# IBM SPSS Complex Samples 20

*Remarque* : Avant d'utiliser ces informations et le produit qu'elles concernent, lisez les informations générales sous Remarques sur p. 286.

Cette version s'applique à IBM® SPSS® Statistics 20 et à toutes les publications et modifications ultérieures jusqu'à mention contraire dans les nouvelles versions.

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

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

# Préface

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

## A propos de IBM Business Analytics

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

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

## Support technique

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

## Support technique pour les étudiants

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

## Service clients

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

## Séminaires de formation

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

## Documents supplémentaires

Les ouvrages *SPSS Statistics : Guide to Data Analysis*, *SPSS Statistics : Statistical Procedures Companion*, et *SPSS Statistics : Advanced Statistical Procedures Companion*, écrits par Marija Norušis et publiés par Prentice Hall, sont suggérés comme documentation supplémentaire. Ces publications présentent les procédures statistiques des modules SPSS Statistics Base, Advanced Statistics et Regression. Que vous soyez novice dans les analyses de données ou prêt à utiliser des applications plus avancées, ces ouvrages vous aideront à exploiter au mieux les fonctionnalités offertes par IBM® SPSS® Statistics. Pour obtenir des informations supplémentaires y compris le contenu des publications et des extraits de chapitres, visitez le site web de l'auteur : *http://www.norusis.com*

# Contenu

# *Partie II: Exemples*

## *13 Assistant d'échantillonnage des échantillons complexes 99*

## *Annexes*

# *Partie I:*
# *Guide de l'utilisateur*

Chapitre
1

# Introduction aux procédures d'échantillons complexes

Les procédures d'analyse effectuées dans les logiciels traditionnels se basent sur les observations d'un fichier de données qui ne représentent qu'un échantillon aléatoire simple de la population ciblée. De nombreuses entreprises et de nombreux chercheurs trouvent cela insuffisant et estiment qu'il est plus économique et pratique d'obtenir des échantillons de façon plus structurée.

L'option Echantillons complexes vous permet de sélectionner un échantillon en fonction d'un plan complexe et d'incorporer les spécifications de plan dans l'analyse de données, garantissant ainsi la validité des résultats.

## Propriétés des échantillons complexes

Un échantillon complexe peut différer d'un échantillon aléatoire simple de plusieurs façons. Dans un échantillon aléatoire simple, des unités d'échantillonnage individuelles sont sélectionnées aléatoirement à probabilité égale et sans remplacement (WOR), directement dans la population entière. A l'opposé, un échantillon complexe spécifique peut avoir une partie ou la totalité des fonctions suivantes :

**Stratification :** Un échantillonnage stratifié nécessite de sélectionner des échantillons indépendamment dans des sous-groupes de population ou des strates qui ne se chevauchent pas. Par exemple, les strates peuvent être des groupes socioéconomiques, des modalités d'emploi, des groupes d'âge ou des groupes ethniques. La stratification vous garantit des tailles d'échantillons adéquates pour les sous-groupes d'intérêt, améliore la précision des estimations globales et permet d'utiliser différentes méthodes d'échantillonnage strate par strate.

**Classification :** L'échantillonnage en grappe nécessite la sélection de groupes d'unités d'échantillonnage ou de classes. Par exemple, les classes peuvent être des écoles, des hôpitaux ou des zones géographiques, et les unités d'échantillonnage peuvent être des étudiants, des patients ou des habitants. La classification est commune dans les plans à plusieurs phases et dans les échantillons de zones (géographiques).

**Phases multiples :** Dans un échantillonnage à plusieurs phases, vous sélectionnez un échantillon de première phase basé sur les classes. Vous créez ensuite un échantillon de deuxième phase en créant des sous-échantillons à partir des classes sélectionnées. Si l'échantillon de la deuxième phase se base sur les sous-classes, vous pouvez ajouter une troisième phase à l'échantillon. Par exemple, dans la première phase d'un sondage, vous pouvez créer un échantillon de villes. Vous pouvez ensuite échantillonner les ménages à partir des villes sélectionnées. Enfin, vous pouvez interroger les individus composant les ménages sélectionnés. Les assistants d'échantillonnage et de préparation d'analyse vous permettent de définir un plan en trois phases.

**Echantillonnage non aléatoire.** Lorsqu'il est difficile d'obtenir une sélection aléatoirement, les unités peuvent être échantillonnées systématiquement (à intervalles fixes) ou séquentiellement.

**Probabilités de sélection inégales :** Lors de l'échantillonnage de classes contenant des nombres d'unités inégaux, vous pouvez utiliser l'échantillonnage de probabilité proportionnelle à la taille (probability proportional to size - PPS) pour qu'une probabilité de sélection de classe soit égale à la proportion d'unités qu'elle contient. L'échantillonnage PPS peut également utiliser des schémas de pondération plus généraux pour sélectionner les unités.

**Echantillonnage sans restriction :** Un échantillonnage sans restricition sélectionne des unités avec remplacement (WR). Par conséquent, une unité individuelle peut être sélectionnée plusieurs fois pour l'échantillon.

**Pondérations d'échantillonnage :** Les pondérations d'échantillonnage sont automatiquement calculées lors de la création d'un plan complexe et correspondent idéalement à la « fréquence » de chaque unité d'échantillonnage dans la population cible. Par conséquent, la somme des pondérations de l'échantillon doit estimer la taille de la population. Les procédures d'analyse d'échantillons complexes nécessitent des pondérations d'échantillonnage pour analyser correctement un échantillon complexe. Ces pondérations doivent être entièrement utilisées dans le module Echantillonnage et ne doivent pas être utilisées avec d'autres procédures d'analyse via la procédure Observations pondérées, qui traite les pondérations comme des réplications d'observations.

## Utilisation des procédures Echantillons complexes

Votre utilisation des procédures Echantillons complexes dépend de vos besoins. Les principaux « types » d'utilisateurs sont ceux qui :

- Planifient et mènent des sondages en fonction de plans complexes et qui analysent éventuellement l'échantillon ultérieurement L'outil principal des sondeurs est l'Assistant d'échantillonnage.
- Analysent les fichiers de données d'échantillons précédemment obtenus en fonction de plans complexes. Avant d'utiliser les procédures d'analyse d'échantillons complexes, vous devez utiliser l'Assistant de préparation d'analyse.

Indépendamment du type d'utilisateur que vous êtes, vous devez fournir les informations sur le plan dans les procédures d'échantillons complexes. Ces informations sont stockées dans un **fichier de plan**, pour que vous puissiez les réutiliser rapidement.

### Fichiers de plan

Un fichier de plan contient les spécifications des échantillons complexes. Il existe deux types de fichiers de plan :

**Plan d'échantillonnage :** Les spécifications fournies dans l'assistant d'échantillonnage définissent un plan d'échantillonnage utilisé pour réaliser un échantillon complexe. Le fichier de plan d'échantillonnage contient ces spécifications. Le fichier de plan d'échantillonnage contient également un plan d'analyse par défaut qui utilise les méthodes d'estimation appropriées pour le plan d'échantillonnage spécifié.

**Plan d'analyse :** Contient des informations nécessaires aux procédures d'analyse d'échantillons complexes. Ces informations permettent de calculer correctement les estimations de variance d'un échantillon complexe. Le plan contient la structure de l'échantillon, les méthodes d'estimation de chaque phase et les références aux variables requises, telles que les pondérations d'échantillon. L'assistant de préparation d'analyse vous permet de créer et de modifier les plans d'analyse.

L'enregistrement de vos spécifications dans un fichier de plan présente plusieurs avantages :

- Un sondeur peut indiquer la première phase d'un plan d'échantillonnage à plusieurs phases et créer les unités de la première phase immédiatement, collecter les informations sur les unités d'échantillonnage de la deuxième phase, puis modifier le plan d'échantillonnage afin d'y inclure la deuxième phase.
- Un analyste qui n'a pas accès au fichier de plan d'échantillonnage peut spécifier un plan d'analyse et faire référence à ce plan depuis chaque procédure d'analyse d'échantillons complexes.
- Le créateur d'échantillons d'usage public à grande échelle peut publier le fichier de plan d'échantillonnage ; ainsi, les instructions sont simplifiées pour les analystes et ceux-ci ne sont pas dans l'obligation de spécifier leurs propres plans d'analyse.

## *Références supplémentaires*

Pour plus d'informations sur les techniques d'échantillonnage, reportez-vous aux textes suivants :

Cochran, W. G. 1977. *Sampling Techniques*, 3rd éd. New York: John Wiley and Sons.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley and Sons.

Kish, L. 1987. *Statistical Design for Research*. New York: John Wiley and Sons.

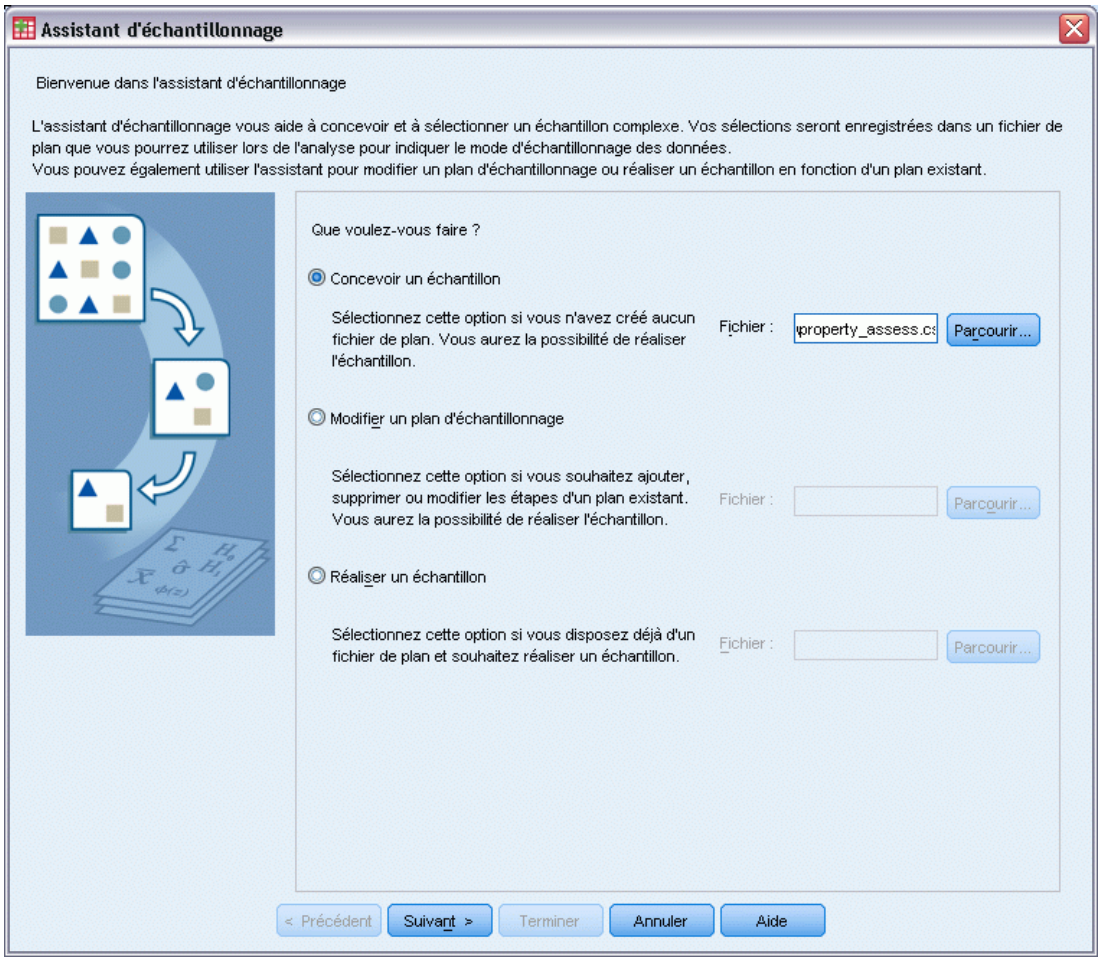
Murthy, M. N. 1967. *Sampling Theory and Methods*. Calcutta, Inde: Statistical Publishing Society.

Särndal, C., B. Swensson, et J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Chapitre
2

# Echantillonnage depuis un plan complexe

Figure 2-1
*Etape Bienvenue de l'assistant d'échantillonnage*




L'assistant d'échantillonnage vous guide lors de la création, de la modification ou de l'exécution d'un fichier de plan d'échantillonnage. Avant d'utiliser l'assistant, pensez à définir précisément la population cible, la liste des unités d'échantillonnage et le plan d'échantillonnage.

## Création d'un plan d'échantillonnage

- A partir des menus, sélectionnez :
  Analyse > Echantillonnage > Sélectionner un échantillon...

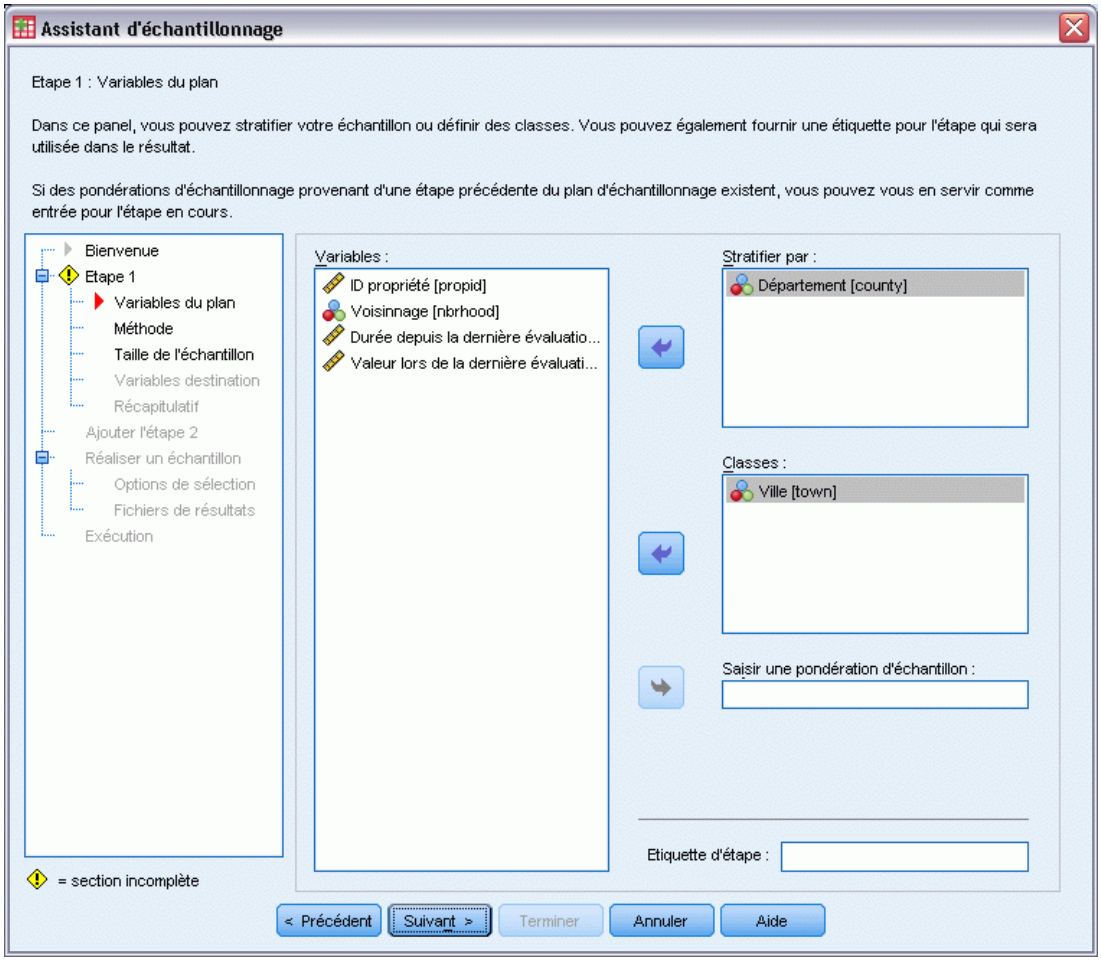
- Sélectionnez Concevoir un échantillon et choisissez le nom de fichier de plan sous lequel enregistrer le plan d'échantillonnage.
- Cliquez sur Suivant pour poursuivre la procédure avec l'assistant.
- A l'étape Variables du plan, vous pouvez éventuellement définir des strates, des classes et saisir des pondérations d'échantillon. Une fois ces éléments définis, cliquez sur Suivant.
- A l'étape Méthode d'échantillonnage, vous pouvez éventuellement choisir une méthode de sélection des éléments.

  Si vous sélectionnez l'échantillonnage de Brewer ou de Murthy avec PPS, vous pouvez cliquer sur Fin pour réaliser l'échantillon. Sinon, cliquez sur Suivant, puis effectuez les opérations suivantes :
- A l'étape Taille de l'échantillon, spécifiez le nombre ou la proportion d'unités à échantillonner.
- Vous pouvez désormais cliquer sur Fin pour réaliser l'échantillon.

  Les étapes suivantes sont facultatives. Elles permettent :
  - Sélectionnez les variables destination à enregistrer.
  - D'ajouter une deuxième ou une troisième étape au plan.
  - Définissez les différentes options de sélection : phases auxquelles les échantillons sont réalisés, générateur de nombre aléatoire. Indiquez également si les valeurs utilisateur manquantes doivent être traitées comme valeurs valides des variables de plan.
  - Sélectionnez l'emplacement d'enregistrement des données de résultat.
  - De coller vos sélections en tant que syntaxe de commande.

# Assistant d'échantillonnage : Variables du plan

Figure 2-2
*Etape Variables de plan de l'assistant d'échantillonnage*




Cette étape vous permet de sélectionner les variables de stratification et de classification, et de définir la saisie des pondérations d'échantillon. Vous pouvez également spécifier une étiquette pour cette étape.

**Stratifier par :** La classification croisée des variables de stratification définit des sous-populations distinctes, ou strates. Des échantillons distincts sont disponibles pour chaque strate. Pour améliorer la précision de vos estimations, les unités comprises dans les strates doivent être homogènes pour les caractéristiques d'intérêt.

**Classes.** Les variables de grappe définissent les groupes d'unités d'observation, ou classes. Les classes sont utiles lorsque l'échantillonnage direct des unités d'observation de la population est coûteux ou impossible. Vous pouvez alors échantillonner des classes de population, puis échantillonner des unités d'observation à partir des classes sélectionnées. Cependant, l'utilisation de classes peut créer des corrélations entre les unités d'échantillonnage. Il peut donc en résulter une perte de précision. Pour minimiser ce risque, les unités comprises dans les classes doivent être hétérogènes pour les caractéristiques d'intérêt. Vous devez définir au moins une variable de

grappe afin de concevoir un plan à plusieurs phases. Les classes sont également nécessaires pour utiliser plusieurs méthodes d'échantillonnage. Pour plus d'informations, reportez-vous à la section Assistant d'échantillonnage : Méthode d'échantillonnage sur p. 8.

**Saisie de pondération d'échantillon.** Si le plan d'échantillonnage actuel fait partie d'un plan d'échantillon plus important, vous devez disposer des pondérations d'échantillon du plan plus important (obtenues lors d'une phase précédente). Vous pouvez spécifier une variable numérique contenant ces pondérations dans la première phase du plan actuel. Les pondérations d'échantillon sont automatiquement calculées pour les étapes suivantes du plan actuel.

**Etiquette d'étape.** Vous pouvez spécifier une étiquette de type chaîne de caractères pour chaque étape. Elle est utilisée dans le résultat pour identifier les informations de chaque étape.

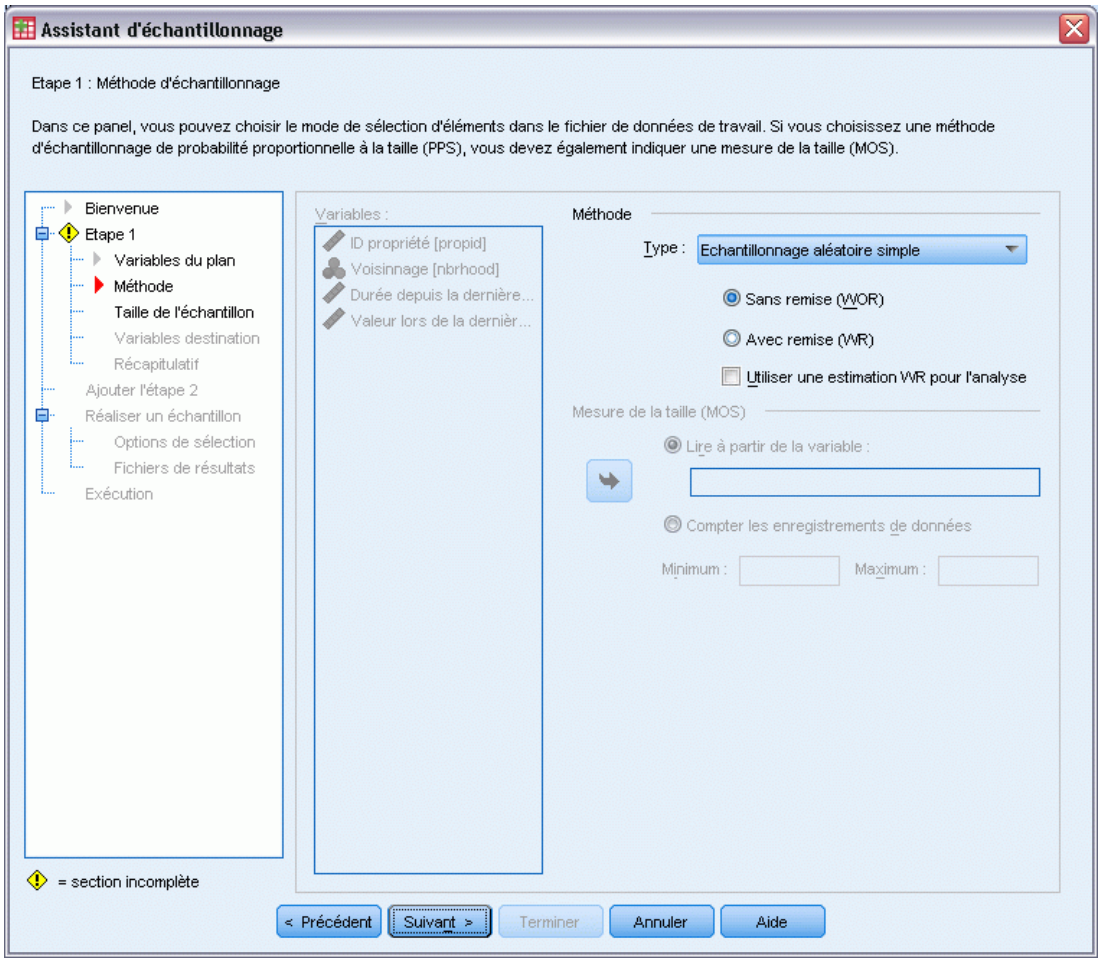
*Remarque :* La liste des variables source compile le contenu obtenu lors de l'exécution des étapes de l'assistant. En d'autres termes, les variables supprimées de la liste source à une étape sont supprimées de toutes les étapes dans la liste. Les variables renvoyées à la liste source apparaissent dans la liste de toutes les étapes.

## *Contrôles d'arbre de navigation de l'assistant d'échantillonnage*

Chaque étape de l'assistant d'échantillonnage dispose à sa gauche d'une légende. Vous pouvez parcourir l'assistant en cliquant sur le nom d'une étape disponible dans la légende. Les étapes sont disponibles tant que les étapes précédentes sont valides. Cela signifie qu'elles le sont si chaque étape précédente a fourni les spécifications minimales requises pour cette étape. Pour plus d'informations sur les raisons de l'invalidité d'une étape, reportez-vous à l'aide de chaque étape.

# *Assistant d'échantillonnage : Méthode d'échantillonnage*

Figure 2-3
*Etape Méthode d'échantillonnage - Assistant d'échantillonnage*




Cette étape vous permet de spécifier le mode de sélection des observations depuis l'ensemble de données actif.

**Méthode :** Les commandes de ce groupe sont utilisées pour choisir une méthode de sélection. Certains types d'échantillonnage vous permettent de choisir entre un échantillonnage avec remplacement (WR) ou sans remplacement (WOR). Pour plus d'informations, reportez-vous aux descriptions des types d'échantillonnage. Certains types de probabilité proportionnelle à la taille (PPS) ne sont disponibles que lorsque les classes ont été définies. Tous les types PPS ne sont disponibles que dans la première phase d'un plan. En outre, les méthodes WR ne sont disponibles que dans l'étape finale d'un plan.

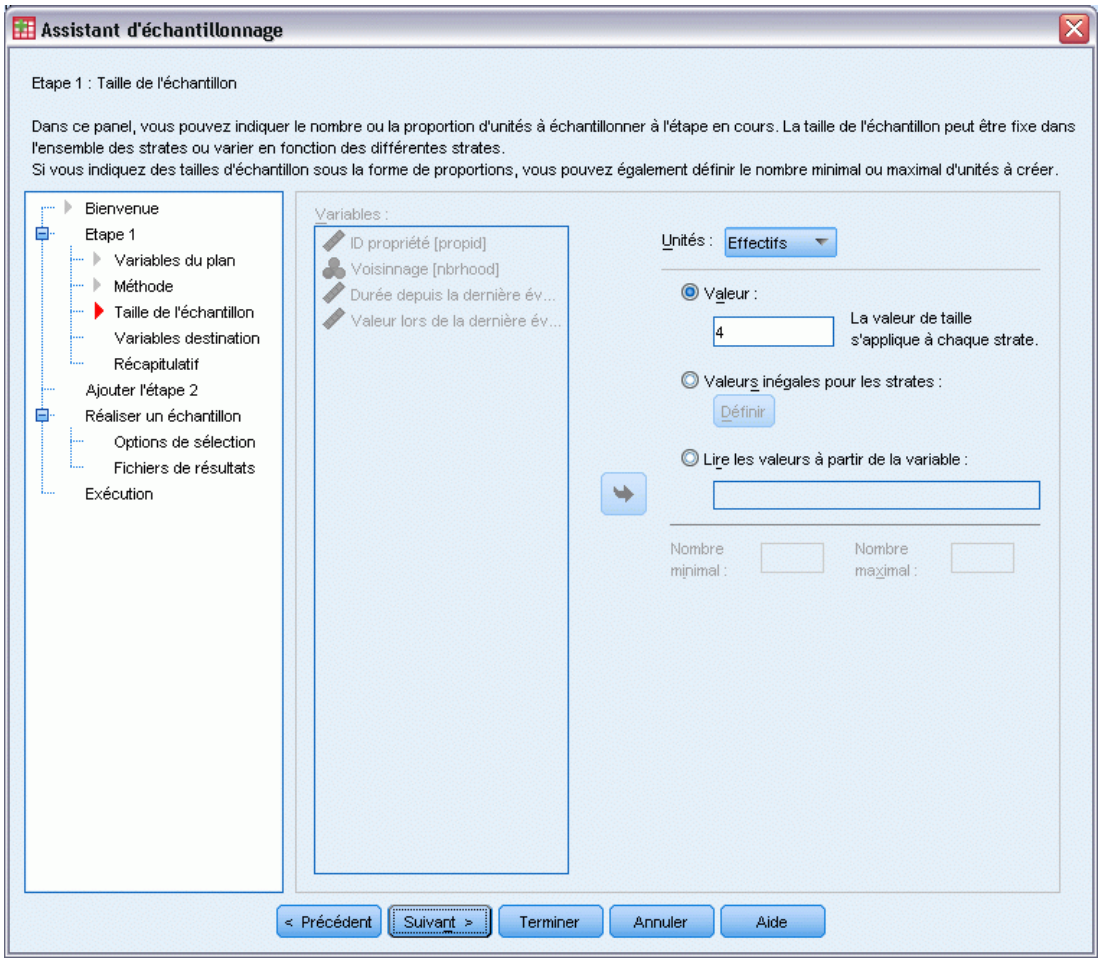
- **Echantillonnage aléatoire simple.** Les unités sont sélectionnées à probabilité égale. Elles peuvent être sélectionnées avec ou sans remplacement.

- **Systématique simple.** Les unités sont sélectionnées à intervalles fixes dans le cadre d'échantillonnage (ou les strates, si celles-ci ont été spécifiées) et extraites sans remplacement. Une unité sélectionnée aléatoirement dans le premier intervalle est définie comme point de départ.
- **Séquentiel simple.** Les unités sont sélectionnées séquentiellement à probabilité égale et sans remplacement.
- **PPS :** Méthode de première phase qui effectue une sélection aléatoire avec probabilité proportionnelle à la taille. N'importe quelle unité peut être sélectionnée avec remplacement ; seules les classes peuvent être échantillonnées sans remplacement.
- **Systématique avec PPS :** Méthode de première phase qui effectue une sélection systématique des unités avec probabilité proportionnelle à la taille. Elles sont sélectionnées sans remplacement.
- **Séquentiel avec PPS :** Méthode de première phase qui effectue une sélection séquentielle des unités avec probabilité proportionnelle à la taille de la classe et sans remplacement.
- **Brewer avec PPS :** Méthode de première phase qui sélectionne deux classes dans chaque strate avec probabilité proportionnelle à la taille de la classe et sans remplacement. Une variable de grappe doit être spécifiée pour que cette méthode puisse être utilisée.
- **PPS Murthy :** Méthode de première phase qui sélectionne deux classes dans chaque strate avec probabilité proportionnelle à la taille de la classe et sans remplacement. Une variable de grappe doit être spécifiée pour que cette méthode puisse être utilisée.
- **Sampford avec PPS :** Méthode de première phase qui sélectionne plus de deux classes dans chaque strate avec probabilité proportionnelle à la taille de la classe et sans remplacement. Il s'agit d'une extension de la méthode Brewer. Une variable de grappe doit être spécifiée pour que cette méthode puisse être utilisée.
- **Utiliser une estimation WR pour l'analyse :** Par défaut, une méthode d'estimation est spécifiée dans le fichier de plan. Celle-ci est adaptée à la méthode d'échantillonnage sélectionnée. Cela vous permet d'utiliser une estimation avec remplacement, même si la méthode d'échantillonnage nécessite une estimation sans remplacement. Cette option n'est disponible que dans la phase 1.

**Mesure de la taille (MOS) :** Si une méthode PPS est sélectionnée, vous devez spécifier une mesure de taille permettant de définir la taille de chaque unité. Ces tailles peuvent être définies dans une variable, ou calculées à partir des données. Vous pouvez éventuellement définir des limites inférieure et supérieure dans la mesure de taille, qui ignorent les valeurs trouvées dans la variable de mesure de taille ou calculées à partir des données. Ces options ne sont disponibles que dans la phase 1.

# *Assistant d'échantillonnage : Taille de l'échantillon :*

Figure 2-4
*Etape Taille de l'échantillon de l'assistant d'échantillonnage*




Cette phase vous permet de spécifier le nombre ou la proportion d'unités à échantillonner dans la phase en cours. La taille de l'échantillon peut être fixe ou varier en fonction des strates. Pour spécifier la taille de l'échantillon, les classes choisies dans les phases précédentes peuvent être utilisées pour définir les strates.

**Unités :** Vous pouvez spécifier une taille d'échantillon exacte ou une proportion d'unités à échantillonner.

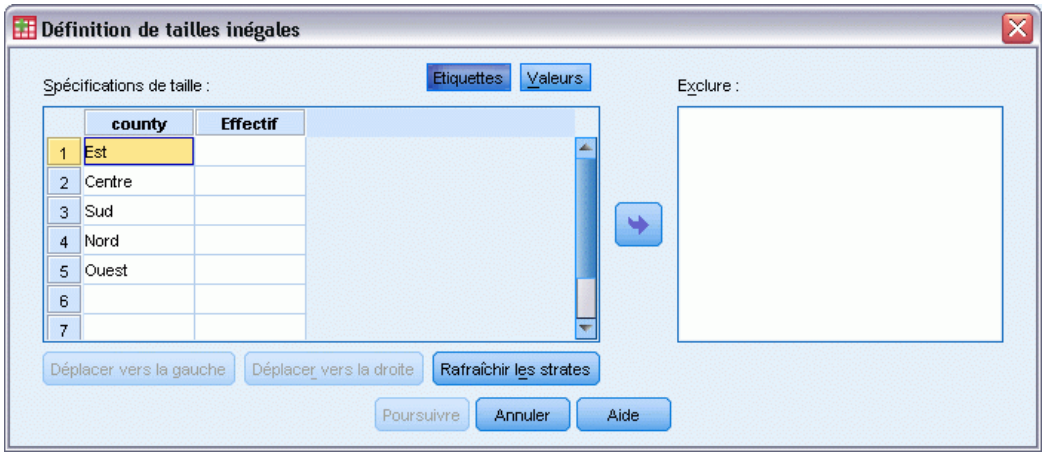
- **Valeur :** Une valeur unique est appliquée à toutes les strates. Si vous sélectionnez Effectifs comme unité de mesure, vous devez saisir un entier positif. Si vous sélectionnez Proportions, vous devez saisir une valeur non négative. A moins qu'il ne s'agisse d'un échantillonnage avec remplacement, les valeurs de proportion ne doivent également pas être supérieures à 1.

- **Valeurs inégales pour les strates :** Vous permet de saisir les tailles pour chaque strate via la boîte de dialogue Définition de tailles inégales.
- **Lire les valeurs à partir de la variable :** Vous permet de sélectionner une variable numérique contenant les valeurs de taille des strates.

Si vous sélectionnez Proportions, vous pouvez éventuellement définir des limites inférieure et supérieure pour le nombre d'unités échantillonnées.

## *Définition de tailles inégales*

Figure 2-5
*Boîte de dialogue Définition de tailles inégales*



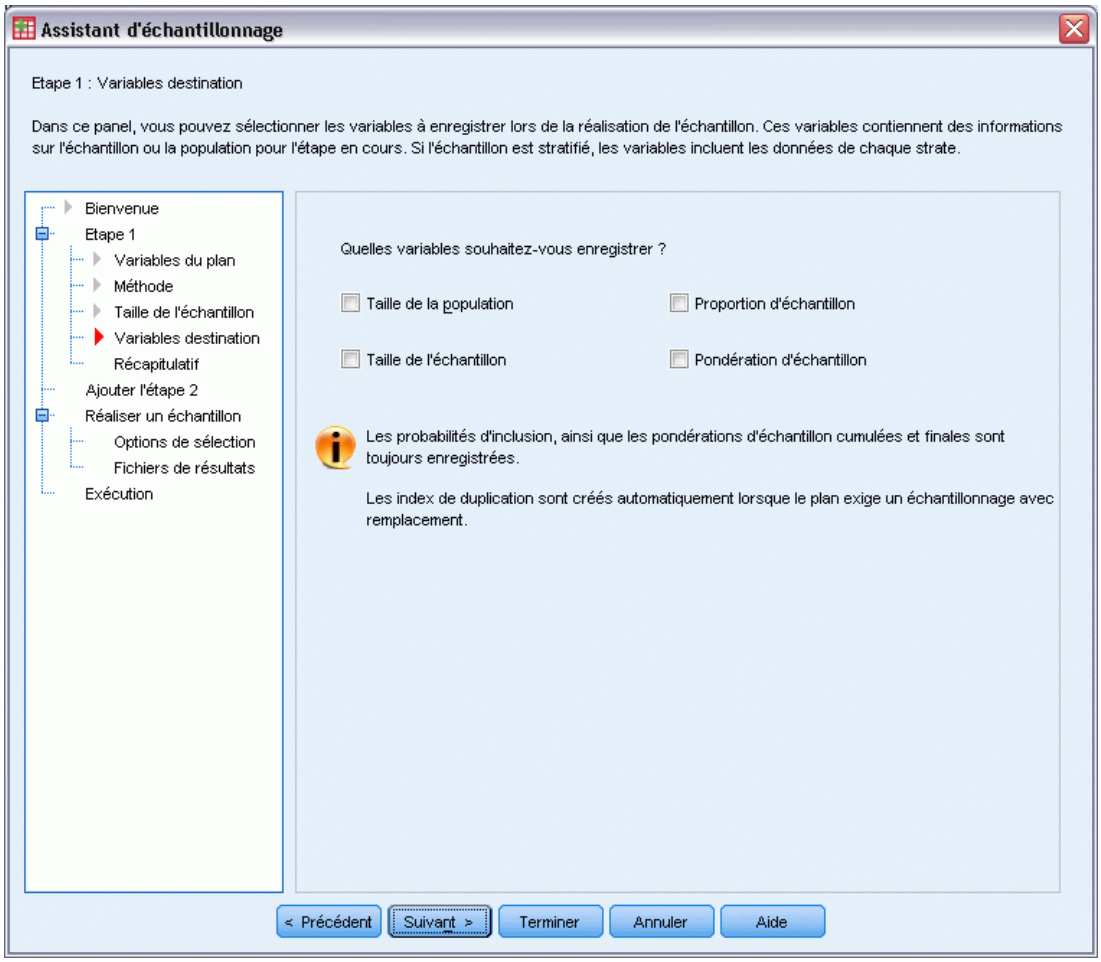

La boîte de dialogue Définition de tailles inégales vous permet de saisir des tailles par strate.

**Grille des spécifications de taille :** La grille affiche les classifications croisées de 5 variables de strates ou de classes maximum avec une combinaison strate/classe par ligne. Les variables de grille sélectionnables incluent toutes les variables de stratification de la phase en cours et des phases précédentes, ainsi que les variables de grappes des phases précédentes. Les variables peuvent être réorganisées dans la grille ou déplacées dans la liste Exclure. Saisissez les tailles dans la colonne située à l'extrême droite. Cliquez sur Etiquettes ou sur Valeurs pour afficher les valeurs d'étiquettes ou les valeurs de données des variables de stratification et de classe dans les cellules de la grille. Les cellules contenant des valeurs non étiquetées affichent toujours des valeurs. Cliquez sur Rafraîchir les strates pour que la grille affiche chaque combinaison de valeurs de données étiquetées des variables dans la grille.

**Exclure :** Pour spécifier les tailles d'un sous-ensemble de combinaisons de strate/classe, déplacez une ou plusieurs valeurs dans la liste Exclure. Ces variables ne sont pas utilisées pour définir les tailles des échantillons.

# *Assistant d'échantillonnage : Variables destination*

Figure 2-6
*Etape Variables destination de l'assistant d'échantillonnage*




Cette étape vous permet de choisir des variables à enregistrer lors de la réalisation de l'échantillon.

**Taille de la population.** Nombre d'unités estimées dans la population d'une phase spécifique. Le nom racine de la variable enregistrée est *PopulationSize_*.

**Proportion d'échantillon :** Taux d'échantillonnage à une phase spécifique. Le nom racine de la variable enregistrée est *SamplingRate_*.

**Taille de l'échantillon :** Nombre d'unités réalisées à une phase spécifique. Le nom racine de la variable enregistrée est *SampleSize_*.

