

- **Centre de cluster :** Par défaut, les cellules affichent les noms/libellés des fonctions et la tendance centrale pour chaque combinaison de cluster/fonction. 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 catégoriels.
- **Distributions absolues :** Affiche des noms/libellés de fonctions et des distributions absolues des fonctions dans chaque cluster. Pour les fonctions catégorielles, l'affichage montre des graphiques à barres où sont superposées des catégories en ordre croissant de la valeur des données. Pour les fonctions continues, l'affichage montre un tracé de densité lissé qui utilise les mêmes points finaux et intervalles pour chaque cluster.

  L'affichage en rouge uni montre la distribution des clusters, alors qu'un affichage plus pâle représente les données générales.
- **Distributions relatives :** Affiche les noms/libellés des fonctions 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 clusters, alors qu'un affichage plus pâle représente les données générales.
- **Vue de base :** Là où il y a beaucoup de clusters, 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 prédicteurs de cluster

La vue de l'importance des prédicteurs affiche l'importance relative de chaque champ dans l'estimation du modèle.

Vue de la taille des clusters

La vue de la taille des clusters affiche un graphique circulaire qui contient chaque cluster. La taille en pourcentage de chaque cluster est affichée sur chaque tranche ; passez la souris sur chaque tranche pour afficher l'effectif de celle-ci.

En dessous du graphique, un tableau répertorie les informations suivantes sur la taille :

- La taille du cluster le plus petit (en effectif et en pourcentage de l'ensemble).
- La taille du cluster le plus grand (en effectif et en pourcentage de l'ensemble).
- Le rapport de taille du cluster le plus grand par rapport au cluster le plus petit.

Vue de la distribution des cellules

La vue de la distribution des cellules affiche un tracé étendu et plus détaillé de la distribution des données pour toutes les cellules de fonction que vous sélectionnez dans le tableau du panneau principal des clusters.

Vue de comparaison des clusters

La vue de comparaison des clusters se compose d'une présentation sous forme de grille avec des fonctions dans les lignes et des clusters sélectionnés dans les colonnes. Cette vue vous aide à mieux comprendre les facteurs qui composent les clusters ; elle vous permet aussi de voir les différences entre les clusters non seulement par comparaison avec les données générales mais aussi en comparant les classes les unes aux autres.

Pour sélectionner des clusters à afficher, cliquez en haut de la colonne de cluster sur le panneau principal des clusters. Cliquez en maintenant la touche Ctrl ou Maj enfoncée pour sélectionner ou désélectionner plus d'un cluster pour la comparaison.

Remarque : Vous pouvez sélectionner jusqu'à cinq clusters à afficher.

Les clusters sont affichés dans l'ordre où ils ont été sélectionnés, alors que l'ordre des champs est déterminé par l'option **Trier les fonctions par**. Lorsque vous sélectionnez **Importance intra-cluster**, les champs sont toujours triés par ordre d'importance générale.

Les tracés d'arrière-plan affichent les distributions générales de chaque fonction :

- Les fonctions catégorielles sont affichées sous forme de tracés de points, où la taille des points indique la catégorie la plus fréquente/modale pour chaque cluster (par fonction).
- Les fonctions continues seront affichées sous forme de boîtes à moustaches, qui affichent les médianes générales et les plages interquartiles.

Des boîtes à moustaches des clusters sélectionnés recouvrent ces vues d'arrière-plan :

- Pour les fonctions continues, des marqueurs en points carrés et des lignes horizontales indiquent la médiane et la plage interquartile de chaque cluster.
- Chaque cluster est représenté par une couleur différente, affichée en haut de la vue.

Navigation dans le visualiseur de cluster

Le visualiseur de cluster est un affichage interactif. Vous pouvez :

- sélectionner un champ ou un cluster pour afficher davantage de détails ;
- comparer des clusters 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 contrôles des barres d'outils. En outre, vous pouvez aussi réinitialiser le visualiseur à ses paramètres par défaut et ouvrir une boîte de dialogue pour spécifier le contenu de la vue des clusters dans le panneau principal.

Les options **Trier les fonctions par**, **Trier les clusters par**, **Cellules** et **Affichage** ne sont disponibles que lorsque vous sélectionnez la vue **Clusters** du panneau principal. Pour plus d'informations, voir «Vue des clusters», à la page 115.

Tableau 2. Icônes de la barre d'outils.

Icône	Rubrique
	Voir Transposer les clusters et les fonctions.
	Voir Trier les fonctions par.
	Voir Trier les clusters par.
	Voir Cellules.

Contrôler l'affichage de la vue des clusters

Pour contrôler ce qui est affiché dans la vue des clusters sur le panneau principal, cliquez sur le bouton **Afficher**. La boîte de dialogue Afficher s'ouvre.

Fonctions : Sélectionné par défaut. Pour masquer toutes les fonctions 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 cluster, mais envoyés au visualiseur de modèle pour évaluer les clusters) ; par défaut, aucun n'est affiché. *Remarque :* Le champ d'évaluation doit être une chaîne comportant plusieurs valeurs. Cette case à cocher n'est pas utilisable si aucun champ d'évaluation n'est disponible.

Descriptions des clusters : Sélectionné par défaut. Pour masquer toutes les cellules de description des clusters, décochez la case.

Tailles des clusters : Sélectionné par défaut. Pour masquer toutes les cellules de taille des clusters, décochez la case.

Nombre maximal de catégories : Spécifiez le nombre maximal de catégorie à afficher dans les graphiques de fonctions catégorielles ; la valeur par défaut est de 20.

Filtrage des enregistrements

Si vous souhaitez en savoir plus sur les observations d'un cluster particulière ou d'un groupe de clusters, vous pouvez sélectionner un sous-ensemble d'enregistrements afin de poursuivre l'analyse en fonction des clusters sélectionnées.

1. Sélectionnez les clusters dans la vue des clusters du visualiseur de cluster. Pour sélectionner plusieurs clusters, cliquez tout en maintenant la touche Ctrl enfoncée.
2. A partir des menus, sélectionnez :

   Générer > **Filtrer les enregistrements...**
3. Entrez le nom d'une variable de filtre. Les enregistrements des clusters sélectionnés 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 le statut du filtre.
4. Cliquez sur **OK**.

Chapitre 25. Analyse de cluster hiérarchique

L'analyse de cluster hiérarchique tente d'identifier les groupes d'observations (ou de variables) relativement homogènes basées sur des caractéristiques sélectionnées, en utilisant un algorithme qui débute avec chaque observation (ou variable) dans un cluster séparée et qui combine les clusters 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 groupes identifiables de spectacles télévisuels qui attirent des audiences similaires à l'intérieur de chaque groupe ? Avec une analyse de cluster hiérarchique, vous pouvez reclasser les spectacles télévisuels (observations) en groupes 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 : Planning des agglomérations, matrice de distances (ou des similarités) et cluster d'affectation pour une seule solution ou un ensemble de solutions. Tracés : dendrogrammes et tracés en stalactite.

Remarques sur les données de l'analyse de cluster hiérarchique

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 clusters. 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 l'analyse de cluster hiérarchique).

Tri par observation : Si des distances ex aequo ou des similitudes se présentent dans les données d'entrée ou entre les clusters mis à jour au cours de l'opération de jointure, la solution de cluster 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 l'analyse de cluster 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 analyse de cluster hiérarchique

1. A partir des menus, sélectionnez :

   Analyse > **Classification** > **Cluster hiérarchique...**
2. 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 libeller les observations.

Méthode d'analyse de cluster hiérarchique

Méthode d'agrégation : Les choix disponibles sont : la Distance moyenne entre groupes, la Distance moyenne dans les groupes, l'Agrégation suivant le saut minimum, l'Agrégation suivant le diamètre, les centroïdes, 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.
- **Effectifs :** 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, la différence 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 les scores *z*, la plage -1 à 1, la plage 0 à 1, l'amplitude maximale de 1, la moyenne de 1 et l'écart type de 1.

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 choix possibles sont Valeurs absolues, Inverser le signe et Rééchelonner entre 0 à 1.

Statistiques de l'analyse de cluster hiérarchique

Planning des agglomérations : Affiche les observations ou les clusters combinés à chaque étape, les distances entre les observations ou les clusters en cours de combinaison, et le dernier niveau de cluster auquel une observation (ou une variable) a rejoint le cluster.

Matrice des distances : Indique les distances ou les similarités entre éléments.

Cluster(s) d'affectation : Affiche le cluster auquel chaque observation appartient lors d'une ou de plusieurs étapes de la combinaison de clusters. Les options disponibles sont Une seule partition, Plusieurs partitions ou Aucune.

Tracés (graphiques) de l'analyse de cluster hiérarchique

Dendrogramme : Affiche un *dendrogramme*. Les dendrogrammes peuvent être utilisés pour évaluer la cohésion des clusters formés et ils fournissent des renseignements sur le nombre approprié de clusters à conserver.

Stalactites : Affiche un *tracé en stalactite*, incluant tous les clusters ou une plage de clusters spécifiée. Les tracés 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 tracé vertical ou horizontal.

Sauvegarde des nouvelles variables de l'analyse de cluster hiérarchique

Cluster(s) d'affectation : Vous permet de sauvegarder les clusters 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.

Fonctions 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 également d'effectuer les actions suivantes :

- 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.

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 26. analyse de cluster des 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 clusters. Vous pouvez indiquer les centres de cluster initiaux si vous connaissez cette information. Vous pouvez choisir entre deux méthodes de classement des observations, soit la mise à jour des centres de cluster de façon itérative, soit la classification seule. Vous pouvez enregistrer l'appartenance à un cluster, les informations de distance et les centres de clusters finaux. Vous pouvez éventuellement indiquer une variable dont les valeurs servent à libeller les sorties par observations. Vous pouvez également demander des statistiques *F* d'analyse de 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 de *cluster de 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 clusters initiaux, tableau ANOVA. Chaque observation : informations de cluster, distance au centre de cluster.

Considérations de données sur l'analyse de cluster de nuées dynamiques

Données : Les variables doivent être quantitatives au niveau intervalle ou rapport. Si vos variables sont binaires ou sont des effectifs, utilisez la procédure d'analyse de cluster hiérarchique.

Ordre des observations et des centres de cluster initiaux : L'algorithme par défaut permettant de choisir les centres de cluster 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 cluster 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 cluster 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 cluster initiaux risque d'affecter la solution s'il existe des distances ex aequo entre les observations et les centres de cluster. 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 d'analyse de cluster 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 de cluster de *nuées dynamiques* (cela peut être fait dans la procédure Descriptives). La procédure suppose que vous avez sélectionné le nombre voulu de clusters et que vous avez inclus toutes les variables pertinentes. Si vous avez choisi un nombre de clusters inadéquat ou omis de variables importantes, vos résultats risquent d'être erronés.

Obtenir une analyse de cluster de nuées dynamiques

1. A partir des menus, sélectionnez :

Analyse > **Classification** > **Cluster de nuées dynamiques...**

2. Sélectionnez les variables à utiliser dans l'analyse de cluster.
3. Spécifiez le nombre de clusters. Le nombre de clusters doit être au moins de deux et ne doit pas être supérieur au nombre d'observations contenues dans le fichier de données.
4. Sélectionnez soit la méthode **Itérer et classer** soit la méthode **Classer seulement**.
5. Vous avez la possibilité de sélectionner une variable d'identification pour libeller les observations.

Efficacité de l'analyse de cluster de nuées dynamiques

La commande d'analyse de cluster de *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 cluster. Sélectionnez **Ecrire les centres finaux dans Fichier**. Ensuite, restaurez la totalité du fichier de données et sélectionnez la méthode**Classer 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 jeu de données. Les jeux 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 jeux de données doit être conforme aux règles de dénomination de variables. Pour plus d'informations, voir .

Itération de l'analyse de cluster de nuées dynamiques

Remarque : Ces options sont disponibles uniquement si vous avez sélectionné la méthode **Itérer et classer** dans la boîte de dialogue Analyse de cluster 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 cluster 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 de cluster 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 de cluster après l'affectation de chaque observation. Si vous ne sélectionnez pas cette option, les nouveaux centres de cluster seront calculés lorsque toutes les observations auront été affectées.

Enregistrement des analyses de cluster de nuées dynamiques

Vous pouvez enregistrer des informations sur la solution relatives aux nouvelles variables à utiliser dans les analyses ultérieures :

Cluster(s) d'affectation : Crée une nouvelle variable indiquant le cluster d'affectation finale de chaque observation. Les valeurs de la nouvelle variable vont de 1 au nombre de clusters.

Distance au centre de cluster : Crée une nouvelle variable indiquant la distance euclidienne entre chaque nouvelle variable et son centre de classification.

Options d'analyses de cluster de nuées dynamiques

Statistiques : Les statistiques suivantes sont disponibles : Centres de cluster initiaux, Tableau ANOVA et Affectations et distances au centre.

- *Centres de cluster initiaux*. Première estimation des moyennes de variables de chacun des clusters. Par défaut, le nombre d'observations assez espacées sélectionné dans les données est égal au nombre de clusters. Les centres de cluster initiaux sont utilisés pour une première classification et sont ensuite mis à jour.
- *Tableau ANOVA*. Affiche un tableau d'analyse de 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 un seul cluster.
- *Affectations et distances au centre*. Affiche pour chaque observation l'affectation de cluster finale et la distance euclidienne entre l'observation et le centre de cluster utilisé pour classer l'observation. Affiche également la distance euclidienne entre les centres de cluster 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 clusters basés sur des distances calculées à partir de toutes les variables n'ayant pas de valeur manquante.

Fonctions supplémentaires de la commande QUICK CLUSTER

La procédure de Cluster de nuées dynamiques utilise la syntaxe de commande `QUICK CLUSTER`. Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Accepter les premières observations *k* comme centres de cluster initiaux, évitant ainsi le passage de données normalement utilisé pour les estimer.
- Spécifier les centres de cluster initiaux directement comme faisant partie de la syntaxe de commande.
- Spécifier les noms des variables enregistrées.

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 27. Tests non paramétriques

Les tests non paramétriques effectuent des hypothèses minimales concernant 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 catégories 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.

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 catégoriels avec seulement deux catégories, le test du khi-deux à tous les autres champs catégoriels 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 caractère 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.

Obtention des tests non paramétriques à un échantillon

A partir des menus, sélectionnez :

Analyse > **Tests non paramétriques** > **Un échantillon...**

1. Cliquez sur **Exécuter**.

Sinon, vous pouvez :

- Spécifier un objectif dans l'onglet Objectif.
- Spécifier les affectations de champ dans l'onglet Champs.
- Spécifier les paramètres d'expert dans l'onglet Paramètres.

Onglet Champs

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 champ 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 des tests

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 catégoriels avec seulement deux catégories valides (non manquantes), le test du khi-deux à tous les autres champs catégoriels 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 indicateur (un champ catégoriel avec seulement deux catégories) est semblable à ce qui est attendu d'une distribution binomiale spécifiée. De plus, vous pouvez demander des intervalles de confiance. Pour en savoir plus sur les paramètres de test, voir «Options du Test binomial».
- **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 catégories d'un champ. Pour en savoir plus sur les paramètres de test, voir «Options du Test de khi-deux», à la page 131.
- **Tester la distribution observée par rapport à la distribution hypothétique (test de Kolmogorov-Smirnov) :** Le test de Kolmogorov-Smirnov s'applique aux champs continus et ordinaux. 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. Pour en savoir plus sur les paramètres de test, voir «Options Kolmogorov-Smirnov», à la page 131.
- **Comparer la médiane à la valeur hypothétique (test de Wilcoxon) :** Le test de Wilcoxon s'applique aux champs continus et ordinaux. 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. Pour en savoir plus sur les paramètres de test, voir «Options Suites en séquence», à la page 131.

Options du Test binomial : Le test binomial est destiné aux champs indicateurs (champs catégoriels avec seulement deux catégories) mais peut s'appliquer à tous les champs en utilisant des règles pour définir le "succès".

Proportion hypothétique : Indique la proportion attendue d'enregistrements définis en tant que "succès" ou *p*. Spécifiez 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 catégoriels : Indique comment le "succès, à savoir les valeurs de données testées par rapport à la proportion hypothétique, est défini pour les champs catégoriels.

- **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 catégoriels 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 : Indique comment le "succès", à savoir les valeurs de données testées para 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 : Toutes les catégories ont la même probabilité : Produit des fréquences égales pour toutes les catégories de l'échantillon. Il s'agit de la valeur par défaut.

Personnaliser la probabilité prévue : Permet de spécifier les fréquences inégales pour une liste de catégories spécifiée. Spécifiez 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 **Catégorie** , spécifiez les valeurs des catégories. Dans la colonne **Fréquence relative** , spécifiez une valeur supérieure à 0 pour chaque catégorie. Les fréquences personnalisées sont traitées comme des rapports. 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 catégorie, 1/3 dans la seconde et 1/2 dans la troisième. Lorsque les probabilités attendues personnalisées sont spécifiées, les valeurs de catégories personnalisées doivent inclure toutes les valeurs de champ dans les données ; sinon, le test n'est pas exécuté pour ce champ.