**Pondération d'échantillon :** L'inverse des probabilités d'inclusion. Le nom racine de la variable enregistrée est *SampleWeight_*.

Certaines variables définies au cours des différentes étapes sont générées automatiquement. Parmi celles-ci, notons :

**Probabilités d'inclusion :** Proportion des unités réalisées à une phase spécifique. Le nom racine de la variable enregistrée est *InclusionProbability_*.

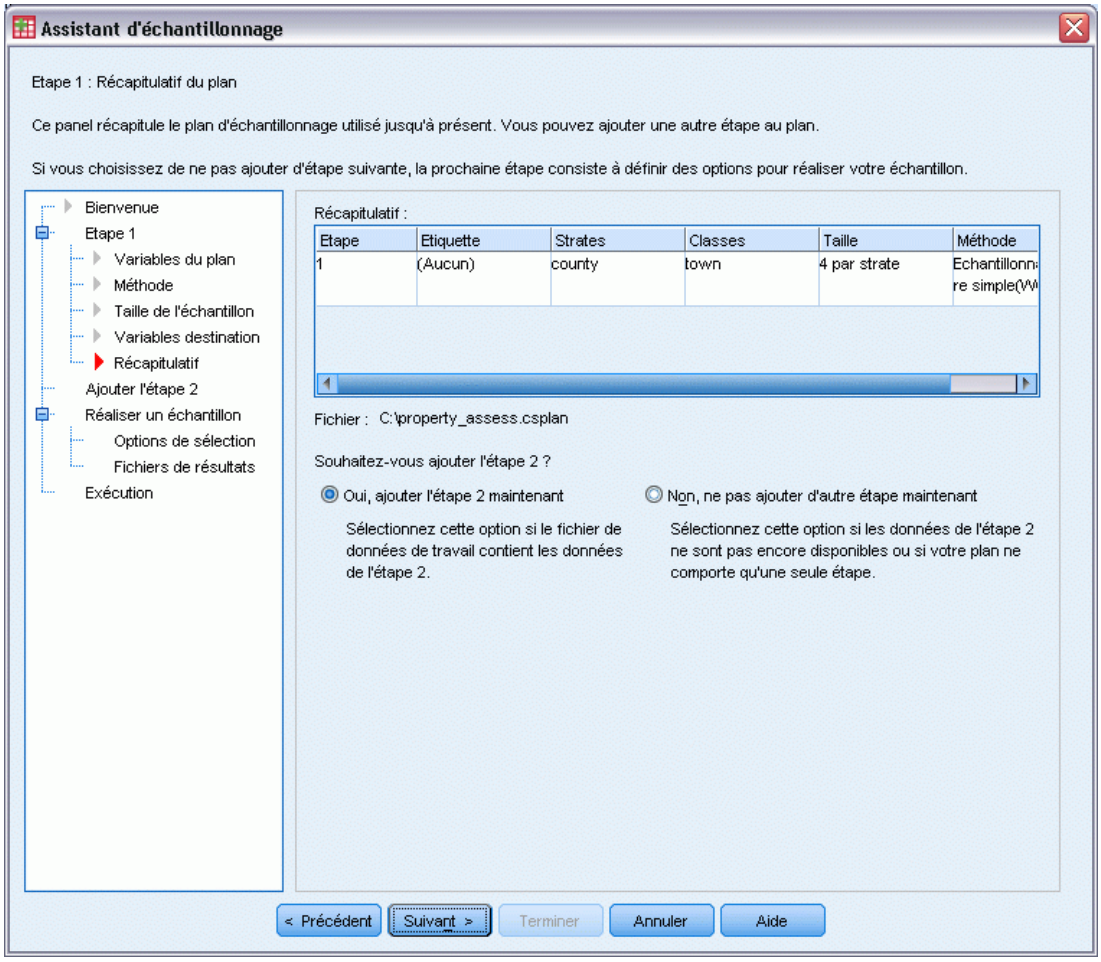
**Pondération cumulée :** Pondération d'échantillon cumulée entre les phases précédentes et la phase actuelle incluse. Le nom racine de la variable enregistrée est *SampleWeightCumulative_*.

**Index :** Identifie les unités sélectionnées plusieurs fois dans une phase spécifique. Le nom racine de la variable enregistrée est *Index_*.

*Remarque :* Les noms racine des variables enregistrées contiennent un suffixe entier faisant référence au numéro de phase—par exemple, *PopulationSize_1_* est la taille de population enregistrée de la phase 1.

# Assistant d'échantillonnage : Récapitulatif du plan

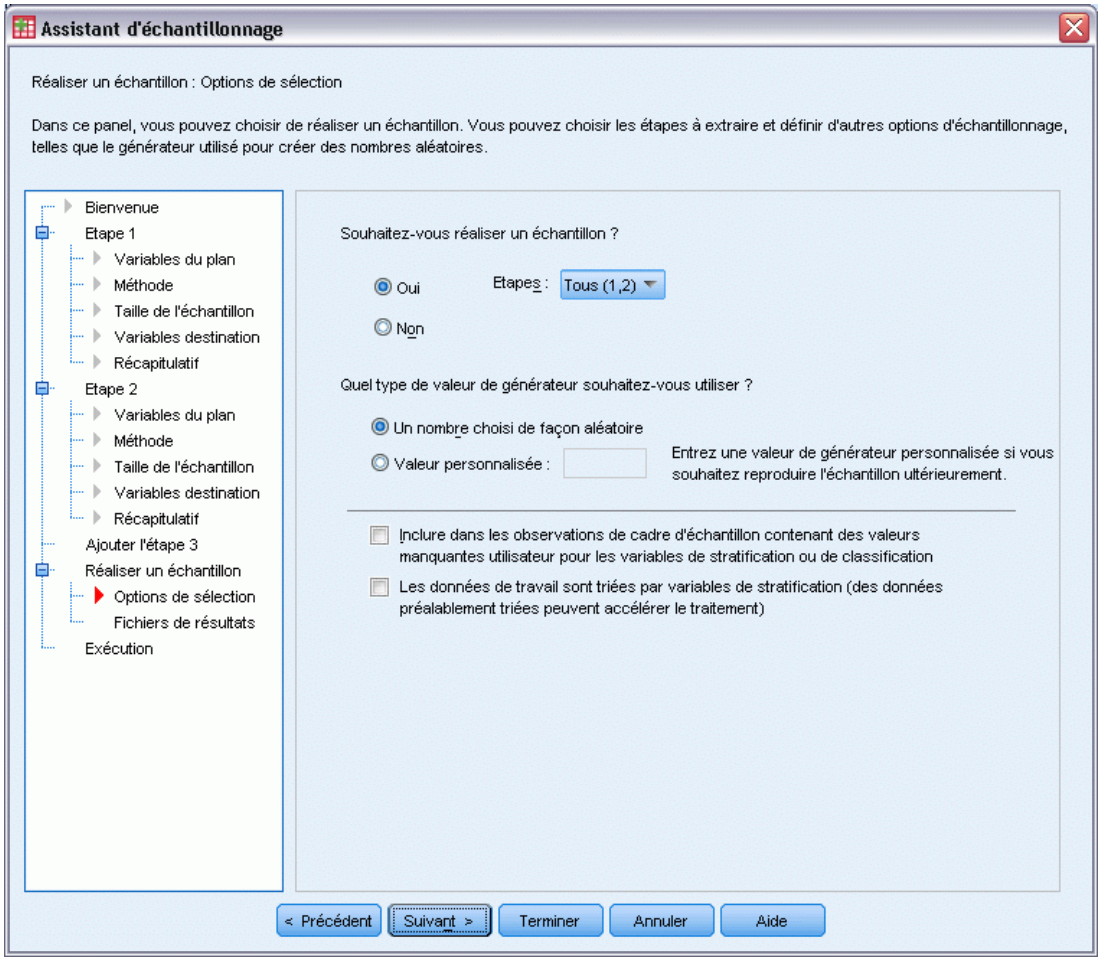
Figure 2-7
*Etape Récapitulatif du plan de l'assistant d'échantillonnage*



Il s'agit de la dernière étape de chaque phase ; elle fournit un récapitulatif des spécifications du plan d'échantillonnage dans la phase actuelle. Vous pouvez ensuite soit passer à la phase suivante (en la créant, si nécessaire), soit définir les options de réalisation de l'échantillon.

## *Assistant d'échantillonnage : Réalisation de l'échantillon : Options de sélection*

Figure 2-8
*Etape Réalisation de l'échantillon : Options de sélection de l'assistant d'échantillonnage*




Cette étape vous permet de choisir de réaliser un échantillon. Vous pouvez également contrôler d'autres options d'échantillonnage telles que la gestion du générateur aléatoire et des valeurs manquantes.

**Réaliser un échantillon :** Outre le fait de choisir de réaliser un échantillon, vous pouvez également choisir d'exécuter une partie du plan d'échantillonnage. Les phases doivent être réalisées dans l'ordre—cela signifie que la phase 2 ne peut pas être réalisée avant la réalisation de la phase 1. Lors de la modification ou de l'exécution d'un plan, vous ne pouvez pas rééchantillonner les phases verrouillées.

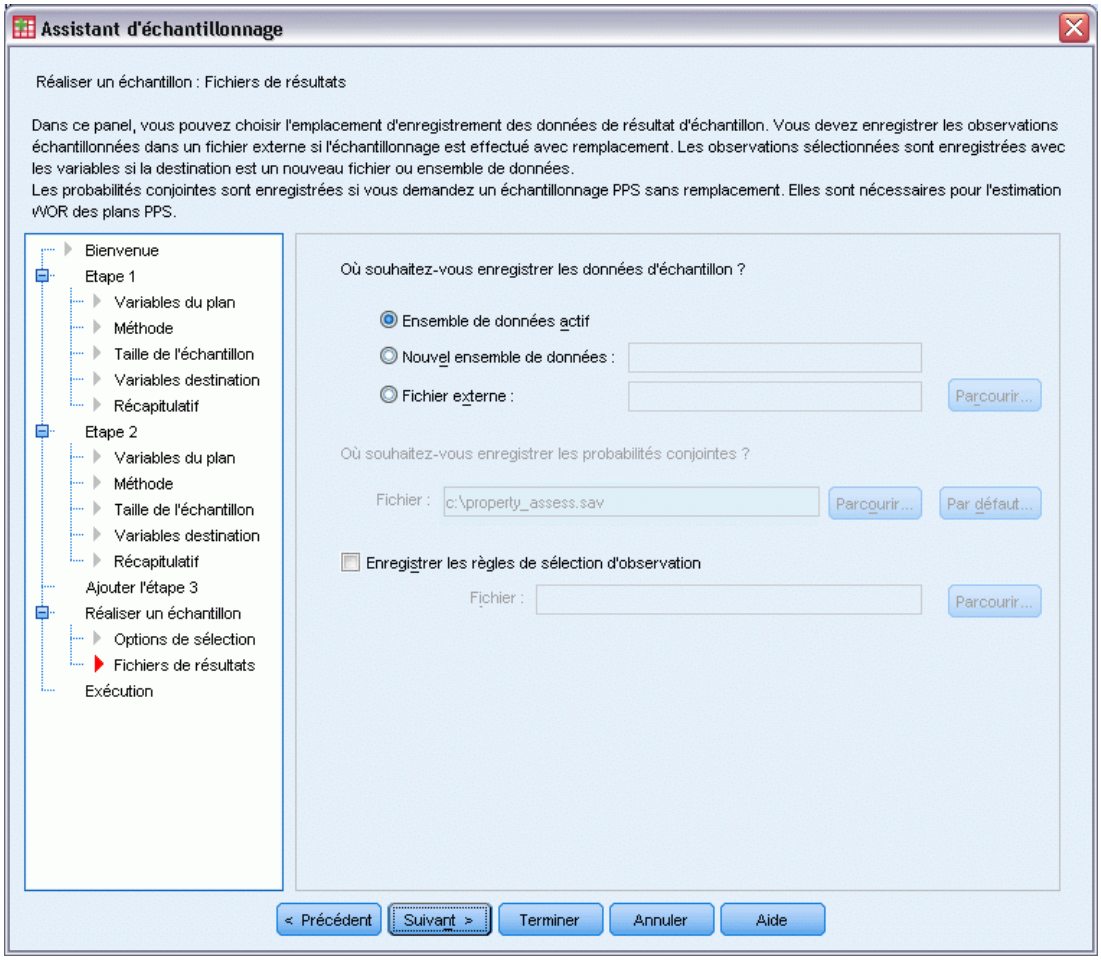
**Graine :** Vous permet d'attribuer une valeur au générateur utilisé pour les nombres aléatoires.

**Inclure les valeurs manquantes spécifiées :** Permet de déterminer si les valeurs manquantes spécifiées par l'utilisateur sont valides. Si c'est le cas, les valeurs manquantes spécifiées par l'utilisateur sont traitées séparément.

**Données déjà triées :** Si votre cadre d'échantillon est prétrié par les valeurs des variables de stratification, cette option vous permet d'accélérer le processus de sélection.

## *Assistant d'échantillonnage : Réalisation de l'échantillon : Fichiers de résultats*

Figure 2-9
*Etape Réalisation de l'échantillon : Fichiers de résultats de l'assistant d'échantillonnage*




Cette étape vous permet de choisir l'emplacement de redirection des observations échantillonnées et des variables de pondération, ainsi que des probabilités conjointes et des règles de sélection des observations.

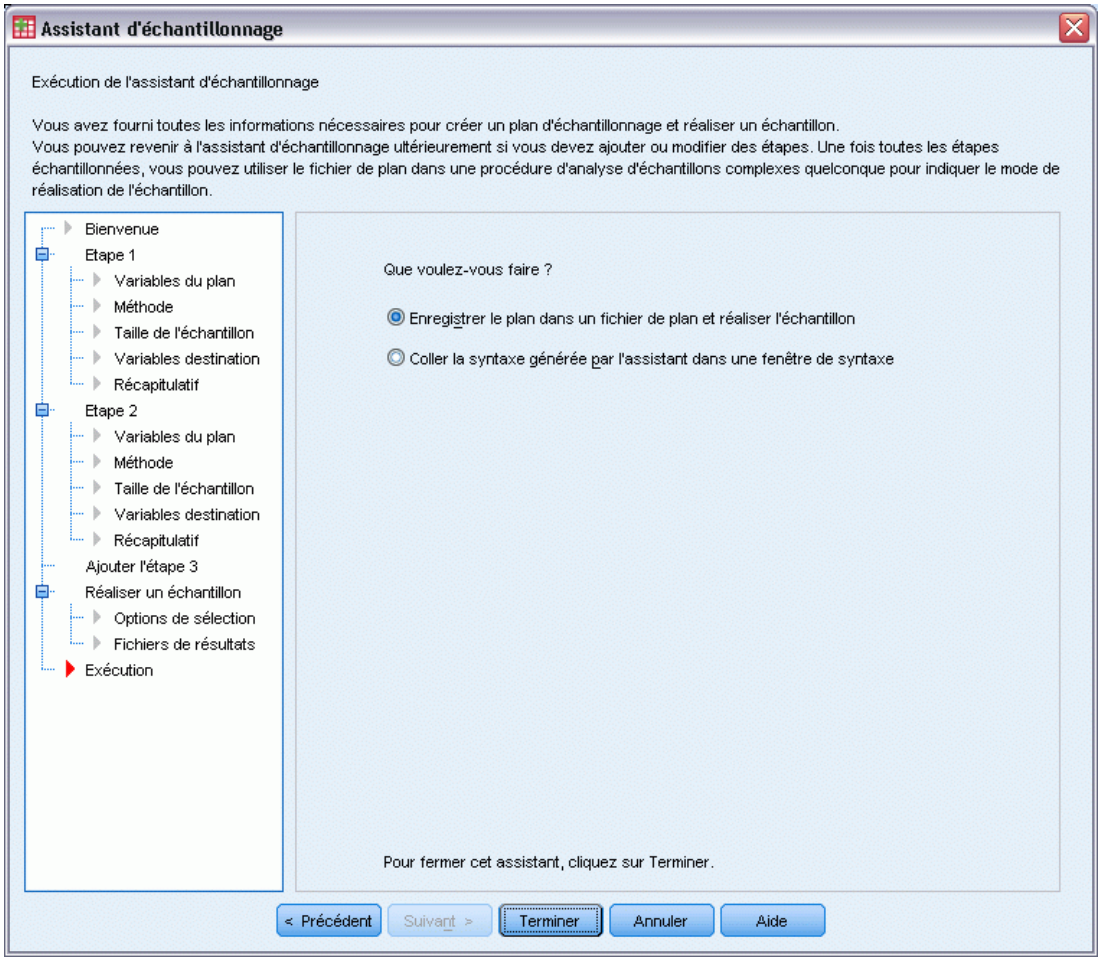
**Données exemple :** Ces options vous permettent de déterminer l'emplacement d'enregistrement de l'échantillon de résultat. Celui-ci peut être ajouté à un nouvel ensemble de données ou sauvegardé vers un fichier de données externe IBM® SPSS® Statistics. Les ensembles de données sont disponibles lors de la session en cours mais ne sont pas disponibles lors des sessions suivantes, sauf si vous les enregistrez clairement comme fichiers de données. Les noms des ensembles de données doivent être conformes aux règles de dénomination de variables. Si vous spécifiez un fichier externe ou un nouvel ensemble de données, les variables destination d'échantillonnage et les variables de l'ensemble de données actif des observations sélectionnées sont enregistrées dans le fichier.

**Probabilités conjointes :** Ces options vous permettent de déterminer l'emplacement d'enregistrement des probabilités conjointes. Elles sont enregistrées dans un fichier de données externe SPSS Statistics. Les probabilités conjointes sont produites si vous spécifiez une méthode PPS WOR, PPS Brewer, PPS Sampford ou PSS Murthy et que vous ne définissez pas d'estimation avec remplacement.

**Règles de sélection d'observation.** Si vous réalisez votre échantillon phase par phase, vous souhaiterez peut-être enregistrer les règles de sélection des observations dans un fichier texte. Elles sont utiles pour réaliser le sous-cadre des phases suivantes.

# Assistant d'échantillonnage : Terminer

Figure 2-10
*Etape Fin de l'assistant d'échantillonnage*




Il s'agit de l'étape finale. Vous pouvez enregistrer le fichier de plan et réaliser l'échantillon maintenant, ou coller vos sélections dans une fenêtre de syntaxe.

Lorsque vous apportez des modifications aux phases du fichier de plan existant, vous pouvez enregistrer le plan modifié dans un nouveau fichier ou remplacer le fichier existant. Lorsque vous ajoutez des phases sans modifier les phases existantes, l'assistant remplace automatiquement le fichier de plan existant. Pour enregistrer le plan dans un nouveau fichier, sélectionnez Coller la syntaxe générée par l'assistant dans une fenêtre de syntaxe et modifiez le nom du fichier dans les commandes de la syntaxe.

# Modification d'un plan d'échantillonnage existant

- A partir des menus, sélectionnez :
  Analyse > Echantillonnage > Sélectionner un échantillon...
- Sélectionnez Modifier un plan d'échantillonnage et choisissez le fichier de plan à modifier.

- Cliquez sur Suivant pour poursuivre la procédure avec l'assistant.

- Révisez le plan d'échantillonnage dans l'étape Récapitulatif du plan, puis cliquez sur Suivant.

  Les étapes suivantes sont pratiquement les mêmes que celles effectuées pour un nouveau plan. Pour plus d'informations, reportez-vous à l'aide de chaque étape.

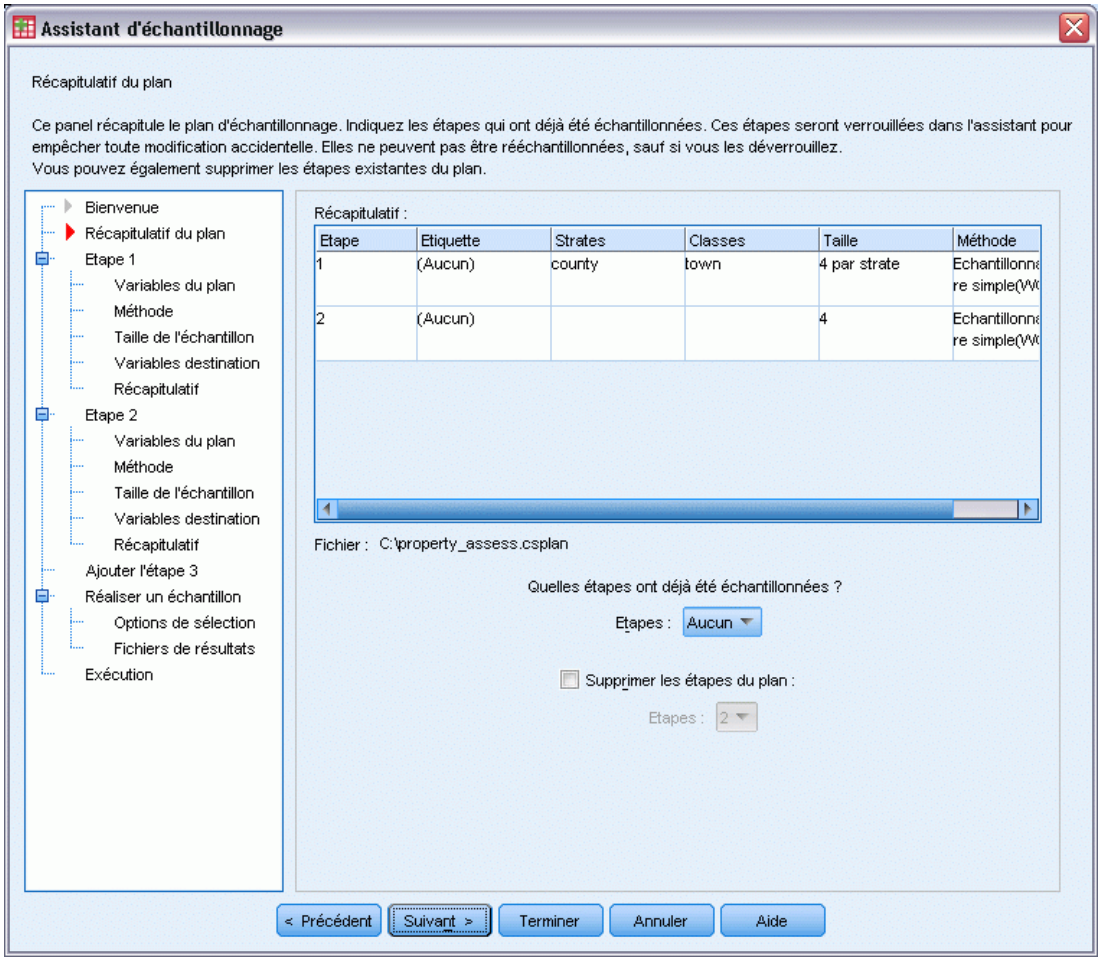
- Allez jusqu'à l'étape Fin et spécifiez un nouveau nom pour le fichier de plan modifié, ou choisissez d'écraser le fichier de plan existant.

  Sinon, vous pouvez :

  - Spécifiez les phases qui ont déjà été échantillonnées.
  - Supprimez les étapes du plan.

## *Assistant d'échantillonnage : Récapitulatif du plan*

Figure 2-11
*Etape Récapitulatif du plan de l'assistant d'échantillonnage*



Cette phase vous permet de réviser le plan d'échantillonnage et d'indiquer les phases qui ont déjà été échantillonnées. Si vous modifiez un plan, vous pouvez également supprimer des phases du plan.

**Etapes précédemment échantillonnées :** Si un cadre d'échantillonnage étendu n'est pas disponible, vous devez exécuter un plan d'échantillonnage multiphase, phase par phase. Sélectionnez dans la liste déroulante les phases qui ont déjà été échantillonnées. Toutes les phases qui ont été exécutées sont verrouillées ; elles ne sont pas disponibles dans l'étape Réalisation de l'échantillon : Options de sélection et ne peuvent être interverties lors de la modification d'un plan.

**Supprimer du plan les étapes :** Vous pouvez supprimer les phases 2 et 3 d'un plan à plusieurs phases.

## *Exécution d'un plan d'échantillonnage existant*

- A partir des menus, sélectionnez :
  Analyse > Echantillonnage > Sélectionner un échantillon...
- Sélectionnez Réaliser un échantillon et sélectionnez un fichier de plan à exécuter.
- Cliquez sur Suivant pour poursuivre la procédure avec l'assistant.
- Révisez le plan d'échantillonnage dans l'étape Récapitulatif du plan, puis cliquez sur Suivant.
- Les étapes individuelles contenant des informations de phase sont ignorées lors de l'exécution d'un plan d'échantillonnage. Vous pouvez passer à l'étape Fin à tout moment.

  Vous pouvez également spécifier les étapes qui ont déjà été échantillonnées.

## *Fonctions supplémentaires des commandes SPLAN et CSSELECT*

Le langage de syntaxe de commande vous permet aussi de :

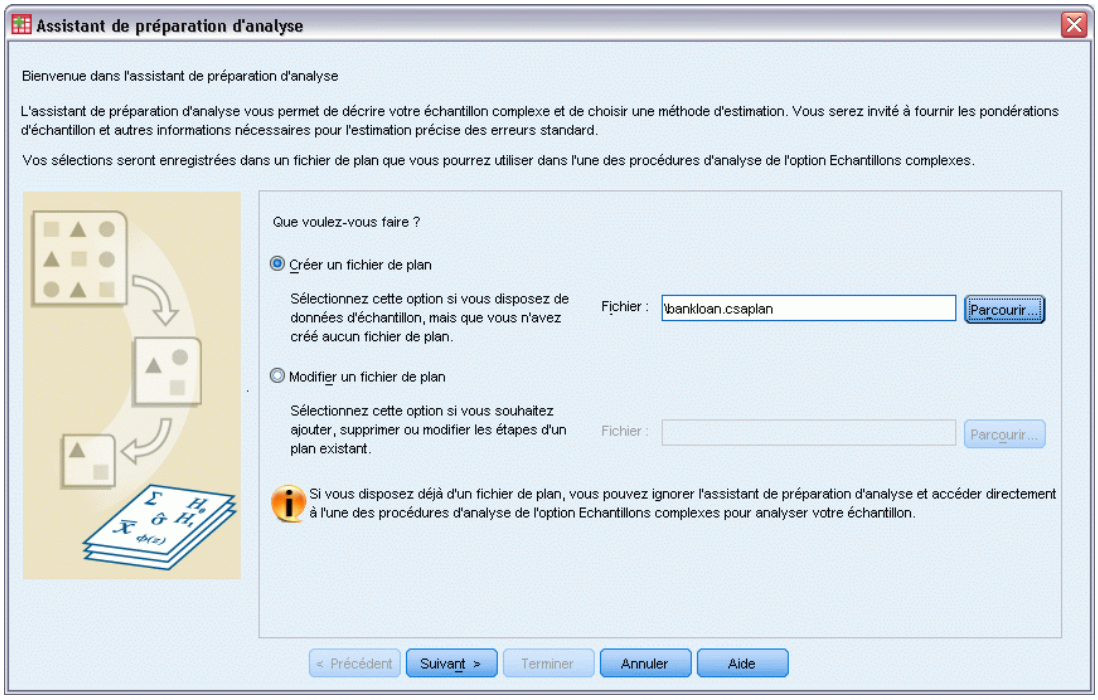
- Spécifiez les noms personnalisés des variables destination.
- Contrôlez le résultat dans le Viewer. Par exemple, vous pouvez supprimer le résumé de chaque phase du plan qui est affiché si un échantillon est réalisé ou modifié, supprimer le récapitulatif de la distribution des observations échantillonnées par strates qui apparaît si le plan d'échantillonnage est exécuté et demander un récapitulatif de traitement des observations.
- Choisissez un sous-ensemble de variables dans l'ensemble de données actif pour effectuer l'enregistrement dans un fichier d'exemple externe ou dans un ensemble de données différent.

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

Chapitre
3

# Préparation d'un échantillon complexe en vue d'une analyse

Figure 3-1
*Etape Bienvenue de l'assistant de préparation d'analyse*




L'assistant de préparation d'analyse vous guide dans la création ou la modification d'un plan d'analyse à utiliser avec les diverses procédures d'analyse des échantillons complexes. Avant d'utiliser l'assistant, vous devez réaliser un échantillon en fonction d'un plan spécifique.

La création d'un nouveau plan est principalement utile lorsque vous n'avez pas accès au fichier de plan d'échantillonnage utilisé pour réaliser l'échantillon (n'oubliez pas que le plan d'échantillonnage contient un plan d'analyse par défaut). Si vous avez accès au fichier de plan d'échantillonnage utilisé pour réaliser l'échantillon, vous pouvez utiliser le plan d'analyse par défaut contenu dans le fichier de plan d'échantillonnage ou ignorer les spécifications d'analyse par défaut et enregistrer vos modifications dans un nouveau fichier.

## Création d'un plan d'analyse

- A partir des menus, sélectionnez :
  Analyse > Echantillonnage > Préparer pour l'analyse...

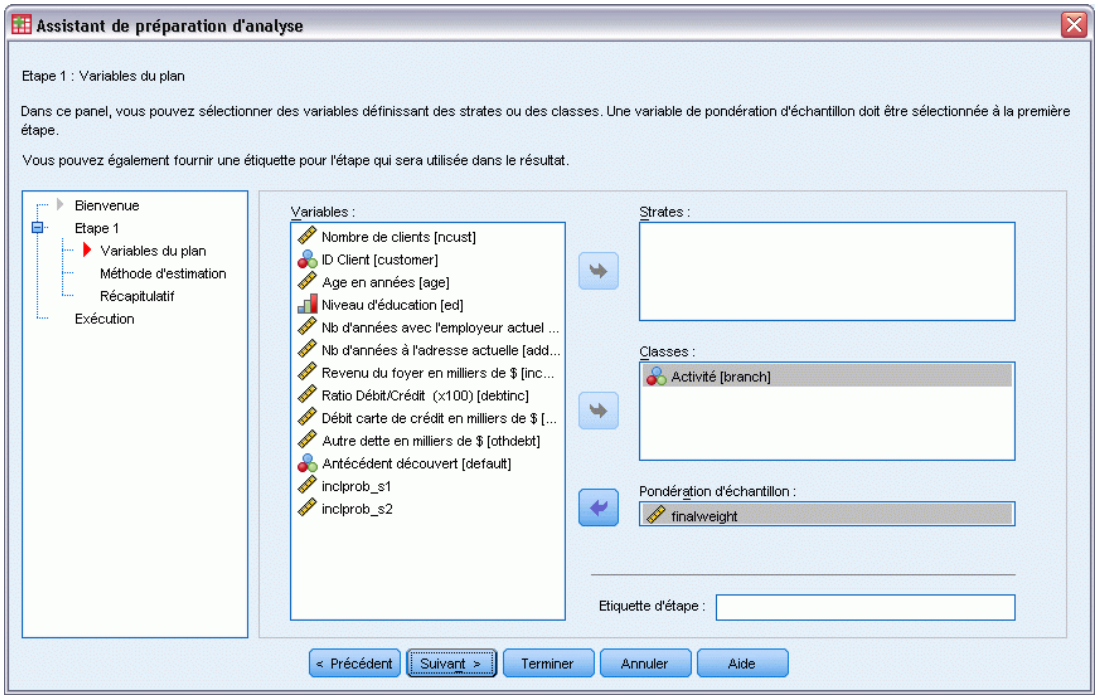
- Sélectionnez Créer un fichier de plan et attribuez un nom au fichier dans lequel enregistrer le plan d'analyse.
- Cliquez sur Suivant pour poursuivre la procédure avec l'assistant.
- A l'étape Variables de plan, spécifiez la variable contenant les pondérations d'échantillon et définissez éventuellement des strates et des classes.
- Vous pouvez alors cliquer sur Fin pour enregistrer le plan.

Les étapes suivantes sont facultatives. Elles permettent :

- De sélectionner la méthode d'estimation des erreurs standard à l'étape Méthode d'estimation.
- De spécifier le nombre d'unités échantillonnées ou la probabilité d'inclusion par unité à l'étape Taille.
- D'ajouter une deuxième ou une troisième étape au plan.
- De coller vos sélections en tant que syntaxe de commande.

## Assistant de préparation d'analyse : Variables du plan

Figure 3-2
*Etape Variables de plan de l'assistant de préparation d'analyse*




Cette étape vous permet d'identifier les variables pour la définition des strates et grappes, et pour celle des pondérations d'échantillon. Vous pouvez également associer une étiquette à cette étape.

**Strates.** La classification croisée des variables de stratification définit des sous-populations distinctes, ou strates. La totalité de votre échantillon représente la combinaison des échantillons indépendants de chaque strate.

**Classes.** Les variables de grappe définissent les groupes d'unités d'observation, ou classes. Les échantillons réalisés en plusieurs étapes sélectionnent les classes des étapes précédentes, puis sous-échantillonnent les unités à partir des classes sélectionnées. Lors de l'analyse d'un fichier de données obtenu par échantillonnage en grappe avec remplacement, vous devez inclure l'index de duplication comme variable de grappe.

**Pondération d'échantillon :** Vous devez fournir des pondérations d'échantillon à la première étape. Les pondérations d'échantillon sont automatiquement calculées pour les étapes suivantes du plan actuel.

**Etiquette d'étape.** Vous pouvez spécifier une étiquette de type chaîne de caractères pour chaque étape. Elle est utilisée dans le résultat pour identifier les informations de chaque étape.

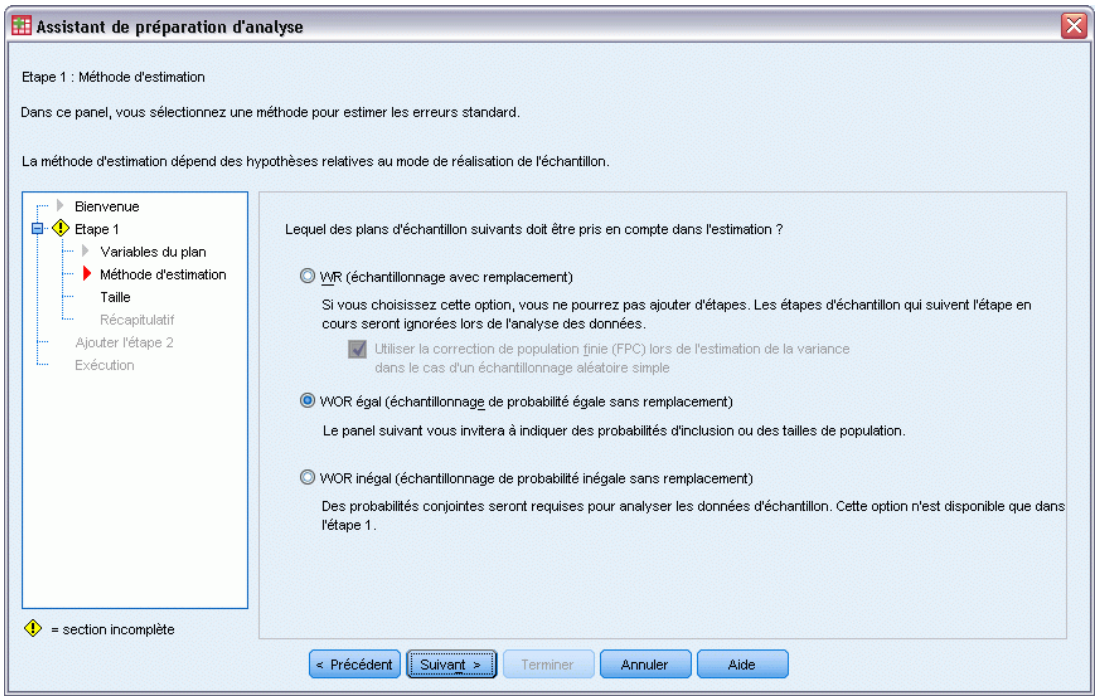
*Remarque :* La liste des variables source compile le contenu obtenu lors de l'exécution des étapes de l'assistant. En d'autres termes, les variables supprimées de la liste source à une étape sont supprimées de toutes les étapes dans la liste. Les variables renvoyées à la liste source apparaissent à toutes les étapes.

## *Contrôles d'arbre de navigation de l'assistant d'analyse*

Chaque étape de l'assistant d'analyse dispose à sa gauche d'une légende. Vous pouvez parcourir l'assistant en cliquant sur le nom d'une étape disponible dans la légende. Les étapes sont disponibles tant que les étapes précédentes sont valides. Cela signifie qu'elles le sont si chaque étape précédente a fourni les spécifications minimales requises pour cette étape. Pour plus d'informations sur les raisons de l'invalidité d'une étape, reportez-vous à l'aide de chaque étape.

# Assistant de préparation d'analyse : Méthode d'estimation

Figure 3-3
*Etape Méthode d'estimation de l'assistant de préparation d'analyse*




Cette étape vous permet d'indiquer une méthode d'estimation pour l'étape.

**WR (échantillonnage avec remplacement)** L'estimation WR n'inclut pas de correction d'échantillonnage d'une population finie (FPC) lors de l'estimation de la variance dans le plan d'échantillonnage complexe. Vous pouvez choisir d'inclure ou d'exclure la correction FPC lors de l'estimation de la variance dans un échantillonnage aléatoire simple (SRS).

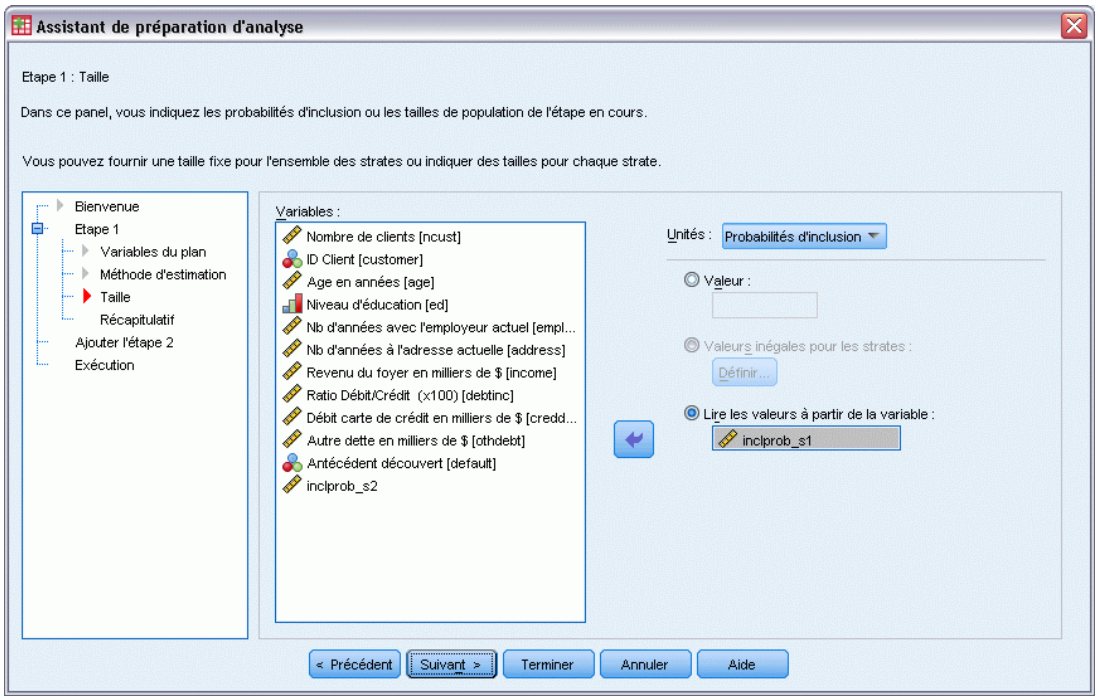
Il est recommandé de ne pas inclure la correction FPC pour l'estimation de la variance SRS lorsque les pondérations d'analyse ont été redimensionnées afin de ne pas s'ajouter à la taille de la population. L'estimation de la variance SRS est utilisée dans le calcul de statistiques telles que l'effet de plan. L'estimation WR ne peut être spécifiée que dans l'étape finale d'un plan. L'assistant ne vous laisse plus ajouter d'autre étape ensuite.

**WOR égal (échantillonnage de probabilité égale sans remplacement).** L'estimation WOR égal inclut la correction de population finie et part du principe que les unités sont échantillonnées à probabilité égale. L'estimation WOR égal peut être spécifiée dans n'importe quelle étape d'un plan.

**WOR inégal (échantillonnage de probabilité inégale sans remplacement).** Outre l'utilisation de la correction de population finie, l'estimation WOR inégal échantillonne les unités (généralement des classes) sélectionnées à probabilité inégale. Cette méthode d'estimation n'est disponible qu'à la première étape.

# *Assistant de préparation d'analyse : Taille*

Figure 3-4
*Etape Taille de l'assistant de préparation d'analyse*




Cette étape est utilisée pour spécifier les probabilités d'inclusion ou les tailles de population de l'étape en cours. Ces tailles peuvent être fixes ou varier suivant les strates. Les classes spécifiées dans les étapes précédentes peuvent être utilisées pour définir les strates dans le cadre de la spécification des tailles. Notez que cette étape est nécessaire seulement si vous avez choisi la méthode d'estimation WOR égal.

**Unités :** Vous pouvez spécifier des tailles de population exactes ou les probabilités d'échantillonnage des unités.

- **Valeur :** Une valeur unique est appliquée à toutes les strates. Si vous définissez l'unité de mesure sur Tailles de population, vous devez saisir un entier non négatif. Si vous définissez l'unité de mesure sur Probabilités d'inclusion, vous devez saisir une valeur comprise entre 0 et 1 inclus.
- **Valeurs inégales pour les strates :** Vous permet de saisir les tailles pour chaque strate via la boîte de dialogue Définition de tailles inégales.
- **Lire les valeurs à partir de la variable :** Vous permet de sélectionner une variable numérique contenant les valeurs de taille des strates.

## *Définition de tailles inégales*

Figure 3-5
*Boîte de dialogue Définition de tailles inégales*

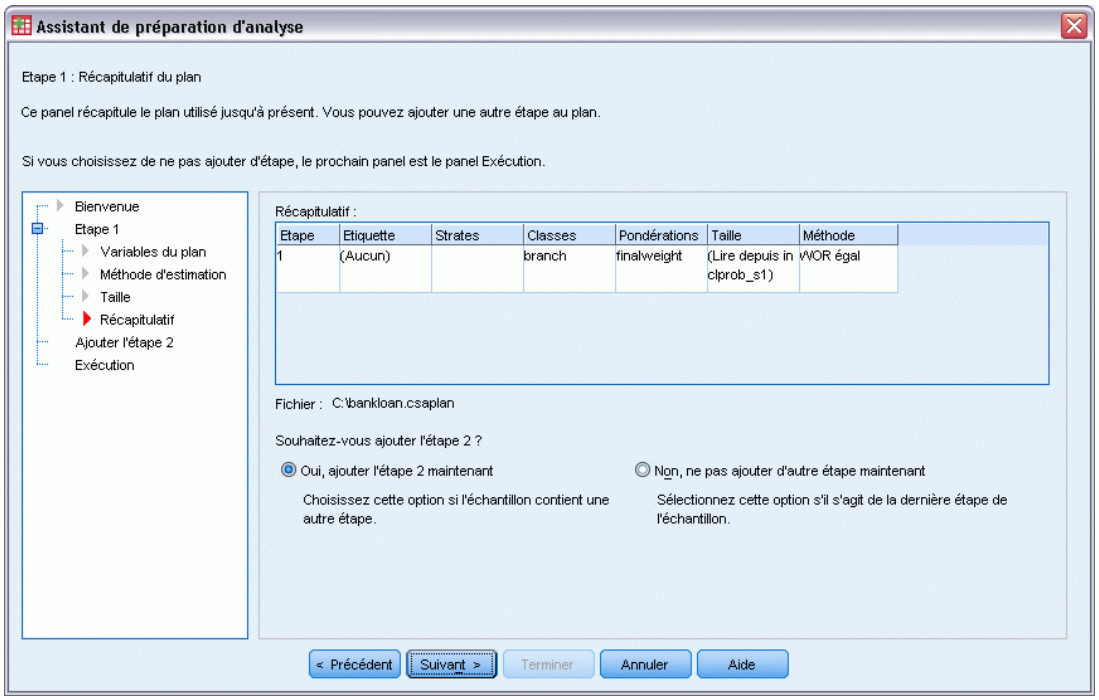

La boîte de dialogue Définition de tailles inégales vous permet de saisir des tailles par strate.

**Grille des spécifications de taille :** La grille affiche les classifications croisées de 5 variables de strates ou de classes maximum avec une combinaison strate/classe par ligne. Les variables de grille sélectionnables incluent toutes les variables de stratification de la phase en cours et des phases précédentes, ainsi que les variables de grappes des phases précédentes. Les variables peuvent être réorganisées dans la grille ou déplacées dans la liste Exclure. Saisissez les tailles dans la colonne située à l'extrême droite. Cliquez sur Etiquettes ou sur Valeurs pour afficher les valeurs d'étiquettes ou les valeurs de données des variables de stratification et de classe dans les cellules de la grille. Les cellules contenant des valeurs non étiquetées affichent toujours des valeurs. Cliquez sur Rafraîchir les strates pour que la grille affiche chaque combinaison de valeurs de données étiquetées des variables dans la grille.

**Exclure :** Pour spécifier les tailles d'un sous-ensemble de combinaisons de strate/classe, déplacez une ou plusieurs valeurs dans la liste Exclure. Ces variables ne sont pas utilisées pour définir les tailles des échantillons.

# *Assistant de préparation d'analyse : Récapitulatif du plan*

Figure 3-6
*Etape Récapitulatif du plan de l'assistant de préparation d'analyse*



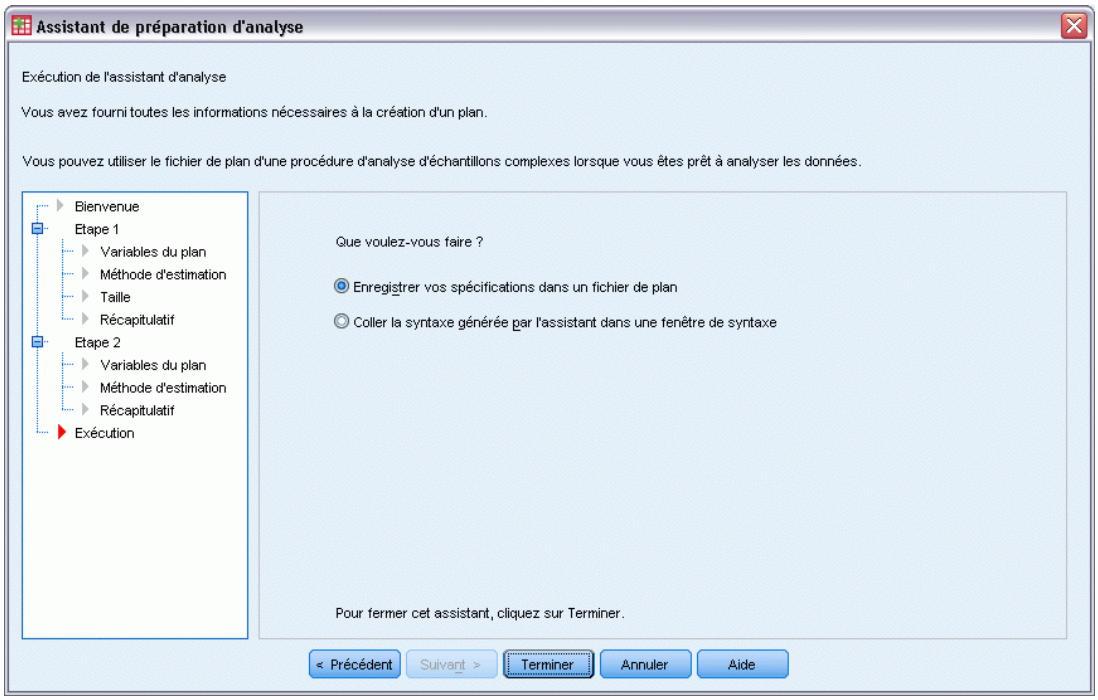

Il s'agit de la dernière étape de chaque phase ; celle-ci fournit un récapitulatif des spécifications du plan d'analyse dans l'étape en cours. Vous pouvez ensuite soit passer à la phase suivante (en la créant, si nécessaire), soit enregistrer les spécifications d'analyse.

Les raisons pour lesquelles vous ne pouvez pas ajouter d'autre étape sont principalement les suivantes :

- A l'étape Variables de plan, aucune variable de grappe n'a été spécifiée.
- Vous avez sélectionné l'estimation WR à l'étape Méthode d'estimation.
- Il s'agit de la troisième étape de l'analyse et l'assistant n'en prend pas en charge davantage.

# Assistant de préparation d'analyse : Terminer

Figure 3-7
*Etape Fin de l'assistant de préparation d'analyse*




Il s'agit de l'étape finale. Vous pouvez enregistrer le fichier de plan maintenant ou collecter vos sélections dans une fenêtre de syntaxe.

Lorsque vous apportez des modifications aux phases du fichier de plan existant, vous pouvez enregistrer le plan modifié dans un nouveau fichier ou remplacer le fichier existant. Lorsque vous ajoutez des phases sans modifier les phases existantes, l'assistant remplace automatiquement le fichier de plan existant. Pour enregistrer le plan dans un nouveau fichier, sélectionnez Coller la syntaxe générée par l'assistant dans une fenêtre de syntaxe et modifiez le nom du fichier dans les commandes de la syntaxe.

# Modification d'un plan d'analyse existant

- A partir des menus, sélectionnez :
  Analyse > Echantillonnage > Préparer pour l'analyse...
- Sélectionnez Modifier un fichier de plan et attribuez un nom au fichier de plan dans lequel enregistrer le plan d'analyse.
- Cliquez sur Suivant pour poursuivre la procédure avec l'assistant.

- Révisez le plan d'analyse dans l'étape Récapitulatif du plan, puis cliquez sur Suivant.

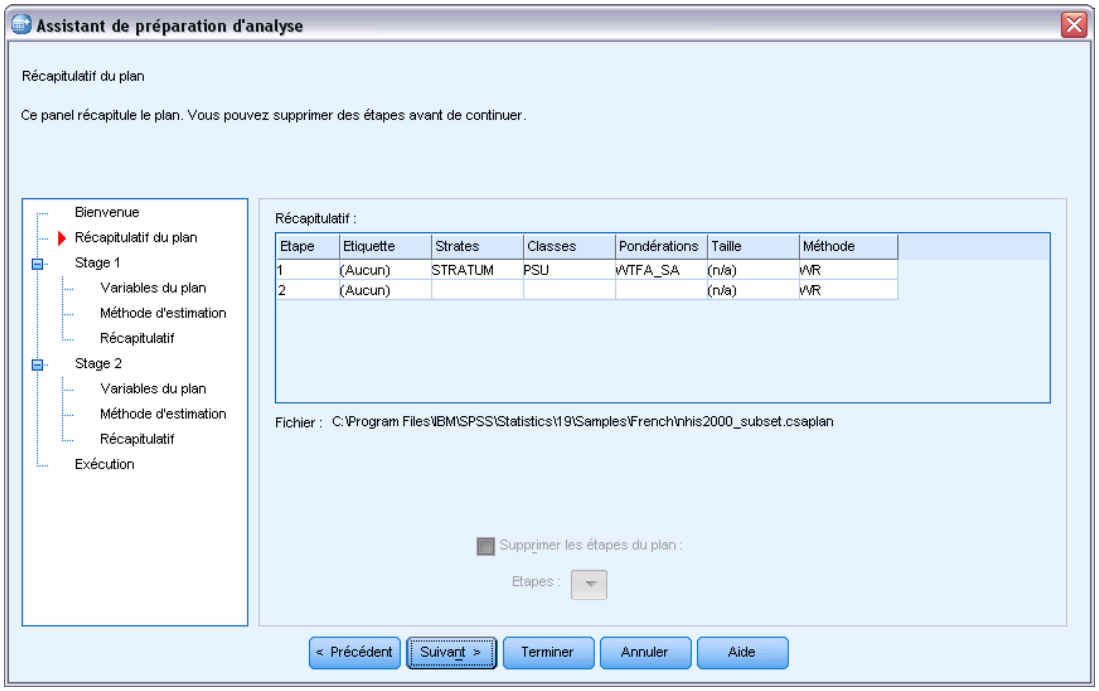
  Les étapes suivantes sont pratiquement les mêmes que celles effectuées pour un nouveau plan. Pour plus d'informations, reportez-vous à l'aide de chaque étape.

- Allez jusqu'à l'étape Fin et spécifiez un nouveau nom pour le fichier de plan modifié ou choisissez d'écraser le fichier de plan existant.

  Vous pouvez éventuellement supprimer des étapes du plan.

# *Assistant de préparation d'analyse : Récapitulatif du plan*

Figure 3-8
*Etape Récapitulatif du plan de l'assistant de préparation d'analyse*




Cette étape vous permet de passer en revue le plan d'analyse et de supprimer des phases du plan.

**Supprimer du plan les étapes :** Vous pouvez supprimer les phases 2 et 3 d'un plan à plusieurs phases. Comme un plan nécessite au moins une phase, vous pouvez modifier la phase 1 du plan, mais pas la supprimer.

Chapitre
4

# Plan d'échantillonnages complexes

Les procédures d'analyse d'échantillons complexes nécessitent des spécifications d'analyse depuis un fichier de plan d'analyse ou d'échantillon, afin de disposer de résultats valides.

Figure 4-1
*Boîte de dialogue Plan d'échantillonnages complexes*


**Plan :** Spécifiez le chemin d'un fichier de plan d'analyse ou d'échantillon.

**Probabilités conjointes :** Pour utiliser l'estimation WOR inégal avec les classes créées à l'aide d'une méthode PPS WOR, vous devez spécifier un fichier distinct ou un ensemble de données ouvert contenant les probabilités conjointes. Le fichier ou l'ensemble de données est créé par l'assistant d'échantillonnage lors de l'échantillonnage.

Chapitre
5

# Echantillons complexes - Fréquences

La procédure Echantillons complexes - Fréquences génère les tableaux de fréquences des variables sélectionnées et affiche des statistiques univariées. Vous pouvez éventuellement classer les statistiques par sous-groupes, définis par une ou plusieurs variables qualitatives.

**Exemple :** La procédure Echantillons complexes - Fréquences vous permet d'obtenir les statistiques tabulaires univariées de la consommation de vitamines des Américains, basées sur les résultats du NHIS (National Health Interview Survey) et dotées d'un plan d'analyse approprié pour ces données d'usage public.

**Statistiques :** La procédure génère pour chaque estimation les estimations des tailles de population et les pourcentages de tableaux, ainsi que les erreurs standard, les intervalles de confiance, les coefficients de variation, les effets de plan, les racines carrées d'effets de plan, les valeurs cumulées et les effectifs non pondérés. En outre, les statistiques Khi-deux et de rapport de vraisemblance sont calculées pour le test d'uniformité des proportions des cellules.

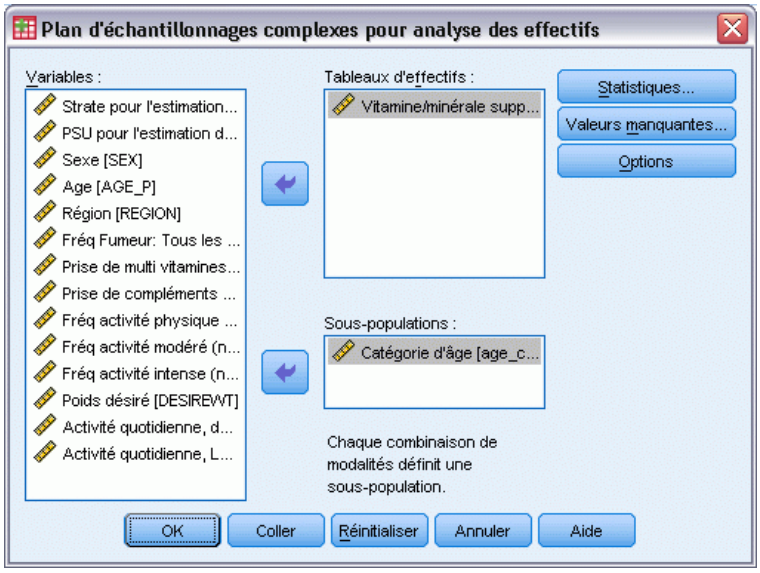
**Données :** Les variables pour lesquelles les tableaux de fréquences sont produits doivent être qualitatives. Les variables de sous-population peuvent être du type chaîne de caractères ou numérique, mais ne doivent pas être qualitatives.

**Hypothèses :** Les observations dans le fichier de données représentent un échantillon provenant d'un plan complexe et devant être analysées en fonction des spécifications du fichier sélectionné dans la Boîte de dialogue Plan d'échantillonnages complexes.

### *Obtention de fréquences d'échantillons complexes*

- A partir des menus, sélectionnez :
  Analyse > Echantillonnage > Fréquences
- Sélectionnez un fichier de plan. Sélectionnez éventuellement un fichier personnalisé de probabilités conjointes.
- Cliquez sur Poursuivre.

Figure 5-1
*Boîte de dialogue Fréquences*

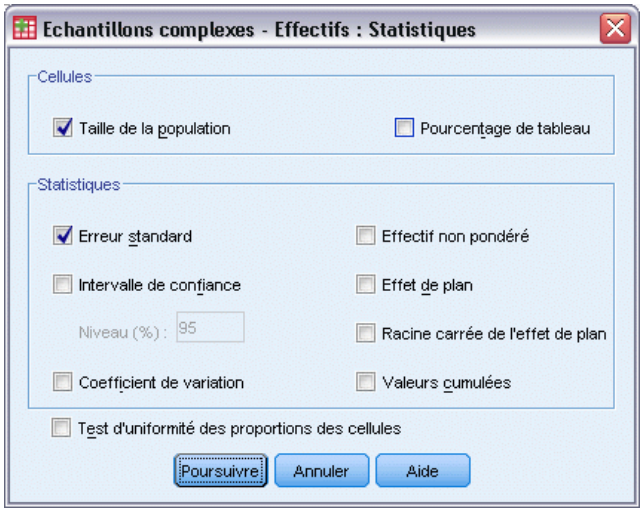



▶ Sélectionnez au moins une variable quantitative.

Vous pouvez éventuellement spécifier les variables définissant les sous-populations. Les statistiques sont calculées séparément pour chaque sous-population.

## *Statistiques de fréquences des échantillons complexes*

Figure 5-2
*Boîte de dialogue Fréquences : Statistiques*




**Cellules :** Ce groupe vous permet de demander les estimations des tailles de population de cellule et les pourcentages en tableau.

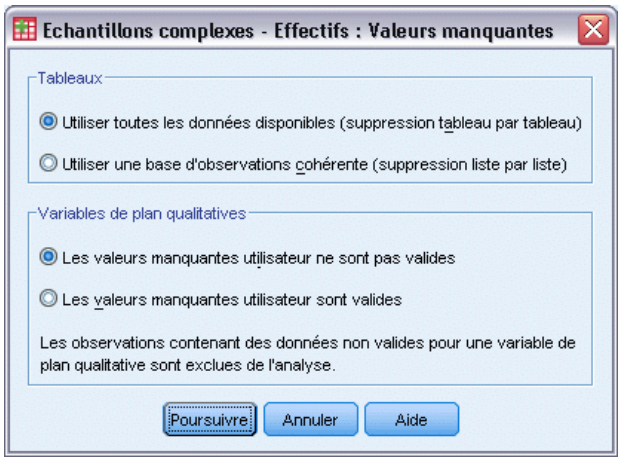
**Statistiques :** Ce groupe génère des statistiques associées à la taille de la population ou au pourcentage en tableau.

- **Erreur standard :** Erreur standard de l'estimation.
- **Intervalle de confiance :** Intervalle de confiance de l'estimation utilisant le niveau spécifié.
- **Coefficient de variation :** Rapport entre l'erreur standard de l'estimation et l'estimation.
- **Effectif non pondéré :** Nombre d'unités utilisées pour calculer l'estimation.
- **Effet de plan :** Rapport entre la variance de l'estimation et la variance, en partant du principe que l'échantillon est un échantillon aléatoire simple. Il s'agit d'une mesure de l'effet de spécification d'un plan complexe, où les valeurs éloignées de 1 indiquent des effets importants.
- **Racine carrée de l'effet de plan :** Il s'agit d'une mesure de l'effet de spécification d'un plan complexe, où les valeurs éloignées de 1 indiquent des effets importants.
- **Valeurs cumulées :** Estimation cumulée pour chaque valeur de la variable.

**Test d'uniformité des proportions des cellules.** Ce test génère des tests du Khi-deux et de rapport de vraisemblance basés sur l'hypothèse que les modalités d'une variable possèdent des fréquences égales. Des tests distincts sont effectués pour chaque variable.

## Valeurs manquantes d'échantillons complexes

Figure 5-3
*Boîte de dialogue Valeurs manquantes*



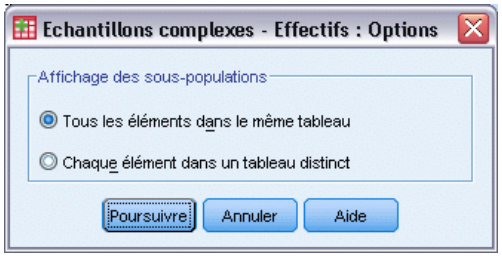

**Tables.** Ce groupe détermine les observations utilisées dans l'analyse.

- **Utilisez toutes les données disponibles.** Les valeurs manquantes sont définies tableau par tableau. Ainsi, les observations utilisées pour calculer les statistiques peuvent varier selon la fréquence ou les tableaux croisés.
- **Utilisez une base d'observations cohérente.** Les valeurs manquantes sont déterminées dans toutes les variables. Par conséquent, les observations utilisées pour calculer les statistiques sont cohérentes dans tous les tableaux.

**Variables de plan qualitatives.** Ce groupe détermine si les valeurs manquantes de l'utilisateur sont valides ou non.

# *Options d'échantillons complexes*

Figure 5-4
*Options*




**Afficher les sous-populations.** Vous pouvez choisir d'afficher les sous-populations dans le même tableau ou dans des tableaux distincts.

Chapitre 6

# Echantillons complexes – Descriptives

La procédure Echantillons complexes – Descriptives affiche les statistiques récapitulatives univariées de plusieurs variables. Vous pouvez éventuellement classer les statistiques par sous-groupes, définis par une ou plusieurs variables qualitatives.

**Exemple :** La procédure Echantillons complexes – Descriptives vous permet d'obtenir les statistiques descriptives univariées des niveaux d'activité des Américains, basées sur les résultats du NHIS (National Health Interview Survey) et dotées d'un plan d'analyse approprié à ces données d'usage public.

**Statistiques :** La procédure calcule pour chaque estimation la moyenne et la somme, ainsi que des tests *t*, les erreurs standard, les intervalles de confiance, les coefficients de variation, les effectifs non pondérés, les tailles de population, les effets de plan et les racines carrées d'effets de plan.

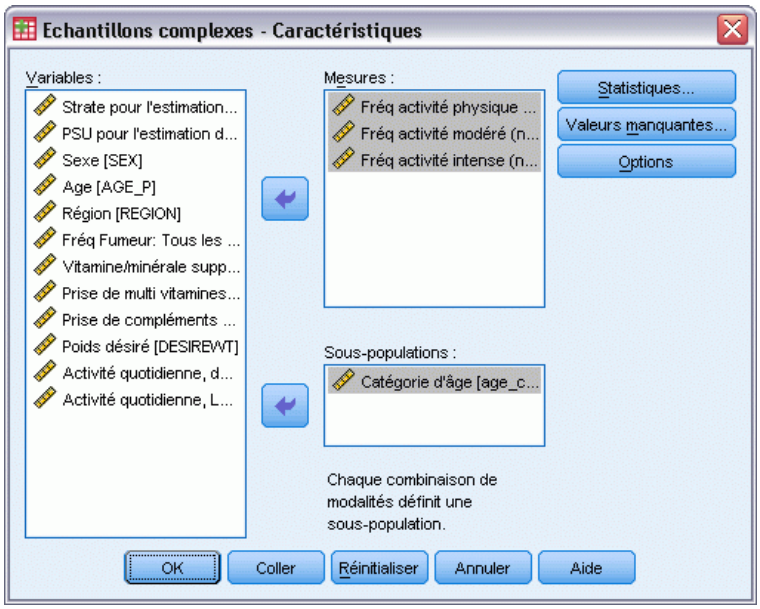
**Données :** Les mesures doivent être des variables d'échelle. Les variables de sous-population peuvent être du type chaîne de caractères ou numérique, mais ne doivent pas être qualitatives.

**Hypothèses :** Les observations dans le fichier de données représentent un échantillon provenant d'un plan complexe et devant être analysées en fonction des spécifications du fichier sélectionné dans la Boîte de dialogue Plan d'échantillonnages complexes.

### *Obtention de descriptives des échantillons complexes*

- A partir des menus, sélectionnez :
  Analyse > Echantillonnage > Descriptives
- Sélectionnez un fichier de plan. Sélectionnez éventuellement un fichier personnalisé de probabilités conjointes.
- Cliquez sur Poursuivre.

Figure 6-1
*Boîte de dialogue Descriptives*

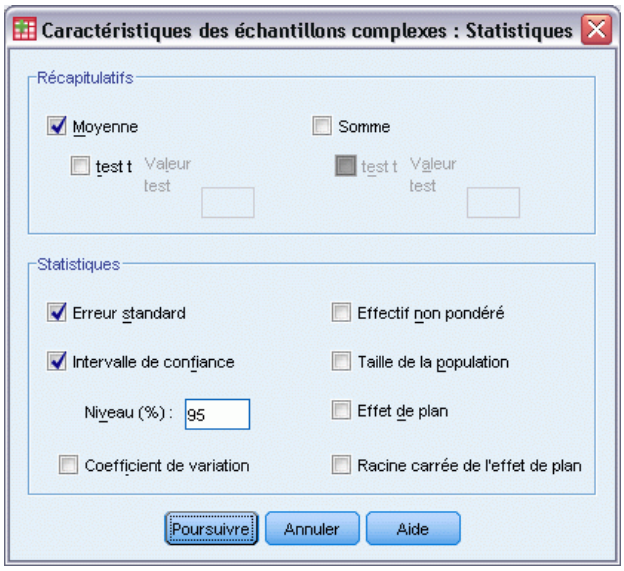



▶ Sélectionnez au moins une variable de mesure.

Vous pouvez éventuellement spécifier les variables définissant les sous-populations. Les statistiques sont calculées séparément pour chaque sous-population.

## *Statistiques des descriptifs des échantillons complexes*

Figure 6-2
*Boîte de dialogue Statistiques Descriptives*



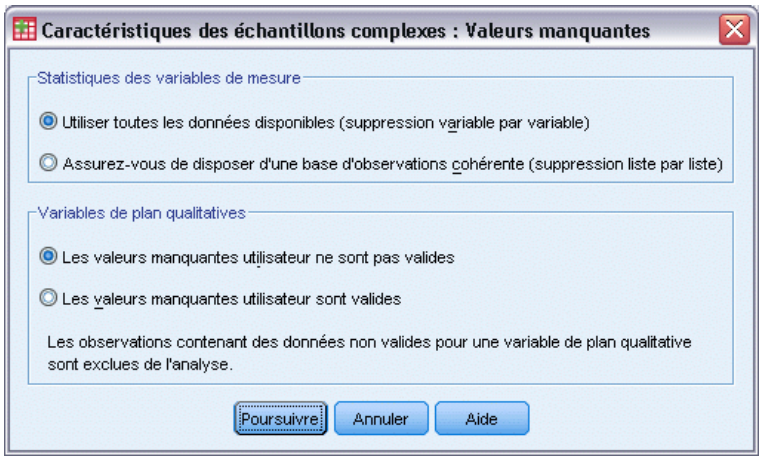
**Principales statistiques :** Ce groupe vous permet de demander les estimations des moyennes et des sommes des variables de mesure. En outre, vous pouvez effectuer des tests *t* sur les estimations en vous basant sur une valeur spécifiée.

**Statistiques :** Ce groupe génère des statistiques associées à la moyenne ou à la somme.

- **Erreur standard :** Erreur standard de l'estimation.
- **Intervalle de confiance :** Intervalle de confiance de l'estimation utilisant le niveau spécifié.
- **Coefficient de variation :** Rapport entre l'erreur standard de l'estimation et l'estimation.
- **Effectif non pondéré :** Nombre d'unités utilisées pour calculer l'estimation.
- **Taille de la population.** Nombre d'unités estimées composant la population.
- **Effet de plan :** Rapport entre la variance de l'estimation et la variance, en partant du principe que l'échantillon est un échantillon aléatoire simple. Il s'agit d'une mesure de l'effet de spécification d'un plan complexe, où les valeurs éloignées de 1 indiquent des effets importants.
- **Racine carrée de l'effet de plan :** Il s'agit d'une mesure de l'effet de spécification d'un plan complexe, où les valeurs éloignées de 1 indiquent des effets importants.

## Valeurs manquantes des descriptifs d'échantillons complexes

Figure 6-3
*Boîte de dialogue Echantillons complexes - Descriptives : Valeurs manquantes*



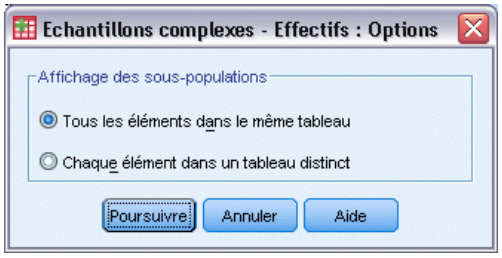

**Statistiques des variables de mesure :** Ce groupe détermine les observations utilisées dans l'analyse.

- **Utilisez toutes les données disponibles.** Les valeurs manquantes sont déterminées variable par variable ; par conséquent, les observations utilisées pour calculer les statistiques peuvent varier suivant les variables de mesure.
- **Assurez-vous de disposer d'une base d'observations cohérente.** Les valeurs manquantes sont déterminées dans toutes les variables ; par conséquent, les observations utilisées pour calculer les statistiques sont cohérentes.

**Variables de plan qualitatives.** Ce groupe détermine si les valeurs manquantes de l'utilisateur sont valides ou non.

# *Options d'échantillons complexes*

Figure 6-4
*Options*




**Afficher les sous-populations.** Vous pouvez choisir d'afficher les sous-populations dans le même tableau ou dans des tableaux distincts.

Chapitre
7

# Tableaux croisés des échantillons complexes

La procédure Echantillons complexes - Tableaux croisés génère les tableaux croisés des paires de variables sélectionnées et affiche des statistiques à deux entrées. Vous pouvez éventuellement classer les statistiques par sous-groupes, définis par une ou plusieurs variables qualitatives.

**Exemple :** La procédure Echantillons complexes - Tableaux croisés vous permet d'obtenir les statistiques de classification croisée fréquence consommation de cigarette/consommation de vitamines des Américains, basées sur les résultats du NHIS (National Health Interview Survey) et dotées d'un plan d'analyse pour ces données d'usage public.

**Statistiques :** La procédure génère pour chaque estimation des tailles de population, des pourcentages en lignes, colonnes et tableaux, des erreurs standard, des intervalles de confiance, des coefficients de variation, des valeurs théoriques, des effets de plan, des racines carrées d'effets de plan, des résidus, des résidus ajustés et des effectifs non pondérés. L'odds ratio, le risque relatif et la différence de risque sont calculés pour les tableaux 2*2. En outre, les statistiques de Pearson et de rapport de vraisemblance sont calculées pour le test d'indépendance des variables de ligne et de colonne.

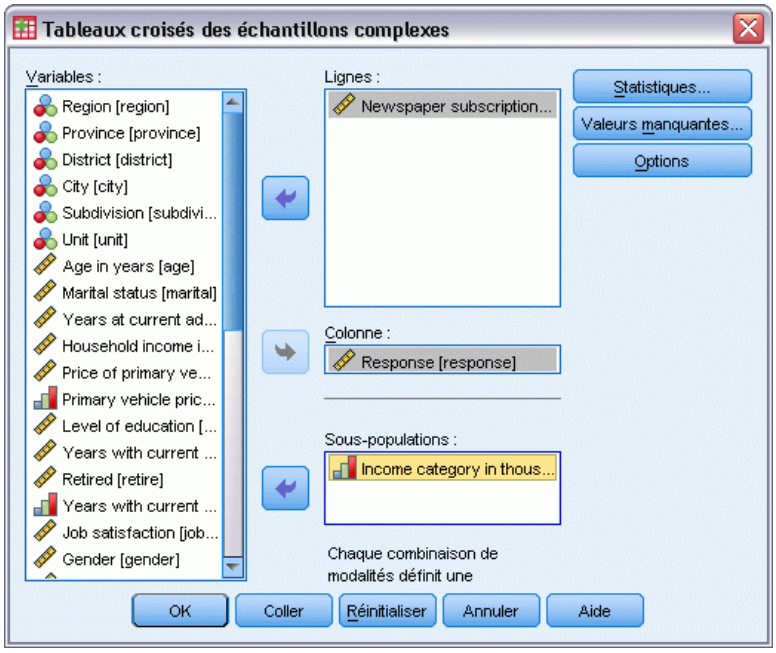
**Données :** Les variables de ligne et de colonne doivent être qualitatives. Les variables de sous-population peuvent être du type chaîne de caractères ou numérique, mais ne doivent pas être qualitatives.

**Hypothèses :** Les observations dans le fichier de données représentent un échantillon provenant d'un plan complexe et devant être analysées en fonction des spécifications du fichier sélectionné dans la Boîte de dialogue Plan d'échantillonnages complexes.

### *Obtention de tableaux croisés des échantillons complexes*

- A partir des menus, sélectionnez :
  Analyse > Echantillonnage > Tableaux croisés
- Sélectionnez un fichier de plan. Sélectionnez éventuellement un fichier personnalisé de probabilités conjointes.
- Cliquez sur Poursuivre.

Figure 7-1
*Boîte de dialogue Tableaux croisés*

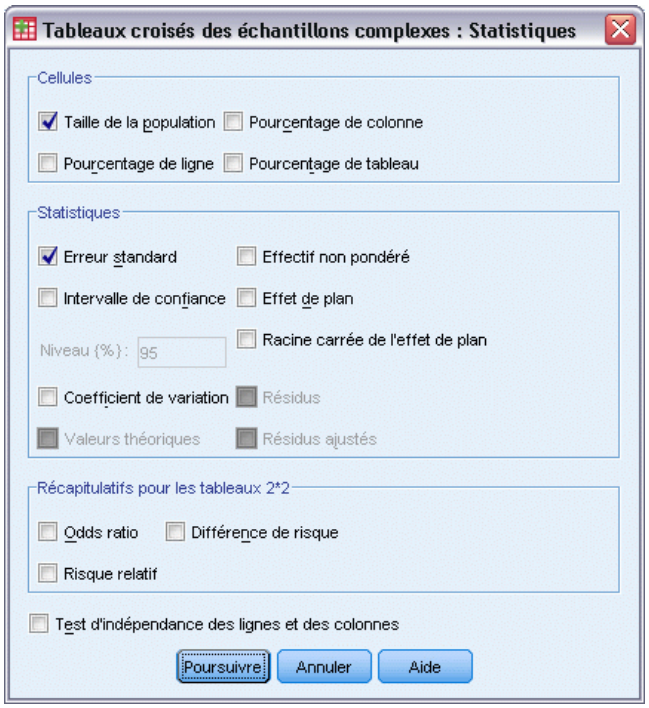



▶ Sélectionnez au moins une variable de ligne et une variable de colonne.

Vous pouvez éventuellement spécifier les variables définissant les sous-populations. Les statistiques sont calculées séparément pour chaque sous-population.

# *Statistiques de tableaux croisés d'échantillons complexes*

Figure 7-2
*Boîte de dialogue Tableaux croisé : Statistiques*




**Cellules :** Ce groupe vous permet de demander les estimations des tailles de population de cellule et les pourcentages de ligne, de colonne et de tableau.

**Statistiques :** Ce groupe génère des statistiques associées à la taille de la population et au pourcentage de ligne, de colonne et de tableau.

- **Erreur standard :** Erreur standard de l'estimation.
- **Intervalle de confiance :** Intervalle de confiance de l'estimation utilisant le niveau spécifié.
- **Coefficient de variation :** Rapport entre l'erreur standard de l'estimation et l'estimation.
- **Effectifs théoriques :** Valeur théorique de l'estimation, sous l'hypothèse de l'indépendance de la variable en ligne et en colonne.
- **Effectif non pondéré :** Nombre d'unités utilisées pour calculer l'estimation.
- **Effet de plan :** Rapport entre la variance de l'estimation et la variance, en partant du principe que l'échantillon est un échantillon aléatoire simple. Il s'agit d'une mesure de l'effet de spécification d'un plan complexe, où les valeurs éloignées de 1 indiquent des effets importants.
- **Racine carrée de l'effet de plan :** Il s'agit d'une mesure de l'effet de spécification d'un plan complexe, où les valeurs éloignées de 1 indiquent des effets importants.

- **Résidus :** 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.
- **Résidus 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.