Options Kolmogorov-Smirnov : Cette boîte de dialogue indique les distributions à tester et les paramètres des distributions hypothétiques.

Normale : L'option **Utiliser des données d'échantillon** s'appuie sur la moyenne et l'écart type observés, tandis que l'option **Personnalisé** vous permet de spécifier des valeurs.

Uniforme : L'option **Utiliser des données d'échantillon** s'appuie sur le minimum et le maximum observés, tandis que l'option **Personnalisé** vous permet de spécifier des valeurs.

Exponentielle : L'option **Moyenne d'échantillon** s'appuie sur la moyenne observée, tandis que l'option **Personnalisé** vous permet de spécifier des valeurs.

Poisson : L'option **Moyenne d'échantillon** s'appuie sur la moyenne observée, tandis que l'option **Personnalisé** vous permet de spécifier des valeurs.

Options Suites en séquence : Les suites en séquence sont destinées aux champs indicateurs (champs catégoriels avec seulement deux catégories) mais peuvent s'appliquer à tous les champs en utilisant des règles pour définir les groupes.

Définir des groupes pour des champs catégoriels : Les options suivantes sont disponibles :

- **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 catégoriels 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 : Indique 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

Niveau de signification : Indique le niveau de signification (alpha) pour tous les tests. Spécifiez une valeur numérique comprise entre 0 et 1. La valeur par défaut est 0.05.

Intervalle de confiance (%) : Indique le niveau de confiance pour tous les intervalles de confiance générés. Spécifiez une valeur numérique comprise entre 0 et 100. La valeur par défaut est 95.

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 de l'utilisateur

Valeurs manquantes de l'utilisateur pour les champs catégoriels : Les champs catégoriels doivent avoir des valeurs valides pour qu'un enregistrement puisse être inclus dans l'analyse. Ces contrôles vous permettent d'indiquer si les valeurs manquantes de l'utilisateur sont considérées comme valides parmi les champs catégoriels. Les valeurs système manquantes et les valeurs manquantes pour les champs continus sont toujours considérées comme non valides.

Fonctions supplémentaires de la commande NPTESTS

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- spécifier des tests à un échantillon, pour échantillon indépendants et pour échantillons liés en n'exécutant la procédure qu'une seule fois.

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

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 de plusieurs tests non paramétriques. Les tests non paramétriques ne supposent pas que vos données suivent la distribution normale.

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.

Obtention des tests non paramétriques pour échantillons indépendants

A partir des menus, sélectionnez :

Analyse > **Tests non paramétriques** > **Echantillons indépendants...**

1. Cliquez sur **Exécuter**.

Sinon, vous pouvez :

- Spécifier un objectif dans l'onglet Objectif.
- Spécifier les affectations de champ dans l'onglet Champs.
- Spécifier les paramètres d'expert dans l'onglet Paramètres.

Onglet Champs

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 et ordinaux avec un rôle prédéfini de Cible ou Les deux seront utilisés comme champs de test. Si un champ unique catégoriel 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 champ 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 ou ordinaux.
- **Groupes :** Sélectionnez un champ catégoriel.

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**.

Choix des tests

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 **étape par étape 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 l'ampleur de la réponse. Ici, l'hypothèse alternative est ordonnée ; le test de Jonckheere-Terpstra est donc le plus approprié. **Du plus petit au plus grand** spécifie l'hypothèse alternative suivante : le paramètre d'emplacement du premier groupe est inférieur ou égal au deuxième, lui-même inférieur ou égal au troisième, et ainsi de suite. **Du plus grand au plus petit** spécifie l'hypothèse alternative suivante : le paramètre d'emplacement du premier groupe est supérieur ou égal au deuxième, lui-même supérieur ou égal au troisième, et ainsi de suite. Pour ces deux options, l'hypothèse alternative suppose aussi que les emplacements ne sont pas tous égaux. En option, vous pouvez demander des comparaisons multiples des *k* échantillons, soit toutes les comparaisons multiples **par paire** soit les comparaisons **étape par étape descendantes** .

- **Comparer les plages entre les groupes :** Cela génère des tests pour échantillons indépendants indiquant si les échantillons ont la même plage. La **Réaction extrême de Moses (2 échantillons)** teste un groupe de contrôles par rapport à un groupe de comparaisons. La première valeur dans l'ordre croissant du champ de regroupement définit le groupe de contrôles 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 regroupé en pool (calculée à partir de tous les enregistrements du jeu 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 **étape par étape 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

Niveau de signification : Indique le niveau de signification (alpha) pour tous les tests. Spécifiez une valeur numérique comprise entre 0 et 1. La valeur par défaut est 0.05.

Intervalle de confiance (%) : Indique le niveau de confiance pour tous les intervalles de confiance générés. Spécifiez une valeur numérique comprise entre 0 et 100. La valeur par défaut est 95.

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 de l'utilisateur

Valeurs manquantes de l'utilisateur pour les champs catégoriels : Les champs catégoriels doivent avoir des valeurs valides pour qu'un enregistrement puisse être inclus dans l'analyse. Ces contrôles vous permettent d'indiquer si les valeurs manquantes de l'utilisateur sont considérées comme valides parmi les champs catégoriels. Les valeurs système manquantes et les valeurs manquantes pour les champs continus sont toujours considérées comme non valides.

Fonctions supplémentaires de la commande NPTESTS

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- spécifier des tests à un échantillon, pour échantillon indépendants et pour échantillons liés en n'exécutant la procédure qu'une seule fois.

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Tests non paramétriques pour échantillons liés

Identifie les 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 du jeu 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 la pondération de chaque sujet est mesurée à intervalles réguliers et stockée dans des champs tels que *Pondération avant le régime*, *Pondération intermédiaire*, et *Pondération après le régime*. Ces champs sont "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 catégorielles lorsque 2 champs sont spécifiés, le Q de Cochran aux données catégorielles 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.

Obtention des tests non paramétriques pour échantillons liés

A partir des menus, sélectionnez :

Analyse > **Tests non paramétriques** > **Echantillons liés...**

1. Cliquez sur **Exécuter**.

Sinon, vous pouvez :

- Spécifier un objectif dans l'onglet Objectif.

- Spécifier les affectations de champ dans l'onglet Champs.
- Spécifier les paramètres d'expert dans l'onglet Paramètres.

Onglet Champs

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 champ 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 des tests

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 catégorielles lorsque 2 champs sont spécifiés, le Q de Cochran aux données catégorielles 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 :** L'option **Test de McNemar (2 échantillons)** peut être appliquée aux champs catégoriels. Cela produit un test pour échantillons liés évaluant si les combinaisons de valeurs entre deux champs indicateurs (champs catégoriels 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. Pour en savoir plus sur les paramètres de test, voir «Test de McNemar : Définition du succès», à la page 137. Le **Q de Cochran (k échantillons)** peut être appliqué aux champs catégoriels. Cela produit un test pour échantillons liés évaluant si les combinaisons de valeurs entre *k* champs indicateurs (champs catégoriels 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 **étape par étape descendantes** . Pour en savoir plus sur les paramètres de test, voir «Q de Cochran : Définition du succès», à la page 137.
- **Tester les modifications au sein des données multinomiales :** L'option **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 est différente de 0. Le test s'applique aux champs ordinaux et continus. 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 appariés. Le test s'applique aux champs ordinaux et continus. S'il y a 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 **étape par étape descendantes** .
- **Comparer les distributions :** L'option **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 **étape par étape descendantes** .

Test de McNemar : Définition du succès : Le test de McNemar concerne les champs indicateurs (champs catégoriels avec seulement deux catégories), mais peut s'appliquer à tous les champs catégoriels en utilisant des règles de définition du "succès".

Définir le succès pour des champs catégoriels : Indique la procédure de définition du "succès" pour les champs catégoriels.

- **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 catégoriels 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 : Définition du succès : Le test Q de Cochran concerne les champs indicateurs (champs catégoriels avec seulement deux catégories), mais peut s'appliquer à tous les champs catégoriels en utilisant des règles de définition du "succès".

Définir le succès pour des champs catégoriels : Indique la procédure de définition du "succès" pour les champs catégoriels.

- **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 catégoriels 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

Niveau de signification : Indique le niveau de signification (alpha) pour tous les tests. Spécifiez une valeur numérique comprise entre 0 et 1. La valeur par défaut est 0.05.

Intervalle de confiance (%) : Indique le niveau de confiance pour tous les intervalles de confiance générés. Spécifiez une valeur numérique comprise entre 0 et 100. La valeur par défaut est 95.

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 de l'utilisateur

Valeurs manquantes de l'utilisateur pour les champs catégoriels : Les champs catégoriels doivent avoir des valeurs valides pour qu'un enregistrement puisse être inclus dans l'analyse. Ces contrôles vous permettent d'indiquer si les valeurs manquantes de l'utilisateur sont considérées comme valides parmi les champs catégoriels. Les valeurs système manquantes et les valeurs manquantes pour les champs continus sont toujours considérées comme non valides.

Fonctions supplémentaires de la commande NPTESTS

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- spécifier des tests à un échantillon, pour échantillon indépendants et pour échantillons liés en n'exécutant la procédure qu'une seule fois.

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Vue du modèle

Vue du modèle

La procédure crée un objet Visualiseur de modèle dans le visualiseur. 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, voir «Récapitulatif d'hypothèses».
- Récapitulatif de l'intervalle de confiance. Pour plus d'informations, voir «Récapitulatif de l'intervalle de confiance», à la page 139.

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, voir «Test à un échantillon», à la page 139.
- 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, voir «Test pour échantillons liés», à la page 140.
- 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, voir «Test pour échantillons indépendants», à la page 141.
- Informations sur les champs catégoriels. Pour plus d'informations, voir «Informations sur les champs catégoriels», à la page 142.
- Informations sur les champs continus. Pour plus d'informations, voir «Informations sur les champs continus», à la page 142.
- Comparaisons appariées. Pour plus d'informations, voir «Comparaisons appariées», à la page 142.
- Sous-ensembles homogènes. Pour plus d'informations, voir «Sous-ensembles homogènes», à la page 143.

Récapitulatif d'hypothèses

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.

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

Le test binomial affiche un graphique à barres empilées et un tableau des tests.

- Le graphique à barres empilées affiche les fréquences observées et théoriques pour les catégories "succès" et "échec" du champ test, les "échecs" étant empilés au dessus des "succès". Lorsque vous passez la souris sur une barre, les pourcentages des catégories s'affichent dans une infobulle. Des différences visibles entre les barres 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

Le test du khi-deux affiche un graphique à barres en cluster et un tableau des tests.

- Le graphique à barres en cluster affiche les fréquences observées et théoriques pour chaque catégorie du champ de test. Lorsque vous passez la souris sur une barre, les fréquences observées et théoriques et leur différence (résidu) s'affichent dans une infobulle. 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

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

La vue Test de suites en séquences affiche un graphique et un tableau des tests.

- Le graphique 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

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 :

Le test de McNemar affiche un graphique à barres en cluster et un tableau des tests.

- Le graphique à barres en cluster affiche les fréquences observées et théoriques pour les cellules hors diagonale du tableau 2×2 défini par les champs de test.
- Le tableau affiche des informations détaillées sur le test.

Test des signes

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

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

La vue Test d'homogénéité marginale affiche un graphique à barres en cluster et un tableau des tests.

- Le graphique à barres en cluster 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

Le test Q de Cochran affiche un graphique à barres empilées et un tableau des tests.

- Le graphique à barres empilées affiche les fréquences observées pour les catégories "succès" et "échec" des champs du test, les "échecs" étant empilés au dessus des "succès". Lorsque vous passez la souris sur une barre, les pourcentages des catégories s'affichent dans une infobulle.
- Le tableau affiche des informations détaillées sur le test.

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, panélisée par les champs du test.
- Le tableau affiche des informations détaillées sur le test.

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

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

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 :

Le test des suites en séquences de Wald-Wolfowitz affiche un graphique à barres empilées 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

Le test de Kruskal-Wallis affiche des boîtes à moustaches et un tableau des tests.

- Des boîtes à moustaches distinctes sont affichées pour chaque catégorie du champ de regroupement. Lorsque vous passez la souris sur une boîte, le rang moyen s'affiche dans une infobulle.
- Le tableau affiche des informations détaillées sur le test.

Test de Jonckheere-Terpstra :

Le test de Jonckheere-Terpstra affiche des boîtes à moustaches et un tableau des tests.

- Des boîtes à moustaches distinctes sont affichées pour chaque catégorie du champ de regroupement.
- Le tableau affiche des informations détaillées sur le test.

Test de réactions extrêmes de Moses

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 catégorie du champ de regroupement. Les libellés des points peuvent être affichés ou masqués en cliquant sur le bouton **ID de l'enregistrement**.
- Le tableau affiche des informations détaillées sur le test.

Test de la médiane

Le test de la médiane affiche des boîtes à moustaches et un tableau des tests.

- Des boîtes à moustaches distinctes sont affichées pour chaque catégorie du champ de regroupement.
- Le tableau affiche des informations détaillées sur le test.

Informations sur les champs catégoriels

La vue Informations sur les champs catégoriels affiche un graphique à barres pour le champ catégoriel sélectionné dans la liste déroulante **Champ(s)**. La liste des champs disponibles est limitée aux champs catégoriels utilisés dans le test actuellement sélectionné dans la vue Récapitulatif d'hypothèses.

- Lorsque vous passez la souris sur une barre, les pourcentages des catégories s'affichent dans une infobulle.

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 appariées

La vue Comparaisons appariées affiche un graphique de réseau des distances et un tableau des comparaisons produits par des tests non paramétriques à échantillon-*k*, lorsque des comparaisons appariées multiples sont requises.

- Le graphique 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 infobulle.
- Le tableau des comparaisons affiche les résultats numériques de toutes les comparaisons appariées. Chaque ligne correspond à une comparaison appariée distincte. Cliquez sur un en-tête de colonne pour trier les lignes de la colonne concernée en fonction de leurs valeurs.

Sous-ensembles homogènes

La vue Sous-ensembles homogènes affiche un tableau des comparaisons produits par des tests non paramétriques à échantillon-*k*, lorsque des comparaisons multiples étape par étape 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 également d'effectuer les actions suivantes :

- spécifier des tests à un échantillon, pour échantillon indépendants et pour échantillons liés en n'exécutant la procédure qu'une seule fois.

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

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 catégories 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 catégorie 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 catégories et calcule une statistique khi-deux. Ce test de qualité de l'ajustement compare les fréquences observées et attendues dans chaque catégorie pour vérifier si toutes les catégories contiennent la même proportion de valeurs ou si chaque catégorie 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 catégorie, les résidus et la statistique du khi-deux.

Test du khi-deux : remarques sur les données

Données : Utilisez des variables catégorielles numériques ordonnées ou désordonnées (niveau de mesure ordinal ou nominal). Pour convertir des variables de chaîne 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 catégorie doivent être au moins égales à 1,20 % des catégories au maximum doivent avoir des fréquences inférieures à 5.

Pour obtenir un test khi-deux

1. A partir des menus, sélectionnez :

   Analyse > **Tests non paramétriques** > **Boîtes de dialogue ancienne version** > **Khi-deux...**
2. Sélectionnez des variables de test. Chaque variable produit un test distinct.
3. Vous pouvez également cliquer sur **Options** pour les statistiques descriptives, les quartiles et le contrôle du traitement des données manquantes.

Valeurs et plages théoriques du test du khi-deux

Plage théorique : Par défaut, chaque valeur distincte de la variable est définie comme catégorie. Pour établir des catégories dans une plage spécifique, sélectionnez **Utiliser la plage indiquée** et fournissez des valeurs entières pour les limites inférieure et supérieure. Des catégories sont établies pour chaque valeur entière comprise dans la plage, 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 catégories ont des valeurs théoriques égales. Les catégories peuvent avoir des proportions attendues définies par l'utilisateur. Sélectionnez **Valeurs**, indiquez une valeur supérieure à 0 pour chaque catégorie 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 catégories 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 catégorie 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.

Options du test du khi-deux

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 25e, 50e et 75e percentiles.

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 la commande NPAR TESTS (test du khi-deux)

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- 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`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Test binomial

La procédure de test binomial compare les fréquences observées des deux catégories 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 niveau 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.

Test binomial : remarques sur les données

Données : Les variables testées doivent être numériques et dichotomiques. Pour convertir des variables de chaîne en variables numériques, utilisez la procédure de recodage automatique, disponible dans le menu Transformer. Une **variable dichotomique** est une variable qui accepte deux valeurs uniquement : *oui* ou *non*, *true* ou *false*, 0 ou 1, etc. La première valeur rencontrée dans le jeu 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

1. A partir des menus, sélectionnez :

   Analyse > **Tests non paramétriques** > **Boîtes de dialogue ancienne version** > **Binomial...**
2. Sélectionnez une ou plusieurs variables numériques à tester.