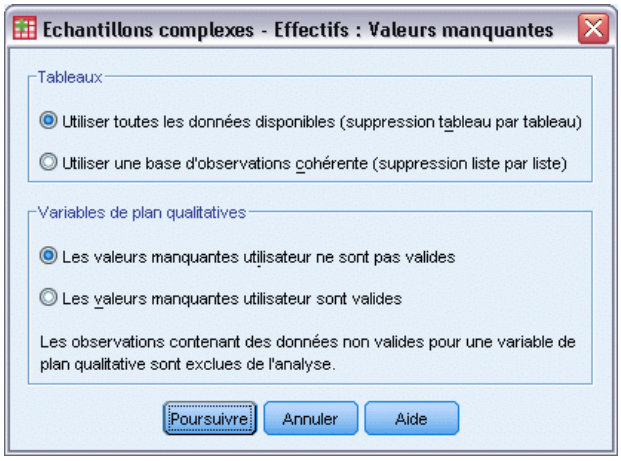
**Récapitulatifs pour les tableaux 2*2 :** Ce groupe génère les statistiques des tableaux dans lesquels les variables de ligne et de colonne possèdent chacune 2 modalités. Chaque groupe est une mesure de la force de l'association entre la présence d'un facteur et la réalisation d'un événement.

- **Rapport de probabilités :** Peut être utilisé comme estimation du risque relatif dans le cas où la réalisation du facteur est rare.
- **Risque relatif :** Rapport entre le risque que survienne un événement en présence du facteur et le risque que survienne l'événement en l'absence du facteur.
- **Différence de risque :** Différence entre le risque que survienne un événement en présence du facteur et le risque que survienne l'événement en l'absence du facteur.

**Test d'indépendance des lignes et des colonnes :** Ce test génère des tests du Khi-deux et de rapport de vraisemblance sur l'hypothèse de l'indépendance entre une ligne et une colonne. Des tests distincts sont effectués pour chaque paire de variables.

## *Valeurs manquantes d'échantillons complexes*

Figure 7-3
*Boîte de dialogue Valeurs manquantes*



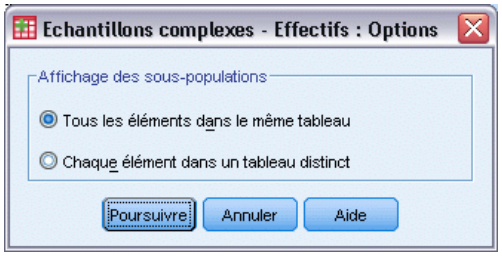

**Tables.** Ce groupe détermine les observations utilisées dans l'analyse.

- **Utilisez toutes les données disponibles.** Les valeurs manquantes sont définies tableau par tableau. Ainsi, les observations utilisées pour calculer les statistiques peuvent varier selon la fréquence ou les tableaux croisés.
- **Utilisez une base d'observations cohérente.** Les valeurs manquantes sont déterminées dans toutes les variables. Par conséquent, les observations utilisées pour calculer les statistiques sont cohérentes dans tous les tableaux.

**Variables de plan qualitatives.** Ce groupe détermine si les valeurs manquantes de l'utilisateur sont valides ou non.

## Options d'échantillons complexes

Figure 7-4
*Options*




**Afficher les sous-populations.** Vous pouvez choisir d'afficher les sous-populations dans le même tableau ou dans des tableaux distincts.

Chapitre
8

# Echantillons complexes – Rapports

La procédure Echantillons complexes – Rapports affiche les statistiques récapitulatives univariées des rapports de variables. Vous pouvez éventuellement classer les statistiques par sous-groupes, définis par une ou plusieurs variables qualitatives.

**Exemple :** La procédure Echantillons complexes – Rapports vous permet d'obtenir les statistiques descriptives du rapport entre la valeur de propriété en cours et la dernière valeur estimée, basée sur les résultats d'un sondage national mené en fonction d'un plan complexe et dotée d'un plan d'analyse approprié pour les données.

**Statistiques :** La procédure génère des estimations de rapport, des tests *t*, des erreurs standard, des intervalles de confiance, des coefficients de variation, des effectifs non pondérés, des tailles de population, des effets de plan et des racines carrées d'effets de plan.

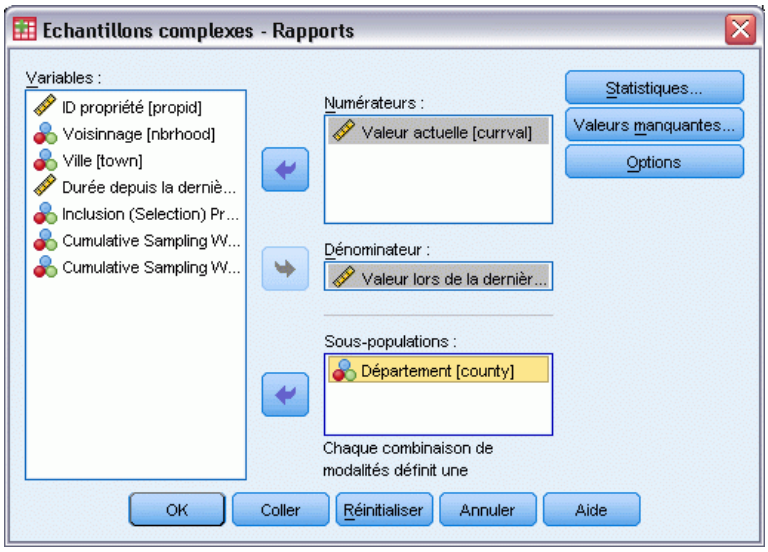
**Données :** Les numérateurs et les dénominateurs doivent être des variables d'échelle positives. Les variables de sous-population peuvent être du type chaîne de caractères ou numérique, mais ne doivent pas être qualitatives.

**Hypothèses :** Les observations dans le fichier de données représentent un échantillon provenant d'un plan complexe et devant être analysées en fonction des spécifications du fichier sélectionné dans la Boîte de dialogue Plan d'échantillonnages complexes.

### *Obtention de rapports d'échantillons complexes*

- A partir des menus, sélectionnez :
  Analyse > Echantillonnage > Ratios...
- Sélectionnez un fichier de plan. Sélectionnez éventuellement un fichier personnalisé de probabilités conjointes.
- Cliquez sur Poursuivre.

Figure 8-1
*Boîte de dialogue Rapports*

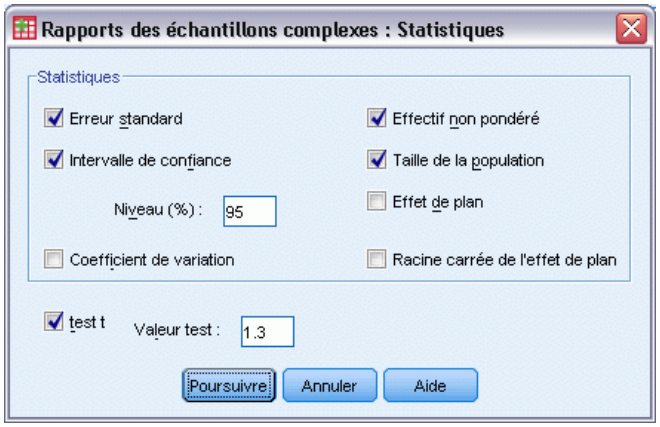



▶ Sélectionnez au moins une variable de numérateur et une variable de dénominateur.

Vous pouvez également indiquer des variables pour définir les sous-groupes pour lesquels les statistiques sont produites.

## *Echantillons complexes – Rapports : Statistiques*

Figure 8-2
*Boîte de dialogue Rapports : Statistiques*




**Statistiques :** Ce groupe produit les statistiques associées à l'estimation du rapport.

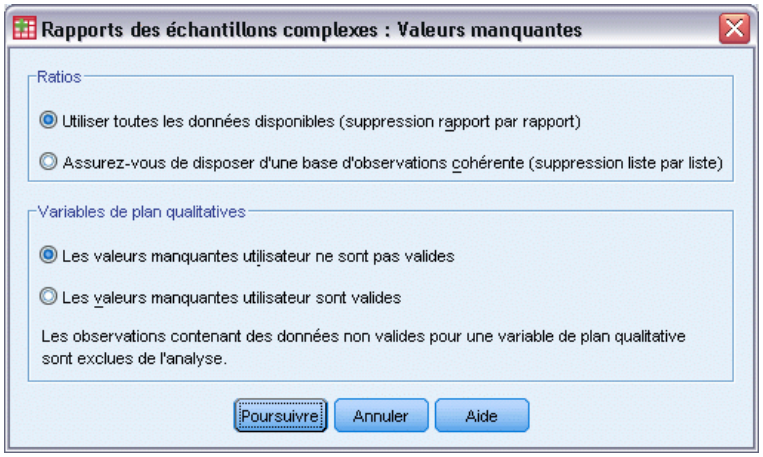
- **Erreur standard :** Erreur standard de l'estimation.
- **Intervalle de confiance :** Intervalle de confiance de l'estimation utilisant le niveau spécifié.
- **Coefficient de variation :** Rapport entre l'erreur standard de l'estimation et l'estimation.
- **Effectif non pondéré :** Nombre d'unités utilisées pour calculer l'estimation.

- **Taille de la population.** Nombre d'unités estimées composant la population.
- **Effet de plan :** Rapport entre la variance de l'estimation et la variance, en partant du principe que l'échantillon est un échantillon aléatoire simple. Il s'agit d'une mesure de l'effet de spécification d'un plan complexe, où les valeurs éloignées de 1 indiquent des effets importants.
- **Racine carrée de l'effet de plan :** Il s'agit d'une mesure de l'effet de spécification d'un plan complexe, où les valeurs éloignées de 1 indiquent des effets importants.

**Test t.** Vous pouvez demander d'effectuer des tests *t* sur les estimations basés sur une valeur spécifiée.

## *Echantillons complexes – Rapports : Valeurs manquantes*

Figure 8-3
*Boîte de dialogue Rapports : Valeurs manquantes*



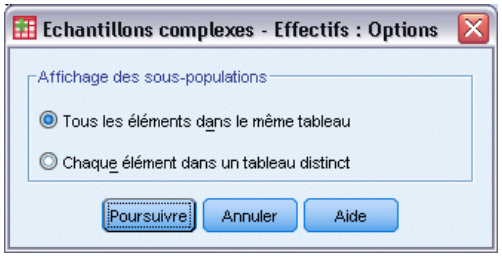

**Ratios :** Ce groupe détermine les observations utilisées dans l'analyse.

- **Utilisez toutes les données disponibles.** Les valeurs manquantes sont définies rapport par rapport. Ainsi, les observations utilisées pour calculer les statistiques peuvent varier d'une paire numérateur-dénominateur à l'autre.
- **Assurez-vous de disposer d'une base d'observations cohérente.** Les valeurs manquantes sont déterminées dans toutes les variables. Par conséquent, les observations utilisées pour le calcul des statistiques sont cohérentes.

**Variables de plan qualitatives.** Ce groupe détermine si les valeurs manquantes de l'utilisateur sont valides ou non.

# Options d'échantillons complexes

Figure 8-4
*Options*




**Afficher les sous-populations.** Vous pouvez choisir d'afficher les sous-populations dans le même tableau ou dans des tableaux distincts.

Chapitre
9

# Modèle linéaire général des échantillons complexes

La procédure relative au modèle linéaire général des échantillons complexes effectue une analyse de régression linéaire, ainsi qu'une analyse de la variance et de la covariance, pour les échantillons réalisés à l'aide de méthodes d'échantillonnage complexes. Vous pouvez également demander une analyse pour une sous-population.

**Exemple :** Une enseigne de supermarchés a interrogé, en fonction d'un plan complexe, un groupe de clients au sujet de leurs habitudes de consommation. Compte tenu des résultats de l'enquête et de la somme dépensée par les clients au cours du mois précédent, l'enseigne souhaite voir si la fréquence des achats est liée aux dépenses mensuelles, et ce en prenant en compte le sexe du client et en intégrant le plan d'échantillonnage.

**Statistiques :** La procédure produit des estimations, des erreurs standard, des intervalles de confiance, des tests *t*, des effets de plan, des racines carrées d'effets de plan pour les paramètres du modèle ; elle fournit également les corrélations et les covariances des estimations des paramètres. Les mesures de qualité d'ajustement et les statistiques descriptives pour les variables dépendantes et indépendantes sont également disponibles. Vous pouvez également demander la moyenne marginale estimée pour les niveaux de facteurs de modèles et les interactions entre facteurs

**Données.** La variable dépendante est quantitative. Les facteurs sont qualitatifs. Les covariables sont des variables quantitatives liées à la variable dépendante. Les variables de sous-population peuvent être du type chaîne de caractères ou numérique, mais ne doivent pas être qualitatives.

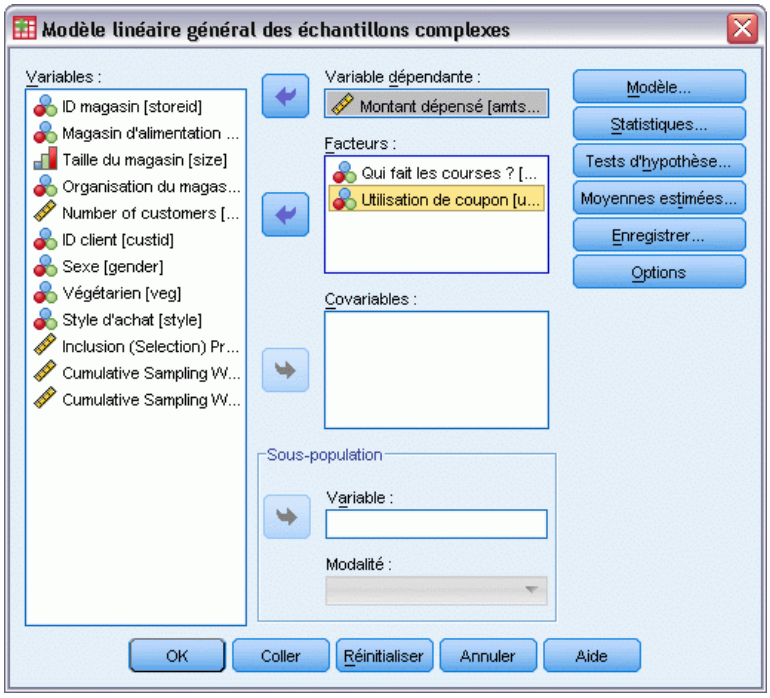
**Hypothèses :** Les observations dans le fichier de données représentent un échantillon provenant d'un plan complexe et devant être analysées en fonction des spécifications du fichier sélectionné dans la Boîte de dialogue Plan d'échantillonnages complexes.

### *Obtention d'un modèle linéaire général des échantillons complexes*

A partir des menus, sélectionnez :
Analyse > Echantillons complexes > Modèle linéaire général...

- Sélectionnez un fichier de plan. Sélectionnez éventuellement un fichier personnalisé de probabilités conjointes.
- Cliquez sur Poursuivre.

Figure 9-1
*Boîte de dialogue Modèle linéaire général*

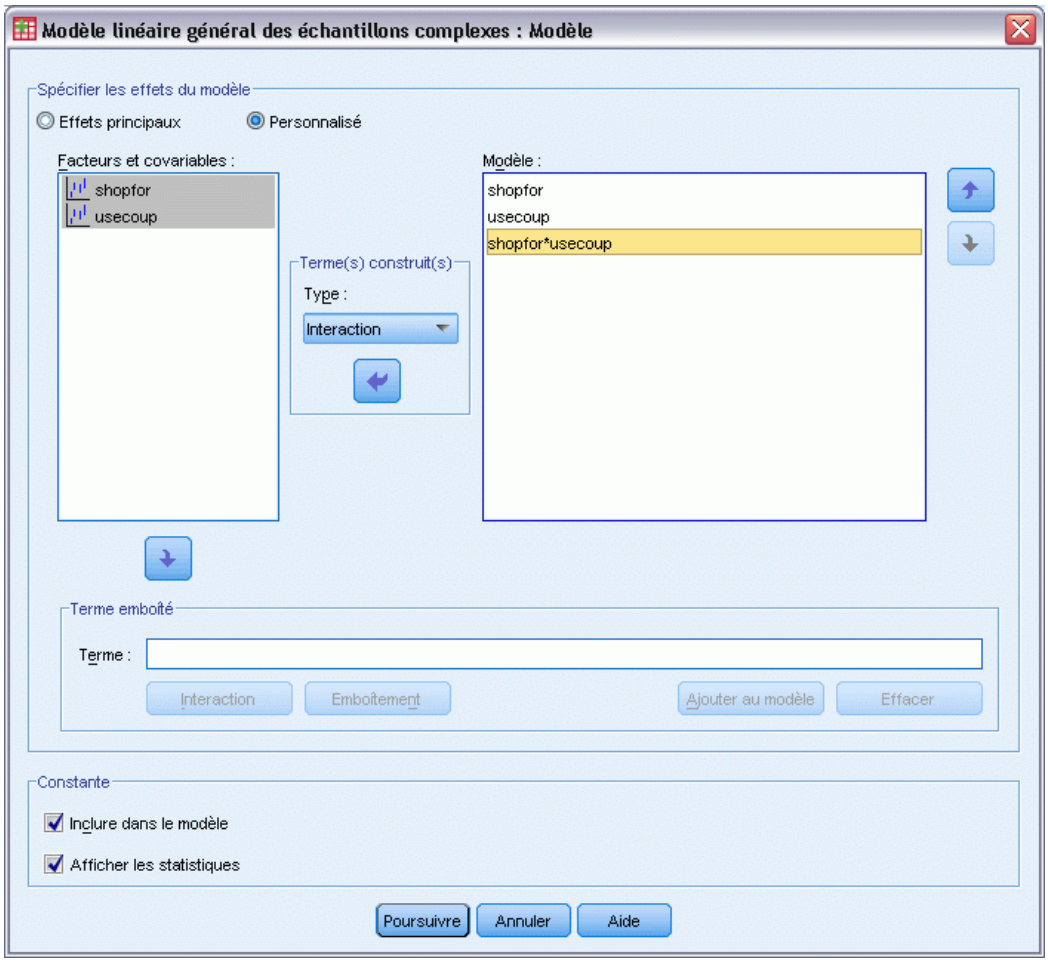



► Sélectionnez une variable dépendante.

Sinon, vous pouvez :

- Sélectionnez des variables pour les facteurs et covariables, en fonction de vos données.
- Indiquez une variable pour définir une sous-population. L'analyse est effectuée uniquement pour la modalité sélectionnée de la variable de sous-population.

Figure 9-2
*Boîte de dialogue Modèle*




**Spécifier les effets du modèle.** Par défaut, la procédure élabore un modèle contenant des effets principaux à l'aide des covariables et facteurs spécifiés dans la boîte de dialogue principale. Vous pouvez également créer un modèle personnalisé contenant des effets d'interaction et des termes emboîtés.

### *Termes non emboîtés*

Pour les facteurs et covariables sélectionnés :

**Interaction :** Crée le terme d'interaction du plus haut niveau pour toutes les variables sélectionnées.

**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.

### *Termes emboîtés*

Dans cette procédure, vous pouvez construire des termes emboîtés pour votre modèle. Les termes emboîtés sont utiles pour modéliser l'effet d'un facteur ou d'une covariable dont les valeurs n'interagissent pas avec les niveaux d'un autre facteur. Par exemple, une chaîne d'épiceries peut suivre les habitudes d'achat de ses clients à divers emplacements de magasin. Puisque chaque client ne fréquente qu'un seul de ces magasins, l'effet *Client* peut être considéré comme étant **emboîté dans** l'effet *Emplacement des magasins*.

En outre, vous pouvez inclure des effets d'interaction, tels que des termes polynomiaux impliquant la même covariable, ou ajouter plusieurs niveaux d'emboîtement au terme emboîté.

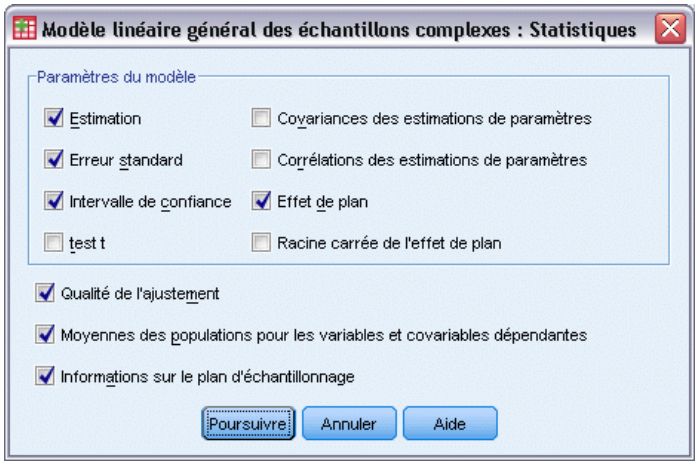
**Limites.** Les termes emboîtés comportent les restrictions suivantes :

- Tous les facteurs d'une interaction doivent être uniques. Ainsi, si *A* est un facteur, la spécification *A*A* n'est pas valide.
- Tous les facteurs d'un effet en cascade doivent être uniques. Ainsi, si *A* est un facteur, la spécification *A(A)* n'est pas valide.
- Aucun effet ne peut être emboîtés dans un effet de covariable. Ainsi, si *A* est un facteur et *X* une covariable, la spécification *A(X)* n'est pas valide.

**Constante.** L'ordonnée est généralement incluse dans le modèle. Si vous partez du principe que les données passent par l'origine, vous pouvez exclure la constante. Même si vous incluez la constante dans le modèle, vous pouvez supprimer les statistiques qui lui sont associées.

## *Modèle linéaire général des échantillons complexes - Statistiques*

Figure 9-3
*Boîte de dialogue Statistiques du modèle linéaire général*




**Paramètres du modèle.** Ce groupe permet de contrôler l'affichage des statistiques associées aux paramètres du modèle.

- **Estimation.** Affiche des estimations des coefficients.
- **Erreur standard :** Affiche l'erreur standard pour chaque estimation de coefficient.

- **Intervalle de confiance :** Affiche l'intervalle de confiance pour chaque estimation de coefficient. Le niveau de confiance de l'intervalle est défini dans la boîte de dialogue Options.
- **Test t.** Affiche un test *t* pour chaque estimation de coefficient. L'hypothèse nulle de chaque test correspond au cas où la valeur du coefficient est 0.
- **Covariances des estimations de paramètres.** Affiche une estimation de la matrice de covariance pour les coefficients du modèle.
- **Corrélations des estimations de paramètres.** Affiche une estimation de la matrice de corrélation pour les coefficients du modèle.
- **Effet de plan :** Rapport entre la variance de l'estimation et la variance, en partant du principe que l'échantillon est un échantillon aléatoire simple. Il s'agit de la mesure d'un effet de la spécification d'un plan complexe, pour lequel des valeurs plus petites indiquent des effets plus importants.
- **Racine carrée de l'effet de plan :** Il s'agit d'une mesure de l'effet de spécification d'un plan complexe, où les valeurs éloignées de 1 indiquent des effets importants.

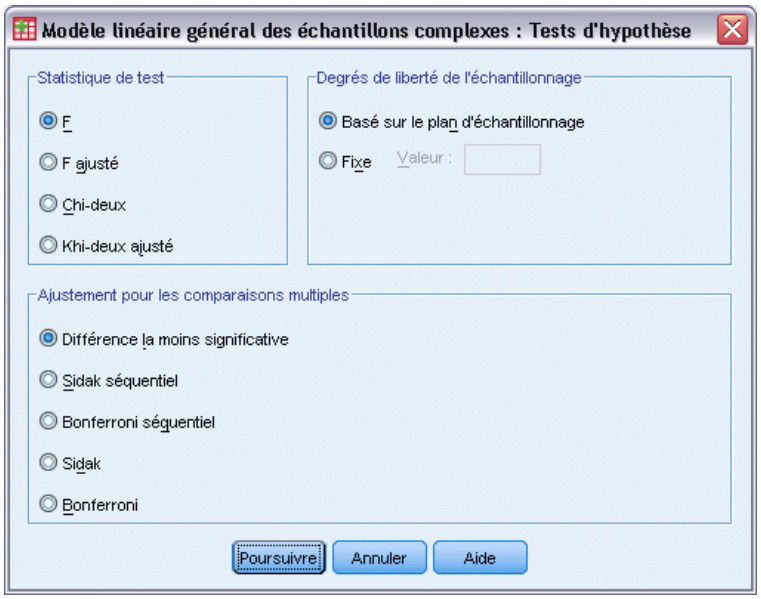
**Qualité de l'ajustement :** Affiche les statistiques de $R^2$ et d'erreur quadratique moyenne.

**Moyennes des populations pour les variables et covariables dépendantes.** Affiche les informations récapitulatives sur la variable dépendante, les covariables et les facteurs.

**Informations sur le plan d'échantillonnage.** Affiche les informations récapitulatives relatives à l'échantillon, y compris les effectifs non pondérés et la taille de la population.

## *Tests d'hypothèse des échantillons complexes*

Figure 9-4
*Boîte de dialogue Tests d'hypothèse*




**Statistique de test.** Ce groupe vous permet de sélectionner le type de statistiques à utiliser pour tester les hypothèses. Vous avez le choix entre *F*, *F* ajusté, Khi-deux et Khi-deux ajusté.

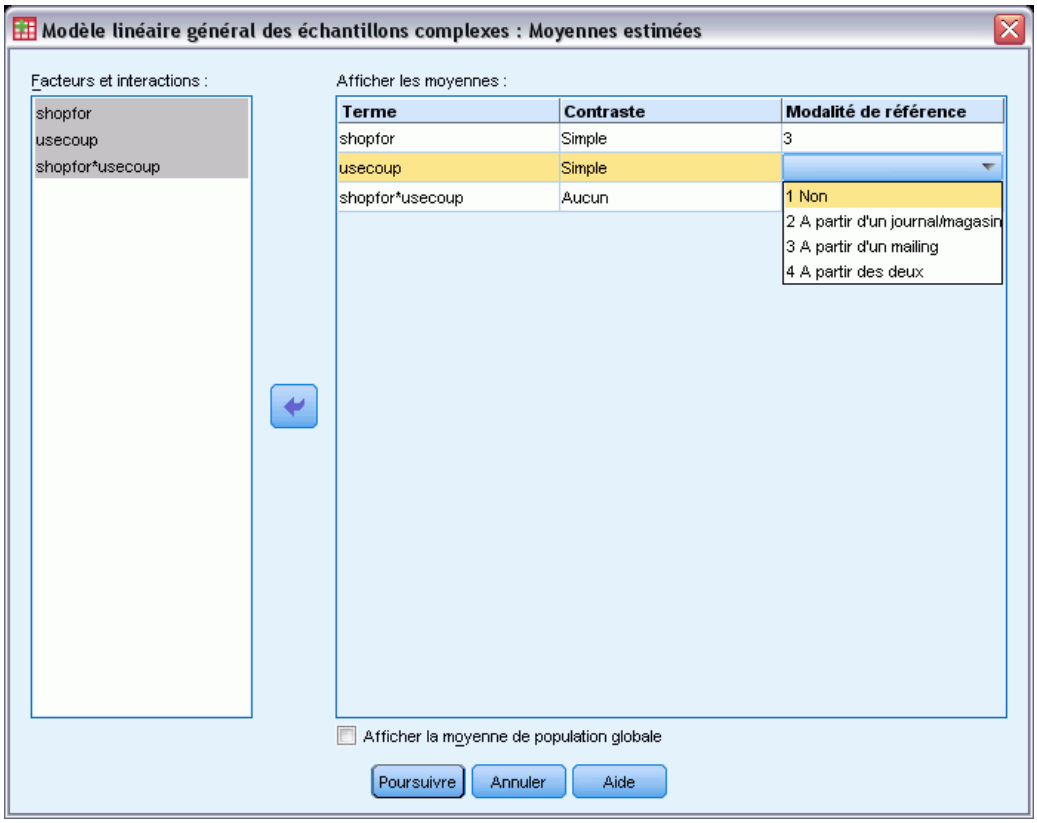
**Degrés de liberté de l'échantillonnage.** Ce groupe permet de contrôler les degrés de liberté du plan d'échantillonnage utilisés pour calculer les valeurs *p* pour toutes les statistiques de test. Si elle est basée sur le plan d'échantillonnage, cette valeur correspond à la différence entre le nombre d'unités d'échantillonnage principales et le nombre de strates présentes à la première étape de l'échantillonnage. Vous pouvez également définir un degré de liberté personnalisé en indiquant un entier positif.

**Ajustement pour les comparaisons multiples.** Lors de l'exécution de tests d'hypothèse avec plusieurs contrastes, vous pouvez ajuster le seuil global de signification à partir des seuils de signification des contrastes inclus. Ce groupe vous permet de choisir la méthode d'ajustement.

- **Différence la moins significative.** Cette méthode ne contrôle pas l'intégralité de la probabilité de rejet des hypothèses qui présentent des contrastes linéaires différents des valeurs d'hypothèse nulles.
- **Procédure de Sidak Séquentielle.** Il s'agit d'une procédure descendante de rejet séquentiel de Bonferroni beaucoup moins stricte en ce qui concerne le rejet des différentes hypothèses mais qui conserve le même niveau global de signification.
- **Bonferroni séquentiel.** Il s'agit d'une procédure descendante de rejet séquentiel de Bonferroni beaucoup moins stricte en ce qui concerne le rejet des différentes hypothèses mais qui conserve le même niveau global de signification.
- **Sidak.** Cette méthode propose des bornes plus petites que l'approche de Bonferroni.
- **Bonferroni.** Cette méthode ajuste le niveau de signification observé lorsque des contrastes multiples sont testés.

# Modèle linéaire général des échantillons complexes - Moyennes estimées

Figure 9-5
*Boîte de dialogue Moyennes estimées du modèle linéaire général*




La boîte de dialogue Moyennes estimées permet d'afficher les moyennes marginales estimées du modèle pour les niveaux de facteurs et les interactions entre facteurs indiqués dans la sous-boîte de dialogue Modèle. Vous pouvez également demander l'affichage de la moyenne de population globale.

**Terme.** Les moyennes estimées sont calculées pour les interactions et les facteurs sélectionnés.

**Contraste.** Le contraste détermine le mode de définition des tests d'hypothèse pour la comparaison des moyennes estimées.

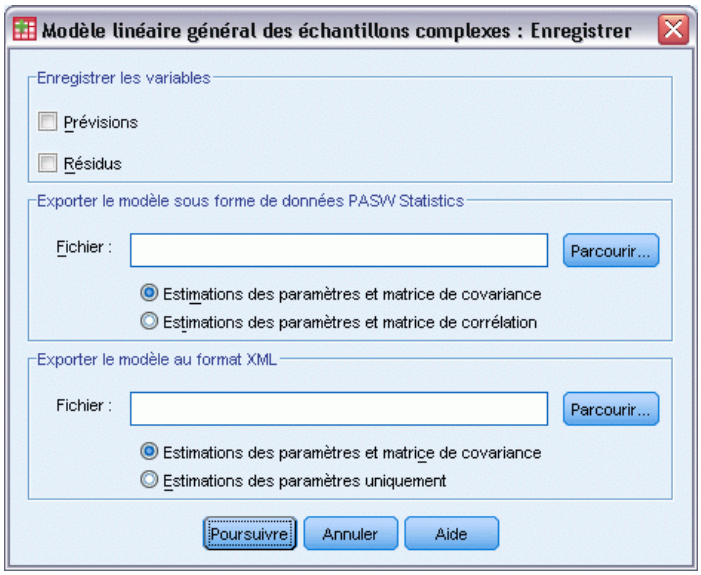
- **Simple.** Compare la moyenne de chaque niveau à celle d'un niveau donné. Ce type de contraste est utile lorsqu'il y a un groupe de contrôle.
- **Ecart.** Compare la moyenne de chaque niveau (hormis une modalité de référence) à la moyenne de tous les niveaux (grande moyenne). Les niveaux du facteur peuvent être de n'importe quel ordre.
- **Différencié d'ordre.** Compare la moyenne de chaque niveau (hormis le premier) à la moyenne des niveaux précédents. (Parfois appelé contraste de 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 précédent.
- **Polynomial.** Compare l'effet linéaire, l'effet quadratique, l'effet cubique, etc. Le premier degré de liberté contient l'effet linéaire sur toutes les modalités, le second degré l'effet quadratique, etc. Ces contrastes servent souvent à estimer les tendances polynomiales.

**Modalité de référence.** Les contrastes simple et d'écart nécessitent une modalité de référence ou un niveau de facteur servant de base de comparaison avec les autres.

## Modèle linéaire général des échantillons complexes - Enregistrement

Figure 9-6
*Boîte de dialogue Enregistrer le modèle linéaire général*




**Enregistrer les variables.** Ce groupe permet d'enregistrer les prévisions et les résidus du modèle en tant que nouvelles variables dans le fichier de travail.

**Exporter le modèle en tant que données SPSS Statistics.** Ecrit un fichier de données dans le format IBM® SPSS® Statisticscontenant la corrélation des paramètres ou la matrice de covariance avec les estimations des paramètres, les erreurs standard, les valeurs de significativité et les degrés de liberté. L'ordre des variables dans le fichier de matrice est le suivant.

- **rowtype_.** Prend les valeurs (et étiquettes de valeurs) suivantes : COV (covariances), CORR (corrélations), EST (estimations des paramètres), SE (erreurs standard), SIG (seuil de signification) et DF (degrés de liberté du plan d'échantillonnage). Il existe une observation distincte avec le type de ligne COV (ou CORR) pour chaque paramètre de modèle et une observation distincte pour chacun des autres types de ligne.

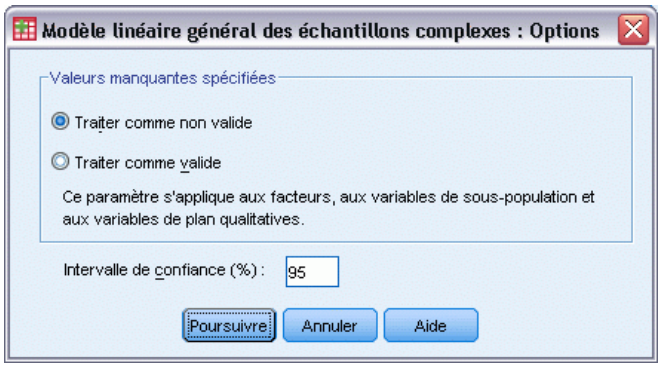
- **varname_.** Prend les valeurs P1, P2, etc., correspondant à une liste triée de tous les paramètres de modèle pour les types de ligne COV ou CORR, avec des étiquettes de valeur correspondant aux chaînes de paramètres affichées dans le tableau Estimations des paramètres. Les cellules sont vides pour les autres types de ligne.
- **P1, P2, ...** Ces variables correspondent à une liste triée de tous les paramètres de modèle, avec des étiquettes de variable correspondant aux chaînes de paramètres affichées dans le tableau Estimations des paramètres, et prennent leurs valeurs en fonction du type de ligne. Pour les paramètres redondants, toutes les covariances et les estimations de paramètres sont définies sur zéro, et l'ensemble des corrélations, erreurs standard, seuils de signification et degrés de liberté résiduels sont définis sur la valeur manquante par défaut.

*Remarque* : Ce fichier n'est pas immédiatement utilisable pour d'autres analyses dans d'autres procédures que la lecture d'un fichier de matrice, sauf si ces procédures acceptent tous les types de ligne exportés ici.

**Exporter le modèle au format XML.** Enregistre les estimations et la matrice de covariance des paramètres, si vous l'avez sélectionnée, 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.

# Modèle linéaire général des échantillons complexes - Options

Figure 9-7
*Boîte de dialogue Options du modèle linéaire général*




**Valeurs manquantes spécifiées.** Toutes les variables de plan, ainsi que la variable dépendante et les covariables, doivent contenir des valeurs valides. Les observations comportant des données non valides pour l'une de ces variables sont supprimées de l'analyse. Vous pouvez ainsi décider si les valeurs manquantes spécifiées par l'utilisateur sont traitées comme étant valides dans la strate, la classe, la sous-population et les variables actives.

**Intervalle de confiance :** Il s'agit du niveau d'intervalle de confiance pour les estimations de coefficient et les moyennes marginales estimées. Spécifiez une valeur supérieure ou égale à 50 et inférieure à 100.

# Fonctionnalités supplémentaires de la commande CSGLM

Le langage de syntaxe de commande vous permet aussi de :

- Spécifier les tests d'effets par rapport à une combinaison linéaire d'effets ou une valeur (à l'aide de la sous-commande `CUSTOM`).
- Donner aux covariables des valeurs autres que leur moyenne lors du calcul de la moyenne marginale estimée (à l'aide de la sous-commande `EMMEANS`).
- Spécifier les mesures pour les contrastes polynomiaux (à l'aide de la sous-commande `EMMEANS`).
- Indiquer une valeur de tolérance pour le contrôle des particularités (à l'aide de la sous-commande `CRITERIA`).
- Créer des noms définis par l'utilisateur pour les variables enregistrées (à l'aide de la sous-commande `SAVE`).
- Générer un tableau des fonctions générales estimées (à l'aide de la sous-commande `PRINT`).

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

Chapitre
10

# Régression logistique des échantillons complexes

La procédure de régression logistique des échantillons complexes effectue une analyse de régression logistique sur une variable dépendante binaire ou multinomiale, pour les échantillons réalisés à l'aide de méthodes d'échantillonnage complexes. Vous pouvez également demander une analyse pour une sous-population.

**Exemple :** Un responsable des prêts a recueilli l'historique des prêts octroyés aux clients à différents guichets, en fonction d'un plan complexe. Lors de l'intégration du plan d'échantillonnage, il souhaite voir si la probabilité de défaut de paiement est liée à l'âge, au parcours professionnel et au montant de la dette.

**Statistiques :** La procédure produit des estimations, des estimations exponentielles, des erreurs standard, des intervalles de confiance, des tests *t*, des effets de plan des racines carrées d'effets de plan pour les paramètres du modèle ; elle fournit également les corrélations et les covariances des estimations des paramètres. Les statistiques de pseudo $R^2$, les tableaux de classement et les statistiques descriptives des variables dépendantes et indépendantes sont également disponibles.

**Données.** La variable dépendante est qualitative. Les facteurs sont qualitatifs. Les covariables sont des variables quantitatives liées à la variable dépendante. Les variables de sous-population peuvent être du type chaîne de caractères ou numérique, mais ne doivent pas être qualitatives.

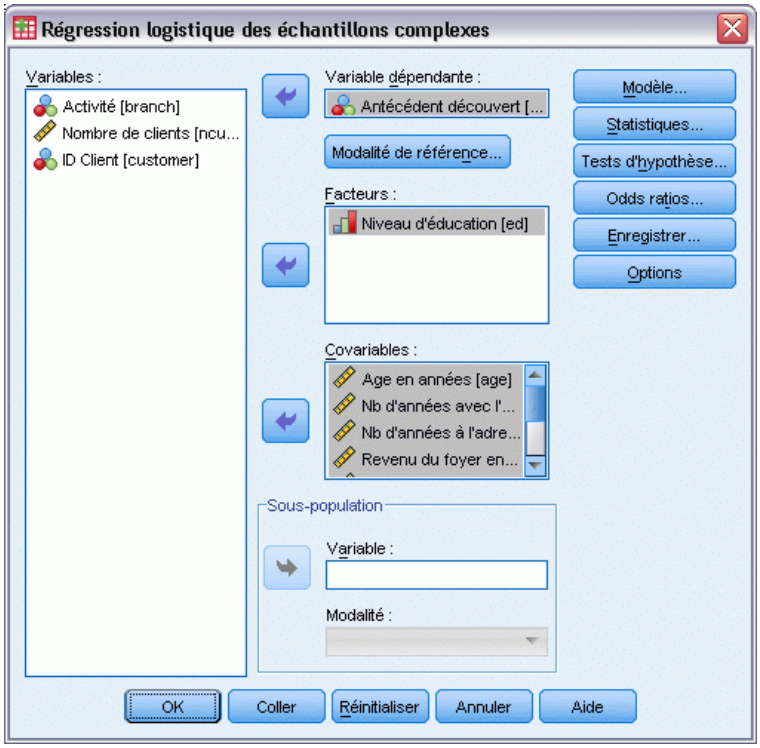
**Hypothèses :** Les observations dans le fichier de données représentent un échantillon provenant d'un plan complexe et devant être analysées en fonction des spécifications du fichier sélectionné dans la Boîte de dialogue Plan d'échantillonnages complexes.

### *Obtention de la régression logistique des échantillons complexes*

A partir des menus, sélectionnez :
Analyse > Echantillons complexes > Régression logistique...

- Sélectionnez un fichier de plan. Sélectionnez éventuellement un fichier personnalisé de probabilités conjointes.
- Cliquez sur Poursuivre.

Figure 10-1
*Boîte de dialogue Régression logistique*



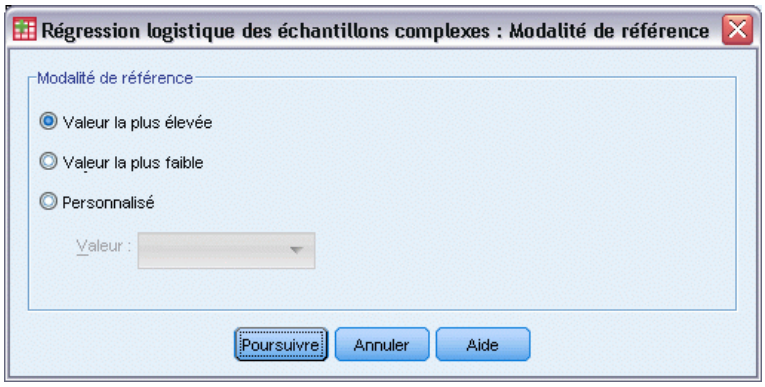

▶ Sélectionnez une variable dépendante.

Sinon, vous pouvez :

- Sélectionnez des variables pour les facteurs et covariables, en fonction de vos données.
- Indiquez une variable pour définir une sous-population. L'analyse est effectuée uniquement pour la modalité sélectionnée de la variable de sous-population.

# Régression logistique des échantillons complexes - Modalité de référence

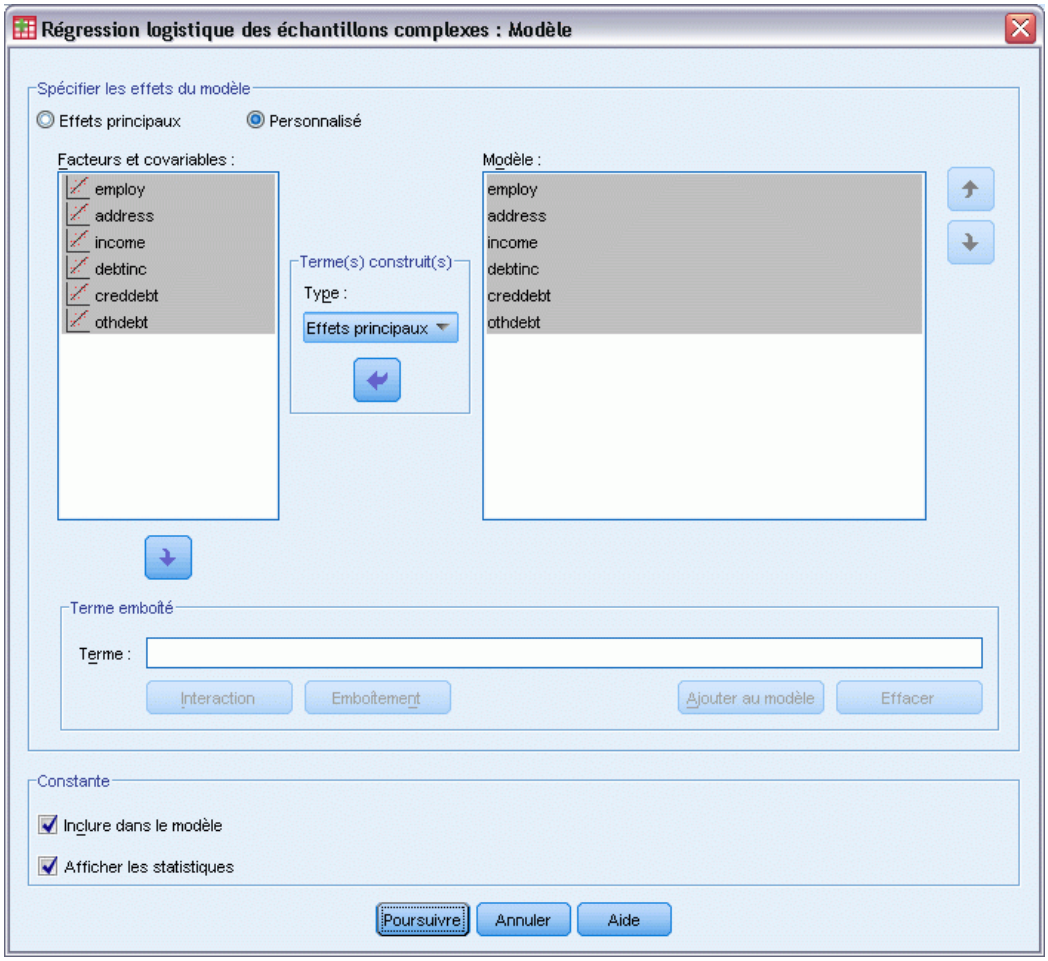
Figure 10-2
*Boîte de dialogue Modalité de référence de la régression logistique*




Par défaut, la procédure de régression logistique des échantillons complexes fait de la modalité ayant la valeur la plus élevée la modalité de référence. Cette boîte de dialogue vous permet de spécifier la valeur la plus élevée, la plus basse ou une modalité personnalisée en tant que modalité de référence.

# Modèle de régression logistique des échantillons complexes

Figure 10-3
*Boîte de dialogue Modèle de régression logistique*




**Spécifier les effets du modèle.** Par défaut, la procédure élabore un modèle contenant des effets principaux à l'aide des covariables et facteurs spécifiés dans la boîte de dialogue principale. Vous pouvez également créer un modèle personnalisé contenant des effets d'interaction et des termes emboîtés.

### *Termes non emboîtés*

Pour les facteurs et covariables sélectionnés :

**Interaction :** Crée le terme d'interaction du plus haut niveau pour toutes les variables sélectionnées.

**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.

#### *Termes emboîtés*

Dans cette procédure, vous pouvez construire des termes emboîtés pour votre modèle. Les termes emboîtés sont utiles pour modéliser l'effet d'un facteur ou d'une covariable dont les valeurs n'interagissent pas avec les niveaux d'un autre facteur. Par exemple, une chaîne d'épiceries peut suivre les habitudes d'achat de ses clients à divers emplacements de magasin. Puisque chaque client ne fréquente qu'un seul de ces magasins, l'effet *Client* peut être considéré comme étant **emboîté dans** l'effet *Emplacement des magasins*.

En outre, vous pouvez inclure des effets d'interaction, tels que des termes polynomiaux impliquant la même covariable, ou ajouter plusieurs niveaux d'emboîtement au terme emboîté.

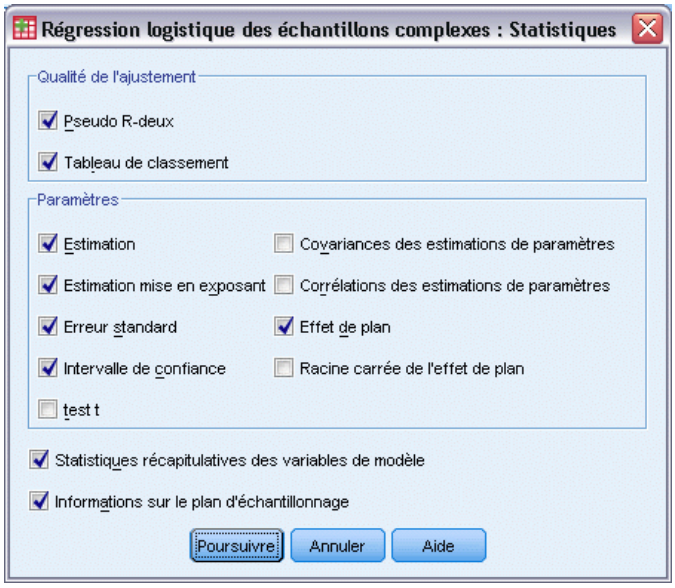
**Limites.** Les termes emboîtés comportent les restrictions suivantes :

- Tous les facteurs d'une interaction doivent être uniques. Ainsi, si *A* est un facteur, la spécification *A*A* n'est pas valide.
- Tous les facteurs d'un effet en cascade doivent être uniques. Ainsi, si *A* est un facteur, la spécification *A(A)* n'est pas valide.
- Aucun effet ne peut être emboîtés dans un effet de covariable. Ainsi, si *A* est un facteur et *X* une covariable, la spécification *A(X)* n'est pas valide.

**Constante.** L'ordonnée est généralement incluse dans le modèle. Si vous partez du principe que les données passent par l'origine, vous pouvez exclure la constante. Même si vous incluez la constante dans le modèle, vous pouvez supprimer les statistiques qui lui sont associées.

## *Régression logistique des échantillons complexes - Statistiques*

Figure 10-4
*Boîte de dialogue Statistiques de régression logistique*



**Qualité de l'ajustement.** Contrôle l'affichage des statistiques qui mesurent les performances globales du modèle.

- **Pseudo R-deux.** La statistique$R^2$ de la régression linéaire n'a pas de véritable équivalent dans les modèles de régression logistique. On trouve, à la place, plusieurs mesures qui essaient de reproduire les propriétés de la statistique $R^2$.
- **Tableau de classement :** Affiche les classifications croisées mises en tableau des modalités observées, classées selon les modalités prévues par le modèle en fonction de la variable dépendante.

**Paramètres.** Ce groupe permet de contrôler l'affichage des statistiques associées aux paramètres du modèle.

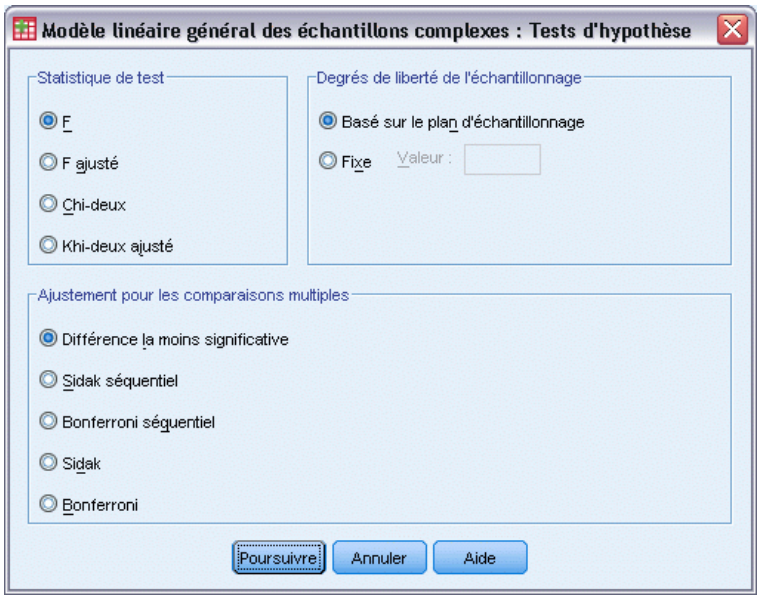
- **Estimation.** Affiche des estimations des coefficients.
- **Estimation mise en exposant.** Affiche la base du logarithme népérien élevée à la puissance des estimations des coefficients. L'estimation présente des propriétés parfaitement adaptées aux tests statistiques ; l'estimation exponentielle, ou exp(B), est, quant à elle, plus facile à interpréter.
- **Erreur standard :** Affiche l'erreur standard pour chaque estimation de coefficient.
- **Intervalle de confiance :** Affiche l'intervalle de confiance pour chaque estimation de coefficient. Le niveau de confiance de l'intervalle est défini dans la boîte de dialogue Options.
- **Test t.** Affiche un test *t* pour chaque estimation de coefficient. L'hypothèse nulle de chaque test correspond au cas où la valeur du coefficient est 0.
- **Covariances des estimations de paramètres.** Affiche une estimation de la matrice de covariance pour les coefficients du modèle.
- **Corrélations des estimations de paramètres.** Affiche une estimation de la matrice de corrélation pour les coefficients du modèle.
- **Effet de plan :** Rapport entre la variance de l'estimation et la variance, en partant du principe que l'échantillon est un échantillon aléatoire simple. Il s'agit de la mesure d'un effet de la spécification d'un plan complexe, pour lequel des valeurs plus petites indiquent des effets plus importants.
- **Racine carrée de l'effet de plan :** Il s'agit d'une mesure de l'effet de spécification d'un plan complexe, où les valeurs éloignées de 1 indiquent des effets importants.

**Statistiques récapitulatives des variables de modèle.** Affiche les informations récapitulatives sur la variable dépendante, les covariables et les facteurs.

**Informations sur le plan d'échantillonnage.** Affiche les informations récapitulatives relatives à l'échantillon, y compris les effectifs non pondérés et la taille de la population.

# Tests d'hypothèse des échantillons complexes

Figure 10-5
*Boîte de dialogue Tests d'hypothèse*




**Statistique de test.** Ce groupe vous permet de sélectionner le type de statistiques à utiliser pour tester les hypothèses. Vous avez le choix entre *F*, *F* ajusté, Khi-deux et Khi-deux ajusté.

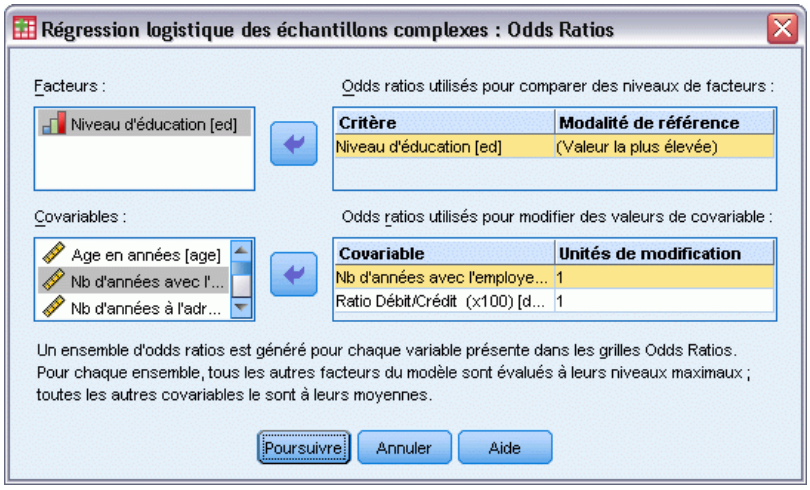
**Degrés de liberté de l'échantillonnage.** Ce groupe permet de contrôler les degrés de liberté du plan d'échantillonnage utilisés pour calculer les valeurs *p* pour toutes les statistiques de test. Si elle est basée sur le plan d'échantillonnage, cette valeur correspond à la différence entre le nombre d'unités d'échantillonnage principales et le nombre de strates présentes à la première étape de l'échantillonnage. Vous pouvez également définir un degré de liberté personnalisé en indiquant un entier positif.

**Ajustement pour les comparaisons multiples.** Lors de l'exécution de tests d'hypothèse avec plusieurs contrastes, vous pouvez ajuster le seuil global de signification à partir des seuils de signification des contrastes inclus. Ce groupe vous permet de choisir la méthode d'ajustement.

- **Différence la moins significative.** Cette méthode ne contrôle pas l'intégralité de la probabilité de rejet des hypothèses qui présentent des contrastes linéaires différents des valeurs d'hypothèse nulles.
- **Procédure de Sidak Séquentielle.** Il s'agit d'une procédure descendante de rejet séquentiel de Bonferroni beaucoup moins stricte en ce qui concerne le rejet des différentes hypothèses mais qui conserve le même niveau global de signification.
- **Bonferroni séquentiel.** Il s'agit d'une procédure descendante de rejet séquentiel de Bonferroni beaucoup moins stricte en ce qui concerne le rejet des différentes hypothèses mais qui conserve le même niveau global de signification.
- **Sidak.** Cette méthode propose des bornes plus petites que l'approche de Bonferroni.
- **Bonferroni.** Cette méthode ajuste le niveau de signification observé lorsque des contrastes multiples sont testés.

# *Régression logistique des échantillons complexes - Odds ratios*

Figure 10-6
*Boîte de dialogue Odds Ratios de la régression logistique*




La boîte de dialogue Odds Ratios permet d'afficher les odds ratios estimés par modèle pour les covariables et les facteurs indiqués. Un groupe d'odds ratios distinct est calculé pour chaque modalité de la variable dépendante, à l'exception de la modalité de référence.

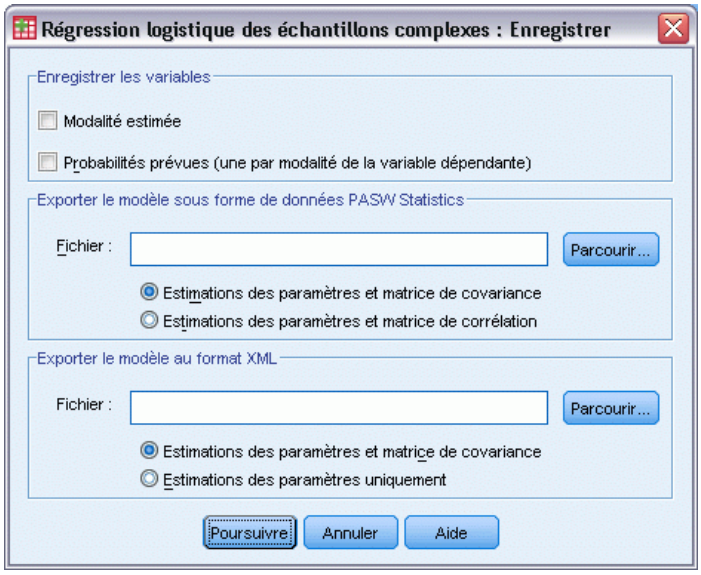
**Facteurs.** Pour chaque facteur sélectionné, affiche l'odds ratio de chaque modalité du facteur par rapport à celui de la modalité de référence indiquée.

**Covariables :** Pour chaque covariable sélectionnée, affiche les odds ratio au niveau de la valeur moyenne de la covariable et des unités de modification indiquées, par rapport à celui de la moyenne.

Lors du calcul des odds ratios d'un facteur ou d'une covariable, la procédure paramètre tous les autres facteurs à leur niveau le plus haut et toutes les autres covariables à leur moyenne. Si un facteur ou une covariable interagit avec les autres variables prédites du modèle, tous les odds ratios dépendent alors non seulement de la modification apportée à la variable indiquée, mais aussi des valeurs des variables avec lesquelles il interagit. Si une covariable spécifiée interagit avec elle-même dans le modèle (par exemple, *âge*âge*), les odds ratios dépendent alors de la modification de la covariable et de la valeur covariable.

# Régression logistique des échantillons complexes - Enregistrement

Figure 10-7
*Boîte de dialogue Enregistrer la régression logistique*




**Enregistrer les variables.** Ce groupe permet d'enregistrer les modalités estimées et les probabilités prévues du modèle en tant que nouvelles variables dans le fichier de travail.

**Exporter le modèle en tant que données SPSS Statistics.** Ecrit un fichier de données dans le format IBM® SPSS® Statisticscontenant la corrélation des paramètres ou la matrice de covariance avec les estimations des paramètres, les erreurs standard, les valeurs de significativité et les degrés de liberté. L'ordre des variables dans le fichier de matrice est le suivant.

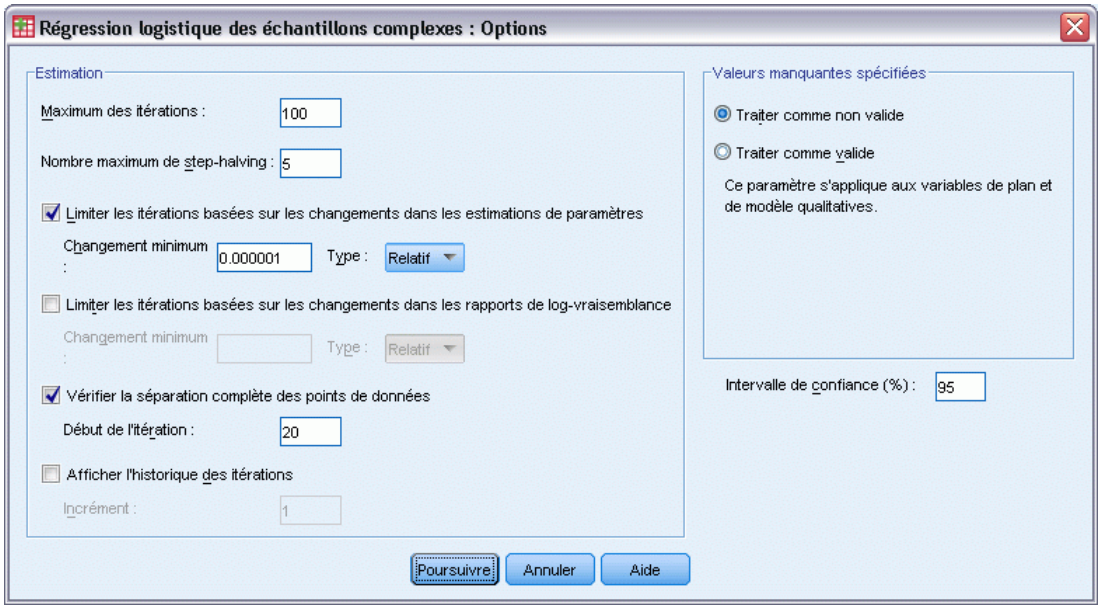
- **rowtype_.** Prend les valeurs (et étiquettes de valeurs) suivantes : COV (covariances), CORR (corrélations), EST (estimations des paramètres), SE (erreurs standard), SIG (seuil de signification) et DF (degrés de liberté du plan d'échantillonnage). Il existe une observation distincte avec le type de ligne COV (ou CORR) pour chaque paramètre de modèle et une observation distincte pour chacun des autres types de ligne.
- **varname_.** Prend les valeurs P1, P2, etc., correspondant à une liste triée de tous les paramètres de modèle pour les types de ligne COV ou CORR, avec des étiquettes de valeur correspondant aux chaînes de paramètres affichées dans le tableau Estimations des paramètres. Les cellules sont vides pour les autres types de ligne.
- **P1, P2, ...** Ces variables correspondent à une liste triée de tous les paramètres de modèle, avec des étiquettes de variable correspondant aux chaînes de paramètres affichées dans le tableau Estimations des paramètres, et prennent leurs valeurs en fonction du type de ligne. Pour les paramètres redondants, toutes les covariances et les estimations de paramètres sont définies sur zéro, et l'ensemble des corrélations, erreurs standard, seuils de signification et degrés de liberté résiduels sont définis sur la valeur manquante par défaut.

*Remarque* : Ce fichier n'est pas immédiatement utilisable pour d'autres analyses dans d'autres procédures que la lecture d'un fichier de matrice, sauf si ces procédures acceptent tous les types de ligne exportés ici.

**Exporter le modèle au format XML.** Enregistre les estimations et la matrice de covariance des paramètres, si vous l'avez sélectionnée, 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.

## Régression logistique des échantillons complexes - Options

Figure 10-8
*Boîte de dialogue Régression logistique : Options*




**Estimation.** Ce groupe permet de contrôler plusieurs critères utilisés dans l'estimation du modèle.

- **Maximum des itérations :** Nombre maximal d'itérations exécutées par l'algorithme. Spécifiez un nombre entier non négatif.
- **Nombre maximum de dichotomies :** A chaque itération, la taille du pas est réduite par un facteur de 0,5 jusqu'à ce que les augmentations de log-vraisemblance ou le nombre maximum de dichotomie soient atteints. Spécifiez un nombre entier positif.
- **Limiter les itérations basées sur les changements dans les estimations de paramètres.** Lorsque cette option est sélectionnée, l'algorithme s'interrompt après une itération dans laquelle la modification relative ou absolue apportée aux estimations de paramètre est inférieure à la valeur spécifiée, qui ne doit pas être négative.
- **Limiter les itérations basées sur les changements dans les rapports de vraisemblance.** Lorsque cette option est sélectionnée, l'algorithme s'interrompt après une itération dans laquelle la modification relative ou absolue apportée à la fonction de log-vraisemblance est inférieure à la valeur spécifiée, qui ne doit pas être négative.

- **Vérifier la séparation complète des points de données.** Lorsque cette option est sélectionnée, l'algorithme effectue les tests permettant de s'assurer que les estimations de paramètre contiennent des valeurs uniques. La séparation s'opère lorsque la procédure peut produire un modèle capable de classer correctement chaque observation.
- **Afficher l'historique des itérations.** Affiche les estimations de paramètre et les statistiques toutes les *n* itérations, en commençant par l'itération 0 (les estimations initiales). Si vous imprimez l'historique des itérations, la dernière itération est toujours imprimée, indépendamment de la valeur de *n*.

**Valeurs manquantes spécifiées.** Toutes les variables de plan, ainsi que la variable dépendante et les covariables, doivent contenir des valeurs valides. Les observations comportant des données non valides pour l'une de ces variables sont supprimées de l'analyse. Vous pouvez ainsi décider si les valeurs manquantes spécifiées par l'utilisateur sont traitées comme étant valides dans la strate, la classe, la sous-population et les variables actives.

**Intervalle de confiance :** Il s'agit du niveau d'intervalle de confiance pour les estimations de coefficient, les estimations de coefficient exponentielles et les odds ratios. Spécifiez une valeur supérieure ou égale à 50 et inférieure à 100.

## Fonctionnalités supplémentaires de la commande CSLOGISTIC

Le langage de syntaxe de commande vous permet aussi de :

- Spécifier les tests d'effets par rapport à une combinaison linéaire d'effets ou une valeur (à l'aide de la sous-commande `CUSTOM`).
- Définir des valeurs pour les autres variables de modèle lors du calcul des odds ratios des facteurs et covariables (à l'aide de la sous-commande `ODDSRATIOS`).
- Indiquer une valeur de tolérance pour le contrôle des particularités (à l'aide de la sous-commande `CRITERIA`).
- Créer des noms définis par l'utilisateur pour les variables enregistrées (à l'aide de la sous-commande `SAVE`).
- Générer un tableau des fonctions générales estimées (à l'aide de la sous-commande `PRINT`).

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

Chapitre
11

# Régression ordinale des échantillons complexes

La procédure de régression ordinale des échantillons complexes effectue une analyse de régression sur une variable dépendante binaire ou ordinale, pour les échantillons réalisés à l'aide de méthodes d'échantillonnage complexes. Vous pouvez également demander une analyse pour une sous-population.

**Exemple :** Des élus étudiant un projet de loi devant l'assemblée législative souhaitent savoir si ce projet est populaire auprès des électeurs et déterminer le lien existant entre cette popularité et la répartition démographique des électeurs. Les enquêteurs conçoivent et mènent des entretiens en fonction d'un plan d'échantillonnage complexe. Avec la procédure de régression ordinale des échantillons complexes, vous pouvez ajuster un modèle concernant la cote de popularité du projet de loi en fonction de la répartition démographique des électeurs.

**Données.** La variable dépendante est ordinale. Les facteurs sont qualitatifs. Les covariables sont des variables quantitatives liées à la variable dépendante. Les variables de sous-population peuvent être du type chaîne de caractères ou numérique, mais ne doivent pas être qualitatives.

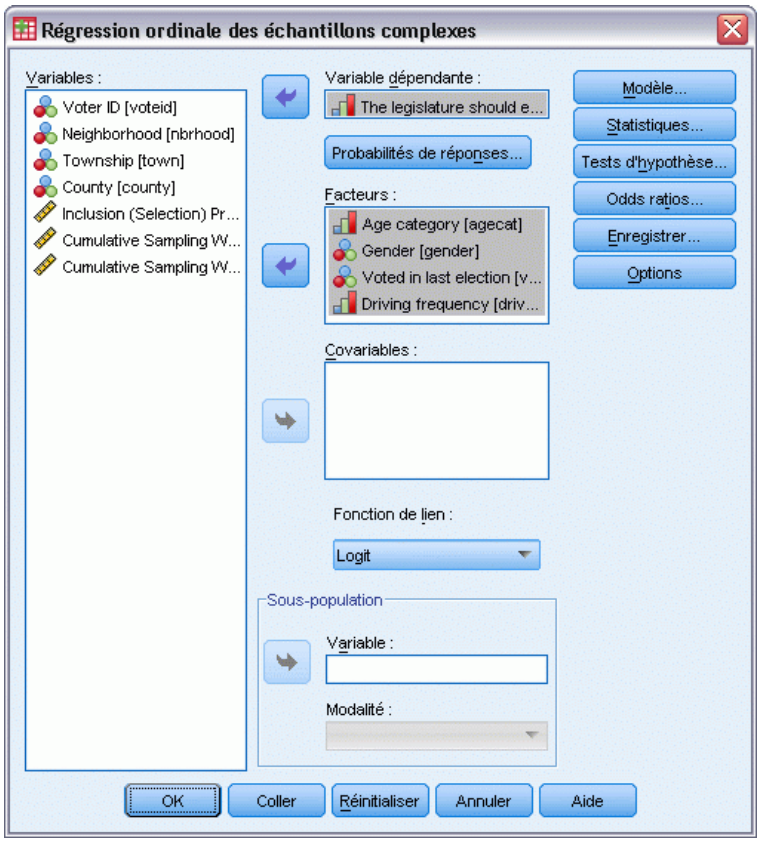
**Hypothèses :** Les observations dans le fichier de données représentent un échantillon provenant d'un plan complexe et devant être analysées en fonction des spécifications du fichier sélectionné dans la Boîte de dialogue Plan d'échantillonnages complexes.

### *Obtention de la régression ordinale des échantillons complexes*

A partir des menus, sélectionnez :
Analyse > Echantillons complexes > Régression ordinale…

- Sélectionnez un fichier de plan. Sélectionnez éventuellement un fichier personnalisé de probabilités conjointes.
- Cliquez sur Poursuivre.

Figure 11-1
*Boîte de dialogue Régression ordinale*




▶ Sélectionnez une variable dépendante.

Sinon, vous pouvez :

- Sélectionnez des variables pour les facteurs et covariables, en fonction de vos données.
- Indiquez une variable pour définir une sous-population. L'analyse n'est effectuée que pour la catégorie sélectionnée de la variable de sous-population, bien que les variances soient correctement estimées sur le fichier de données entier.
- Sélectionnez une fonction de lien.

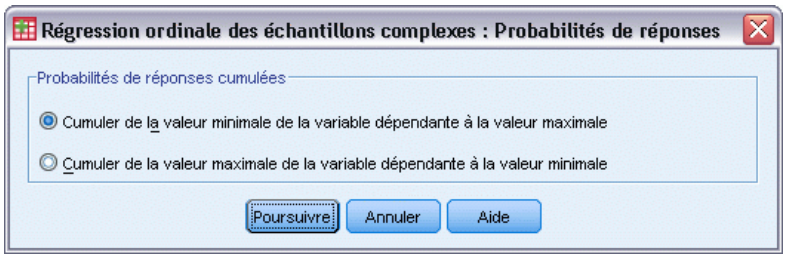
**Fonction de lien.** La fonction de lien consiste en une transformation des probabilités cumulées permettant d'estimer le modèle. Les cinq fonctions de lien disponibles sont récapitulées dans le tableau suivant.

| Fonction | Forme | Application standard |
|---|---|---|
| Logit | $\log( \xi / (1-\xi) )$ | Modalités réparties de façon égale |
| Log-log complémentaire | $\log(-\log(1-\xi))$ | Modalités supérieures les plus probables |
| Log-log négatif | $-\log(-\log(\xi))$ | Modalités inférieures plus probables |

| Fonction | Forme | Application standard |
|---|---|---|
| Probit | $\Phi^{-1}(\xi)$ | Variable de latence normalement répartie |
| Cauchit (Cauchy inverse) | $\tan(\pi(\xi-0,5))$ | Variable de latence avec de nombreux extrema |

## Probabilités des réponses de régression ordinale des échantillons complexes

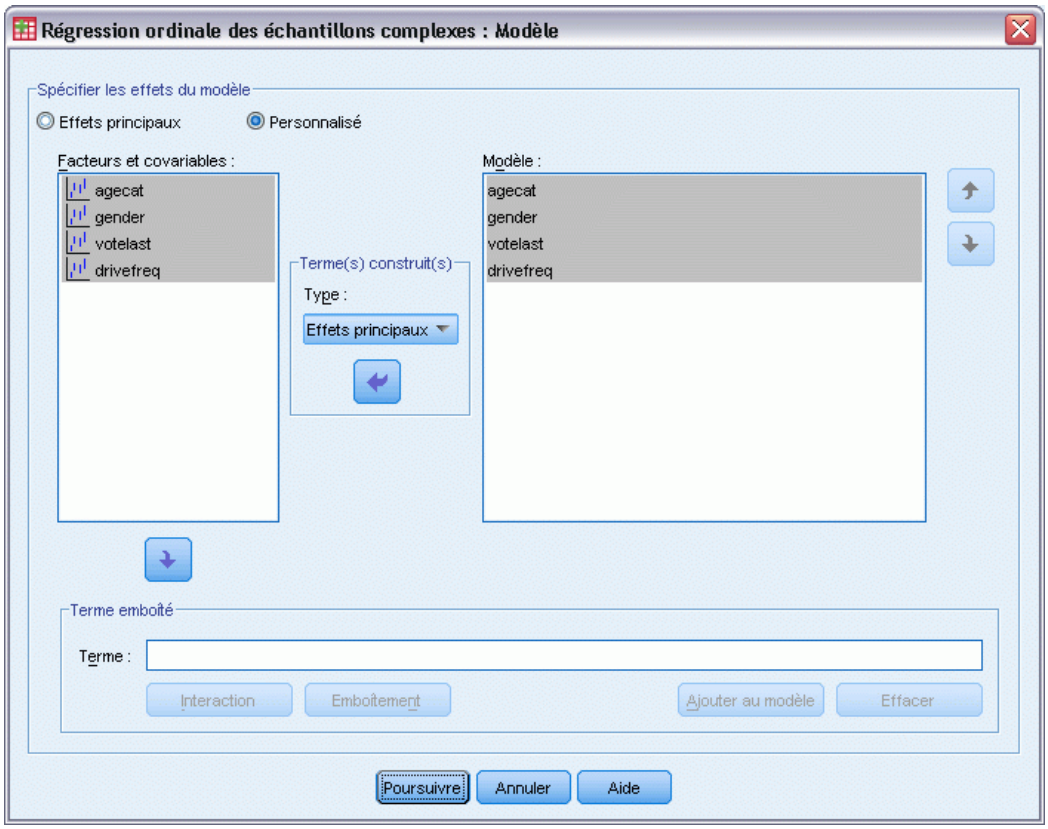
Figure 11-2
*Boîte de dialogue Probabilités de réponse de régression ordinale*




La boîte de dialogue Probabilités de réponses permet d'indiquer si la probabilité cumulée d'une réponse (à savoir la probabilité d'appartenir à une modalité particulière de la variable dépendante et d'inclure cette modalité) augmente avec les valeurs croissantes ou décroissantes de la variable dépendante.

# Modèle de régression ordinale des échantillons complexes

Figure 11-3
*Boîte de dialogue Modèle de régression ordinale*




**Spécifier les effets du modèle.** Par défaut, la procédure élabore un modèle contenant des effets principaux à l'aide des covariables et facteurs spécifiés dans la boîte de dialogue principale. Vous pouvez également créer un modèle personnalisé contenant des effets d'interaction et des termes emboîtés.

### *Termes non emboîtés*

Pour les facteurs et covariables sélectionnés :

**Interaction :** Crée le terme d'interaction du plus haut niveau pour toutes les variables sélectionnées.

**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.

### *Termes emboîtés*

Dans cette procédure, vous pouvez construire des termes emboîtés pour votre modèle. Les termes emboîtés sont utiles pour modéliser l'effet d'un facteur ou d'une covariable dont les valeurs n'interagissent pas avec les niveaux d'un autre facteur. Par exemple, une chaîne d'épiceries peut suivre les habitudes d'achat de ses clients à divers emplacements de magasin. Puisque chaque client ne fréquente qu'un seul de ces magasins, l'effet *Client* peut être considéré comme étant **emboîté dans** l'effet *Emplacement des magasins*.

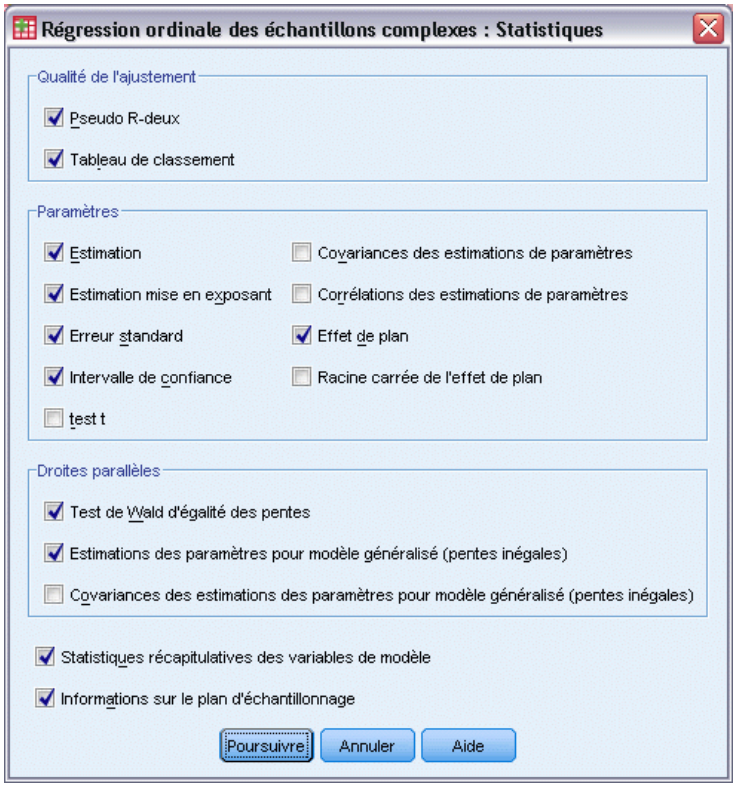
En outre, vous pouvez inclure des effets d'interaction, tels que des termes polynomiaux impliquant la même covariable, ou ajouter plusieurs niveaux d'emboîtement au terme emboîté.