3. Vous pouvez également cliquer sur **Options** pour les statistiques descriptives, les quartiles et le contrôle du traitement des données manquantes.

Options du 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 25e, 50e et 75e percentiles.

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.

Fonctions supplémentaires de la commande NPAR TESTS (Test binomial)

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Sélectionner des groupes spécifiques (et en exclure d'autres) lorsqu'une variable comporte plus de deux catégories (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`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

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.

Suites en séquences : Remarques sur les Données

Données : Les variables doivent être numériques. Pour convertir des variables de chaîne 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

1. A partir des menus, sélectionnez :

   Analyse > **Tests non paramétriques** > **Boîtes de dialogue ancienne version** > **Séquences...**
2. Sélectionnez une ou plusieurs variables numériques à tester.
3. 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

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 25e, 50e et 75e percentiles.

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 la commande NPAR TESTS (Suites en séquences)

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- 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`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

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.

Remarques sur les données du test de Kolmogorov-Smirnov pour un échantillon

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 la plage 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

1. A partir des menus, sélectionnez :

   Analyse > **Tests non paramétriques** > **Boîtes de dialogue ancienne version** > **K-S à 1 échantillon...**
2. Sélectionnez une ou plusieurs variables numériques à tester. Chaque variable produit un test distinct.
3. 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

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 25e, 50e et 75e percentiles.

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`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

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, en moyenne, 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 et suites de Wald-Wolfowitz.

Tests pour deux échantillons indépendants : remarques sur les données

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

1. A partir des menus, sélectionnez :

Analyse > **Tests non paramétriques** > **Boîtes de dialogue ancienne version** > **2 échantillons indépendants**

2. Sélectionnez une ou plusieurs variables numériques.
3. 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 score du groupe 1 précède un score du groupe 2, ainsi que le nombre de fois qu'un score du groupe 2 précède un score 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 extrêmes peuvent occasionnellement et facilement fausser la plage d'amplitude, 5 % des observations de contrôle sont tronquées automatiquement à chaque extrémité.

Définition de deux groupes d'échantillons indépendants

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.

Options des tests pour deux échantillons indépendants

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 25e, 50e et 75e percentiles.

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 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 tronquées pour le test de Moses (avec la sous-commande MOSES).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur 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.

Tests pour deux échantillons liés : remarques sur les données

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

1. A partir des menus, sélectionnez :

   Analyse > **Tests non paramétriques** > **Boîtes de dialogue ancienne version** > **2 échantillons liés...**
2. 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 catégorielles, 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.

Options des tests pour deux échantillons liés

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 25e, 50e et 75e percentiles.

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 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.

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

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.

Tests pour Plusieurs Echantillons Indépendants : Remarques sur les Données

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

1. A partir des menus, sélectionnez :

   Analyse > **Tests non paramétriques** > **Boîtes de dialogue ancienne version** > **K échantillons indépendants...**
2. Sélectionnez une ou plusieurs variables numériques.
3. Sélectionnez une variable de regroupement et cliquez sur **Définir plage** 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 l'ampleur 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 Plage

Pour définir la plage, entrez des valeurs entières pour **Minimum** et **Maximum** qui correspondent à la catégorie 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

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 25e, 50e et 75e percentiles.

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 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`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

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.

Tests pour plusieurs échantillons liés de considérations de données

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

1. A partir des menus, sélectionnez :

   Analyse > **Tests non paramétriques** > **Boîtes de dialogue ancienne version** > **K échantillons liés...**
2. 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 à deux facteurs 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 est basée sur ce classement.

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

Vous pouvez choisir les statistiques.

- **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 25e, 50e et 75e percentiles.

Fonctions supplémentaires de la commande NPAR TESTS (K échantillons liés)

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 28. Analyse des réponses multiples

Analyse des réponses multiples

Deux procédures sont proposées pour l'analyse des jeux comportant plusieurs variables dichotomiques ou plusieurs catégories. La procédure Fréquences multiréponses affiche les tables 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 jeux de réponses multiples.

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 au cours des 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 catégories 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 catégorie 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 jeux de réponses multiples

La procédure de définition de jeux de réponses multiples regroupe des variables élémentaires dans des jeux de dichotomies ou de catégories multiples, pour lesquels vous pouvez obtenir des tables de fréquences et des tableaux croisés. Vous pouvez définir jusqu'à 20 jeux de réponses multiples. Chaque vecteur doit avoir un nom unique. Pour éliminer un jeu, sélectionnez-le dans la liste des jeux de réponses multiples 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 élémentaires sous forme de dichotomies ou de catégories. Pour utiliser des variables dichotomiques, sélectionnez **Variables dichotomiques** afin de créer un jeu de dichotomies multiples. Entrez une valeur entière dans Valeur comptée. Chaque variable ayant au moins une occurrence de la valeur comptée devient une catégorie du jeu de dichotomies multiples. Sélectionnez **Catégories** pour créer un jeu de catégories multiples ayant la même plage de valeurs que les variables qui le composent. Entrez des nombres entiers pour le minimum et le maximum de la plage des catégories

du jeu de catégories multiples. La procédure totalise chaque valeur entière contenue dans la plage pour toutes les variables qui le composent. Les catégories vides ne sont pas tabulées.

A chaque jeu de réponses multiples doit être attribué un nom unique de 7 caractères maximum. La procédure ajoute un signe dollar ($) devant le nom que vous avez attribué. Vous ne pouvez pas utiliser les noms réservés suivants : *casenum*, *sysmis*, *jdate*, *date*, *time*, *length* et *width*. Le nom du jeu de réponses multiples doit uniquement être utilisé dans les procédures multiréponses. Vous ne pouvez pas faire référence aux noms d'ensemble de réponses multiples dans les autres procédures. A titre facultatif, vous pouvez entrer un libellé de variable décrivant le jeu de réponses multiples. Ce libellé peut comporter jusqu'à 40 caractères.

Définir des jeux de réponses multiples

1. A partir des menus, sélectionnez :

   Analyse > **Réponses multiples** > **Définir des jeux de variables...**
2. Sélectionnez deux ou plusieurs variables.
3. Si vos variables sont codées comme dichotomies, indiquez la valeur que vous souhaitez calculer. Si elles sont codées comme catégories, définissez leur plage.
4. Entrez un nom unique pour chaque jeu de réponses multiples.
5. Cliquez sur **Ajouter** pour ajouter le jeu de réponses multiples à la liste des jeux définis.

Tableaux d'effectifs des réponses multiples

La procédure Tableaux d'effectifs des réponses multiples produit des tables de fréquences pour les jeux de réponses multiples. Vous devez d'abord définir un ou plusieurs jeux de réponses multiples (voir " Définir les jeux de réponses multiples ").

Pour les jeux de dichotomies multiples, les noms de catégorie apparaissant dans la sortie proviennent de libellés de variable définis pour les variables élémentaires du groupe. Si les libellés de variable ne sont pas définis, les noms de variables servent de libellés. Pour les jeux de catégories multiples, les libellés des catégories proviennent des libellés de valeurs de la première variable du groupe. Si les catégories manquantes de la première variable sont présentes pour d'autres variables du groupe, définissez un libellé de valeurs pour les catégories 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 jeu de dichotomies multiples. Ceci s'applique seulement aux jeux de réponses multiples définis comme jeux de dichotomies. Par défaut, une observation est considérée manquante pour un jeu 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 catégories :** Cela permet d'exclure les observations ayant des valeurs manquantes pour toute variable provenant du tableau croisé du jeu des catégories multiples. Ceci s'applique seulement aux jeux de réponses multiples définis comme des jeux de catégories. Par défaut, une observation est considérée manquante pour un jeu de catégories multiples si aucune de ses composantes n'a de valeurs valides à l'intérieur de la plage définie.

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 jeux de réponses multiples : jeu de catégories multiples ou jeu 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 **jeu de dichotomies multiples**, chacune des trois variables du jeu devient une catégorie de la variable de groupe. Les effectifs et les pourcentages correspondant aux trois compagnies aériennes s'affichent dans une table 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 **jeu de catégories 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 élémentaires. 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 une table de fréquences.

Statistiques : Tables de fréquences contenant des effectifs, des pourcentages de réponses, des pourcentages de cas, le nombre de cas valables, et le nombre de cas manquants.

Gestion des données des fréquences de réponses multiples

Données : Utilisez des jeux de réponses multiples.

Hypothèses : Les effectifs et pourcentages représentent une description utile des données de n'importe quelle distribution.

Procédures apparentées : La procédure Définir Jeux de réponses multiples vous permet de définir des jeux de réponses multiples.

Pour obtenir des tableaux de fréquences de réponses multiples

1. A partir des menus, sélectionnez :
   Analyse > **Réponses multiples** > **Effectifs...**
2. Sélectionnez un ou plusieurs jeux de réponses multiples.

Tableaux croisés des réponses multiples

La procédure Tableaux croisés de réponses multiples classe, par tableaux croisés, des jeux de réponses multiples définis, des variables élémentaires 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 jeux de réponses multiples (veuillez consulter " Pour Définir des jeux de réponses multiples ").

Pour les jeux de dichotomies multiples, les noms de catégorie apparaissant dans la sortie proviennent de libellés de variable définis pour les variables élémentaires du groupe. Si les libellés de variable ne sont pas définis, les noms de variables servent de libellés. Pour les jeux de catégories multiples, les libellés des catégories proviennent des libellés de valeurs de la première variable du groupe. Si les catégories manquantes de la première variable sont présentes pour d'autres variables du groupe, définissez un libellé de valeurs pour les catégories manquantes. La procédure affiche les libellés de catégorie 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 libellés.

Exemple : Les jeux de dichotomies multiples et les jeux de catégories 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 catégories 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.

Remarques sur les Données de Tableaux croisés de réponses multiples

Données : Utilisez des jeux de réponses multiples ou des variables catégorielles numériques.

Hypothèses : Les effectifs et pourcentages offrent une description utile des données qui suivent tout type de distribution.

Procédures apparentées : La procédure Définir Jeux de réponses multiples vous permet de définir des jeux de réponses multiples.

Pour obtenir des tableaux croisés des réponses multiples

1. A partir des menus, sélectionnez :
   Analyse > **Réponses multiples** > **Tableaux croisés...**
2. Sélectionnez une ou plusieurs variables numériques ou jeux de réponses multiples pour chaque dimension de tableau croisé.
3. Définissez la plage de chaque variable élémentaire.

Sinon, vous pouvez obtenir un tableau croisé bilatéral pour chaque catégorie de variable de contrôle ou chaque jeu de réponses multiples. Sélectionnez un ou plusieurs éléments pour la liste de couche(s).

Définir Plages Tableaux croisés De réponses multiples

Les plages des valeurs doivent être définies pour toute variable élémentaire de tableaux croisés. Entrez les valeurs entières de catégories minimum et maximum que vous souhaitez tabuler. Les catégories se situant en dehors de la plage sont exclues de l'analyse. Les valeurs se situant à l'intérieur de la plage inclusive sont supposées être des nombres entiers (les nombres non entiers sont tronqués).

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 jeux de catégories multiples. Vous pouvez aussi baser les pourcentages cellules sur les réponses. Pour les jeux de dichotomies multiples, le nombre de réponses est égal au nombre de valeurs comptées à travers les observations. Pour les jeux de catégories multiples, le nombre de réponses correspond au nombre de valeurs comprises dans la plage 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 jeu de dichotomies multiples. Ceci s'applique seulement aux jeux de réponses multiples définis comme jeux de dichotomies. Par défaut, une observation est considérée manquante pour un jeu 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 catégories :** Cela permet d'exclure les observations ayant des valeurs manquantes pour toute variable provenant du tableau croisé du jeu des catégories multiples. Ceci s'applique seulement aux jeux de réponses

multiples définis comme des jeux de catégories. Par défaut, une observation est considérée manquante pour un jeu de catégories multiples si aucune de ses composantes n'a de valeurs valides à l'intérieur de la plage définie.

Par défaut, lorsque vous croisez deux jeux de catégories 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 jeux de dichotomies multiples ou les variables élémentaires.

Fonctions supplémentaires de la commande MULT RESPONSE

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Obtenir des tableaux croisés ayant jusqu'à cinq dimensions (avec la sous-commande `BY`).
- Modifier les options de formatage de la sortie, y compris la suppression des libellés de valeurs (avec la sous-commande `FORMAT`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 29. Rapports de Résultats

Rapports 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 fréquences 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 Rapport en lignes et Rapport en colonnes vous permettent de contrôler la présentation des données.

Rapport en lignes

Rapport en lignes produit des rapports 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 la division où chaque employé travaille. Vous pourriez générer un rapport fournissant les informations individuelles sur les employés (liste) divisées par magasin et par division (variables d'agrégation), avec les statistiques récapitulatives (par exemple, salaire moyen) pour chaque magasin, division et par division dans chaque magasin.

Colonnes de données : Donne la liste des variables de rapport pour lesquelles vous voulez obtenir des listes d'observations ou des statistiques récapitulatives, et contrôle le format d'affichage des colonnes de données.

Colonne de rupture : Donne la liste des variables d'agrégation optionnels qui divisent le rapport en groupes et contrôle les statistiques récapitulatives et les formats d'affichage des colonnes de rupture. Pour les variables d'agrégation multiples, il y aura un groupe séparé pour chaque catégorie de chaque variable d'agrégation à l'intérieur des catégories de la variable d'agrégation précédent dans la liste. Les variables d'agrégation doivent être des variables catégorielles discrètes qui divisent les observations en un nombre limité de catégories significatives. Les valeurs individuelles de chaque variable d'agrégation apparaissent, triées, dans une colonne séparée à gauche des colonnes de données.

Rapport : Contrôle les caractéristiques globales du rapport, 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 libellés de valeurs) des variables de Variables en colonnes pour chaque observation. Cela produit une liste , qui peut être nettement plus longue qu'un rapport.

Aperçu : N'affiche que la première page du rapport. Cette option est utile pour avoir un aperçu du format de votre rapport sans traiter le rapport entier.

Les données sont déjà triées : Pour les rapports avec variables d'agrégation, le fichier de données doit être trié par valeur des variables d'agrégation avant de générer le rapport. Si votre fichier de données est déjà trié par valeur des variables d'agrégation, vous pouvez gagner du temps de traitement en sélectionnant cette option. Cette option est particulièrement utile après avoir vu un aperçu du rapport.

Obtention d'un rapport récapitulatif : Récapitulatifs en lignes

1. A partir des menus, sélectionnez :

   Analyse > **Rapports** > **Rapports récapitulatifs en lignes...**
2. Sélectionnez une ou plusieurs variables pour les Colonnes de données. Une colonne est générée dans le rapport pour chaque variable sélectionnée.
3. Pour les rapports triés et affichés par sous-groupe, sélectionnez une ou plusieurs variables pour les critères d'agrégation.
4. Pour les rapports avec statistiques récapitulatives de sous-groupe définies par des variables d'agrégation, sélectionnez la variable d'agrégation dans la liste Variables de rupture et cliquez sur **Récapitulatif** dans le groupe Variables de rupture pour spécifier les mesures récapitulatives.
5. Pour les rapports avec statistiques récapitulatives globales, cliquez sur **Récapitulatif** pour spécifier les mesures récapitulatives.

Format des Colonnes de données/Rupture des rapports

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 libellés de valeurs. Format colonne de données contrôle le format des colonnes de données du côté droit de la page du rapport. Format colonne de rupture contrôle le format des Colonnes de rupture du côté gauche.

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 retours à la ligne 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 libellés de données dans la colonne. L'alignement des valeurs ou des libellés 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 libellés de valeurs définis. Les valeurs de données sont affichées pour toutes les valeurs qui ne possèdent pas de libellé de valeur défini. (Non disponible pour les colonnes de données dans les rapports en colonnes)

Lignes récapitulatives des rapports pour/Lignes récapitulatives Finales

Les deux boîtes de dialogue Lignes récapitulatives contrôlent l'affichage des statistiques récapitulatives pour les groupes de rupture et pour l'ensemble du rapport. Lignes récapitulatives contrôle les statistiques de sous-groupe pour chaque catégorie définie par la ou les variables d'agrégation. Lignes récapitulatives Finales contrôle les statistiques globales affichées à la fin du rapport.

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'une plage donnée de valeurs, l'écart type, kurtosis, la variance et l'asymétrie.

Options de rupture de rapport

Les options de rupture contrôle l'espacement et la pagination des informations de catégorie de rupture.

Contrôle de page : Contrôle l'espacement et la pagination des catégories de la variable d'agrégation sélectionné. Vous pouvez spécifier un nombre de lignes vides entre les catégories de rupture ou commencer chaque catégorie de rupture sur une nouvelle page.

Lignes à sauter avant fonctions élémentaires : Contrôle le nombre de lignes vides entre les libellés ou les données des catégories de rupture et les statistiques récapitulatives. Cette option est particulièrement utile pour les rapports combinés incluant des listes d'observations individuelles et des statistiques récapitulatives pour les catégories de rupture ; dans ces rapports, vous pouvez insérer des espaces entre les listes d'observations et les statistiques récapitulatives.