**Limites.** Les termes emboîtés comportent les restrictions suivantes :

- Tous les facteurs d'une interaction doivent être uniques. Ainsi, si *A* est un facteur, la spécification *A*A* n'est pas valide.
- Tous les facteurs d'un effet en cascade doivent être uniques. Ainsi, si *A* est un facteur, la spécification *A(A)* n'est pas valide.
- Aucun effet ne peut être emboîtés dans un effet de covariable. Ainsi, si *A* est un facteur et *X* une covariable, la spécification *A(X)* n'est pas valide.

## *Régression ordinale des échantillons complexes - Statistiques*

Figure 11-4
*Boîte de dialogue Statistiques de régression ordinale*



**Qualité de l'ajustement.** Contrôle l'affichage des statistiques qui mesurent les performances globales du modèle.

- **Pseudo R-deux.** La statistique$R^2$ de la régression linéaire n'a pas de véritable équivalent dans les modèles de régression ordinale. On trouve, à la place, plusieurs mesures qui essaient de reproduire les propriétés de la statistique $R^2$.
- **Tableau de classement :** Affiche les classifications croisées mises en tableau des modalités observées, classées selon les modalités prévues par le modèle en fonction de la variable dépendante.

**Paramètres.** Ce groupe permet de contrôler l'affichage des statistiques associées aux paramètres du modèle.

- **Estimation.** Affiche des estimations des coefficients.
- **Estimation mise en exposant.** Affiche la base du logarithme népérien élevée à la puissance des estimations des coefficients. L'estimation présente des propriétés parfaitement adaptées aux tests statistiques ; l'estimation exponentielle, ou exp(B), est, quant à elle, plus facile à interpréter.
- **Erreur standard :** Affiche l'erreur standard pour chaque estimation de coefficient.
- **Intervalle de confiance :** Affiche l'intervalle de confiance pour chaque estimation de coefficient. Le niveau de confiance de l'intervalle est défini dans la boîte de dialogue Options.
- **Test t.** Affiche un test *t* pour chaque estimation de coefficient. L'hypothèse nulle de chaque test correspond au cas où la valeur du coefficient est 0.
- **Covariances des estimations de paramètres.** Affiche une estimation de la matrice de covariance pour les coefficients du modèle.
- **Corrélations des estimations de paramètres.** Affiche une estimation de la matrice de corrélation pour les coefficients du modèle.
- **Effet de plan :** Rapport entre la variance de l'estimation et la variance, en partant du principe que l'échantillon est un échantillon aléatoire simple. Il s'agit de la mesure d'un effet de la spécification d'un plan complexe, pour lequel des valeurs plus petites indiquent des effets plus importants.
- **Racine carrée de l'effet de plan :** Il s'agit d'une mesure, exprimée en unités comparables à celles de l'erreur standard, de l'effet de spécification d'un plan complexe où les valeurs éloignées de 1 indiquent des effets importants.

**Droites parallèles.** Ce groupe permet de demander des statistiques associées à un modèle avec courbes non parallèles où une courbe de régression distincte est ajustée pour chaque modalité de réponse (sauf la dernière).

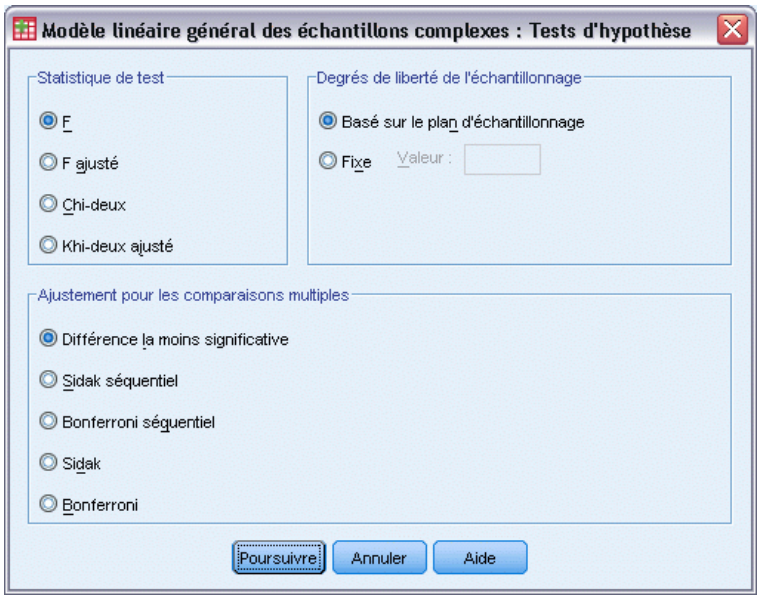
- **Wald.** Produit un test de l'hypothèse nulle selon laquelle les paramètres de régression sont égaux pour toutes les réponses cumulées. Le modèle avec courbes non parallèles est estimé et le test de Wald des paramètres égaux est appliqué.
- **Estimations des paramètres :** Affiche les estimations des coefficients et des erreurs standard du modèle avec courbes non parallèles.
- **Covariances des estimations de paramètres.** Affiche une estimation de la matrice de covariance pour les coefficients du modèle avec courbes non parallèles.

**Statistiques récapitulatives des variables de modèle.** Affiche les informations récapitulatives sur la variable dépendante, les covariables et les facteurs.

**Informations sur le plan d'échantillonnage.** Affiche les informations récapitulatives relatives à l'échantillon, y compris les effectifs non pondérés et la taille de la population.

## *Tests d'hypothèse des échantillons complexes*

Figure 11-5
*Boîte de dialogue Tests d'hypothèse*




**Statistique de test.** Ce groupe vous permet de sélectionner le type de statistiques à utiliser pour tester les hypothèses. Vous avez le choix entre *F*, *F* ajusté, Khi-deux et Khi-deux ajusté.

**Degrés de liberté de l'échantillonnage.** Ce groupe permet de contrôler les degrés de liberté du plan d'échantillonnage utilisés pour calculer les valeurs *p* pour toutes les statistiques de test. Si elle est basée sur le plan d'échantillonnage, cette valeur correspond à la différence entre le nombre d'unités d'échantillonnage principales et le nombre de strates présentes à la première étape de l'échantillonnage. Vous pouvez également définir un degré de liberté personnalisé en indiquant un entier positif.

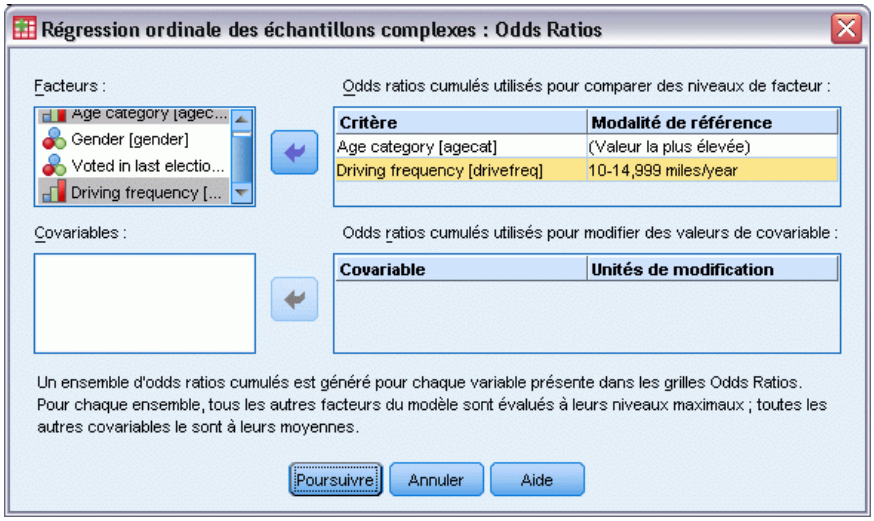
**Ajustement pour les comparaisons multiples.** Lors de l'exécution de tests d'hypothèse avec plusieurs contrastes, vous pouvez ajuster le seuil global de signification à partir des seuils de signification des contrastes inclus. Ce groupe vous permet de choisir la méthode d'ajustement.

- **Différence la moins significative.** Cette méthode ne contrôle pas l'intégralité de la probabilité de rejet des hypothèses qui présentent des contrastes linéaires différents des valeurs d'hypothèse nulles.
- **Procédure de Sidak Séquentielle.** Il s'agit d'une procédure descendante de rejet séquentiel de Bonferroni beaucoup moins stricte en ce qui concerne le rejet des différentes hypothèses mais qui conserve le même niveau global de signification.

- **Bonferroni séquentiel.** Il s'agit d'une procédure descendante de rejet séquentiel de Bonferroni beaucoup moins stricte en ce qui concerne le rejet des différentes hypothèses mais qui conserve le même niveau global de signification.
- **Sidak.** Cette méthode propose des bornes plus petites que l'approche de Bonferroni.
- **Bonferroni.** Cette méthode ajuste le niveau de signification observé lorsque des contrastes multiples sont testés.

## *Régression ordinale des échantillons complexes - Odds ratios*

Figure 11-6
*Boîte de dialogue Odds ratios de régression ordinale*




La boîte de dialogue Odds Ratios permet d'afficher les odds ratios cumulés estimés par modèle pour les covariables et les facteurs indiqués. Cette fonction n'est disponible que pour les modèles utilisant la fonction de lien logit. Un odds ratio cumulé unique est calculé pour toutes les modalités de la variable dépendante, à l'exception de la dernière. Le modèle d'odds proportionnel émet le postulat selon lequel toutes les modalités sont égales.

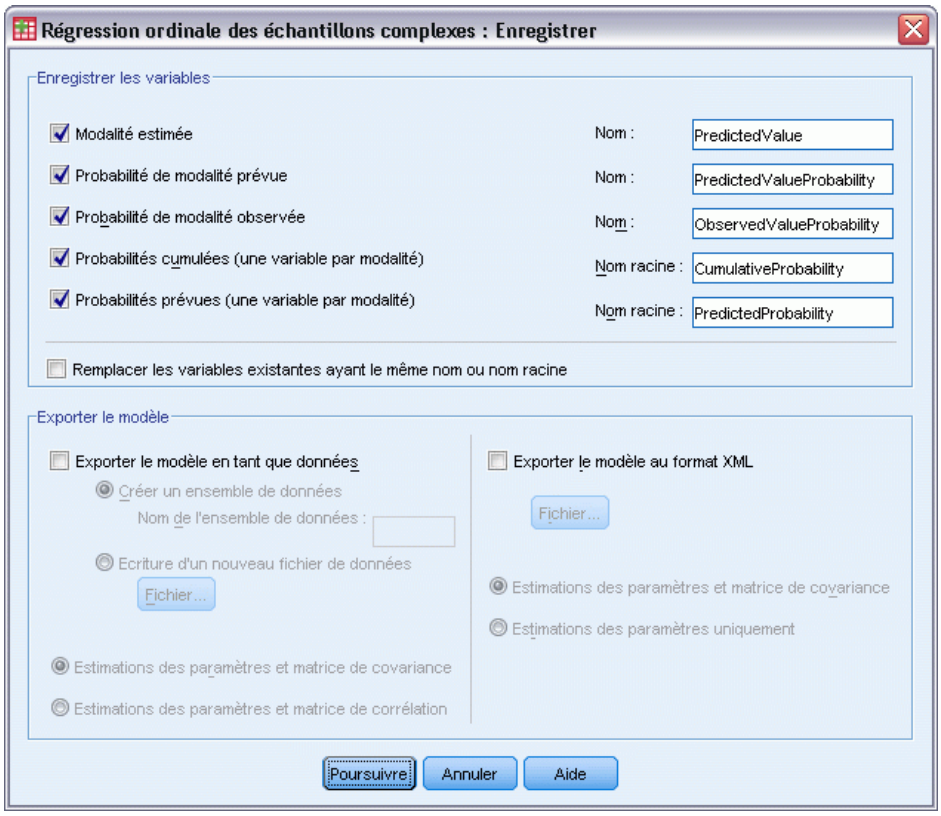
**Facteurs.** Pour chaque facteur sélectionné, affiche l'odds ratio cumulé de chaque modalité du facteur par rapport à celui de la modalité de référence indiquée.

**Covariables :** Pour chaque covariable sélectionnée, affiche les odds ratios cumulés au niveau de la valeur moyenne de la covariable et des unités de modification indiquées, par rapport à celui de la moyenne.

Lors du calcul des odds ratios d'un facteur ou d'une covariable, la procédure paramètre tous les autres facteurs à leur niveau le plus haut et toutes les autres covariables à leur moyenne. Si un facteur ou une covariable interagit avec les autres variables prédites du modèle, tous les odds ratios dépendent alors non seulement de la modification apportée à la variable indiquée, mais aussi des valeurs des variables avec lesquelles il interagit. Si une covariable spécifiée interagit avec elle-même dans le modèle (par exemple, *âge*âge*), les odds ratios dépendent alors de la modification de la covariable et de la valeur covariable.

# Régression ordinale des échantillons complexes - Enregistrement

Figure 11-7
*Boîte de dialogue Enregistrement de régression ordinale*




**Enregistrer les variables.** Ce groupe permet d'enregistrer les probabilités prévues du modèle, la probabilité de modalité prévue, la probabilité de modalité observée, les probabilités cumulées et les probabilités prévues en tant que nouvelles variables dans le fichier de données actif.

**Exporter le modèle en tant que données SPSS Statistics.** Ecrit un fichier de données dans le format IBM® SPSS® Statisticscontenant la corrélation des paramètres ou la matrice de covariance avec les estimations des paramètres, les erreurs standard, les valeurs de significativité et les degrés de liberté. L'ordre des variables dans le fichier de matrice est le suivant.

- **rowtype_.** Prend les valeurs (et étiquettes de valeurs) suivantes : COV (covariances), CORR (corrélations), EST (estimations des paramètres), SE (erreurs standard), SIG (seuil de signification) et DF (degrés de liberté du plan d'échantillonnage). Il existe une observation distincte avec le type de ligne COV (ou CORR) pour chaque paramètre de modèle et une observation distincte pour chacun des autres types de ligne.

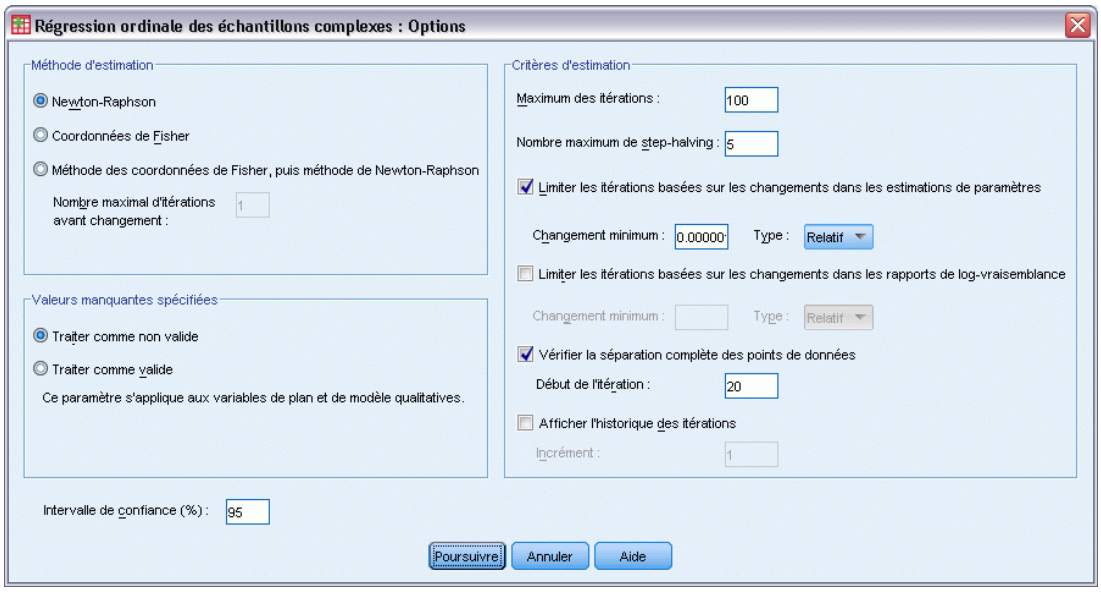
- **varname_.** Prend les valeurs P1, P2, etc., correspondant à une liste triée de tous les paramètres de modèle pour les types de ligne COV ou CORR, avec des étiquettes de valeur correspondant aux chaînes de paramètres affichées dans le tableau Estimations des paramètres. Les cellules sont vides pour les autres types de ligne.
- **P1, P2, ...** Ces variables correspondent à une liste triée de tous les paramètres de modèle, avec des étiquettes de variable correspondant aux chaînes de paramètres affichées dans le tableau Estimations des paramètres, et prennent leurs valeurs en fonction du type de ligne. Pour les paramètres redondants, toutes les covariances et les estimations de paramètres sont définies sur zéro, et l'ensemble des corrélations, erreurs standard, seuils de signification et degrés de liberté résiduels sont définis sur la valeur manquante par défaut.

*Remarque* : Ce fichier n'est pas immédiatement utilisable pour d'autres analyses dans d'autres procédures que la lecture d'un fichier de matrice, sauf si ces procédures acceptent tous les types de ligne exportés ici.

**Exporter le modèle au format XML.** Enregistre les estimations et la matrice de covariance des paramètres, si vous l'avez sélectionnée, 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.

# *Régression ordinale des échantillons complexes - Options*

Figure 11-8
*Boîte de dialogue Régression ordinale : Options*




**Méthode d'estimation.** Vous pouvez sélectionner une méthode d'estimation de paramètre. Vous avez le choix entre la méthode de Newton-Raphson, les coordonnées de Fisher ou une méthode hybride dans laquelle les itérations des coordonnées de Fisher sont effectuées avant le passage à la méthode de Newton-Raphson. En cas de convergence durant la phase des coordonnées de Fisher de la méthode hybride, avant que le nombre maximal d'itérations de Fisher soit atteint, l'algorithme passe à la méthode de Newton-Raphson.

**Estimation.** Ce groupe permet de contrôler plusieurs critères utilisés dans l'estimation du modèle.

- **Maximum des itérations :** Nombre maximal d'itérations exécutées par l'algorithme. Spécifiez un nombre entier non négatif.
- **Nombre maximum de dichotomies :** A chaque itération, la taille du pas est réduite par un facteur de 0,5 jusqu'à ce que les augmentations de log-vraisemblance ou le nombre maximum de dichotomie soient atteints. Spécifiez un nombre entier positif.
- **Limiter les itérations basées sur les changements dans les estimations de paramètres.** Lorsque cette option est sélectionnée, l'algorithme s'interrompt après une itération dans laquelle la modification relative ou absolue apportée aux estimations de paramètre est inférieure à la valeur spécifiée, qui ne doit pas être négative.
- **Limiter les itérations basées sur les changements dans les rapports de vraisemblance.** Lorsque cette option est sélectionnée, l'algorithme s'interrompt après une itération dans laquelle la modification relative ou absolue apportée à la fonction de log-vraisemblance est inférieure à la valeur spécifiée, qui ne doit pas être négative.
- **Vérifier la séparation complète des points de données.** Lorsque cette option est sélectionnée, l'algorithme effectue les tests permettant de s'assurer que les estimations de paramètre contiennent des valeurs uniques. La séparation s'opère lorsque la procédure peut produire un modèle capable de classer correctement chaque observation.
- **Afficher l'historique des itérations.** Affiche les estimations de paramètre et les statistiques toutes les *n* itérations, en commençant par l'itération 0 (les estimations initiales). Si vous imprimez l'historique des itérations, la dernière itération est toujours imprimée, indépendamment de la valeur de *n*.

**Valeurs manquantes spécifiées.** Les variables de plan d'échelle, ainsi que la variable dépendante et les covariables, doivent contenir des valeurs valides. Les observations comportant des données non valides pour l'une de ces variables sont supprimées de l'analyse. Vous pouvez ainsi décider si les valeurs manquantes spécifiées par l'utilisateur sont traitées comme étant valides dans la strate, la classe, la sous-population et les variables actives.

**Intervalle de confiance :** Il s'agit du niveau d'intervalle de confiance pour les estimations de coefficient, les estimations de coefficient exponentielles et les odds ratios. Spécifiez une valeur supérieure ou égale à 50 et inférieure à 100.

## Fonctionnalités supplémentaires de la commande CSORDINAL

Le langage de syntaxe de commande vous permet aussi de :

- Spécifier les tests d'effets par rapport à une combinaison linéaire d'effets ou une valeur (à l'aide de la sous-commande `CUSTOM`).
- Donner aux autres variables de modèle des valeurs autres que leur moyenne lors du calcul des odds ratios cumulés des facteurs et covariables (à l'aide de la sous-commande `ODDSRATIOS`).
- Utiliser des valeurs non étiquetées comme modalités de référence personnalisées pour les facteurs lorsque les odds ratios sont demandés (à l'aide de la sous-commande `ODDSRATIOS`).
- Indiquer une valeur de tolérance pour le contrôle des particularités (à l'aide de la sous-commande `CRITERIA`).

- Générer un tableau des fonctions générales estimées (à l'aide de la sous-commande `PRINT`).
- Enregistrer plus de 25 variables de probabilité (à l'aide de la sous-commande `SAVE`).

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

Chapitre
12

# Régression de Cox des échantillons complexes

La procédure de la régression de Cox des échantillons complexes effectue une analyse de survie pour les échantillons réalisés à l'aide de méthodes d'échantillonnage complexes. Vous pouvez également demander une analyse pour une sous-population.

**Exemples :** Une administration chargée de l'application de la loi s'inquiète des taux de récidive dans sa juridiction. L'une des mesures de récidive est le temps qui s'écoule avant la deuxième arrestation des délinquants. L'agence souhaite modéliser le temps s'écoulant jusqu'à la deuxième arrestation à l'aide d'une régression de Cox, mais craint que l'hypothèse des hasards proportionnels ne soit pas valide sur l'ensemble des tranches d'âge.

Des chercheurs en médecine étudient les temps de survie des patients qui quittent un programme de rééducation à la suite d'un accident ischémique. Il est possible d'avoir plusieurs observations par sujet, étant donné que les antécédents des patients changent à mesure que des événements non mortels significatifs sont constatés et que l'heure de ces événements est enregistrée. L'échantillon est également tronqué à gauche dans le sens où les temps de survie observés sont « augmentés » par la durée de la rééducation. En effet, si le risque commence au moment de l'accident ischémique, seuls les patients survivant au programme de rééducation sont inclus dans l'échantillon.

**Temps de survie.** La procédure applique la régression de Cox à l'analyse des temps de survie—c'est-à-dire le temps qui s'écoule avant qu'un événement se produise. Il existe deux façons d'indiquer le temps de survie, en fonction de l'heure de début de l'intervalle :

- **Temps = 0.** En général, vous aurez des informations complètes concernant le début de l'intervalle pour chaque sujet et vous aurez simplement une variable contenant les heures de fin (ou créerez une variable unique avec des heures de fin à partir des variables de date et d'heure ; reportez-vous aux paragraphes ci-dessous).
- **Varie en fonction du sujet.** Cette solution est adaptée lorsque vous avez une **troncation à gauche**, également appelée **entrée différée**. Par exemple, si vous analysez les temps de survie de patients quittant un programme de rééducation suite à une attaque, vous pouvez considérer que le risque commence au moment de l'attaque. Cependant, si votre échantillon ne comprend que des patients ayant survécu au programme de rééducation, il est tronqué à gauche car les temps de survie observés sont « augmentés » par la durée de la rééducation. Pour représenter cette information, indiquez l'heure à laquelle ils ont quitté le programme de rééducation comme heure d'entrée dans l'étude.

**Variables de date et d'heure.** Vous ne pouvez pas utiliser les variables de date et d'heure pour définir directement le début et la fin de l'intervalle. Si vous disposez de variables de ce type, vous devez les utiliser pour créer des variables contenant des temps de survie. En l'absence de troncation à gauche, il suffit de créer une variable contenant des heures de fin, basée sur la différence entre la date d'entrée dans l'étude et la date d'observation. En présence d'une troncation à gauche, créez une variable contenant des heures de début, basée sur la différence entre

la date de début de l'étude et la date d'entrée, et une variable contenant des heures de fin, basée sur la différence entre la date de début de l'étude et la date d'observation.

**Etat des événements.** Vous avez besoin d'une variable pour enregistrer si le sujet a vécu l'événement étudié dans l'intervalle. Les sujets qui n'ont pas connu cet événement sont censurés à droite.

**Identificateur de sujet.** Vous pouvez facilement incorporer des variables prédites chronologiques de constante « par morceau » en scindant les observations d'un sujet sur plusieurs observations. Par exemple, si vous analysez les temps de survie des patients suite à l'attaque, les variables représentant leurs antécédents médicaux doivent être utiles en tant que variables prédites. Au fil du temps, ils peuvent vivre des événements médicaux majeurs modifiant leurs antécédents médicaux. Le tableau ci-dessous montre comment structurer un tel ensemble de données : *ID patient* est l'identificateur de sujet, *Fin du programme* définit les intervalles observés, *Etat* enregistre les événements médicaux majeurs, et *Antécédents d'attaques cardiaques* et *Antécédents d'hémorragies* sont des variables prédites chronologiques de constante « par morceau ».

| *ID patient* | *Fin du programme* | *Statut* | *Antécédents d'attaques cardiaques* | *Antécédents d'hémorragies* |
|---|---|---|---|---|
| 1 | 5 | Attaque cardiaque | Non | Non |
| 1 | 7 | Hémorragie | Oui | Non |
| 1 | 8 | Décédé | Oui | Oui |
| 2 | 24 | Décédé | Non | Non |
| 3 | 8 | Attaque cardiaque | Non | Non |
| 3 | 15 | Décédé | Oui | Non |

**Hypothèses :** Les observations dans le fichier de données représentent un échantillon provenant d'un plan complexe et devant être analysées en fonction des spécifications du fichier sélectionné dans la Boîte de dialogue Plan d'échantillonnages complexes.

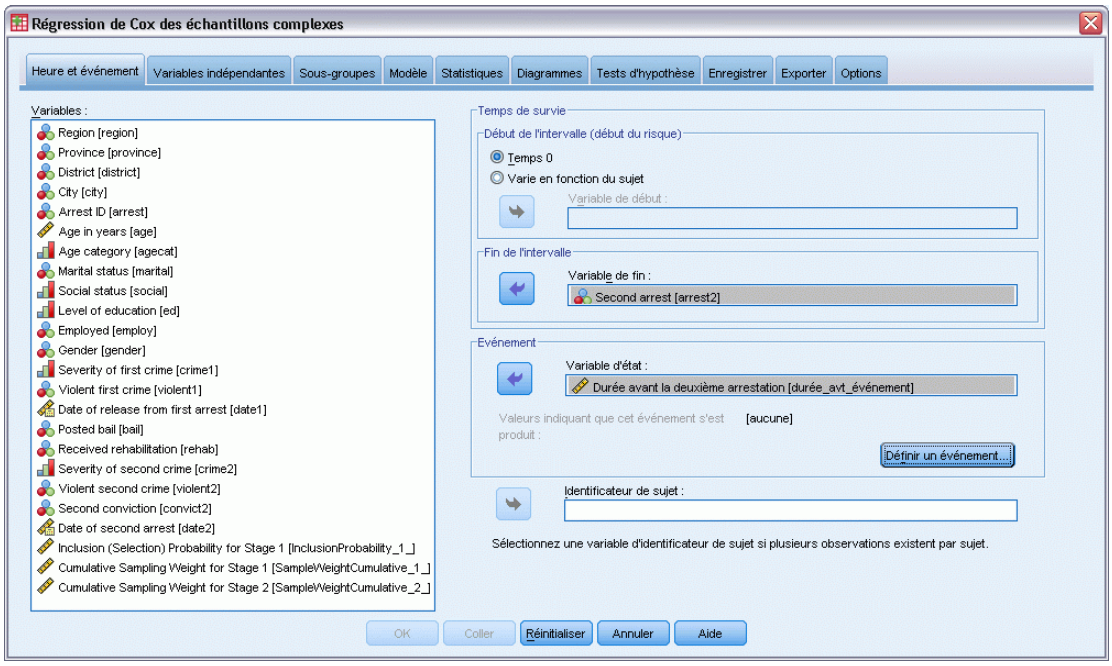
En général, les modèles de régression de Cox partent du principe que les hasards sont proportionnels, c'est-à-dire que le rapport des hasards d'une observation à une autre ne doit pas varier dans le temps. Si cette hypothèse n'est pas valable, il peut être nécessaire d'ajouter au modèle des variables prédites chronologiques.

**Analyse Kaplan-Meier.** Si vous ne sélectionnez aucune variable prédite (ou que vous n'entrez aucune variable prédite sélectionnée dans le modèle) et que vous choisissez la méthode de limite du produit pour calculer la courbe de survie de base dans l'onglet Options, la procédure effectue une analyse de survie de type Kaplan-Meier.

### *Pour obtenir une régression de Cox des échantillons complexes*

- A partir des menus, sélectionnez :
  Analyse > Echantillons complexes > Modèle de Cox
- Sélectionnez un fichier de plan. Sélectionnez éventuellement un fichier personnalisé de probabilités conjointes.
- Cliquez sur Poursuivre.

Figure 12-1
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Heure et événement*

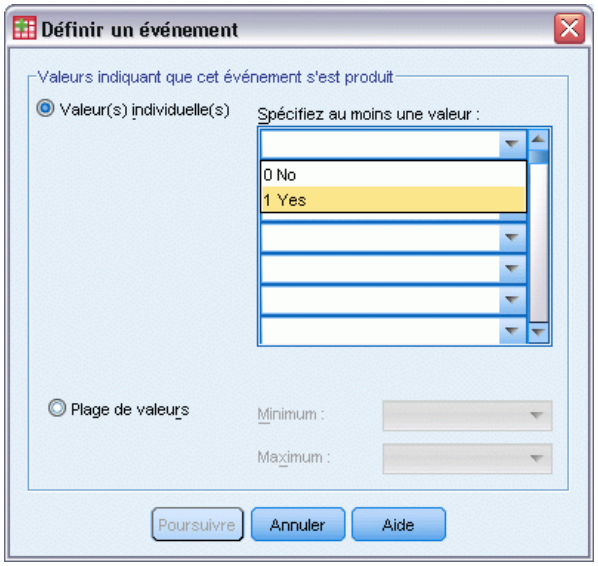



- Spécifiez le temps de survie en sélectionnant les heures d'entrée et de sortie de l'étude.
- Sélectionnez une variable d'état de l'événement.
- Cliquez sur Définir l'événement et définissez au moins une valeur d'événement.

  Vous pouvez également sélectionner un identificateur de sujet.

# *Définir l'événement*

Figure 12-2
*Boîte de dialogue Définir un événement*

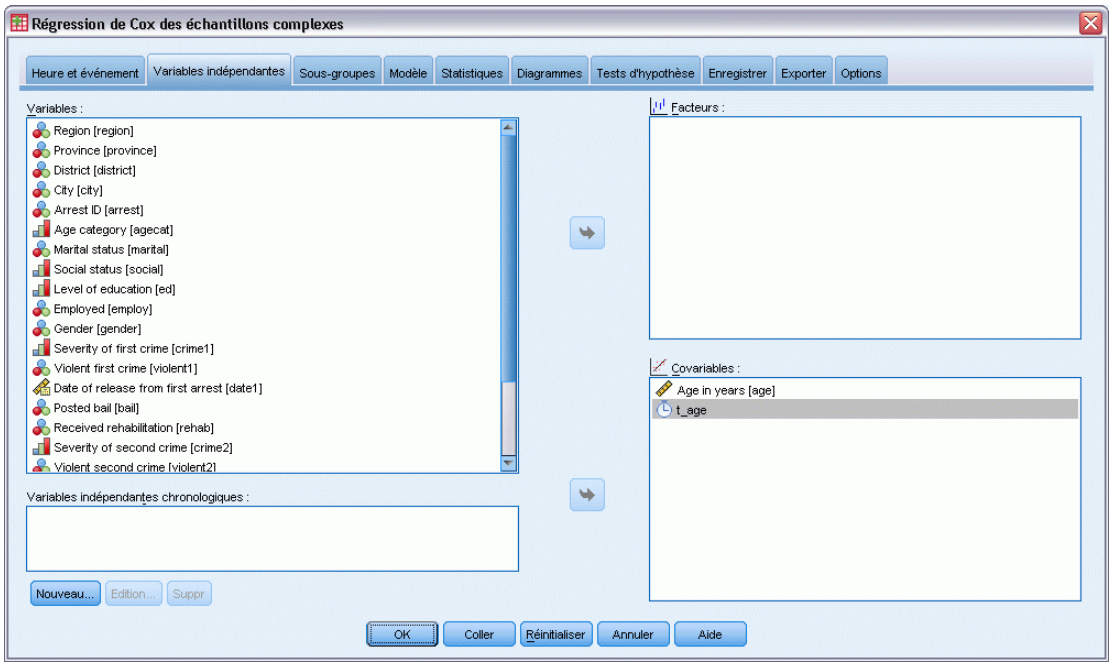



Spécifiez les valeurs indiquant qu'un événement final s'est produit.

- **Valeur(s) individuelle(s).** Spécifiez une ou plusieurs valeurs en les entrant dans la grille ou en les sélectionnant à partir d'une liste de valeurs contenant des étiquettes de valeur définies.
- **Plage de valeurs.** Spécifiez une plage de valeurs en entrant les valeurs minimum et maximum, ou en sélectionnant des valeurs à partir d'une liste contenant des étiquettes de valeur définies.

# Variables prédites

Figure 12-3
*Boîte de dialogue Régression de Cox, onglet Variables prédites*




L'onglet Variables prédites vous permet d'indiquer les covariables et facteurs utilisés pour élaborer des effets de modèle.

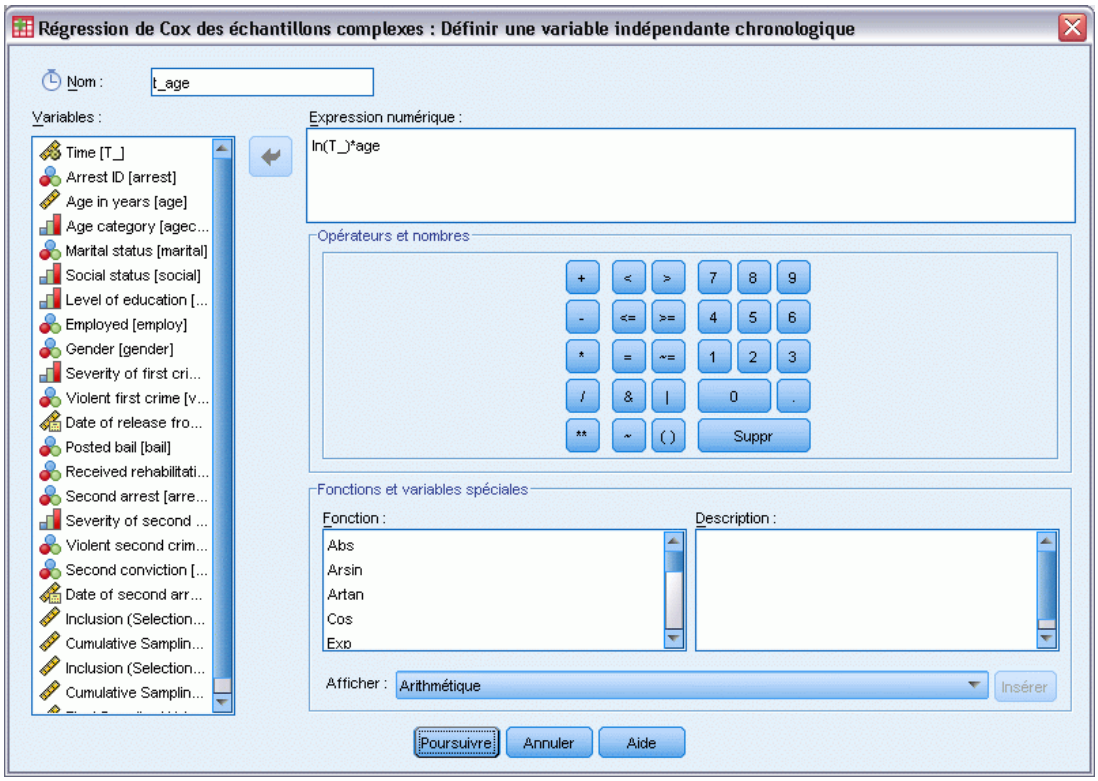
**Facteurs.** Les facteurs sont des variables prédites catégorielles de type numérique ou chaîne.

**Covariables :** Les covariables sont des variables prédites d'échelle ; elles doivent être numériques.

**Variables prédites chronologiques.** Dans certains cas, l'hypothèse des hasards proportionnels n'est pas valable. Les taux de probabilité varient dans le temps. Les valeurs de l'une ou plusieurs de vos variables prédites sont différentes à différentes dates. Dans de tels cas, vous devez spécifier des variables prédites chronologiques. Pour plus d'informations, reportez-vous à la section Définir une variable prédite chronologique sur p. 85. Les variables prédites chronologiques peuvent être sélectionnées en tant que facteurs ou covariables.