Options du rapport

Options du rapport contrôle le traitement et l'affichage des valeurs manquantes et la numérotation des pages du rapport.

Exclure les observations avec des valeurs manquantes : Elimine (du rapport) toute observation avec des valeurs manquantes pour l'une des variables du rapport.

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 système manquantes* et les *valeurs manquantes de l'utilisateur*.

Paginer à partir de : Vous permet de spécifier un numéro pour la première page du rapport.

Présentation du rapport

Présentation du rapport contrôle la largeur et la longueur de chaque page du rapport, l'emplacement du rapport sur la page et l'insertion de lignes vides et de libellés.

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 rapport.

Colonne de rupture : Contrôle l'affichage des colonnes de rupture. Si des variables d'agrégation multiples sont spécifiés, ils peuvent être affichés en colonnes séparées ou dans la première colonne. Placer toutes les variables d'agrégation dans la première colonne produit un rapport 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 rapport, et l'alignement vertical des titres de colonnes.

Lignes de variables en colonnes et libellés de rupture : Contrôle l'emplacement des informations de Variables en colonnes (valeurs de données et/ou statistiques récapitulatives) par rapport aux libellés de rupture au début de chaque catégorie de rupture. La première ligne des informations de Variables en colonnes peut commencer soit sur la même ligne que le libellé de catégorie de rupture, soit un nombre donné de lignes après ce libellé. (Non disponible pour les rapports en colonnes)

Titres du rapport

Titres du rapport contrôle le contenu et l'emplacement des titres et pieds de page du rapport. 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.

Si vous insérez des variables dans les titres ou les pieds de page, le libellé de valeur actuel ou la valeur de la variable est affichée dans le titre ou le pied de page. Dans les titres, le libellé de valeur correspondant à la valeur de la variable au début de la page est affiché. Dans les pieds de page, le libellé de valeur correspondant à la valeur de la variable à la fin de la page est affiché. S'il n'y a aucun libellé 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 rapports.

Rapport en colonnes

Rapport en colonnes produit des rapports 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 la division où chaque employé travaille. Vous pourriez générer un rapport fournissant des statistiques récapitulatives sur les salaires (par exemple moyenne, minimum, maximum) pour chaque division.

Colonnes de données : Fournit la liste des variables du rapport 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 variables d'agrégation optionnelles qui divisent le rapport en groupes et contrôle les formats d'affichage des colonnes de rupture. Pour les variables d'agrégation multiples, il y aura un groupe séparé pour chaque catégorie de chaque variable d'agrégation à l'intérieur des catégories de la variable d'agrégation précédent dans la liste. Les variables d'agrégation doivent être des variables catégorielles discrètes qui divisent les observations en un nombre limité de catégories significatives.

Rapport : Contrôle les caractéristiques globales du rapport, y compris l'affichage des valeurs manquantes, la numérotation des pages et les titres.

Aperçu : N'affiche que la première page du rapport. Cette option est utile pour avoir un aperçu du format de votre rapport sans traiter le rapport entier.

Les données sont déjà triées : Pour les rapports avec variables d'agrégation, le fichier de données doit être trié par valeur des variables d'agrégation avant de générer le rapport. Si votre fichier de données est déjà trié par valeur des variables d'agrégation, vous pouvez gagner du temps de traitement en sélectionnant cette option. Cette option est particulièrement utile après avoir vu un aperçu du rapport.

Obtention d'un rapport récapitulatif : Récapitulatifs en colonnes

1. A partir des menus, sélectionnez :

   Analyse > **Rapports** > **Rapports récapitulatifs en colonnes...**
2. Sélectionnez une ou plusieurs variables pour les Colonnes de données. Une colonne est générée dans le rapport pour chaque variable sélectionnée.
3. Pour modifier la mesure récapitulative d'une variable, sélectionnez la variable dans la liste Variables en colonnes et cliquez sur **Récapitulatif**.
4. 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.
5. 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 des colonnes de donnée.
6. Pour les rapports triés et affichés par sous-groupe, sélectionnez une ou plusieurs variables pour les critères d'agrégation.

Fonction récapitulative des colonnes de données

Lignes récapitulatives contrôle les statistiques récapitulatives affichées pour la variable de colonne de données sélectionnée.

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'une plage donnée de valeurs, l'écart type, kurtosis, 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 colonnes de données.

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.

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.

1re colonne - 2e 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.

1re colonne / 2e 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.

% 1re colonne / 2e 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 rapport

Les options de format des colonnes de données et de rupture pour les Rapports en colonnes sont identiques à celles décrites pour les Rapports en lignes.

Rapport en Colonnes : Options de rupture

Options de rupture contrôle l'affichage des sous-totaux, l'espacement et la pagination des catégories de rupture.

Sous-total : Contrôle l'affichage des sous-totaux pour les catégories de rupture.

Contrôle de page : Contrôle l'espacement et la pagination des catégories de la variable d'agrégation sélectionné. Vous pouvez spécifier un nombre de lignes vides entre les catégories de rupture ou commencer chaque catégorie de rupture sur une nouvelle page.

Lignes à sauter avant sous-total : Contrôle le nombre de lignes vides entre les données des catégories de rupture et les sous-totaux.

Options des Rapports en Colonnes

Options contrôle l'affichage des totaux généraux, l'affichage des valeurs manquantes et la pagination dans les rapports en colonnes.

Total général : Affiche et libelle un total général pour chaque colonne ; affiché au bas de la colonne.

Valeurs manquantes : Vous pouvez exclure les valeurs manquantes du rapport ou sélectionner un caractère unique indiquant les valeurs manquantes dans le rapport.

Présentation du rapport en colonnes

Les options de présentation pour les Rapports en colonnes sont identiques à celles présentées pour les Rapports en lignes.

Fonctions supplémentaires de la commande REPORT

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Afficher différentes fonctions récapitulatives dans les colonnes d'une ligne récapitulative unique.
- Insérer des lignes récapitulatives dans les colonnes de données 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 rapport.
- 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 rapport avec syntaxe, d'approcher le rapport généré à partir des boîtes de dialogue, de copier et coller la syntaxe correspondante, puis de préciser cette syntaxe afin d'obtenir le rapport exact que vous souhaitez.

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

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.

Remarques sur les données de l'analyse de fiabilité

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 motif des calculs), utilisez Analyse factorielle ou Positionnement multidimensionnel. Pour identifier des groupes homogènes de variables, utilisez l'analyse de cluster hiérarchique pour classer les variables.

Obtenir une analyse de fiabilité

1. A partir des menus, sélectionnez :
   Analyse > **Echelle** > **Analyse de la fiabilité...**
2. Sélectionnez deux variables (éléments) au moins en tant que composants potentiels d'une échelle additive.
3. Sélectionnez un modèle dans la liste déroulante Modèle.

Statistiques de l'analyse de fiabilité

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.

- **Elément :** Produit des statistiques descriptives pour les éléments sur les observations.
- **Echelle :** Produit des statistiques descriptives pour les échelles.
- **Echelle sans l'élément :** Affiche les statistiques récapitulatives en comparant chaque élément à l'échelle composée des autres éléments. Les statistiques incluent la moyenne et la variance de l'échelle si l'élément a été supprimé de l'échelle, la corrélation entre l'élément et l'échelle composée des autres éléments, et l'alpha de Cronbach si l'élément 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 des moyennes d'élément. Les moyennes d'élément minimale, maximale et moyenne sont affichées, ainsi que le rapport de la moyenne maximale à la moyenne minimale.
- *Variances*. Statistiques récapitulatives 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 éléments. Les covariances entre éléments minimale, maximale et moyenne sont affichées, ainsi que la plage 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 éléments. Les corrélations entre éléments minimale, maximale et moyenne sont affichées, ainsi que la plage 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 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'éléments sont fixes, sélectionnez **Aléatoire à deux facteurs** lorsque les effets de population et les effets d'éléments 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.

Fonctions supplémentaires de la commande RELIABILITY

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- 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 de la bipartition.

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

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 réduction de données (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 classements subjectives de dissimilarité entre objets ou concepts. D'autre part, le positionnement multidimensionnel peut gérer les données 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 classements 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, vous disposez de la matrice des données, du positionnement optimisé des données de la matrice, de la contrainte S (de Young), de la contrainte S (de Kruskal), de RSQ, des coordonnées de stimulus, de la contrainte moyenne et de RSQ pour chaque stimulus (modèles RMDS). Pour les modèles des différences individuelles (INDSCAL), vous disposez des pondérations de sujet et de l'indice de singularité pour chaque objet. Pour les matrices situées dans les modèles de positionnement multidimensionnel répliqués, vous disposez de la contrainte et de RSQ pour chaque stimulus. Pour les tracés, vous disposez des coordonnées de stimulus (à deux ou trois dimensions), du nuage de points des disparités par rapport aux distances.

Remarques sur les données du positionnement multidimensionnel

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 réduction de données, 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 de cluster de *nuées dynamiques* ou une analyse de la classification hiérarchique.

Obtenir une analyse par positionnement multidimensionnel

1. A partir des menus, sélectionnez :

   Analyse > **Echelle** > **Positionnement multidimensionnel...**
2. Sélectionnez au moins quatre variables numériques pour l'analyse.
3. Dans le groupe Distances, sélectionnez **Données en matrice(s)** ou **Calculées à partir des données**.
4. 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 de 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

Si votre jeu 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

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, Différence 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

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 la plage. 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 de sujet négative**, si cela convient à vos données.

Positionnement multidimensionnel : Options

Vous pouvez spécifier les options de votre analyse par positionnement multidimensionnel.

Affichage : Vous permet d'afficher les différents types d'affichage. Les options possibles sont Tracés des stimuli, Tracés 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 S-stress**, le **Minimum du s-stress** et le **Maximum des itérations**.

Traiter les dissimilarités inférieures à n comme des valeurs manquantes : Ces distances sont exclues de l'analyse.

Fonctions supplémentaires de la commande ALSCAL

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- 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 rapport.
- 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.

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Chapitre 32. Statistiques de rapport

La procédure Statistiques de rapport permet d'obtenir la liste exhaustive des statistiques récapitulatives qui servent à décrire le rapport entre deux variables d'échelle.

Vous pouvez trier la sortie sur la base des valeurs d'une variable de regroupement, dans l'ordre croissant ou décroissant. Vous pouvez supprimer le rapport des statistiques de rapport 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 sorties, 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, déviation absolue moyenne (AAD), plage, valeurs minimale et maximale, et index de concentration calculés pour une plage ou un pourcentage défini par l'utilisateur dans le rapport médian.

Remarque sur les données des statistiques de rapport

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 rapport

1. A partir des menus, sélectionnez :
   Analyse > **Statistiques descriptives** > **Rapport...**
2. Sélectionnez la variable du numérateur.
3. Sélectionnez la variable du dénominateur.

Eventuellement :

- Sélectionner une variable de regroupement et préciser l'ordre de présentation des groupes dans le résultat.
- Choisissez si vous souhaitez afficher les résultats dans le visualiseur 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 rapport

Tendance centrale : Les mesures de tendance centrale sont des statistiques qui décrivent la distribution des rapports.

- **Médiane :** La valeur telle que le nombre de rapports inférieurs à cette valeur est identique au nombre de rapports supérieurs à cette valeur.
- **Moyenne :** Résultat de la somme des rapports divisée par le nombre total de rapports.
- **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 rapports 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 :** La déviation absolue moyenne est égal à la somme des déviations absolues des rapports relatifs à la médiane, divisée par le nombre total de rapports.
- **COD :** Le coefficient de dispersion résulte de l'expression de la déviation moyenne absolue 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 la déviation 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 rapports relatifs à la moyenne divisée par le nombre total des rapports moins un.
- **Plage :** Résultat de la soustraction du rapport minimal au rapport maximal.
- **Minimum :** Le minimum est le plus petit rapport.
- **Maximum :** Le maximum est le plus grand rapport.

Index de concentration : Le coefficient de concentration mesure le pourcentage des rapports compris dans un intervalle. Vous pouvez le calculer de deux manières :

- **Rapports 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.
- **Rapports 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 oeuvre qu'une seule variable à deux catégories et utilisées pour la classification des sujets.

Exemple : Une banque envisage de classer correctement ses clients en catégories, 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. Tracés : 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.

Remarques sur les données de la courbe ROC

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 catégories. La variable d'état peut être d'un type quelconque et indique la véritable catégorie à laquelle un sujet appartient. La valeur de la variable d'état indique la catégorie à considérer comme *positive*.

Hypothèses : Les nombres croissants d'une échelle d'indicateurs confirment que le sujet appartient à une catégorie, tandis que les nombres décroissants d'une échelle confirment qu'il appartient à une autre catégorie. L'utilisateur doit choisir la direction *positive*. On suppose également que la *véritable* catégorie à laquelle chaque sujet appartient est connue.

Pour obtenir une courbe ROC

1. A partir des menus, sélectionnez :
   Analyse > **Courbe ROC...**
2. Sélectionnez une ou plusieurs variables de probabilité de test.
3. Sélectionnez une variable d'état.
4. Identifiez la valeur *positive* de la variable d'état.

Options de la courbe ROC

Vous pouvez indiquer les options suivantes pour votre analyse ROC :

Classification : Permet de spécifier si la valeur de césure doit être incluse ou exclue lors d'une classification *positive*. Ce paramètre n'a pas de conséquence sur la sortie.

Direction du test : Permet de spécifier la direction de l'échelle en fonction de la catégorie *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 la plage 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 d'IBM 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.

Vous pouvez également simuler des données en l'absence d'un modèle prédictif en indiquant les distributions de probabilité pour les variables que vous souhaitez simuler. Chaque observation de données générée consiste en l'ensemble des valeurs simulées pour les variables spécifiées.

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 de la sortie 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.

Conception d'une simulation basée sur un fichier de modèle

1. A partir des menus, sélectionnez :

   Analyse > **Simulation...**
2. Cliquez sur **Sélectionner un fichier de modèle SPSS**, puis cliquez sur **Poursuivre**.
3. Ouvrez le fichier de modèle.

   Le fichier de modèle est un fichier XML qui contient le fichier PMML Model créé à partir d'IBM SPSS Statistics ou d'IBM SPSS Modeler. Pour plus d'informations, voir «Onglet Modèle», à la page 182.
4. Sur l'onglet Simulation (dans le Générateur de simulation), spécifiez les distributions de probabilité des entrées simulées et les valeurs des entrées fixes. Si le jeu de données actif contient des données historiques relatives aux entrées simulées, cliquez sur **Ajuster tout** pour déterminer automatiquement la distribution la mieux adaptée à chaque entrée, ainsi que les corrélations entre celles-ci. Pour chaque entrée simulée non ajustée aux données historiques, vous devez explicitement indiquer une distribution en sélectionnant un type et en entrant les paramètres requis.
5. 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.

Les options suivantes sont disponibles :

- Modifier l'emplacement du plan de simulation enregistré.
- Spécifier les corrélations connues existant entre les entrées simulées.
- Calculer automatiquement un tableau de contingence des associations entre les entrées catégorielles et utiliser ces associations lorsque les données sont générées pour ces entrées.
- 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.
- Spécifier des options avancées telles que le nombre maximum d'observations à générer ou un échantillonnage des extrémités.
- Personnaliser la sortie.
- Enregistrer les données simulées dans un fichier de données.

Pour concevoir une simulation basée sur des équations personnalisées

1. A partir des menus, sélectionnez :

   Analyse > **Simulation...**
2. Cliquez sur **Entrer les Equations** puis cliquez sur **Poursuivre**.
3. Cliquez sur **Nouvelle équation** sur l'onglet Modèle (du Générateur de simulation) afin de définir chaque équation dans votre modèle prédictif.
4. Cliquez sur l'onglet Simulation et spécifiez les distributions de probabilité des entrées simulées ainsi que les valeurs des entrées fixes. Si le jeu de données actif contient des données historiques relatives aux entrées simulées, cliquez sur **Ajuster tout** pour déterminer automatiquement la distribution la mieux adaptée à chaque entrée, ainsi que les corrélations entre celles-ci. Pour chaque entrée simulée non ajustée aux données historiques, vous devez explicitement indiquer une distribution en sélectionnant un type et en entrant les paramètres requis.
5. 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.

Les options suivantes sont disponibles :

- Modifier l'emplacement du plan de simulation enregistré.
- Spécifier les corrélations connues existant entre les entrées simulées.

- Calculer automatiquement un tableau de contingence des associations entre les entrées catégorielles et utiliser ces associations lorsque les données sont générées pour ces entrées.
- 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.
- Spécifier des options avancées telles que le nombre maximum d'observations à générer ou un échantillonnage des extrémités.
- Personnaliser la sortie.
- Enregistrer les données simulées dans un fichier de données.