## *Définir une variable prédite chronologique*

Figure 12-4
*Boîte de dialogue Régression de Cox - Définir une variable prédite chronologique*




La boîte de dialogue Définir une variable prédite chronologique vous permet de créer une variable prédite dépendant de la variable de temps intégrée *T_*. Vous pouvez l'utiliser pour définir des prédicteurs chronologiques de deux façons :

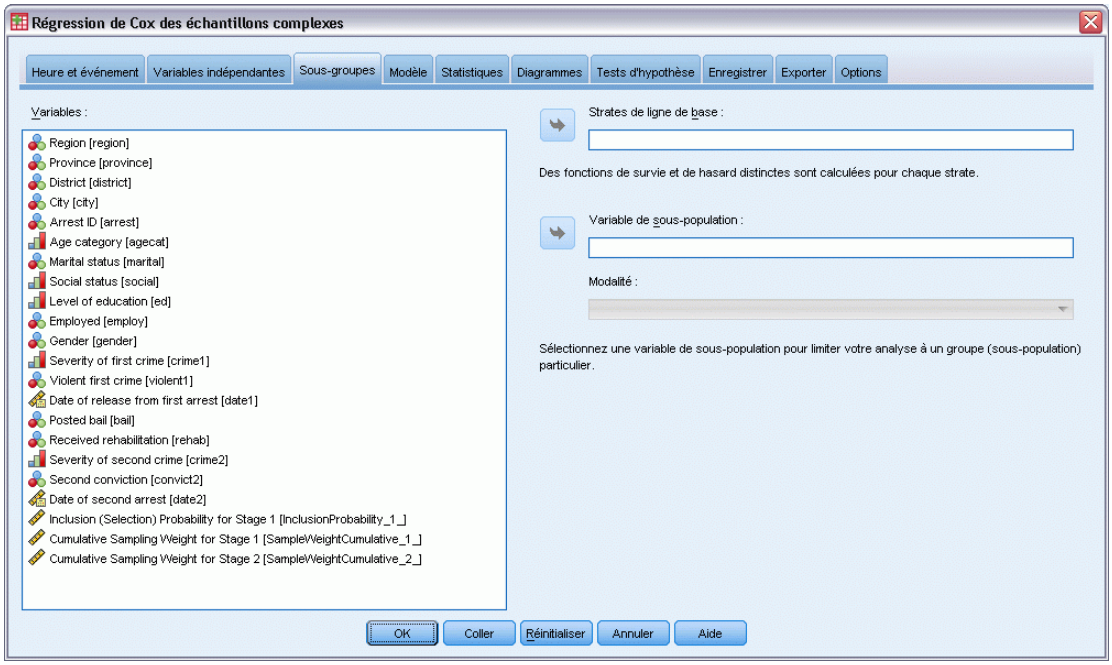
- Si vous voulez estimer un modèle de régression de Cox étendue autorisant des hasards non proportionnels, vous pouvez définir votre variable prédite chronologique sous forme de fonction de la variable de temps *T_* et de la covariable en question. Exemple : le simple produit de la variable de temps et de la variable prédite. Vous pouvez également définir des fonctions plus complexes.
- Certaines variables peuvent avoir différentes valeurs à des périodes différentes sans pour autant être liées au temps (chronologiques). Dans ce cas, vous devez définir une **variable prédite chronologique segmentée** à l'aide d'une expression logique. Les expressions logiques prennent la valeur 1 si elles sont vraies, 0 si elles sont fausses. A l'aide d'une série d'expressions logiques, vous pouvez créer votre variable prédite chronologique à partir d'un ensemble de mesures. Par exemple, si votre pression artérielle est mesurée une fois par semaine pendant les quatre semaines de votre étude, (mesures identifiées par *PA1* to *PA4*), vous pouvez définir votre valeur prédite chronologique sous la forme (*T_* < 1) * *PA1* + (*T_* >= 1 & *T_* < 2) * *PA2* + (*T_* >= 2 & *T_* < 3) * *PA3* + (*T_* >= 3 & *T_* < 4) * *PA4*. Notez qu'un seul des termes entre parenthèses est égal à 1 pour chaque cas, tandis que les autres termes sont égaux à 0. Cette fonction peut être interprétée ainsi : « Si le temps est inférieur à une semaine, utilisez *PA1*. S'il est supérieur à une semaine mais inférieur à deux, utilisez *PA2*, et ainsi de suite. »

*Remarque :* Si votre variable prédite chronologique segmentée est constante dans les segments, comme dans l'exemple de pression artérielle ci-dessus, il vous sera plus facile d'indiquer la variable prédite chronologique de constante « par morceau » en scindant les sujets sur plusieurs observations. Reportez-vous à la section consacrée aux Identificateurs de sujet dans Régression de Cox des échantillons complexes sur p. 80 pour plus d'informations.

Dans la boîte de dialogue Définir une variable prédite chronologique, vous pouvez utiliser des commandes de construction de fonction pour construire l'expression pour le prédicteur chronologique ou bien vous pouvez la saisir directement dans la zone de texte Expression numérique. Notez que les constantes alphanumériques doivent être saisies entre guillemets ou apostrophes, tandis que les constantes numériques doivent être en format Américain avec un point en tant que séparateur décimal. Le nom attribué à la variable est celui que vous spécifiez. Cette variable doit être incluse comme facteur ou covariable dans l'onglet Variables prédites.

## Sous-groupes

Figure 12-5
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Sous-groupes*

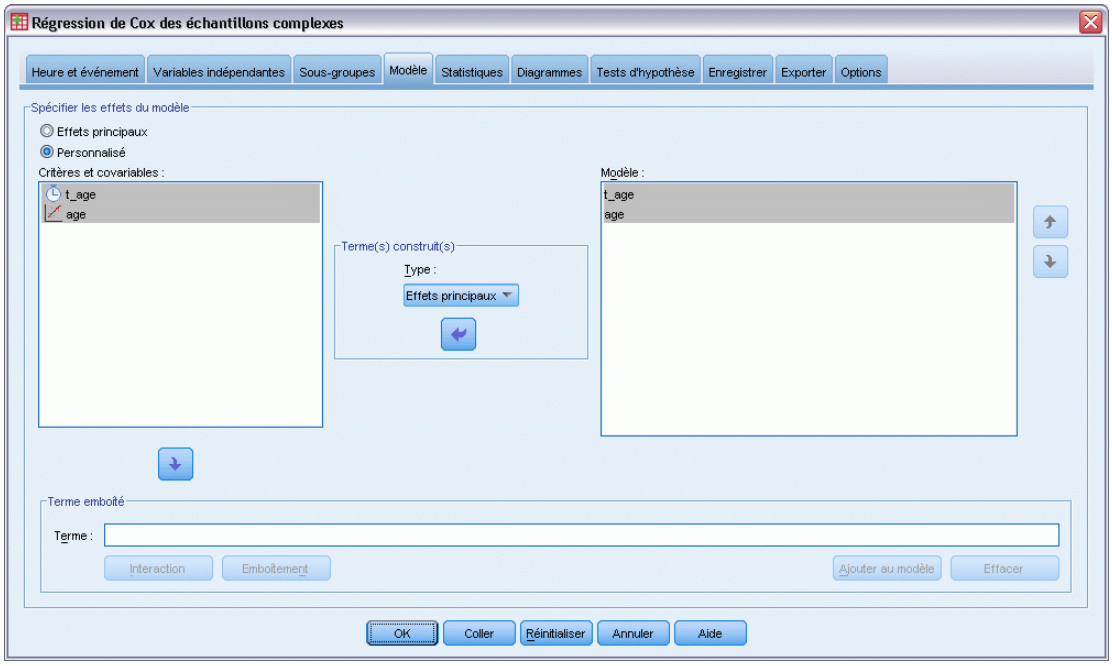



**Strates de ligne de base.** Une fonction de base de hasard et de survie distincte est calculée pour chaque valeur de cette variable, alors qu'un ensemble unique de coefficients du modèle est estimé sur l'ensemble des strates.

**Variable de sous-population.**Indiquez une variable pour définir une sous-population. L'analyse est effectuée uniquement pour la modalité sélectionnée de la variable de sous-population.

# Modèle

Figure 12-6
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Modèle///boîte de dialogue*




**Spécifier les effets du modèle.** Par défaut, la procédure élabore un modèle contenant des effets principaux à l'aide des covariables et facteurs spécifiés dans la boîte de dialogue principale. Vous pouvez également créer un modèle personnalisé contenant des effets d'interaction et des termes emboîtés.

### *Termes non emboîtés*

Pour les facteurs et covariables sélectionnés :

**Interaction :** Crée le terme d'interaction du plus haut niveau pour toutes les variables sélectionnées.

**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.

### *Termes emboîtés*

Dans cette procédure, vous pouvez construire des termes emboîtés pour votre modèle. Les termes emboîtés sont utiles pour modéliser l'effet d'un facteur ou d'une covariable dont les valeurs n'interagissent pas avec les niveaux d'un autre facteur. Par exemple, une chaîne d'épiceries peut suivre les habitudes d'achat de ses clients à divers emplacements de magasin. Puisque chaque

client ne fréquente qu'un seul de ces magasins, l'effet *Client* peut être considéré comme étant **emboîté dans** l'effet *Emplacement des magasins*.

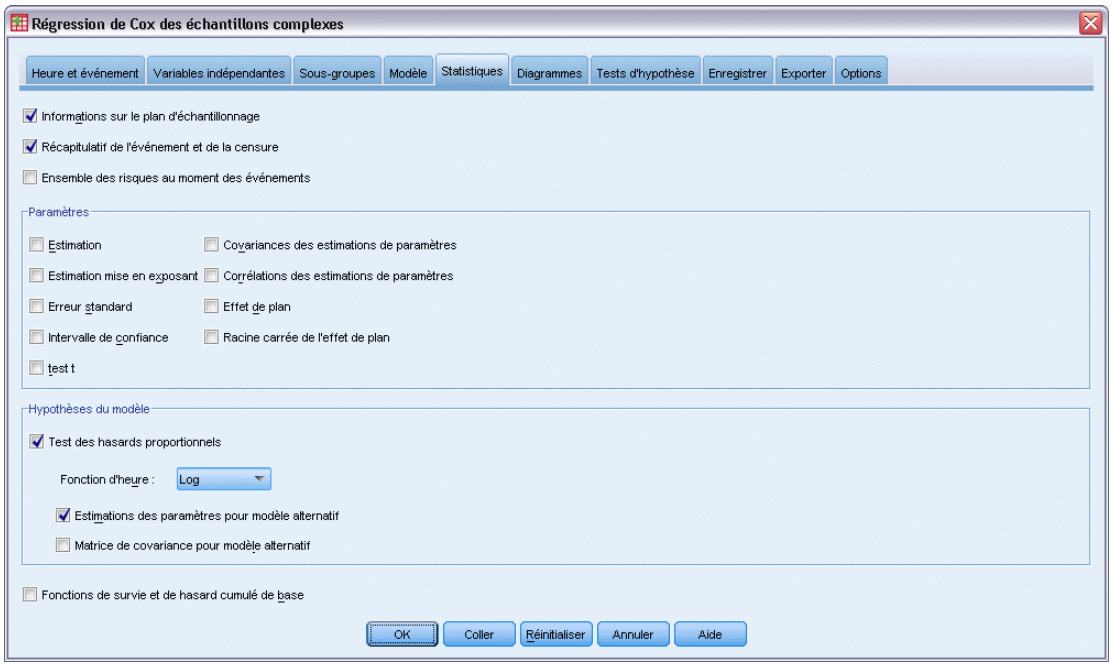
En outre, vous pouvez inclure des effets d'interaction, tels que des termes polynomiaux impliquant la même covariable, ou ajouter plusieurs niveaux d'emboîtement au terme emboîté.

**Limites.** Les termes emboîtés comportent les restrictions suivantes :

- Tous les facteurs d'une interaction doivent être uniques. Ainsi, si *A* est un facteur, la spécification *A*A* n'est pas valide.
- Tous les facteurs d'un effet en cascade doivent être uniques. Ainsi, si *A* est un facteur, la spécification *A(A)* n'est pas valide.
- Aucun effet ne peut être emboîtés dans un effet de covariable. Ainsi, si *A* est un facteur et *X* une covariable, la spécification *A(X)* n'est pas valide.

## Statistiques

Figure 12-7
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Statistiques*




**Informations sur le plan d'échantillonnage.** Affiche les informations récapitulatives relatives à l'échantillon, y compris les effectifs non pondérés et la taille de la population.

**Récapitulatif de l'événement et de la censure.** Affiche des informations récapitulatives relatives au nombre et au pourcentage d'observations censurées.

**Ensemble des risques au moment des événements.** Affiche le nombre d'événements et le nombre à risque au moment de chaque événement dans chacune des strates de ligne de base.

**Paramètres.** Ce groupe permet de contrôler l'affichage des statistiques associées aux paramètres du modèle.

- **Estimation.** Affiche des estimations des coefficients.
- **Estimation mise en exposant.** Affiche la base du logarithme népérien élevée à la puissance des estimations des coefficients. L'estimation présente des propriétés parfaitement adaptées aux tests statistiques ; l'estimation exponentielle, ou exp(B), est, quant à elle, plus facile à interpréter.
- **Erreur standard :** Affiche l'erreur standard pour chaque estimation de coefficient.
- **Intervalle de confiance :** Affiche l'intervalle de confiance pour chaque estimation de coefficient. Le niveau de confiance de l'intervalle est défini dans la boîte de dialogue Options.
- **Test t :** Affiche un test *t* pour chaque estimation de coefficient. L'hypothèse nulle de chaque test correspond au cas où la valeur du coefficient est 0.
- **Covariances des estimations de paramètres.** Affiche une estimation de la matrice de covariance pour les coefficients du modèle.
- **Corrélations des estimations de paramètres.** Affiche une estimation de la matrice de corrélation pour les coefficients du modèle.
- **Effet de plan :** Rapport entre la variance de l'estimation et la variance, en partant du principe que l'échantillon est un échantillon aléatoire simple. Il s'agit de la mesure d'un effet de la spécification d'un plan complexe, pour lequel des valeurs plus petites indiquent des effets plus importants.
- **Racine carrée de l'effet de plan :** Il s'agit d'une mesure de l'effet de spécification d'un plan complexe, où les valeurs éloignées de 1 indiquent des effets importants.

**Hypothèses du modèle.** Ce groupe vous permet de générer un test de l'hypothèse des hasards proportionnels. Le test compare le modèle ajusté à un modèle alternatif incluant des variables prédites chronologiques *x**_TF* pour chaque variable prédite *x*, *_TF* étant la fonction d'heure indiquée.

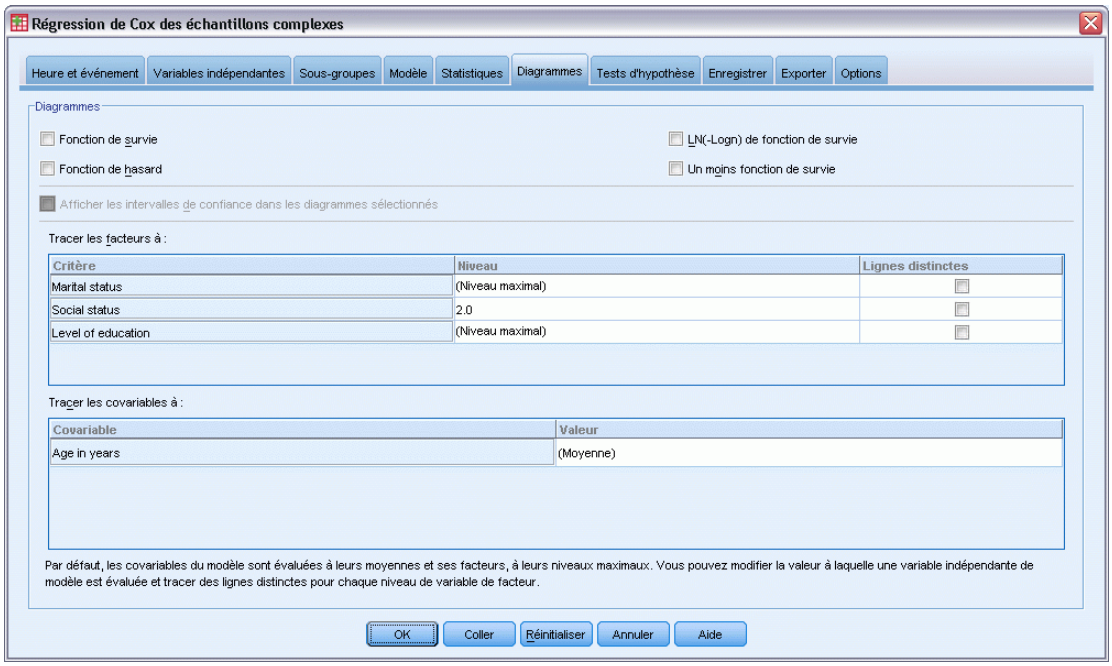
- **Fonction d'heure.** Indique la forme de *_TF* pour le modèle alternatif. Pour la fonction d'**identité**, *_TF*=*T_*. Pour la fonction **log**, *_TF*=log(*T_*). Pour **Kaplan-Meier**, *_TF*=1−$S_{KM}$(*T_*), où $S_{KM}$(.) est l'estimation de Kaplan-Meier de la fonction de survie. Pour le **rang**, *_TF* est le rang de *T_* parmi les heures de fin observées.
- **Estimations des paramètres pour modèle alternatif.** Affiche l'estimation, l'erreur standard et l'intervalle de confiance pour chaque paramètre du modèle alternatif.
- **Matrice de covariance pour modèle alternatif.** Affiche la matrice de covariances estimées entre les paramètres du modèle alternatif.

**Fonctions de survie et de hasard cumulé de base.** Affiche la fonction de survie de base et la fonction de hasards cumulés de base, ainsi que leurs erreurs standard.

*Remarque* : Si des variables prédites chronologiques définies dans l'onglet Variables prédites sont incluses dans le modèle, cette option n'est pas disponible.

# Diagrammes

Figure 12-8
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Diagrammes*



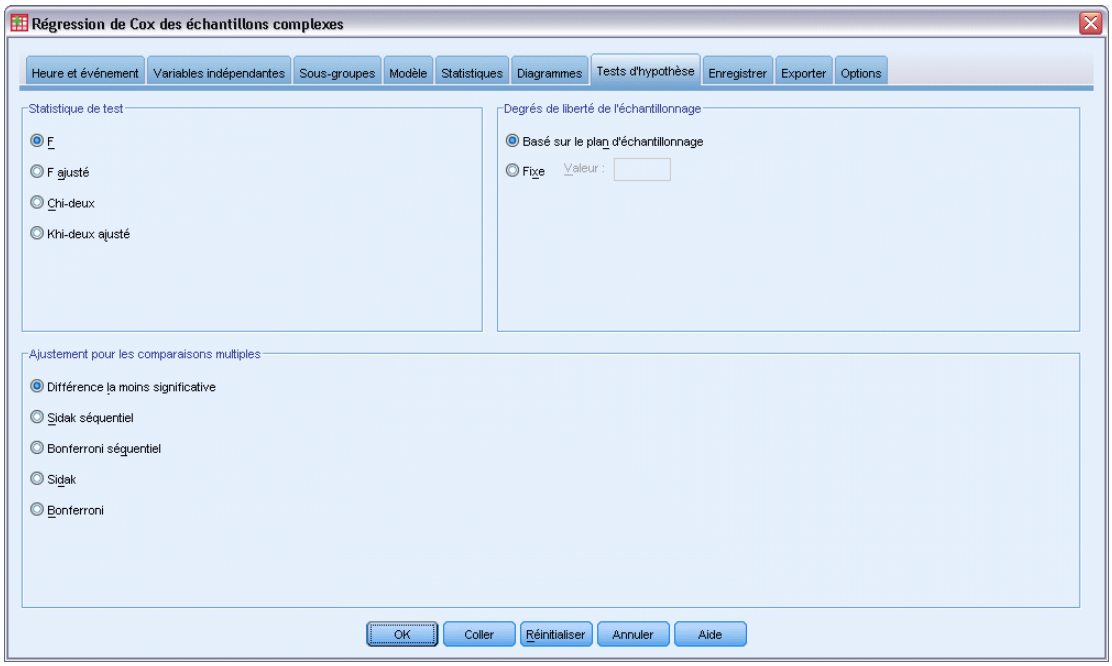

L'onglet Diagrammes vous permet de demander des diagrammes des fonctions de hasard, de survie, LN (-Logn) de la fonction de survie et un moins survie. Vous pouvez également choisir de tracer les intervalles de confiance le long des fonctions spécifiées. Le niveau de confiance est défini dans l'onglet Options.

**Modèles de variables prédites.** Vous pouvez indiquer un modèle de valeurs de variable prédite à utiliser pour les diagrammes demandés et le fichier de survie exporté dans l'onglet Exporter. Notez que ces options ne sont pas disponibles si les variables prédites chronologiques définies dans l'onglet Variables prédites sont incluses dans le modèle.

- **Tracer les facteurs à.** Par défaut, chaque facteur est évalué à son niveau maximum. Entrez ou sélectionnez un autre niveau, si vous le souhaitez. Vous pouvez également choisir de tracer des courbes distinctes pour chaque niveau d'un facteur unique en cochant la case de ce facteur.
- **Tracer les covariables à.** Chaque covariable est évaluée à sa moyenne. Entrez ou sélectionnez une autre valeur, si vous le souhaitez.

# Tests d'hypothèse

Figure 12-9
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Tests d'hypothèse*




**Statistique de test.** Ce groupe vous permet de sélectionner le type de statistiques à utiliser pour tester les hypothèses. Vous avez le choix entre *F*, *F* ajusté, Khi-deux et Khi-deux ajusté.

**Degrés de liberté de l'échantillonnage.** Ce groupe permet de contrôler les degrés de liberté du plan d'échantillonnage utilisés pour calculer les valeurs *p* pour toutes les statistiques de test. Si elle est basée sur le plan d'échantillonnage, cette valeur correspond à la différence entre le nombre d'unités d'échantillonnage principales et le nombre de strates présentes à la première étape de l'échantillonnage. Vous pouvez également définir un degré de liberté personnalisé en indiquant un entier positif.

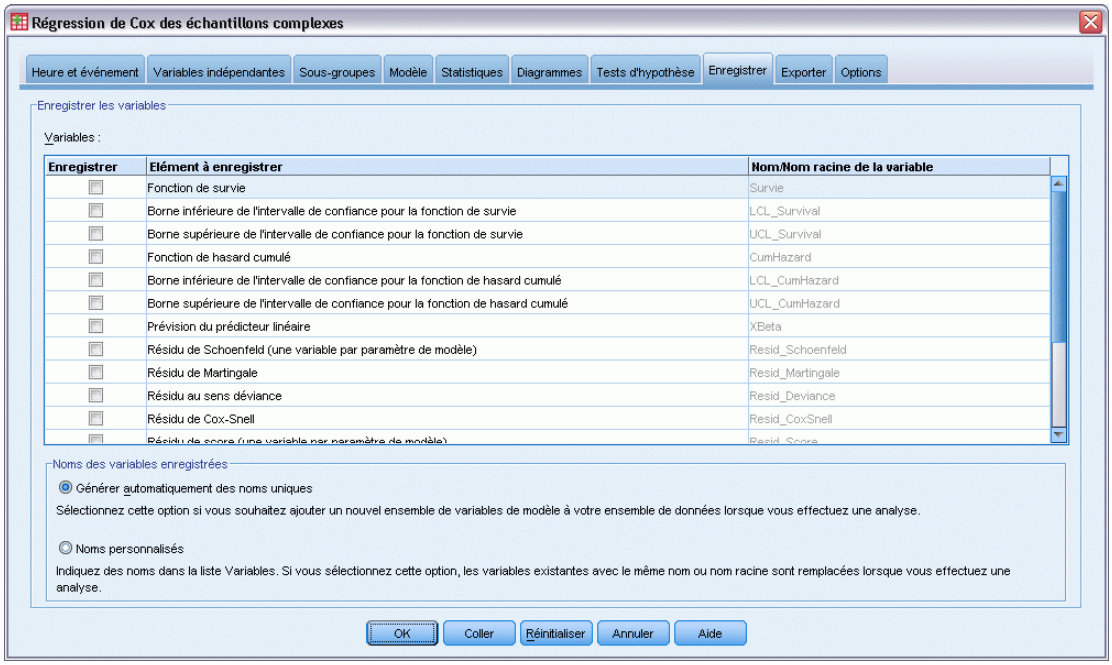
**Ajustement pour les comparaisons multiples.** Lors de l'exécution de tests d'hypothèse avec plusieurs contrastes, vous pouvez ajuster le seuil global de signification à partir des seuils de signification des contrastes inclus. Ce groupe vous permet de choisir la méthode d'ajustement.

- **Différence la moins significative.** Cette méthode ne contrôle pas l'intégralité de la probabilité de rejet des hypothèses qui présentent des contrastes linéaires différents des valeurs d'hypothèse nulles.
- **Procédure de Sidak Séquentielle.** Il s'agit d'une procédure descendante de rejet séquentiel de Bonferroni beaucoup moins stricte en ce qui concerne le rejet des différentes hypothèses mais qui conserve le même niveau global de signification.
- **Bonferroni séquentiel.** Il s'agit d'une procédure descendante de rejet séquentiel de Bonferroni beaucoup moins stricte en ce qui concerne le rejet des différentes hypothèses mais qui conserve le même niveau global de signification.

- **Sidak.** Cette méthode propose des bornes plus petites que l'approche de Bonferroni.
- **Bonferroni.** Cette méthode ajuste le niveau de signification observé lorsque des contrastes multiples sont testés.

# Enregistrer

Figure 12-10
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Enregistrer*




**Enregistrer les variables.** Ce groupe vous permet d'enregistrer les variables liées au modèle dans l'ensemble de données actif pour les utiliser ultérieurement dans les diagnostics et les rapports des résultats. Notez qu'elles ne sont pas disponibles quand des variables prédites chronologiques sont incluses dans le modèle.

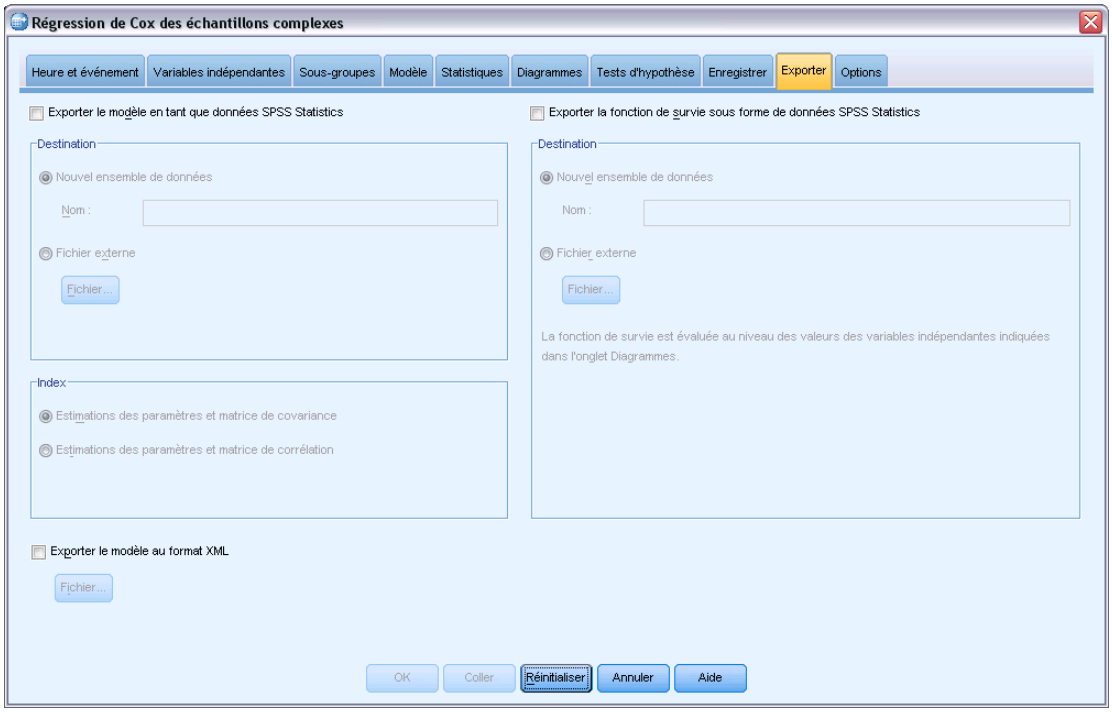
- **Fonction de survie.** Enregistre la probabilité de survie (valeur de la fonction de survie) au moment observé et les valeurs des variables prédites pour chaque observation.
- **Borne inférieure de l'intervalle de confiance pour la fonction de survie.** Enregistre la borne inférieure de l'intervalle de confiance pour la fonction de survie au moment observé et les valeurs des variables prédites pour chaque observation.
- **Borne supérieure de l'intervalle de confiance pour la fonction de survie.** Enregistre la borne supérieure de l'intervalle de confiance pour la fonction de survie au moment observé et les valeurs des variables prédites pour chaque observation.
- **Fonction de hasard cumulé.** Enregistre le hasard cumulé, ou −ln(survie), au moment observé et les valeurs des variables prédites pour chaque observation.
- **Borne inférieure de l'intervalle de confiance pour la fonction de hasard cumulé.** Enregistre la borne inférieure de l'intervalle de confiance pour la fonction de hasard cumulé au moment observé et les valeurs des variables prédites pour chaque observation.

- **Borne supérieure de l'intervalle de confiance pour la fonction de hasard cumulé.** Enregistre la borne supérieure de l'intervalle de confiance pour la fonction de hasard cumulé au moment observé et les valeurs des variables prédites pour chaque observation.
- **Prévision de la variable prédite linéaire.** Enregistre la combinaison linéaire des coefficients de régression pour les heures des variables prédites corrigées de la valeur de référence. La variable prédite linéaire correspond au rapport entre la fonction de hasard et de hasard de base. Dans le modèle des hasards proportionnels, cette valeur est constante dans le temps.
- **Résidu de Schoenfeld.** Pour chaque observation non censurée et chaque paramètre non redondant du modèle, le résidu de Schoenfeld correspond à la différence entre la valeur observée de la variable prédite associée au paramètre du modèle et la valeur théorique de la variable prédite pour les observations de l'ensemble des risques au moment où les événements sont observés. Les résidus de Schoenfeld peuvent être utilisés pour évaluer l'hypothèse des hasards proportionnels. Par exemple, pour une variable prédite *x*, les diagrammes des résidus de Schoenfeld pour la variable prédite chronologique $x*\ln(T\_)$ par rapport au temps doivent montrer une ligne horizontale au niveau de 0 si les hasards proportionnels sont valables. Une variable distincte est enregistrée pour chaque paramètre non redondant du modèle. Les résidus de Schoenfeld sont uniquement calculés pour les observations non censurées.
- **Résidu de Martingale.** Pour chaque observation, le résidu de Martingale correspond à la différence entre la censure observée (0 si elle est censurée, 1 si ce n'est pas le cas) et la prévision d'un événement pendant le temps d'observation.
- **Résidus au sens déviance.** Les résidus au sens déviance sont des résidus de Martingale « ajustés » pour paraître plus symétriques au niveau de 0. Les diagrammes des résidus au sens déviance par rapport aux variables prédites ne doivent révéler aucun modèle.
- **Résidu de Cox-Snell.** Pour chaque observation, le résidu de Cox-Snell correspond à la prévision d'un événement pendant la période d'observation, ou la censure observée moins le résidu de Martingale.
- **Résidu de score.** Pour chaque observation et chaque paramètre non redondant du modèle, le résidu de score est la contribution de l'observation à la première dérivée de la pseudo-vraisemblance. Une variable distincte est enregistrée pour chaque paramètre non redondant du modèle.
- **Résidu de la différence de Bêta.** Pour chaque observation et chaque paramètre non redondant du modèle, le résidu de la différence de Bêta se rapproche du changement de la valeur de l'estimation des paramètres lorsque l'observation est supprimée du modèle. Il est possible que les observations ayant des résidus de la différence de Bêta relativement élevés aient un effet d'annulation sur l'analyse. Une variable distincte est enregistrée pour chaque paramètre non redondant du modèle.
- **Résidus agrégés.** Quand plusieurs observations représentent un sujet unique, le résidu agrégé d'un sujet est simplement la somme des résidus de l'observation correspondante sur l'ensemble des observations appartenant au même sujet. Pour le résidu de Schoenfeld, la version agrégée est semblable à celle de la version non agrégée parce que le résidu de Schoenfeld est défini uniquement pour les observations non censurées. Ces résidus ne sont disponibles que si un identificateur de sujet est spécifié dans l'onglet Heure et événement.

**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.

# Exporter

Figure 12-11
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Exporter*




**Exporter le modèle en tant que données SPSS Statistics.** Ecrit un fichier de données dans le format IBM® SPSS® Statisticscontenant la corrélation des paramètres ou la matrice de covariance avec les estimations des paramètres, les erreurs standard, les valeurs de significativité et les degrés de liberté. L'ordre des variables dans le fichier de matrice est le suivant.

- **rowtype_.** Prend les valeurs (et étiquettes de valeurs) suivantes : COV (covariances), CORR (corrélations), EST (estimations des paramètres), SE (erreurs standard), SIG (seuil de signification) et DF (degrés de liberté du plan d'échantillonnage). Il existe une observation distincte avec le type de ligne COV (ou CORR) pour chaque paramètre de modèle et une observation distincte pour chacun des autres types de ligne.
- **varname_.** Prend les valeurs P1, P2, etc., correspondant à une liste triée de tous les paramètres de modèle pour les types de ligne COV ou CORR, avec des étiquettes de valeur correspondant aux chaînes de paramètres affichées dans le tableau Estimations des paramètres. Les cellules sont vides pour les autres types de ligne.
- **P1, P2, ...** Ces variables correspondent à une liste triée de tous les paramètres de modèle, avec des étiquettes de variable correspondant aux chaînes de paramètres affichées dans le tableau Estimations des paramètres, et prennent leurs valeurs en fonction du type de ligne. Pour les

paramètres redondants, toutes les covariances et les estimations de paramètres sont définies sur zéro, et l'ensemble des corrélations, erreurs standard, seuils de signification et degrés de liberté résiduels sont définis sur la valeur manquante par défaut.

*Remarque* : Ce fichier n'est pas immédiatement utilisable pour d'autres analyses dans d'autres procédures que la lecture d'un fichier de matrice, sauf si ces procédures acceptent tous les types de ligne exportés ici.

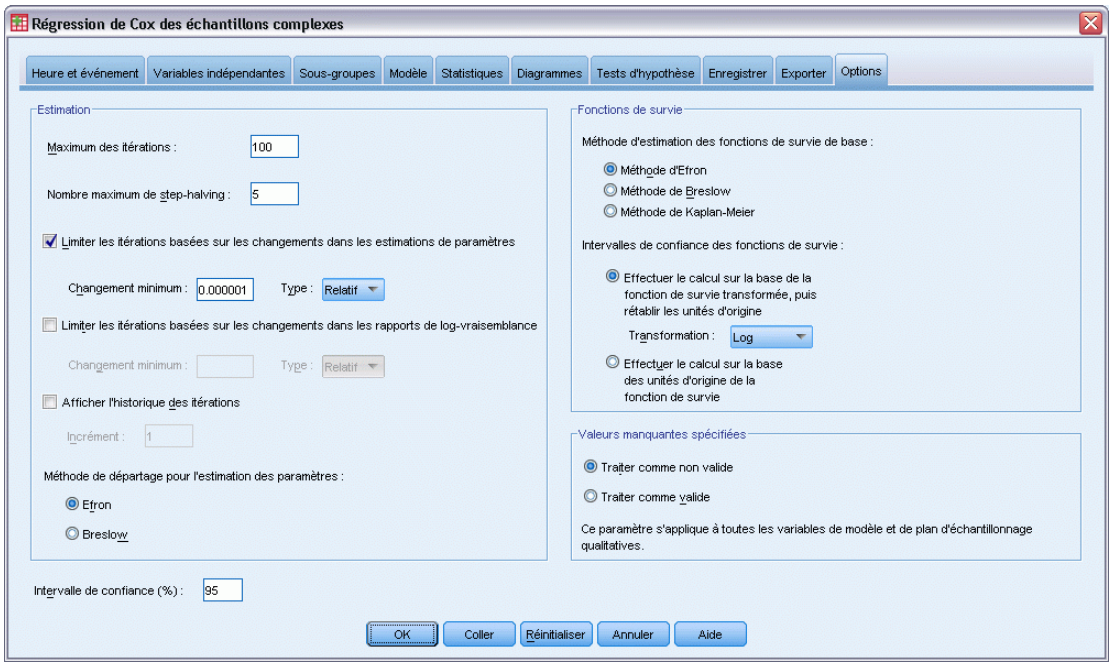
**Exporter la fonction de survie sous forme de données SPSS Statistics.** Ecrit un ensemble de données dans un format SPSS Statistics contenant la fonction de survie, l'erreur standard de la fonction de survie, les bornes supérieure et inférieure de l'intervalle de confiance de la fonction de survie, et la fonction des hasards cumulés pour chaque échec ou moment de l'événement, évaluées à la ligne de base et aux modèles de variables prédites spécifiées dans l'onglet Diagramme. L'ordre des variables dans le fichier de matrice est le suivant.

- **Variable de strate de ligne de base.** Des tableaux de survie distincts sont générés pour chaque valeur de la variable de strate.
- **Variable de temps de survie.** Moment de l'événement. Une observation distincte est créée à chaque moment de l'événement.
- **Sur_0, LCL_Sur_0, UCL_Sur_0.** Fonction de survie de base et bornes supérieure et inférieure de son intervalle de confiance.
- **Sur_R, LCL_Sur_R, UCL_Sur_R.** Fonction de survie évaluée au modèle de « référence » (consultez le tableau des valeurs de modèle dans les résultats) et les bornes supérieure et inférieure de son intervalle de confiance.
- **Sur_#.#, LCL_Sur_#.#, UCL_Sur_#.#, …** Fonction de survie évaluée à chacun des modèles de variables prédites spécifiés dans l'onglet Diagrammes et les bornes supérieure et inférieure de leurs intervalles de confiance. Consultez le tableau des valeurs de modèle dans les résultats pour faire correspondre les modèles au nombre #.#.
- **Haz_0, LCL_Haz_0, UCL_Haz_0.** Fonction de base de hasard cumulé et les bornes supérieure et inférieure de son intervalle de confiance.
- **Haz_R, LCL_Haz_R, UCL_Haz_R.** Fonction de hasard cumulé évaluée au modèle de « référence » (consultez le tableau des valeurs de modèle dans les résultats) et les bornes supérieure et inférieure de son intervalle de confiance.
- **Haz_#.#, LCL_Haz_#.#, UCL_Haz_#.#, …** Fonction de hasard cumulé évaluée à chacun des modèles de variables prédites spécifiés dans l'onglet Diagrammes et les bornes supérieure et inférieure de leurs intervalles de confiance. Consultez le tableau des valeurs de modèle dans les résultats pour faire correspondre les modèles au nombre #.#.

**Exporter le modèle au format XML.** Enregistre toutes les informations nécessaires pour prévoir la fonction de survie, y compris les estimations des paramètres et la fonction de survie de base, 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.

# *Options*

Figure 12-12
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Options*




**Estimation.** Ces commandes spécifient des critères pour l'estimation des coefficients de régression.

- **Maximum des itérations :** Nombre maximal d'itérations exécutées par l'algorithme. Spécifiez un nombre entier non négatif.
- **Nombre maximum de dichotomies :** A chaque itération, la taille du pas est réduite par un facteur de 0,5 jusqu'à ce que les augmentations de log-vraisemblance ou le nombre maximum de dichotomie soient atteints. Spécifiez un nombre entier positif.
- **Limiter les itérations basées sur les changements dans les estimations de paramètres.** Lorsque cette option est sélectionnée, l'algorithme s'interrompt après une itération dans laquelle la modification relative ou absolue apportée aux estimations de paramètre est inférieure à la valeur spécifiée, qui doit être positive.
- **Limiter les itérations basées sur les changements dans les rapports de vraisemblance.** Lorsque cette option est sélectionnée, l'algorithme s'interrompt après une itération dans laquelle la modification relative ou absolue apportée à la fonction de log-vraisemblance est inférieure à la valeur spécifiée, qui doit être positive.
- **Afficher l'historique des itérations.** Affiche l'historique des itérations pour les estimations des paramètres et la pseudo-log-vraisemblance, et imprime la dernière évaluation du changement des estimations des paramètres et de la pseudo-log-vraisemblance. Le tableau de l'historique des itérations imprime toutes les *n* itérations en commençant par l'itération *0* (les estimations

initiales), *n* étant la valeur de l'incrément. Si l'historique des itérations est requis, la dernière itération est toujours affichée quel que soit *n*.