Pour concevoir une simulation sans modèle prédictif

1. A partir du menu, sélectionnez :

   Analyse > **Simulation...**
2. Cliquez sur **Créer des données simulées**, puis sur **Poursuivre**.
3. Sur l'onglet Modèle (dans le Générateur de simulation), sélectionnez les champs à simuler. Vous pouvez sélectionner des champs à partir du jeu de données actif ou définir de nouveaux champs en cliquant sur **Nouveau**.
4. Cliquez sur l'onglet Simulation et spécifiez les distributions de probabilité des champs à simuler. Si le jeu de données actif contient des données historiques relatives à ces champs, cliquez sur **Ajuster tout** pour déterminer automatiquement la distribution la mieux adaptée aux données, ainsi que les corrélations entre les champs. Pour les champs non ajustés aux données historiques, vous devez explicitement indiquer une distribution en sélectionnant un type de distribution et en entrant les paramètres requis.
5. Cliquez sur **Exécuter** pour exécuter la simulation. Par défaut, les données simulées sont sauvegardées dans le nouveau jeu de données défini dans les paramètres d'enregistrement. De plus, le plan de simulation, qui spécifie les détails de la simulation, est enregistré à l'emplacement indiqué dans les paramètres d'enregistrement.

Les options suivantes sont disponibles :

- Modifier l'emplacement des données simulées ou du plan de simulation enregistré.
- Spécifier les corrélations entre les champs simulés.
- Calculer automatiquement un tableau de contingence des associations entre les champs catégoriels et utiliser ces associations lorsque les données sont générées pour ces champs.
- Spécifier l'analyse de sensibilité pour vérifier l'effet de la variation d'un paramètre de distribution d'un champ simulé.
- Spécifier des options avancées telles que le nombre d'observations à générer.

Exécution d'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 :

1. A partir des menus, sélectionnez :

   Analyse > **Simulation...**
2. Cliquez sur **Ouvrir un plan de simulation existant**.
3. Vérifiez que la case **Ouvrir dans le Générateur de simulation** n'est pas cochée et cliquez sur **Poursuivre**.
4. Ouvrez le plan de simulation.
5. Cliquez sur **Exécuter** dans la boîte de dialogue Exécuter la simulation.

Pour exécuter la simulation à partir du Générateur de simulation :

1. A partir des menus, sélectionnez :

   Analyse > **Simulation...**
2. Cliquez sur **Ouvrir un plan de simulation existant**.
3. Sélectionnez la case **Ouvrir dans le Générateur de simulation** et cliquez sur **Poursuivre**.
4. Ouvrez le plan de simulation.
5. Modifiez les paramètres requis dans l'onglet Simulation.
6. Cliquez sur **Exécuter** pour exécuter la simulation.

Vous pouvez au besoin 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 la sortie.
- 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. Elle 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.
- Concevoir et exécuter une simulation qui génère des données en l'absence d'un modèle prédictif.
- Exécuter une simulation basée sur un plan de simulation existant, et éventuellement modifier des paramètres du plan.

Onglet Modèle

Pour les simulations basées sur un modèle prédictif, l'onglet Modèle indique la source du modèle. Pour les simulations qui n'incluent pas de modèle prédictif, l'onglet Modèle spécifie les champs à simuler.

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 qui contient le modèle PMML créé à partir d'IBM SPSS Statistics ou d'IBM SPSS Modeler. Les modèles prédictifs sont créés par des procédures telles que la régression linéaire et les arbres de décisions dans IBM SPSS Statistics, et ils peuvent être exportés vers un fichier de modèle. Vous pouvez sélectionne 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
- Modèle linéaire général
- Régression logistique binaire
- Régression logistique multinomiale
- Régression multinomiale ordinale
- Régression de Cox
- Arbre

- Arbre boosté (C5)
- Discriminant
- Cluster en deux étapes
- Cluster de nuées dynamiques
- Réseau de neurones
- Ensemble de règles (Liste de décision)

Remarque :

- Les modèles PMML qui possèdent plusieurs champs cible (variables) ou scissions ne sont pas pris en charge par les simulations.
- Les valeurs des entrées de chaîne dans les modèles de régression binaire sont limitées à 8 octets. Si vous ajustez de telles chaînes dans le jeu de données actif, veillez à ce que les valeurs dans les données ne dépassent pas 8 octets de longueur. Les valeurs de données qui dépassent 8 octets sont exclues de la distribution catégorielle associée pour l'entrée, et sont affichées comme n'ayant pas de correspondance dans la table de sortie Catégories sans correspondance.

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
  ```

Créer des données simulées sans modèle : Sélectionnez cette option pour simuler des données sans modèle prédictif. Indiquez les champs à simuler en les sélectionnant dans le jeu de données actif ou en cliquant sur **Nouveau** pour définir de nouveaux champs.

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 du jeu 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 libellés de valeur, et choisir si une sortie est générée 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'affichent sur l'onglet Modèle en même temps que l'équation.

1. 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.

2. 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 du jeu 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). Entrez tous les paramètres indiqués par un point d'interrogation. Le groupe de fonctions libellé **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.

- Si une entrée devant être utilisée dans une équation n'existe pas dans le jeu de données actif, vous devez la définir pour pouvoir l'utiliser dans l'équation.
- Si vous simulez des données sans modèle prédictif, vous devez définir l'ensemble des entrées simulées qui ne figurent pas dans le jeu de données actif.

Nom : Spécifiez le nom d'une cible ou d'une entrée.

Cible : Vous pouvez spécifier le niveau de mesure d'une cible. Le niveau de mesure par défaut est continu. Vous pouvez également indiquer si une sortie sera créée pour la cible. Par exemple, pour un ensemble d'équations couplées, il se peut que vous soyez intéressé uniquement par la génération de la sortie de la cible de l'équation finale, par conséquent dans un tel cas vous supprimerez la sortie 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 n de mesure détermine l'ensemble par défaut des distributions prises en compte lors de la recherche de la distribution la mieux adaptée à l'entrée (cliquez sur **Ajustement** ou **Ajuster tout** dans l'onglet Simulation). Par exemple, si le niveau de mesure est continu, la distribution normale (appropriée aux données continues) sera prise en compte, mais pas la distribution binomiale.

Remarque : Sélectionnez un niveau de mesure de chaîne pour les entrées de chaîne. Les entrées de chaîne à simuler sont limitées à la distribution catégorielle.

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.

Libellés de valeurs : Vous pouvez spécifier des libellés de valeurs pour les cibles, les entrées simulées et les entrées fixes. Les libellés de valeur sont utilisés dans les graphiques et les tableaux de sortie.

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. Pour les entrées catégorielles, vous pouvez spécifier que les associations qui existent entre ces entrées du jeu de données actif soient utilisées lorsque les données sont générées pour ces entrées.
- Spécifier les options avancées telles que l'échantillonnage des extrémités et les critères d'ajustement des distributions aux données historiques.
- Personnaliser la sortie.
- Spécifier l'emplacement d'enregistrement du plan de simulation et éventuellement enregistrer les données simulées.

Champs simulés

Pour exécuter une simulation, chaque champ d'entrée doit être spécifié en tant que champ fixe ou simulé. Les entrées simulées sont celles dont les valeurs sont incertaines et seront générées à partir d'une distribution de probabilité spécifique. Lorsque les données historiques sont disponibles pour les entrées à simuler, les distributions les mieux adaptées peuvent être déterminées automatiquement, ainsi que les corrélations entre ces entrées. 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.

Pour les simulations basées sur un modèle prédictif, chaque prédicteur du modèle représente un champ d'entrée pour la simulation. Pour les simulations qui n'incluent pas de modèle prédictif, les champs spécifiés sur l'onglet Modèle représentent les entrées de la simulation.

Ajustement automatique des distributions et calcul automatique des corrélations pour les entrées simulées : Si le jeu de données actif contient des données historiques relatives aux entrées à simuler, vous pouvez déterminer automatiquement les distributions les mieux adaptées à ces entrées, ainsi que les corrélations entre celles-ci. Les étapes à suivre sont les suivantes :

1. Vérifiez que chaque entrée que vous souhaitez simuler est mise en correspondance avec le champ approprié dans le jeu de données actif. Les entrées sont répertoriées dans la colonne Entrée et la colonne Ajuster à affiche le champ correspondant dans le jeu de données actif. Vous pouvez mettre les entrées en correspondance avec un champ différent du jeu 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 du jeu de données actif. Par défaut, les entrées sont mises en correspondance avec des champs du jeu de données dont le nom, le niveau de mesure et le type (numérique ou chaîne) correspondent. Si le jeu 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.
2. 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 tracé de la distribution superposé à un histogramme (graphique à barres) 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, voir «Détails de l'ajustement», à la page 187.

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 du jeu de données actif sont calculées automatiquement.

Remarque :

- Les observations comportant des valeurs manquantes pour n'importe quelle entrée simulée sont exclues de l'ajustement de la distribution, du calcul des corrélations et du calcul du tableau de contingence facultatif (pour les entrées avec une distribution Catégorielle). Vous pouvez également indiquer si les valeurs manquantes de l'utilisateur des entrées comportant une distribution catégorielle sont traitées comme étant valides. Par défaut, elles sont traitées comme étant manquantes. Pour plus d'informations, voir la rubrique «Options avancées», à la page 189.
- 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 catégorielle des données historiques.

Spécification manuelle des distributions : Vous pouvez spécifier manuellement la distribution de probabilité de n'importe quelle entrée simulée en sélectionnant une distribution 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 tracé de la distribution s'appuyant sur les paramètres spécifiés s'affiche à côté de la grille Paramètres. Voici quelques remarques sur des distributions particulières :

- **Catégorielle :** La distribution catégorielle décrit un champ d'entrée comportant un nombre fixe de valeurs, appelées catégories. Chaque catégorie possède une probabilité associée, de sorte que la somme des probabilités de toutes les catégories est égale à 1. Pour entrer une catégorie, cliquez sur la colonne de gauche dans la grille Paramètres et indiquez la valeur de catégorie. Entrez la probabilité associée à la catégorie dans la colonne de droite.

  Remarque : Les entrées catégorielles d'un modèle PMML possèdent des catégories qui sont définies à partir du modèle et qui ne sont pas modifiables.
- **Binomiale négative - Echecs :** Décrit la distribution du nombre d'échecs dans une séquence d'essais 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 essai donné.
- **Binomiale négative - Essais :** Décrit la distribution du nombre d'essais 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 essai donné.
- **Plage :** 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 Plage 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 (essais et échecs, 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 : Le verrouillage d'une entrée, 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 ou une valeur fixe 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 entrées. A cette fin, les spécifications 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 sensibilité, 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, voir «Analyse de sensibilité», à la page 188.

Icônes de statut d'ajustement

Les icônes de la colonne Ajuster à indiquent le statut d'ajustement de chaque champ d'entrée.

Tableau 3. Icônes de statut.

Icône	Description
	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 le jeu de données actif. Aucune action n'est nécessaire sauf si vous souhaitez réajuster la distribution de l'entrée à le jeu de données actif.
	La distribution la mieux adaptée a été remplacée par une distribution automatique de la boîte de dialogue Ajuster les détails.
	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 par la qualité de l'ajustement, la distribution la plus appropriée étant répertoriée en tête de liste. Vous pouvez remplacer la distribution la mieux adaptée en sélectionnant le bouton radio correspondant à la distribution de votre

choix dans la colonne Utilisation. En sélectionnant un bouton radio dans la colonne Utilisation, vous affichez également un tracé de la distribution superposé à un histogramme (graphique à barres) 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 (essais et échecs, Normale, Poisson et Uniforme. Pour plus d'informations, voir . Pour la distribution catégorielle, les noms des paramètres sont les catégories et les valeurs de 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, les 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 Poisson et Binomiale, ne le sont pas. Vous pouvez choisir un sous-ensemble des distributions par défaut en sélectionnant les distributions 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 comme (Mesure)**, puis en sélectionnant les distributions dans la colonne Réajuster. Cliquez sur **Exécuter le réajustement** pour réajuster l'ensemble de distributions personnalisé.

Remarque :

- Les observations comportant des valeurs manquantes pour n'importe quelle entrée simulée sont exclues de l'ajustement de la distribution, du calcul des corrélations et du calcul du tableau de contingence facultatif (pour les entrées avec une distribution Catégorielle). Vous pouvez également indiquer si les valeurs manquantes de l'utilisateur des entrées comportant une distribution catégorielle sont traitées comme étant valides. Par défaut, elles sont traitées comme étant manquantes. Pour plus d'informations, voir la rubrique «Options avancées», à la page 189.
- 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 catégorielle des données historiques.

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 Catégorielle et Plage, les probabilités des catégories ou des intervalles, respectivement, peuvent être modifiées mais les valeurs des catégories et les points finaux des intervalles ne peuvent l'être. Sélectionnez une catégorie ou un intervalle dans la liste déroulante et spécifiez l'ensemble des probabilités dans la grille des valeurs de paramètres par itération. Les probabilités des autres catégories 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 champs d'entrée à simuler sont souvent corrélés, comme par exemple, la taille et la pondération. 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 le jeu 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 le jeu 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.

Utiliser un tableau de contingence à entrées multiples pour les entrées comportant une distribution catégorielle : Pour les entrées comportant une distribution catégorielle, vous pouvez automatiquement calculer un tableau de contingence à entrées multiple à partir du jeu de données actif qui décrit les associations entre ces entrées. Le tableau de contingence est ensuite utilisé lorsque les données sont générées pour ces entrées. Si vous choisissez d'enregistrer le plan de simulation, le tableau de contingence est enregistré dans le fichier de plan et est utilisé lorsque vous exécutez le plan.

- **Calculer le tableau de contingence à partir du jeu de données actif :** Si vous travaillez avec un plan de simulation existant qui contient déjà un tableau de contingence, il vous est possible de recalculer le tableau de contingence à partir du jeu de données actif. Cette action remplace le tableau de contingence du fichier de plan chargé.
- **Utiliser le tableau de contingence à partir du plan de simulation chargé :** Par défaut, lorsque vous chargez un plan de simulation qui contient un tableau de contingence, le tableau du plan est utilisé. Vous avez la possibilité de recalculer le tableau de contingence à partir du jeu de données actif à l'aide de l'option **Calculer le tableau de contingence à partir du jeu de données actif**.

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 percentile, par ex. le 99e percentile. 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 Percentile dans la liste déroulante **Type**, saisissez une valeur dans la zone de texte Percentile.

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 le jeu de données actif. Si votre jeu 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 le jeu 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 : Pour les entrées continues, la distribution empirique est la fonction de distribution cumulée des données historiques. Cette option vous permet de spécifier le nombre de casiers 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 une valeur de départ 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 629111597.

Valeurs manquantes de l'utilisateur pour les entrées comportant une distribution catégorielle : Option indiquant si les valeurs manquantes de l'utilisateur pour les entrées comportant une distribution catégorielle sont traitées comme étant valides. Les valeurs système manquantes et les valeurs manquantes des utilisateurs pour tous les autres types d'entrée sont traitées comme étant non valides. Toutes les entrées doivent avoir des valeurs valides pour qu'une observation puisse être incluse dans l'ajustement de la distribution, le calcul des corrélations et le calcul d'un tableau de contingence optionnel.

Fonctions de densité

Ces paramètres vous permettent de personnaliser la sortie des fonctions de densité de probabilité et des fonctions de distribution cumulée pour les cibles continues, ainsi que les graphiques à barres des valeurs prédites pour les cibles catégorielles.

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 catégorielles (cibles dont le niveau de mesure est nominal ou ordinal), un graphique à barres est généré, qui affiche le pourcentage d'observations situées dans chaque catégorie de la cible. Des options supplémentaires sont disponibles pour les cibles catégorielles des modèles PMML, dont le paramètre Valeurs de catégorie à rapporter, décrit ultérieurement.

Pour les modèles de cluster en deux étapes et les modèles de cluster en nuées dynamiques, un graphique à barres représentant l'appartenance aux clusters 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.

Positions des curseurs : Vous pouvez spécifier les positions initiales des lignes de référence sur les graphiques FDP et FDC. Les valeurs saisies dans les lignes inférieure et supérieure indiquent les positions le long de l'axe horizontal, et non les percentiles. Vous pouvez supprimer la ligne inférieure en sélectionnant **-Infini** ou la ligne supérieure en sélectionnant **Infini**. Par défaut, les lignes sont positionnées sur les 5e et 95e percentiles. Lorsque plusieurs fonctions de distribution sont affichées sur un même graphique (en raison de cibles ou de résultats multiples provenant des itérations de l'analyse de sensibilité), la valeur par défaut indique la distribution associée à la première itération ou à la première cible.

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 référence peuvent être ajoutées à plus et moins un nombre d'écarts types de la moyenne d'une cible.
- **Percentiles :** Des lignes de références peuvent être ajoutées à une ou deux valeurs de percentiles de la distribution d'une cible, en entrant des valeurs dans les zones de texte Haut et Bas. Par exemple, une valeur de 95 dans la zone de texte Haut représente le 95e percentile, qui est la valeur en dessous de laquelle se trouvent 95 % des observations. De même, une valeur de 5 dans la zone de texte Bas représente le 5e percentile, 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 distribution sont affichées sur un même graphique (en raison de cibles ou de résultats multiples provenant des itérations de l'analyse de sensibilité), les lignes de référence sont appliquées à la distribution de la première itération ou de la première cible. Vous pouvez ajouter des lignes de référence à d'autres distributions depuis la boîte de dialogue Options de graphique, accessible à partir du graphique FDP ou FDC.

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 graphique 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 graphique distinct.

Valeurs de catégorie à rapporter : Pour les modèles PMML comportant des cibles catégorielles, le résultat du modèle est un ensemble de probabilités prédites, une pour chaque catégorie, que la valeur de la cible

tombe dans chaque catégorie. La catégorie présentant la probabilité la plus élevée est considérée comme la catégorie prédite et est utilisée pour générer le graphique à barres décrit dans le paramètre **Fonction de densité de probabilité** sus-mentionné. Sélectionnez **Catégorie prédite** pour générer le graphique à barres. Sélectionnez **Probabilités prédites** pour générer les histogrammes de la distribution des probabilités prédites de chaque catégorie 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 graphique à barres de la catégorie prédite d'une cible catégorielle est représenté sous forme de graphique à barres en cluster qui comprend les résultats de toutes les itérations. Vous pouvez choisir de regrouper les catégories ou les itérations.

Sortie

Graphiques tornado : Les graphiques Tornado sont des graphiques à barres 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 graphique tornado des coefficients de corrélation existant entre une cible donnée et chacune de ses entrées simulées. Ce type de graphique tornado ne prend pas en charge les cibles avec un niveau de mesure nominal ou ordinal ni les entrées simulées avec une distribution catégorielle.
- **Contribution à la variance :** Cette option crée un graphique 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 graphique tornado ne prend pas en charge les cibles avec des niveaux de mesure ordinaux ou nominaux ni les entrées simulées avec l'une des distributions suivantes : catégorielle, Bernoulli, binomiale, Poisson ou binomiale négative.
- **Sensibilité de la cible à modifier :** Cette option crée un graphique 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 graphique tornado ne prend pas en charge les cibles avec des niveaux de mesure ordinaux ou nominaux ni les entrées simulées avec l'une des distributions suivantes : catégorielle, Bernoulli, binomiale, Poisson ou binomiale négative.

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 graphique.

Nuages de points comparant les cibles et les entrées : Les nuages de points comparant les cibles et les entrées simulées sont disponibles à la fois pour les cibles continues et les cibles catégorielles, et comprennent les dispersions des cibles ayant à la fois des entrées continues et catégorielles associées. Les nuages de points utilisant une cible ou une entrée catégorielle sont affichées sous la forme d'une carte thermique.

Créer un tableau des valeurs de percentiles : Pour les cibles continues, vous pouvez obtenir un tableau des percentiles des distributions cible spécifiés. Les quartiles (25e, 50e et 75e percentiles) divisent les observations en quatre groupes de taille égale. Si vous souhaitez un nombre égal de groupes différent de quatre, sélectionnez **Intervalles** et indiquez le nombre souhaité. Sélectionnez **Percentiles personnalisés** pour indiquer des percentiles personnalisés, par ex. le 99e percentile.

Statistiques descriptives des distributions cible : Cette option crée des tableaux de statistiques descriptives pour les cibles continues et catégorielles, 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 percentiles de la distribution cible. Pour les cibles catégorielles, le tableau comprend le pourcentage d'observations qui se situent dans

chaque catégorie de la cible. Pour les cibles catégorielles des modèles PMML, le tableau comprend également la probabilité moyenne de chaque catégorie de la cible. Pour les entrées continues, le tableau comprend la moyenne, l'écart type et les valeurs minimale et maximale.

Corrélations et tableau de contingence pour les entrées : Cette option affiche un tableau des coefficients de corrélation entre les entrées simulées. Lorsque des entrées comportant une distribution catégorielle sont générées à partir d'un tableau de contingence, le tableau de contingence des données générées pour ces entrées est également affiché.

Entrées simulées à inclure dans la sortie : Par défaut, toutes les entrées simulées sont incluses dans les sorties. Vous pouvez exclure certaines entrées simulées des sorties. Dans ce cas, elles seront exclues des graphiques tornado, des nuages de points et des sorties sous forme de tableau.

Limiter les plages des cibles continues : Vous pouvez déterminer la plage des valeurs vides d'une ou de plusieurs cibles continues. Les valeurs non comprises dans cette plage sont exclues de toutes les sorties et analyses associées aux cibles. Pour définir une limite inférieure, sélectionnez **Inférieur** dans la colonne Limite et saisissez une valeur dans la colonne Minimum. Pour définir une limité supérieure, sélectionnez **Supérieur** dans la colonne Limite et saisissez une valeur dans la colonne Maximum. Si vous souhaitez définir les deux limites, sélectionnez **Les deux** dans la colonne Limite et saisissez une valeur dans les colonnes Minimum et Maximum.

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).

Enregistrement

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 sortie des graphiques et tableaux, paramètres d'options avancées pour l'ajustement, la distribution empirique et la valeur de départ aléatoire.

Enregistrer les données simulées en tant que 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 nouveau jeu 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 libellées 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 la sortie.
- 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 une entrée pour chaque champ d'entrée défini dans le plan de simulation. Ces données comprennent le nom de l'entrée et le type de distribution de probabilité associée à l'entrée, ainsi qu'un tracé de la courbe de distribution associée. Chaque entrée se voit également affecter une icône de statut (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 le jeu 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 sensibilité, 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, voir «Analyse de sensibilité», à la page 188.

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 du jeu de données actif :

1. Vérifiez que chaque entrée du modèle est mise en correspondance avec le champ approprié dans le jeu de données actif. Chaque entrée simulée est ajustée au champ du jeu 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 de statut représente une coche et un point d'interrogation, tel qu'illustré ci-dessous.

2. Au besoin, modifiez les champs de mise en correspondance en sélectionnant l'option **Ajuster à un champ dans le jeu de données**, puis en choisissant le champ dans la liste.
3. Cliquez sur **Tout ajuster**.

Pour chaque entrée ajustée, la distribution la mieux adaptée s'affiche en même temps qu'un tracé de la distribution superposé à un histogramme (graphique à barres) 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 de statut, voit «Champs simulés», à la page 185.

Modification des distributions de probabilité

Vous pouvez modifier la distribution de probabilité d'une entrée simulée et éventuellement changer une entrée simulée en entrée fixe, et vice-versa.

1. Sélectionnez l'entrée et choisissez **Définir manuellement la distribution**.
2. Sélectionnez le type de distribution et spécifiez les paramètres correspondants. Pour changer une entrée simulée en entrée fixe, sélectionnez Fixe dans la liste déroulante **Type**.

Une fois les paramètres d'une distribution définis, le tracé représentant la distribution (affiché dans le jeu 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, voir «Champs simulés», à la page 185.

Inclure les valeurs manquantes de l'utilisateur pour les entrées catégorielles lors de l'ajustement : Option indiquant si les valeurs manquantes de l'utilisateur pour les entrées comportant une distribution catégorielle sont traitées comme étant valides lorsque vous réajustez les données dans le jeu de données actif. Les valeurs système manquantes et les valeurs manquantes des utilisateurs pour tous les autres types d'entrée sont traitées comme étant non valides. Toutes les entrées doivent avoir des valeurs valides pour qu'une observation puisse être incluse dans l'ajustement de la distribution et le calcul des corrélations.

Onglet Sortie

L'onglet Sortie vous permet de personnaliser les sorties générées 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 graphique à barres est généré, qui affiche le pourcentage d'observations situées dans chaque catégorie 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.

Graphiques tornado : Les graphiques Tornado sont des graphiques à barres 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 graphique tornado des coefficients de corrélation existant entre une cible donnée et chacune de ses entrées simulées.
- **Contribution à la variance :** Cette option crée un graphique 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 graphique 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.

Nuages de points comparant les cibles et les entrées : Cette option génère des nuages de points des cibles par rapport aux entrées simulées.

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 percentiles de cette distribution et ils divisent les observations en quatre groupes de taille égale.

Corrélations et tableau de contingence pour les entrées : Cette option affiche un tableau des coefficients de corrélation entre les entrées simulées. Une tableau de contingence des associations entre les entrées comportant une distribution catégorielle est affiché lorsque le plan de simulation indique la génération des données catégorielles à partir d'un tableau de contingence.

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 graphique. Ce paramètre s'applique aux graphiques 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 graphique.

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 sortie.

Enregistrer les données simulées en tant que 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 nouveau jeu 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 libellées par le numéro d'itération.

Si vous avez besoin de personnaliser davantage les sorties, exécutez plutôt votre simulation dans le Générateur de simulation. Pour plus d'informations, voir «Exécution d'une simulation à partir d'un plan de simulation», à la page 181.

Utilisation de la sortie graphique créée par la simulation

Plusieurs graphiques générés par la simulation offrent des fonctions interactives vous permettant de personnaliser leur affichage. Ces fonctions interactives sont disponibles en activant (par double-clic) l'objet de graphique dans le visualiseur de sortie. Tous les graphiques de simulation sont des visualisations graphiques.

Graphiques de fonction de densité de probabilité pour les cibles continues : Ce type de graphique présente deux lignes de référence verticales pouvant être déplacées par glissement, qui divisent le graphique en zones distinctes. Le tableau en dessous du graphique affiche la probabilité que la cible se trouve dans chacune des zones. Si plusieurs fonctions de densité sont affichées sur le même graphique, 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 fonctions supplémentaires sont disponibles en cliquant sur le bouton **Options de graphique** situé sur le graphique. Vous pouvez notamment définir explicitement les positions des curseurs, ajouter des lignes de référence fixes et modifier l'affichage du graphique, pour passer par exemple d'une courbe continue à un histogramme ou vice-versa. Pour plus d'informations, voir «Options de graphique», à la page 197.

Graphiques de fonction de distribution cumulée pour les cibles continues : Ce type de graphique présente les mêmes deux lignes de référence verticales et le même tableau associé décrits ci-dessus pour les graphiques de fonction de densité de probabilité. Il offre également un accès à la boîte de dialogue Options de graphique 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, voir «Options de graphique», à la page 197.

Graphiques à barres pour les cibles catégorielles avec itérations d'analyse de sensibilité : Pour les cibles catégorielles avec itérations d'analyse de sensibilité, les résultats de la catégorie cible prédite sont affichés sous la forme d'un graphique à barres en cluster qui comprend les résultats de toutes les itérations. Le graphique comprend une liste déroulante qui vous permet de regrouper par catégorie ou par itération. Pour les modèles de cluster en deux étapes et les modèles de cluster en nuées dynamiques, il est possible de regrouper par cluster 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 graphique unique, sous la forme d'une boîte à moustache juxtaposée. Le graphique comprend une liste déroulante qui vous permet de regrouper par cible ou par itération.

Options de graphique

La boîte de dialogue Options de graphique vous permet de personnaliser l'affichage des graphiques 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 graphique de fonction de densité de probabilité. Elle vous permet de faire basculer l'affichage d'une courbe continue à un histogramme. Cette fonction n'est pas disponible lorsque plusieurs fonctions de densité sont affichées sur le même graphique. 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 graphique 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. Lorsque plusieurs fonctions sont affichées sur un même graphique (en raison de cibles ou de résultats multiples provenant des itérations de l'analyse de sensibilité), vous pouvez indiquer les fonctions auxquelles les lignes sont appliquées.

- **Sigmas :** Des lignes de référence peuvent être ajoutées à plus et moins un nombre d'écarts types de la moyenne d'une cible.
- **Percentiles :** Des lignes de références peuvent être ajoutées à une ou deux valeurs de percentiles de la distribution d'une cible, en entrant des valeurs dans les zones de texte Haut et Bas. Par exemple, une valeur de 95 dans la zone de texte Haut représente le 95e percentile, qui est la valeur en dessous de laquelle se trouvent 95 % des observations. De même, une valeur de 5 dans la zone de texte Bas représente le 5e percentile, qui est la valeur en dessous de laquelle se trouvent 5 % des observations.
- **Positions personnalisées :** Des lignes de référence peuvent être ajoutées à des valeurs spécifiques le long de l'axe horizontal.

Lignes de référence des libellés : Cette option permet de contrôler si les libellés sont appliqués aux lignes de référence sélectionnées.

Les lignes de référence peuvent être supprimées en décochant l'option associée dans la boîte de dialogue Options de graphique avant de cliquer sur **Poursuivre**.

Chapitre 35. Modélisation géospatiale

Les techniques de modélisation géospatiales ont été conçues pour découvrir les motifs dans les données qui incluent un composant géospatial (carte). L'assistant de modélisation géospatiale fournit des méthodes d'analyse des données géospatiales avec et sans composant temporel.

Rechercher des associations en fonction de l'événement et des données géospatiales (Règles d'association géospatiales)

A l'aide des règles d'association géospatiales, vous pouvez rechercher des motifs dans les données en fonction des propriétés spatiales et non spatiales. Par exemple, vous pouvez identifier des motifs dans les données de crime par le biais de l'emplacement et des attributs démographiques. A partir de ces motifs, vous pouvez ensuite générer des règles qui établissent des prévisions quant à l'endroit supposé de certains crimes.

Effectuer des prévisions à l'aide des séries temporelles et des données géospatiales (prévision spatio-temporelle)

Une prévision spatio-temporelle utilise des données qui contiennent des données d'emplacement, des champs d'entrée pour la prévision (prédicteurs), un ou plusieurs champs Heure et un champ cible. Chaque emplacement comporte de nombreuses lignes dans les données qui représentent les valeurs de chaque prédicteur et la cible dans chaque intervalle de temps.

Utilisation de l'assistant de modélisation géospatiale

1. A partir des menus, sélectionnez :

   Analyser > **Modélisation spatio-temporelle** > **Modélisation spatiale**
2. Suivez les instructions indiquées dans l'assistant.

Sélection de cartes

La modélisation géospatiale peut utiliser une ou plusieurs sources de données de carte. Ces dernières contiennent des informations qui définissent les zones géographiques et d'autres fonctions géographiques, telles que des routes ou des rivières. Un grand nombre de sources de cartes contiennent également des données démographiques ou descriptives, ainsi que des données d'événement, par exemple un rapport sur un crime ou un taux de chômage. Vous pouvez vous servir d'un fichier de spécification de carte défini précédemment ou alors définir des spécifications de carte et les enregistrer pour une utilisation ultérieure.

Charger une spécification de carte

Charge un fichier de spécification de carte (.mplan) précédemment défini. Les sources de données de carte que vous définissez ici peuvent être enregistrées dans un fichier de spécification de carte. Dans le cadre d'une prévision spatio-temporelle, si vous sélectionnez un fichier de spécification de carte qui identifie plusieurs cartes, le système vous invite à sélectionner celle de votre choix dans le fichier.

Ajouter un fichier de carte

Ajoutez un fichier de forme (.shp) ESRI ou une archive .zip qui contient un fichier de forme ESRI.

- Un fichier .dbf correspondant doit se trouver au même emplacement que le fichier .shp et tous deux doivent posséder le même nom racine.
- Si le fichier est une archive .zip, les fichiers .shp et .dbf doivent posséder le même nom racine que l'archive.
- S'il n'y a pas de fichier de projection (.prj) correspondant, le système vous invite à sélectionner un système de projection.

Relation
Pour les règles d'association géospatiales, cette colonne indique la manière dont les événements sont reliés aux fonctions dans la carte. Ce paramètre n'est pas disponible pour la prévision spatio-temporelle.

Déplacer vers le haut, Déplacer vers le bas
L'ordre des couches des éléments de la carte est déterminé par l'ordre dans lequel ils apparaissent dans la liste. La première carte de la liste correspond à la couche du bas.

Sélection d'une carte

Dans le cadre d'une prévision spatio-temporelle, si vous sélectionnez un fichier de spécification de carte qui identifie plusieurs cartes, le système vous invite à sélectionner celle de votre choix dans le fichier. La prévision temporelle spatiale ne prend pas en charge les cartes multiples.

Relation géospatiale

Pour les règles d'association géospatiales, la boîte de dialogue Relation géospatiale définit la manière dont les événements sont reliés aux fonctions dans la carte.

- Ce paramètre ne s'applique qu'aux règles d'association géospatiales.
- Il n'affecte que les sources de données associées à des cartes qui sont spécifiées en tant que données de contexte lors de l'étape de sélection des sources de données.

Relation

Proche
L'événement se produit près d'une zone ou d'un point donné sur la carte.

Dans L'événement se produit dans une zone donnée sur la carte.

Contient
La zone d'événement contient un objet de contexte de carte.

Intersection
Emplacement correspondant à un point d'intersection de lignes ou de régions provenant de différentes cartes.

Croisé Lorsque plusieurs cartes co-existent, points de croisement des différentes lignes (routes, rivières, chemins de fer).

Au nord de, Au sud de, A l'est de, A l'ouest de
L'événement se produit dans la zone nord, sud, est ou ouest d'un point donné sur la carte.

Définition d'un système de coordonnées

Si un fichier de projection (.prj) n'est pas associé à la carte ou si vous définissez deux champs d'une source de données comme ensemble de coordonnées, vous devez définir le système de coordonnées.

Système géographique par défaut (longitude et latitude)
Le système de coordonnées est celui de la longitude et latitude.

Système cartésien simple (X et Y)
Le système de coordonnées est celui des coordonnées X et Y simple.

Utiliser un ID connu (WKID)
"ID connu" pour les projections standard.

Utiliser un nom de système de coordonnées
Le système de coordonnées est basé sur la projection citée. Le nom est mentionné entre parenthèses.

Définition de projection

Si le système de projection ne peut être déterminé à partir des informations fournies avec la carte, vous devez le spécifier. La cause la plus courante de cette situation est le fait qu'il n'y a pas de fichier de projection (.prj) associé à la carte ou que le fichier de projection ne peut être utilisé.

- **Une ville, une région ou un pays (projection de Mercator)**
- **Un grand pays, plusieurs pays ou des continents (projection de Winkel Tripel)**
- **Une zone proche de l'équateur (projection de Mercator)**
- **Une zone proche de l'un des pôles (projection stéréographique)**

La projection de Mercator est une projection couramment utilisée dans beaucoup de cartes. Cette projection traite le globe comme un cylindre qui est roulé sur une surface plane. La projection de Mercator déforme la taille et la forme des grands objets. Cette distorsion augmente au fur et à mesure que vous vous éloignez de l'équateur et que vous vous rapprochez des pôles. Les projections de Winkel Tripel et stéréographique font des ajustements pour compenser le fait qu'une carte représente une portion de sphère en trois dimensions affichée en deux dimensions.

Projection et système de coordonnées

Si vous sélectionnez plusieurs cartes et que celles-ci possèdent différents systèmes de coordonnées et projections, vous devez sélectionner la carte avec le système de projection que vous souhaitez utiliser. Ce système de projection sera ensuite utilisé pour toutes les cartes lorsqu'elles sont associées dans la sortie.

Sources de données

Une source de données peut être un fichier dBase fourni avec le fichier de forme, un fichier de données IBM SPSS Statistics ou un jeu de données ouvert dans la session en cours.

Données de contexte. Les données de contexte identifient les fonctions sur la carte. Elles peuvent aussi contenir des champs pouvant être utilisés comme entrées pour le modèle. Pour utiliser un fichier de contexte dBase (.dbf) qui est associé à un fichier de forme carte (.shp), le fichier de contexte dBase doit se trouver au même emplacement que le fichier de forme et posséder le même nom racine. Ainsi, si le fichier de forme s'appelle `geodata.shp`, le fichier dBase doit s'appeler `geodata.dbf`

Données d'événement. Les données d'événement comportent des informations sur les événements qui se produisent, notamment des crimes ou des accidents. Cette option est disponible uniquement pour les règles d'association géospatiales.

Densité de point : Intervalle de temps et données de coordonnées pour les estimations de la densité par noyau. Cette option n'est disponible que pour la prévision spatio-temporelle.

Ajouter : Ouvre une boîte de dialogue dans laquelle ajouter des sources de données. Une source de données peut être un fichier dBase fourni avec le fichier de forme, un fichier de données IBM SPSS Statistics ou un jeu de données ouvert dans la session en cours.

Associer : Ouvre une boîte de dialogue permettant de spécifier les identificateurs (coordonnées ou clés) utilisés pour associer les données aux cartes. Chaque source de données doit contenir un ou plusieurs identificateurs permettant d'associer les données à la carte. Les fichiers dBase qui sont livrés avec le fichier de forme contiennent généralement un champ qui est automatiquement utilisé comme identificateur par défaut. Pour les autres sources de données, vous devez spécifier les champs qui sont utilisés comme identificateurs.

Valider la clé : Ouvre une boîte de dialogue permettant de valider la correspondance des clés entre la carte et la source de données.

Règles d'association géospatiales

- Au moins une source de données doit être une source de données d'événement.
- Toutes les sources de données d'événement doivent utiliser les mêmes identificateurs d'association de règle : coordonnées ou valeurs de clés.
- Si les sources de données d'événement sont associées aux cartes avec des valeurs clés, toutes les sources d'événement doivent utiliser le même type de fonction de carte (par exemple, polygones, points, lignes).

Prévision temporelle spatiale

- Une source de données contextuelle doit exister.
- S'il existe une seule source de données (un fichier de données sans carte associée), elle doit inclure les valeurs de coordonnées.
- Si vous disposez de deux sources de données, une source de données doit correspondre aux données contextuelles et l'autre, aux données de densité des points.
- Vous ne pouvez inclure plus de deux sources de données.

Ajout d'une source de données

Une source de données peut être un fichier dBase fourni avec les fichiers de forme et de contexte, un fichier de données IBM SPSS Statistics ou un jeu de données ouvert dans la session en cours.

Vous pouvez ajouter plusieurs fois une même source de données si sous souhaitez utiliser une association spatiale différente avec chacune d'entre elle.

Association de carte et données

Chaque source de données doit contenir un ou plusieurs identificateurs permettant d'associer les données à la carte.

Coordonnées
: La source de données comporte des champs qui représentent des données cartésiennes. Sélectionnez ceux qui représentent les coordonnées X et Y. Pour les règles d'association géospatiales, vous pouvez également disposer de coordonnées Z.

Valeurs de clés
: Les valeurs de clés figurant dans les champs de la source de données correspondent aux clés de cartes sélectionnées. Ainsi, une carte des régions peut avoir un identificateur de nom (clé de carte) fournissant un libellé à chaque région. L'identificateur correspond à un champ de données qui contient également le nom des régions (clé de données). Les champs sont mis en corrélation afin de mapper les clés en fonction de l'ordre dans lequel elles apparaissent dans les deux listes.

Valider les clés

La boîte de dialogue Valider les clés fournit un récapitulatif de la correspondance des enregistrements entre la carte et la source de données, basé sur les clés d'identificateur sélectionnées. Si des valeurs de clés de données ne correspondent pas, vous pouvez les faire correspondre manuellement aux valeurs de clés de cartes.

Règles d'association géospatiales

Pour les règles d'association géospatiales, après défini les cartes et les sources de données, les étapes restantes de l'assistant sont les suivantes :

- Si plusieurs sources de données d'événement existent, définissez leurs modalités de fusion.
- Sélectionnez les champs à utiliser en tant que conditions et prévisions dans l'analyse.

Eventuellement, vous pouvez aussi :

- Sélectionner des options de sortie différentes.
- Enregistrer un fichier de modèle d'évaluation.
- Créer de nouveaux champs pour les prévisions et les règles des sources de données utilisées dans le modèle.
- Personnaliser les paramètres de génération de règles d'association.
- Personnaliser les paramètres de regroupement et d'agrégation.

Définition de champs de données d'événement

Pour les règles d'association géospatiales, s'il existe plusieurs sources de données d'événement, celles-ci sont fusionnées.

- Par défaut, seuls les champs communs à toutes les sources de données d'événement sont inclus.
- Vous pouvez afficher la liste des champs communs, des champs spécifiques à une source de données ou des champs de toutes les sources de données, puis sélectionner ceux que vous voulez inclure.
- Pour les champs communs, les options **Type** et **Mesure** doivent être identiques pour toutes les sources de données. En cas de conflit, vous pouvez préciser le type et le niveau de mesure de chaque champ commun.

Sélection des champs

La liste des champs disponibles inclut les champs provenant des sources de données d'événement et ceux provenant des sources de données contextuelles.

- Vous pouvez contrôler la liste des champs affichés en sélectionnant une source de données dans la liste **Sources de données**.
- Vous devez sélectionner au moins deux champs, dont l'un doit être une condition et l'autre une prévision. Vous pouvez atteindre cet objectif de plusieurs manières, notamment en sélectionnant deux champs pour la liste **Les deux (Condition et Prévision)**.
- Les valeurs de prévision des règles d'association reposent sur les valeurs des champs de condition. Par exemple, dans la règle "Si x=1 et y=2, alors z=3", les valeurs x et y sont des conditions et la valeur z est la prévision.

Sortie

Tables de règles
: Chaque table de règles affiche les règles et valeurs principales de confiance, support de règle, lift, support de condition et déployabilité. Chaque table est triée en fonction des valeurs du critère sélectionné. Vous pouvez afficher toutes les règles ou le plus grand **Nombre** de règles, en fonction du critère sélectionné.

Nuage de mots pour tri
: Liste des règles principales en fonction des valeurs du critère sélectionné. La taille du texte indique l'importance relative de la règle. L'objet de sortie interactive contient les règles principales de confiance, de support de règle, de lift, de support de condition et de déployabilité. Les critères sélectionnés déterminent la liste de règles qui est affichée par défaut. Vous pouvez choisir un critère différent de façon interactive dans la sortie. L'option **Nombre maximal de règles à afficher** détermine le nombre de règles qui sont affichées dans la sortie.

Carte
: Carte et diagramme à barres interactifs des principales règles, en fonction du critère sélectionné. Chaque objet de sortie interactive contient les règles principales de confiance, de support de règle, de lift, de support de condition et de déployabilité. Le critère sélectionné détermine la liste de règles qui est affichée par défaut. Vous pouvez choisir un critère différent de façon interactive dans la sortie. L'option **Nombre maximal de règles à afficher** détermine le nombre de règles qui sont affichées dans la sortie.

Tables d'informations sur le modèle

Transformations de champs (zones).
Décrit les transformations qui sont appliquées aux champs utilisés dans l'analyse.

Récapitulatif des enregistrements
Nombre et pourcentage d'enregistrements inclus et exclus.

Statistiques de règles
Statistiques récapitulatives sur le support de condition, la confiance, le support de règle, le lift et la déployabilité. Ces statistiques incluent la moyenne, les valeurs minimum et maximum et l'écart type.

Eléments les plus fréquents
Eléments qui se produisent le plus fréquemment. Un élément est inclus dans une condition ou une prévision d'une règle. Par exemple, âge < 18 ou genre=féminin.

Champs les plus fréquents
Champs qui figurent le plus fréquemment dans les règles.

Entrées exclues
Champs qui sont exclus de l'analyse et raison de l'exclusion.

Critère des tables de règles, nuages de mots et cartes

Confiance
Pourcentage des prévisions de règles correctes.

Support de règle
Pourcentage des observations pour lesquelles la règle est true (vraie). Par exemple, si la règle est "Si x=1 et y=2, alors z=3," le support de règle correspond au pourcentage réel d'observations dans les données pour lequel x=1, y=2 et z=3.

Lift Le lift détermine la mesure dans laquelle les prévisions définies par la règle sont plus efficaces que celles définies de manière aléatoire. Il s'agit du rapport entre les prévisions correctes et l'occurrence générale de la prévision. La valeur doit être supérieure à 1. Par exemple, si la prévision se produit dans 20 % des cas et que la confiance de la prévision est de 80 %, la valeur de lift est de 4.

Support de condition
Pourcentage des observations pour lequel la condition de règle existe. Par exemple, si la règle est "Si x=1 et y=2, alors z=3," le support de condition correspond à la proportion d'observations dans les données pour lesquelles x=1 et y=2.

Déployabilité
Pourcentage des prédictions incorrectes quand les conditions sont vérifiées. La déployabilité se calcule ainsi : (confiance à 1) multiplié par support de condition, ou support de condition moins support de règle.

Enregistrer

Enregistrer la carte et les données de contexte en tant que spécification de carte
Enregistrez les spécifications de carte dans un fichier externe (.mplan). Vous pouvez ensuite charger ce fichier dans l'assistant à des fins d'analyse ultérieure. Vous pouvez également l'utiliser avec la commande `SPATIAL ASSOCIATION RULES`.

Copier les fichiers de carte et de données dans la spécification
Les données issues des fichiers de forme carte, des fichiers de données externes et des jeux de données utilisés dans la spécification de carte sont enregistrées dans le fichier de spécification de carte.

Evaluation
Enregistre les meilleures valeurs de règle, valeurs de confiance des règles et valeurs d'ID numérique des règles en tant que nouveaux champs dans la source de données spécifiée.

Sources de données à évaluer
Sources de données dans lesquelles sont créés les nouveaux champs. La source de données est ouverte dans la session en cours, si ce n'est pas déjà fait. Vous devez explicitement enregistrer le fichier modifié pour pouvoir enregistrer les nouveaux champs.

Valeurs cible
Créez de nouveaux champs pour les champs cible sélectionnés (prévision).

- Deux nouveaux champs sont créés pour chaque cible : la valeur prévue et la valeur de confiance.
- Pour les champs cible continus (échelle), la prévision est une chaîne décrivant une plage de valeurs. Une valeur au format "(valeur1, valeur2]" signifie "supérieure à la valeur1 et inférieure ou égale à la valeur2".

Nombre de meilleures règles
Créez de nouveaux champs pour le nombre de meilleures règles spécifié. Trois nouveaux champs sont créés pour chaque règle, à savoir, la valeur de règle, la valeur de confiance et la valeur d'ID numérique de la règle.

Préfixe de nom
Préfixe à utiliser pour les nouveaux noms de champs.

Construction de règle

Les paramètres de construction de règle définissent les critères des règles d'association générées.

Eléments par règle
Nombre de valeurs de champs pouvant être incluses dans les prévisions et conditions de règles. Le nombre total d'éléments ne doit pas dépasser 10. Par exemple, la règle "Si x=1 et y=2, alors z=3" comporte deux éléments de condition et un élément de prévision.

Nombre maximal de prévisions
Nombre maximal de valeurs de champs pouvant se produire dans les prévisions d'une règle.

Nombre maximal de conditions
Nombre maximal de valeurs de champs qui peuvent se produire dans les conditions d'une règle.

Exclure la paire
Empêche les paires de champs indiquées d'être incluses dans la même règle.

Critères de règles

Confiance
Niveau de confiance minimal qu'une règle doit avoir pour être incluse dans la sortie. La confiance correspond au pourcentage de prévisions correctes.

Support de règle
Support de règle minimal qu'une règle doit avoir pour être incluse dans la sortie. La valeur représente le pourcentage d'observations pour lequel la règle est vérifiée (true) dans les données observées. Par exemple, si la règle est "Si x=1 et y=2, alors z=3," le support de règle correspond au pourcentage réel d'observations dans les données pour lesquelles x=1, y=2 et z=3.

Support de condition
Support de condition minimal qu'une règle doit avoir pour être incluse dans la sortie. La valeur représente le pourcentage d'observations pour lequel la condition existe. Par exemple, si la règle est "Si x=1 et y=2, alors z=3", le support de condition correspond au pourcentage d'observations dans les données pour lesquelles x=1 et y=2.

Lift Lift minimal qu'une règle doit avoir pour être incluse dans la sortie. Le lift détermine la

mesure dans laquelle les prévisions définies par la règle sont plus efficaces que celles définies de manière aléatoire. Il s'agit du rapport entre les prévisions correctes et l'occurrence générale de la prévision. Ainsi, si la prévision se produit dans 20 % des cas et que la confiance de la prévision est de 80 %, la valeur de lift est de 4.

Traiter de manière identique
Identifie les paires de champs pouvant être traitées comme un même champ.

Regroupement et agrégation

- L'agrégation est nécessaire lorsque les enregistrements dans les données sont plus nombreux que les fonctions dans la carte. Par exemple, vous possédez des enregistrements de données pour des pays individuels, mais vous disposez d'une carte des Etats.
- Vous pouvez spécifier la méthode de mesure récapitulative d'agrégation pour les champs ordinaux et continus. Les champs nominaux sont agrégés en fonction de la valeur modale.

Continu
Pour les champs continus (échelle), la mesure récapitulative peut être moyenne, médiane ou somme.

Ordinal
Pour les champs ordinaux, la mesure récapitulative peut être médiane, mode, plus élevée ou plus faible.

Nombre de regroupements
Définit le nombre de regroupements pour les champs continus (échelle). Les champs continus sont toujours rassemblés ou "regroupés" dans des plages de valeurs. Par exemple : inférieur ou égal à 5, supérieur à 5 et inférieur ou égal à 10, ou supérieur à 10.

Agréger la carte
Applique l'agrégation à la fois aux données et aux cartes.

Paramètres personnalisés pour les champs spécifiques
Vous pouvez remplacer la mesure récapitulative par défaut et le nombre de regroupements pour des champs spécifiques.

- Cliquez sur l'icône permettant d'ouvrir la boîte de dialogue **Sélecteur de champs** et sélectionnez un champ à ajouter à la liste.
- Dans la colonne **Agrégation**, sélectionnez une mesure récapitulative.
- Pour les champs continus, cliquez sur le bouton dans la colonne **Regroupements** pour spécifier le nombre de regroupements de votre choix dans le champ de la boîte de dialogue **Regroupements**.

Prévision spatio-temporelle

Pour la prévision spatio-temporelle, après avoir défini les cartes et les sources de données, les étapes restantes de l'assistant sont les suivantes :

- Indiquez le champ cible, les champs Heure et les prédicteurs facultatifs.
- Définissez les intervalles de temps ou les périodes cycliques des champs Heure.

Eventuellement, vous pouvez aussi :

- Sélectionner des options de sortie différentes.
- Personnaliser les paramètres de génération de modèle.
- Personnaliser les paramètres d'agrégation.
- Enregistrer les prévisions dans un jeu de données de la session en cours ou dans un fichier de données au format IBM SPSS Statistics.

Sélection des champs

La liste des champs disponibles inclut les champs provenant des sources de données sélectionnées. Vous pouvez contrôler la liste des champs affichés en sélectionnant une source de données dans la liste **Sources de données**.

Cible Le champ cible est obligatoire. La cible est le champ pour lequel les valeurs sont prédites.

- Le champ cible doit être un champ numérique continu (échelle).
- S'il existe deux sources de données, la cible correspond aux estimations de densité du noyau et la "Densité" s'affiche comme nom de cible. Vous ne pouvez pas changer cette sélection.

Prédicteurs

Vous pouvez spécifier un ou plusieurs champs prédicteurs. Ce paramètre est facultatif.

Champs Heure

Vous devez sélectionner un ou plusieurs champs représentant des plages de temps ou sélectionner l'option **Périodes cycliques**.

- S'il existe deux sources de données, vous devez sélectionner les champs Heure de ces deux sources. Ces champs doivent représenter le même intervalle.
- Pour les périodes cycliques, vous devez spécifier les champs qui définissent des cycles périodiques dans le panneau Intervalles de temps de l'assistant.

Intervalles de temps

Les options de ce panneau dépendent du choix de l'option **Champs Heure** ou **Période cyclique** effectué à l'étape de sélection des champs.

Champs Heure

Champs Heure sélectionnés : Si vous sélectionnez un ou plusieurs champs Heure au cours de l'étape de sélection des champs, ceux-ci s'affichent dans cette liste.

Intervalle de temps : Sélectionnez l'intervalle de temps approprié dans la liste. En fonction de l'intervalle de temps, vous pouvez également indiquer d'autres paramètres, tels que l'intervalle entre les observations (incrément) et la valeur de début. Cet intervalle de temps s'applique à tous les champs horaires sélectionnés.

- La procédure suppose que toutes les observations (enregistrements) représentent des intervalles réguliers.
- Selon l'intervalle de temps sélectionné, la procédure peut détecter des observations manquantes ou plusieurs observations d'un même intervalle de temps à agréger. Par exemple, si l'intervalle de temps est exprimé en jours et que la date 2014-10-27 est suivie de la date 2014-10-29, il existe une observation manquante pour la date 2014-10-28. Si l'intervalle de temps est exprimé en mois, plusieurs dates du même mois sont agrégées.
- Pour certains intervalles de temps, le paramètre supplémentaire peut définir des ruptures dans les intervalles régulièrement espacés. Par exemple, si l'intervalle de temps est toujours exprimé en jours, mais que seuls les jours de la semaine sont valides, vous pouvez indiquer qu'une semaine comporte cinq jours et qu'elle commence le lundi.
- Si le champ Heure sélectionné ne représente pas un champ de format de date ni d'heure, l'intervalle de temps est automatiquement défini sur **Périodes** et ne peut pas être modifié.

Champs de cycle

Si vous sélectionnez **Période cyclique** à l'étape de sélection des champs, vous devez indiquer les champs qui définissent les périodes cycliques. Une période cyclique identifie une variation cyclique répétitive, telle que le nombre de mois dans une année ou le nombre de jours de la semaine.

- Vous pouvez spécifier jusqu'à trois champs qui définissent les périodes cycliques.

- Le premier champ de cycle représente le niveau le plus haut du cycle. Par exemple, s'il existe une variation par année, trimestre et mois, le champ qui représente l'année est le champ du premier cycle.
- La longueur du cycle des champs de premier et deuxième cycle correspond à la périodicité au niveau suivant. Par exemple, si les champs de cycle sont année, trimestre et mois, la longueur du premier cycle est 4 et celle du deuxième est 3.
- La valeur de début des champs des deuxième et troisième cycles correspond à la première valeur de chacune des périodes cycliques.
- La longueur du cycle et les valeurs de début doivent être des entiers positifs.

Agrégation

- Si vous sélectionnez des **Prédicteurs** à l'étape de sélection des champs, vous pouvez choisir la méthode récapitulative d'agrégation des prédicteurs.
- L'agrégation est nécessaire lorsqu'un intervalle de temps défini comporte plusieurs enregistrements. Par exemple, si l'intervalle de temps est exprimé en mois, plusieurs dates d'un même mois sont agrégées.
- Vous pouvez spécifier la méthode récapitulative d'agrégation pour les champs continus et ordinaux. Les champs nominaux sont agrégés en fonction de la valeur modale.

Continu
Pour les champs continus (échelle), la mesure récapitulative peut être moyenne, médiane ou somme.

Ordinal
Pour les champs ordinaux, la mesure récapitulative peut être médiane, mode, plus élevée ou plus faible.

Paramètres personnalisés pour les champs spécifiques
Vous pouvez remplacer la mesure récapitulative d'agrégation pour des prédicteurs spécifiques.

- Cliquez sur l'icône permettant d'ouvrir la boîte de dialogue **Sélecteur de champs** et sélectionnez un champ à ajouter à la liste.
- Dans la colonne **Agrégation**, sélectionnez une mesure récapitulative.

Sortie

Cartes

Valeurs cible
Carte de valeurs pour le champ cible sélectionné.

Corrélation
Carte des corrélations.

Clusters
Carte mettant en évidence les clusters d'emplacements similaires.

Seuil de similitude d'emplacement
Similitude qui est requise pour créer des clusters. La valeur doit être un nombre supérieur à zéro et inférieur à 1.

Spécifier le nombre maximal de clusters
Nombre maximal de clusters à afficher.

Tableaux et graphiques d'évaluation de modèle

Spécifications de modèles
Récapitulatif des spécifications qui sont utilisées pour exécuter l'analyse, y compris les champs cible, d'entrée et d'emplacement.

Récapitulatif d'informations temporelles
Identifie les champs d'heure et les intervalles de temps utilisés dans le modèle.

Test des effets dans la structure moyenne
La sortie contient la valeur de statistiques de test, les degrés de liberté et le niveau de confiance pour le modèle corrigé et pour chaque effet.

Structure de la moyenne des coefficients du modèle
La sortie inclut la valeur de coefficient, l'erreur standard, la valeur de statistiques de test, le niveau d'importance et les intervalles de confiance pour chaque terme de modèle.

Coefficients autorégressifs
La sortie inclut la valeur de coefficient, l'erreur standard, la valeur de statistiques de test, le niveau d'importance et les intervalles de confiance pour chaque décalage.

Tests de covariance spatiale
Pour les modèles paramétriques de variogramme, affiche les résultats du test de la qualité d'ajustement pour la structure de covariance spatiale. Ces résultats permettent de déterminer s'il convient de modéliser la structure de covariance spatiale de façon paramétrique ou d'utiliser un modèle non paramétrique.

Covariance spatiale paramétrique
Affiche les estimations de covariance spatiale paramétrique pour les modèles paramétriques de variogramme.

Options de modèle

Paramètres du modèle

Inclure automatiquement une constante
Incluez la constante dans le modèle.

Nombre maximal de décalages d'autorégression
Nombre maximal de décalages d'autorégression. Cette valeur doit être un entier compris entre 1 et 5.

Covariance spatiale
Indique la méthode d'estimation de la covariance spatiale.

Paramétrique
La méthode d'estimation est paramétrique. Elle peut être de type **Gaussien**, **Exponentiel** ou **Exponentiel de puissance**. Pour ce dernier type, vous pouvez spécifier la valeur **Puissance**.

Non paramétrique
La méthode d'estimation est non paramétrique.

Enregistrer

Enregistrer la carte et les données de contexte en tant que spécification de carte
Enregistrez les spécifications de carte dans un fichier externe (.mplan). Vous pouvez ensuite charger ce fichier dans l'assistant à des fins d'analyse ultérieure. Vous pouvez également l'utiliser avec la commande `SPATIAL TEMPORAL PREDICTION`.

Copier les fichiers de carte et de données dans la spécification
Les données issues des fichiers de forme carte, des fichiers de données externes et des jeux de données utilisés dans la spécification de carte sont enregistrées dans le fichier de spécification de carte.

Evaluation
Enregistre les prévisions, la variance et les bornes de confiance inférieure et supérieure dans le fichier de données sélectionné.

- Vous pouvez enregistrer les prévisions dans un jeu de données ouvert dans la session en cours ou dans un fichier de données au format IBM SPSS Statistics.

- Le fichier de données ne peut pas être une source de données utilisée dans le modèle.
- Le fichier de données doit inclure tous les champs d'heure et les prédicteurs utilisés dans le modèle.
- Les valeurs d'heure doivent être supérieures aux valeurs d'heure utilisées dans le modèle.

Avancé

Nombre maximal d'observations avec valeurs manquantes (%)
Pourcentage maximal d'observations comportant des valeurs manquantes.

Niveau d'importance
Niveau d'importance permettant de déterminer si un modèle paramétrique de variogramme est approprié. Cette valeur doit être supérieure à 0 et inférieure à 1. La valeur par défaut est 0,05. Le niveau d'importance est utilisé dans le test de la qualité d'ajustement pour la structure de covariance spatiale. La statistique de qualité d'ajustement permet de déterminer s'il faut utiliser un modèle paramétrique ou non paramétrique.

Facteur d'incertitude (%)
Le facteur d'incertitude est une valeur exprimée en pourcentage qui représente la croissance de l'incertitude pour les prévisions futures. Les bornes supérieure et inférieure de l'incertitude des prévisions augmentent d'après ce pourcentage à chaque étape dans le futur.

Terminer

Dans l'étape finale de l'assistant de modélisation géospatiale, vous pouvez soit exécuter le modèle, soit coller la syntaxe de commande générée dans une fenêtre de syntaxe. Vous pouvez ensuite modifier et enregistrer la syntaxe ainsi générée pour une utilisation ultérieure.

Remarques

Le présent document peut contenir des informations ou des références concernant certains produits, logiciels ou services IBM non annoncés dans ce pays. Pour plus de détails, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial IBM. Toute référence à un produit, logiciel ou service IBM n'implique pas que seul ce produit, logiciel ou service puisse être utilisé. Tout autre élément fonctionnellement équivalent peut être utilisé, s'il n'enfreint aucun droit d'IBM. Il est de la responsabilité de l'utilisateur d'évaluer et de vérifier lui-même les installations et applications réalisées avec des produits, logiciels ou services non expressément référencés par IBM.

IBM peut détenir des brevets ou des demandes de brevet couvrant les produits mentionnés dans le présent document. La remise de ce document ne vous donne aucun droit de licence sur ces brevets ou demandes de brevet. Si vous désirez recevoir des informations concernant l'acquisition de licences, veuillez en faire la demande par écrit à l'adresse suivante :

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Pour le Canada, veuillez adresser votre courrier à :

IBM Director of Commercial Relations
IBM Canada Ltd
3600 Steeles Avenue East
Markham, Ontario
L3R 9Z7 Canada

Les informations sur les licences concernant les produits utilisant un jeu de caractères double octet peuvent être obtenues par écrit à l'adresse suivante :

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

Le paragraphe suivant ne s'applique ni au Royaume-Uni, ni dans aucun pays dans lequel il serait contraire aux lois locales : LE PRESENT DOCUMENT EST LIVRE "EN L'ETAT" SANS AUCUNE GARANTIE EXPLICITE OU IMPLICITE. IBM DECLINE NOTAMMENT TOUTE RESPONSABILITE RELATIVE A CES INFORMATIONS EN CAS DE CONTREFAÇON AINSI QU'EN CAS DE DEFAUT D'APTITUDE A L'EXECUTION D'UN TRAVAIL DONNE. Certaines juridictions n'autorisent pas l'exclusion des garanties implicites, auquel cas l'exclusion ci-dessus ne vous sera pas applicable.

Le présent document peut contenir des inexactitudes ou des coquilles. Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. IBM peut, à tout moment et sans préavis, modifier les produits et logiciels décrits dans ce document.

Les références à des sites Web non IBM sont fournies à titre d'information uniquement et n'impliquent en aucun cas une adhésion aux données qu'ils contiennent. Les éléments figurant sur ces sites Web ne font pas partie des éléments du présent produit IBM et l'utilisation de ces sites relève de votre seule responsabilité.

IBM pourra utiliser ou diffuser, de toute manière qu'elle jugera appropriée et sans aucune obligation de sa part, tout ou partie des informations qui lui seront fournies.

Les licenciés souhaitant obtenir des informations permettant : (i) l'échange des données entre des logiciels créés de façon indépendante et d'autres logiciels (dont celui-ci), et (ii) l'utilisation mutuelle des données ainsi échangées, doivent adresser leur demande à :

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Ces informations peuvent être soumises à des conditions particulières, prévoyant notamment le paiement d'une redevance.

Le logiciel sous licence décrit dans ce document et tous les éléments sous licence disponibles s'y rapportant sont fournis par IBM conformément aux dispositions du Livret Contractuel IBM, des Conditions Internationales d'Utilisation de Logiciels IBM, des Conditions d'Utilisation du Code Machine ou de tout autre contrat équivalent.

Les données de performance indiquées dans ce document ont été déterminées dans un environnement contrôlé. Par conséquent, les résultats peuvent varier de manière significative selon l'environnement d'exploitation utilisé. Certaines mesures évaluées sur des systèmes en cours de développement ne sont pas garanties sur tous les systèmes disponibles. En outre, elles peuvent résulter d'extrapolations. Les résultats peuvent donc varier. Il incombe aux utilisateurs de ce document de vérifier si ces données sont applicables à leur environnement d'exploitation.

Les informations concernant des produits non IBM ont été obtenues auprès des fournisseurs de ces produits, par l'intermédiaire d'annonces publiques ou via d'autres sources disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances ni leur compatibilité. Elle ne peut recevoir aucune réclamation concernant des produits non IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits.

Toute instruction relative aux intentions d'IBM pour ses opérations à venir est susceptible d'être modifiée ou annulée sans préavis, et doit être considérée uniquement comme un objectif.

Le présent document peut contenir des exemples de données et de rapports utilisés couramment dans l'environnement professionnel. Ces exemples mentionnent des noms fictifs de personnes, de sociétés, de marques ou de produits à des fins illustratives ou explicatives uniquement. Toute ressemblance avec des noms de personnes, de sociétés ou des données réelles serait purement fortuite.

LICENCE DE COPYRIGHT :

Le présent logiciel contient des exemples de programmes d'application en langage source destinés à illustrer les techniques de programmation sur différentes plateformes d'exploitation. Vous avez le droit de copier, de modifier et de distribuer ces exemples de programmes sous quelque forme que ce soit et sans paiement d'aucune redevance à IBM, à des fins de développement, d'utilisation, de vente ou de distribution de programmes d'application conformes aux interfaces de programmation des plateformes pour lesquels ils ont été écrits ou aux interfaces de programmation IBM. Ces exemples de programmes n'ont pas été rigoureusement testés dans toutes les conditions. Par conséquent, IBM ne peut garantir expressément ou implicitement la fiabilité, la maintenabilité ou le fonctionnement de ces programmes. Les exemples de programmes sont fournis "EN L'ETAT", sans garantie d'aucune sorte. IBM ne sera en aucun cas responsable des dommages liés à l'utilisation des exemples de programmes.

Toute copie totale ou partielle de ces programmes exemples et des oeuvres qui en sont dérivées doit comprendre une notice de copyright, libellée comme suit :

© (nom de votre société) (année). Des segments de code sont dérivés des Programmes exemples d'IBM Corp.

Marques

IBM, le logo IBM et ibm.com sont des marques d'International Business Machines Corp. dans de nombreux pays. Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés. La liste actualisée de toutes les marques d'IBM est disponible sur la page Web "Copyright and trademark information" à l'adresse www.ibm.com/legal/copytrade.shtml.

Adobe, le logo Adobe, PostScript et le logo PostScript sont des marques d'Adobe Systems Incorporated aux Etats-Unis et/ou dans certains 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 d'Intel Corporation ou de ses filiales aux Etats-Unis et/ou dans certains autres pays.

Linux est une marque de Linus Torvalds aux Etats-Unis et/ou dans certains autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques de Microsoft Corporation aux Etats-Unis et/ou dans certains autres pays.

UNIX est une marque enregistrée de The Open Group aux Etats-Unis et/ou dans certains autres pays.

Java ainsi que toutes les marques et tous les logos incluant Java sont des marques d'Oracle et/ou de ses sociétés affiliées.

Index

A

E

F

G

H

I

J

K

L

M

N

P

Q

R

S

T

U

V

W

Z