- **Méthode de départage pour l'estimation des paramètres.** Lorsque vous observez des heures d'échec ex aequo, l'une de ces méthodes est utilisée pour départager les ex aequo. La méthode d'Efron nécessite plus de calculs.

**Fonctions de survie.** Ces commandes spécifient des critères pour les calculs impliquant la fonction de survie.

- **Méthode d'estimation des fonctions de survie de base.** La méthode de **Breslow** (ou de Nelson-Aalan ou empirique) estime le hasard cumulé de la fonction de base par une fonction d'échelon non décroissante avec des échelons aux heures d'échec observées, puis calcule la survie de base par la relation survie=exp(−hasard cumulé). La méthode d'**Efron** nécessite plus de calculs et se limite à la méthode de Breslow quand il n'y a pas d'ex aequo. La méthode de **limite du produit** estime la survie de base par une fonction continue à droite non croissante. S'il n'y a pas de variables prédites dans le modèle, cette méthode se limite à l'estimation de Kaplan-Meier.
- **Intervalles de confiance des fonctions de survie.** L'intervalle de confiance peut être calculé de trois manières : dans les unités d'origine, via une transformation log ou une transformation LN (-Logn). Seule la transformation LN (-Logn) garantit que les bornes de l'intervalle de confiance seront comprises entre 0 et 1, mais la transformation log semble généralement donner de « meilleurs » résultats.

**Valeurs manquantes spécifiées.** Toutes les variables doivent avoir des valeurs valides pour qu'une observation puisse être incluse dans l'analyse. Ces commandes vous permettent de décider si les valeurs manquantes spécifiées sont considérées comme valides parmi les modèles qualitatifs (y compris les variables qualitatives, d'événement, de strate et de sous-population) et les variables du plan d'échantillonnage.

**Intervalle de confiance (%).** Il s'agit du niveau d'intervalle de confiance utilisé pour les estimations de coefficient, les estimations de coefficient exponentielles, les estimations de fonction de survie et les estimations de fonction de hasard cumulé. Spécifiez une valeur supérieure ou égale à 0, et inférieure à 100.

## Fonctionnalités supplémentaires de la commande CSCOXREG

Le langage de commande vous permet aussi de :

- Effectuer des tests d'hypothèse personnalisés (à l'aide de la sous-commande `CUSTOM` et de `/PRINT LMATRIX`).
- Spécification de la tolérance (à l'aide de `/CRITERIA SINGULAR`).
- Tableau de la fonction générale estimée (à l'aide de `/PRINT GEF`).
- Modèles de variables prédites multiples (à l'aide de plusieurs sous-commandes `PATTERN`).
- Nombre maximum de variables enregistrées quand un nom de racine est spécifié (à l'aide de la sous-commande `SAVE`). La boîte de dialogue prend en compte la valeur `CSCOXREG` par défaut de 25 variables.

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

# *Partie II: Exemples*

Chapitre
13

# Assistant d'échantillonnage des échantillons complexes

L'assistant d'échantillonnage vous guide lors de la création, de la modification ou de l'exécution d'un fichier de plan d'échantillonnage. Avant d'utiliser l'assistant, pensez à définir précisément la population cible, la liste des unités d'échantillonnage et le plan d'échantillonnage.

## Obtention d'un échantillon à partir d'un cadre d'échantillonnage complet

Une agence d'état est chargée d'assurer l'application d'impôts fonciers équitables dans chaque comté. Les impôts étant calculés à partir de la valeur estimée de la propriété, l'agence souhaite évaluer un échantillon de propriétés dans chaque comté afin de s'assurer que les enregistrements de chaque département sont mis à jour uniformément. Cependant, les ressources permettant d'obtenir des évaluations actuelles étant limitées, il est important d'utiliser judicieusement les informations disponibles. L'agence décide d'utiliser la méthode de l'échantillonnage complexe pour sélectionner un échantillon de propriétés.

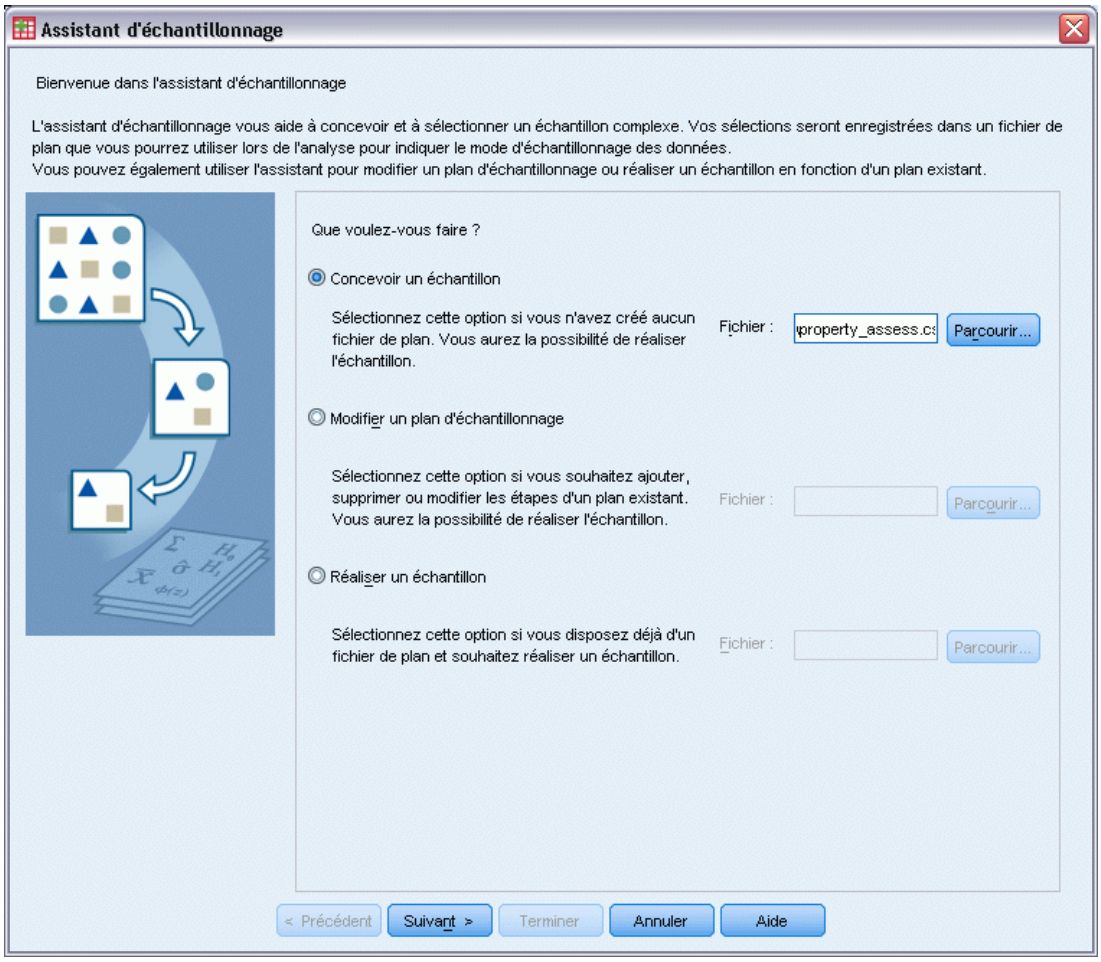
Les propriétés sont regroupées dans le fichier *property_assess_cs.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*. Utilisez l'assistant d'échantillonnage des échantillons complexes pour sélectionner un échantillon.

### Utilisation de l'assistant

- Pour exécuter l'assistant d'échantillonnage des échantillons complexes, sélectionnez dans le menu l'option suivante :

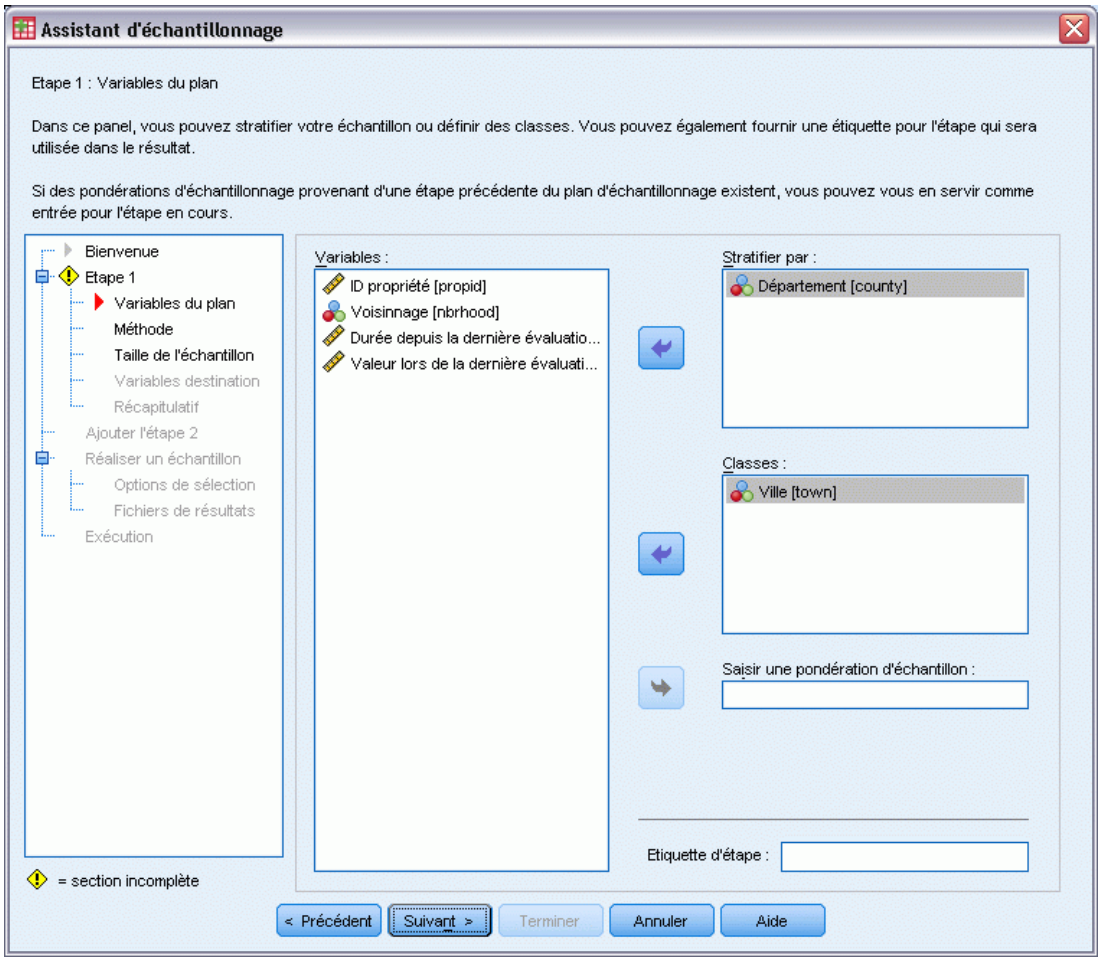
  Analyse > Echantillons complexes > Sélectionner un échantillon...

Figure 13-1
*Etape Bienvenue de l'assistant d'échantillonnage*




- Sélectionnez Concevoir un échantillon, accédez à l'emplacement auquel vous souhaitez enregistrer le fichier, et saisissez property_assess.csplan comme nom du fichier de plan.
- Cliquez sur Suivant.

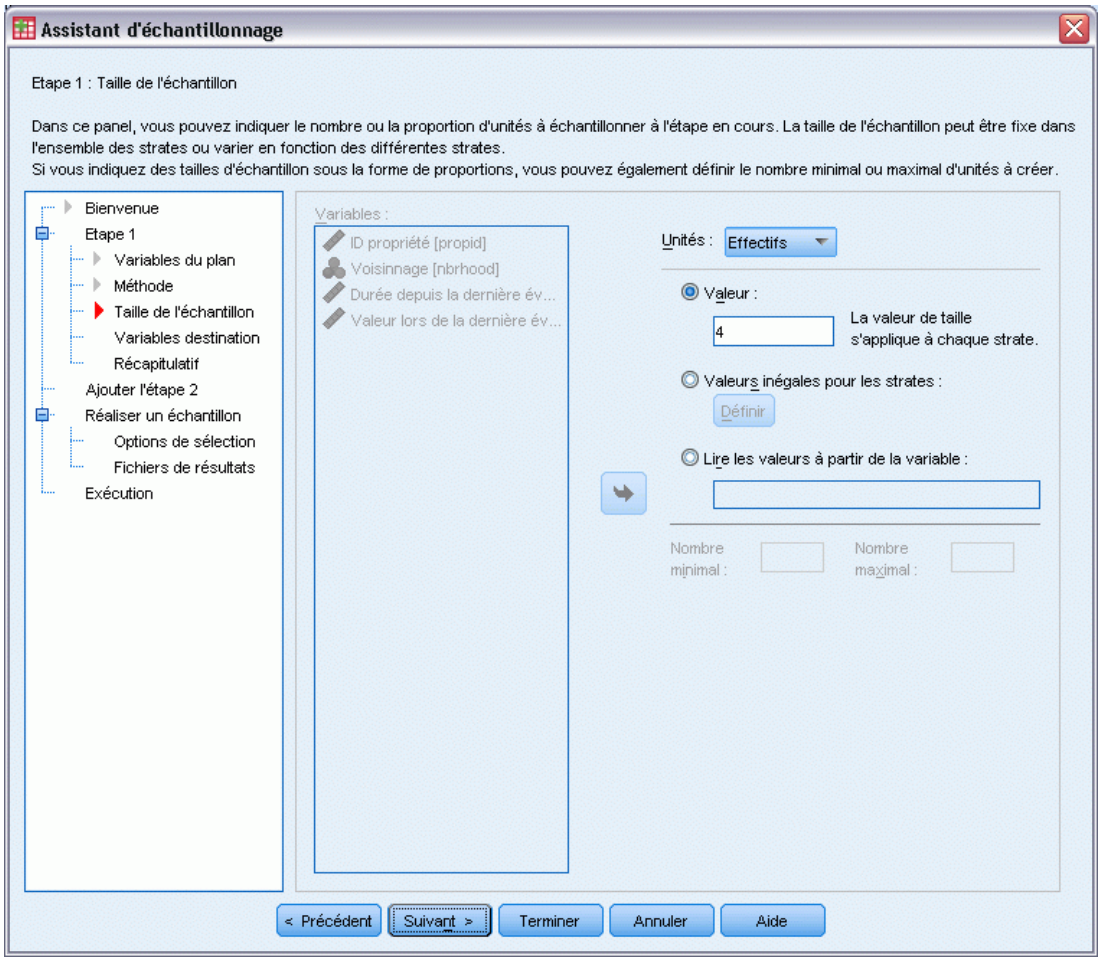
Figure 13-2
*Etape Variables de plan de l'assistant d'échantillonnage (phase 1)*




- Sélectionnez *Comté* comme variable de stratification.
- Sélectionnez *Commune* comme variable de grappe.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Méthode d'échantillonage.

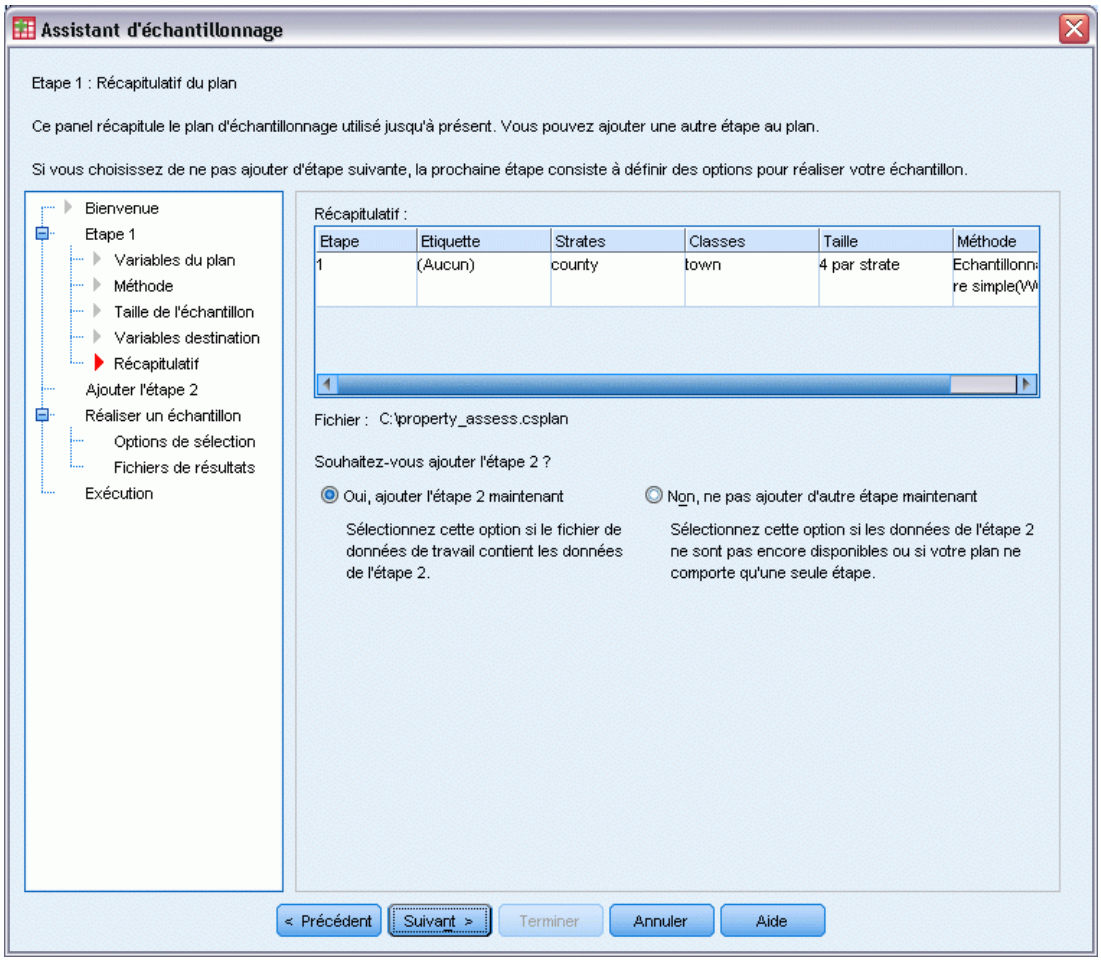
  Cette structure de plan indique que des échantillons indépendants sont réalisés pour chaque comté. Au cours de cette phase, les communes sont tracées en tant qu'unité d'échantillonnage principale à l'aide de la méthode par défaut, l'échantillonnage aléatoire simple.

Figure 13-3
*Etape Taille de l'échantillon de l'assistant d'échantillonnage (phase 1)*



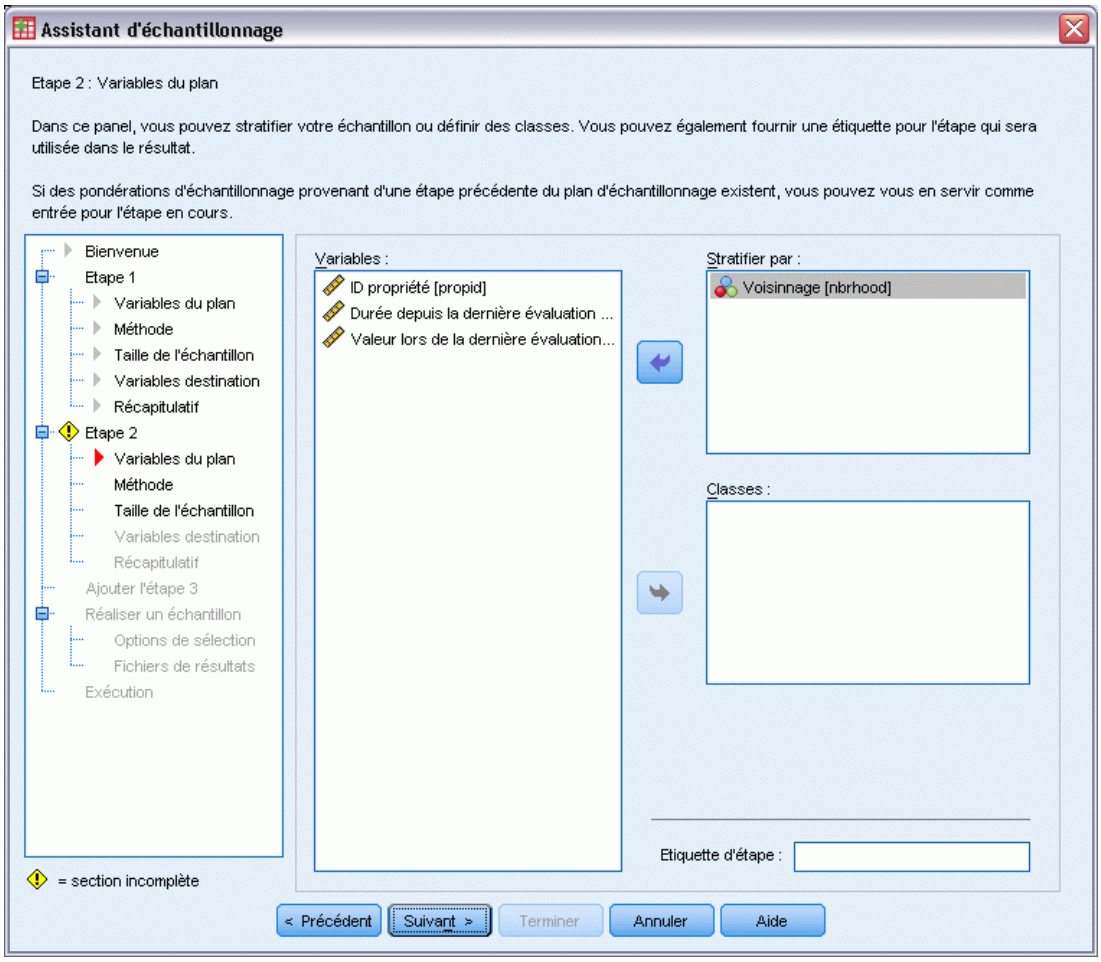

- Sélectionnez Effectifs dans la liste déroulante Unités.
- Saisissez 4 comme valeur du nombre d'unités à sélectionner au cours de cette phase.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Variables destination.

Figure 13-4
*Etape Récapitulatif du plan de l'assistant d'échantillonnage (phase 1)*




- Sélectionnez Oui, ajouter l'étape 2 maintenant.
- Cliquez sur Suivant.

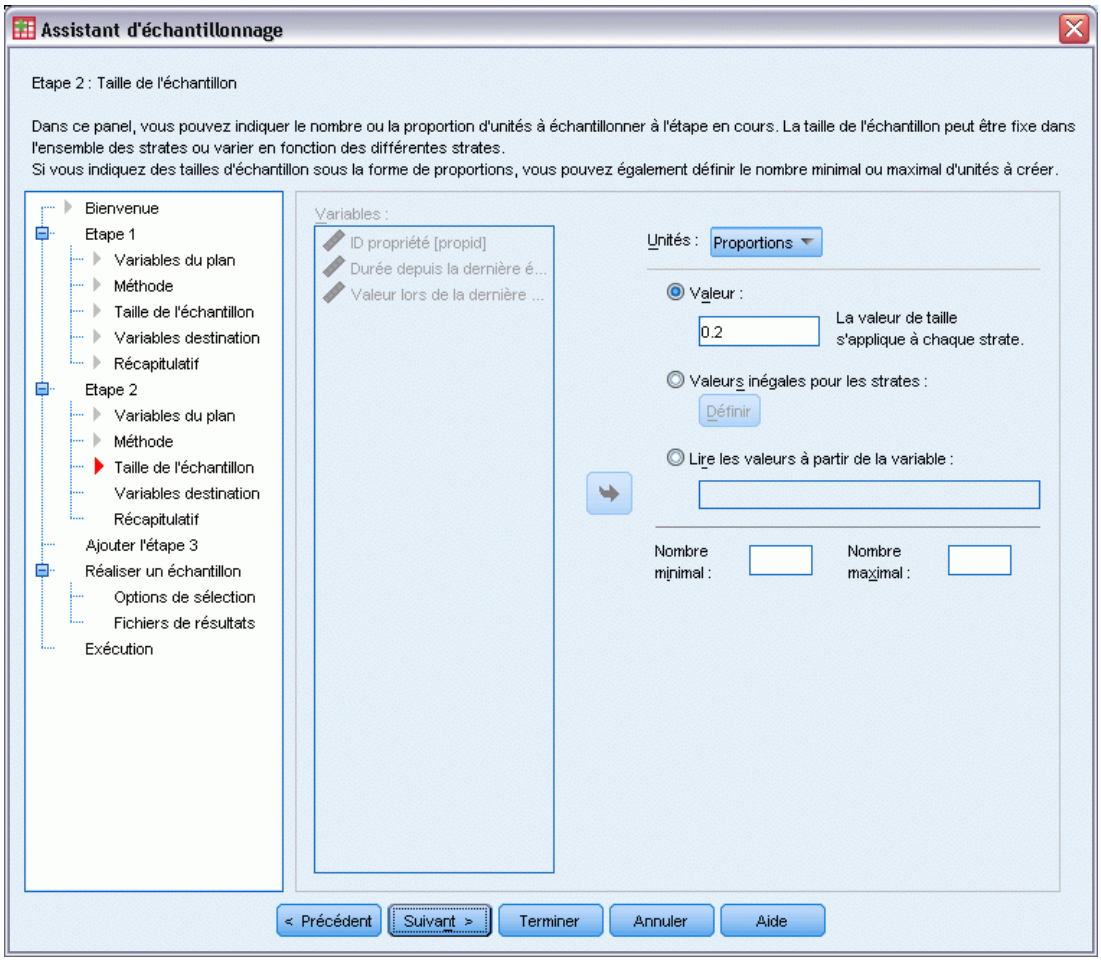
Figure 13-5
*Etape Variables de plan de l'assistant d'échantillonnage (phase 2)*




- Sélectionnez *Voisinage* comme variable de stratification.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Méthode d'échantillonage.

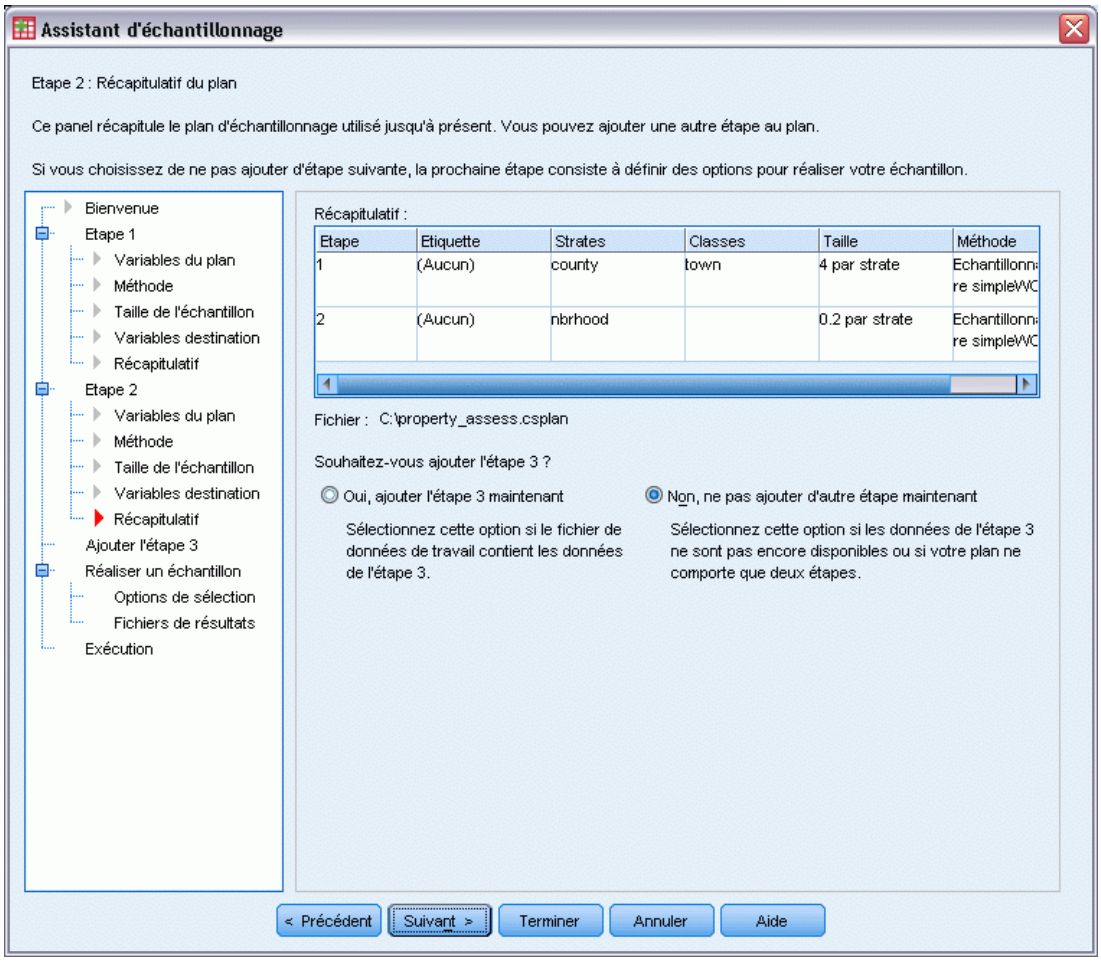
  Cette structure de plan indique que des échantillons indépendants sont réalisés pour chaque quartier des communes tracées au cours de la phase 1. Au cours de cette phase, les propriétés sont tracées en tant qu'unité d'échantillonnage principale à l'aide de la méthode Echantillonnage aléatoire simple.

Figure 13-6
*Etape Taille de l'échantillon de l'assistant d'échantillonnage (phase 2)*



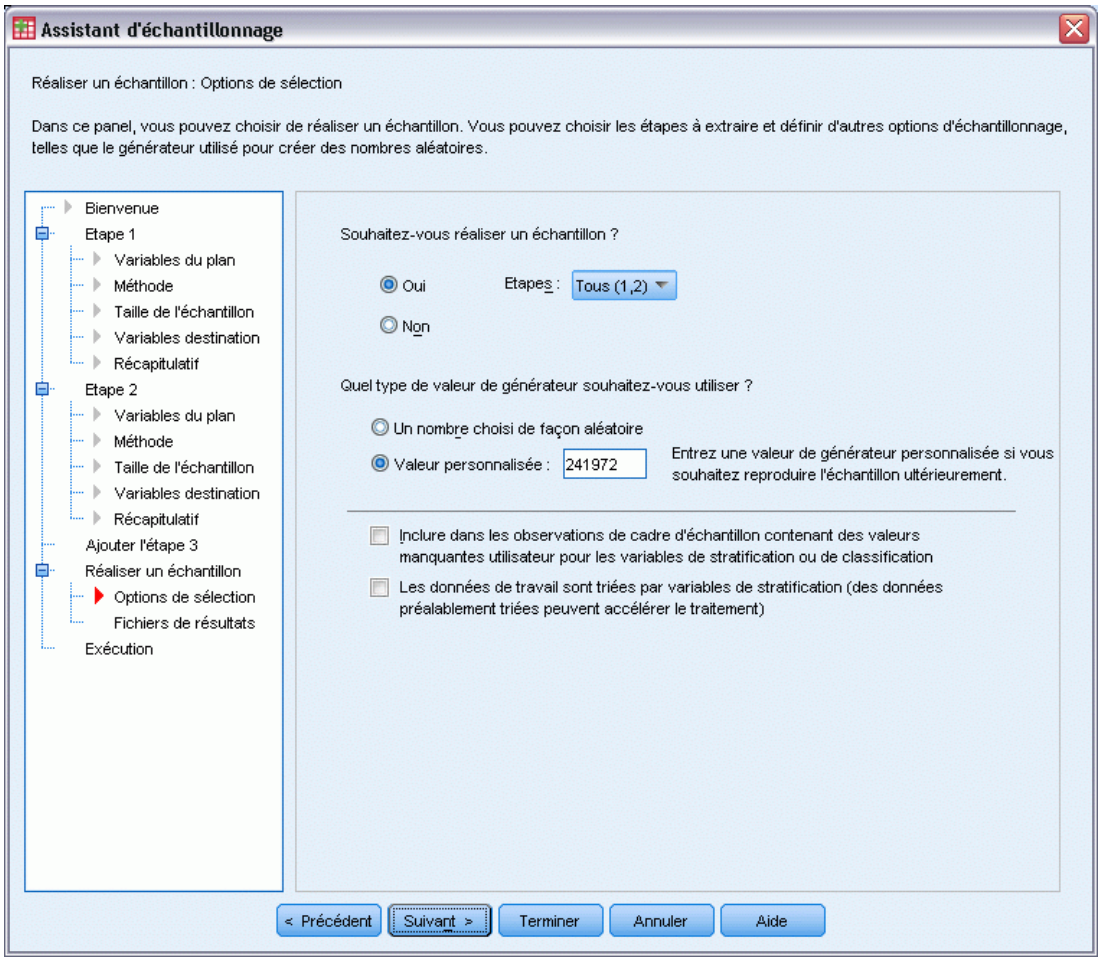

- Sélectionnez Proportions dans la liste déroulante Unités.
- Entrez 0,2 comme valeur de la proportion d'unités à échantillonner dans chaque strate.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Variables destination.

Figure 13-7
*Etape Récapitulatif du plan de l'assistant d'échantillonnage (phase 2)*




▶ Consultez le plan d'échantillonnage, puis cliquez sur Suivant.

Figure 13-8
*Etape Réalisation de l'échantillon : Options de sélection de l'assistant d'échantillonnage*

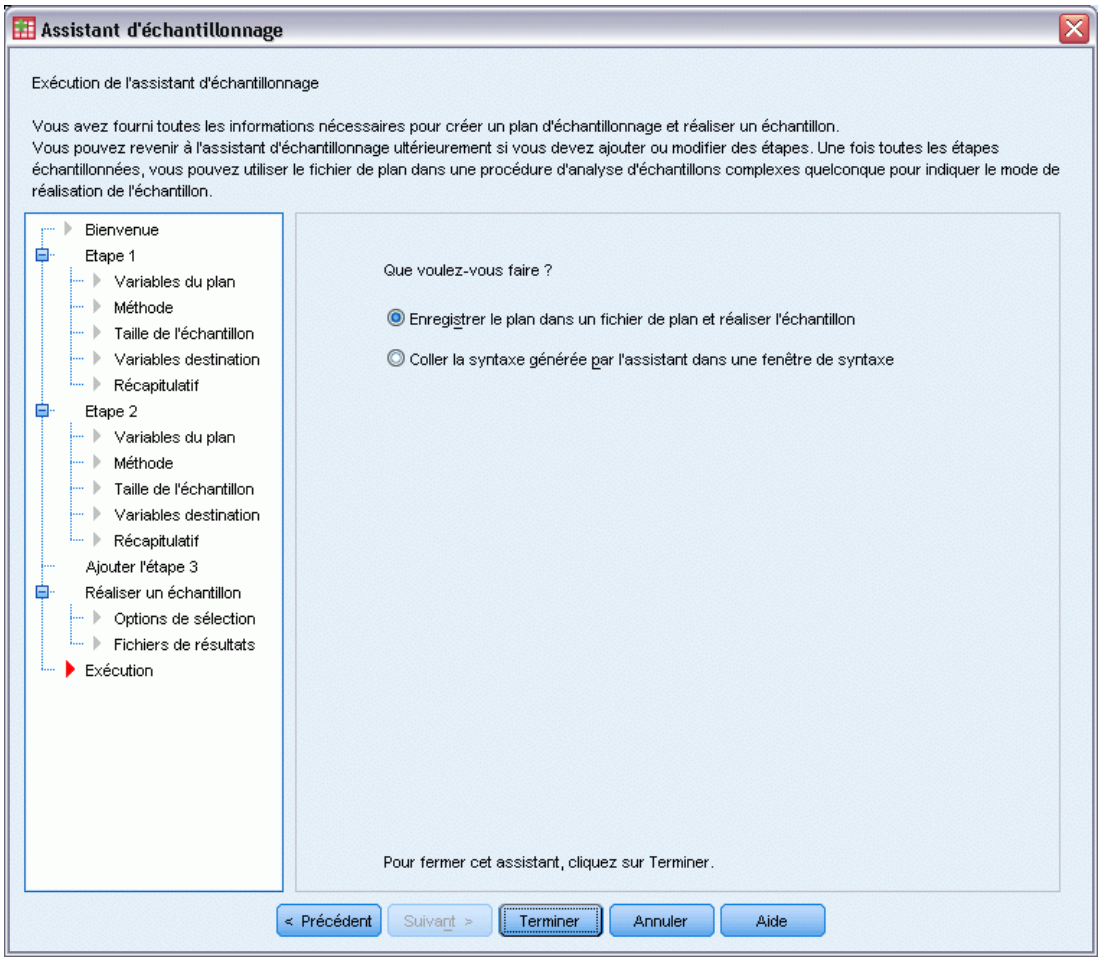



- Sélectionnez Valeur personnalisée comme type de générateur aléatoire à utiliser, puis entrez la valeur 241972.

  L'utilisation d'une valeur personnalisée vous permet de répliquer précisément les résultats de cet exemple.

- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Réalisation de l'échantillon : Fichiers de résultats.

Figure 13-9
*Etape Fin de l'assistant d'échantillonnage*




▶ Cliquez sur Terminer.

Ces sélections permettent de générer le fichier de plan d'échantillonnage *property_assess.csplan* et de réaliser un échantillon en fonction de ce plan.

## Récapitulatif du plan

Figure 13-10
*Récapitulatif du plan*

| | | | Etape 1 | Etape 2 |
|---|---|---|---|---|
| Variables du plan | Stratification | 1 | Département | Voisinnage |
| | Grappe | 1 | Ville | |
| Informations sur l'échantillon | Méthode de sélection | | Echantillonnage aléatoire simple sans remise | Echantillonnage aléatoire simple sans remise |
| | Nombre d'unités échantillonnées | | 4 | |
| | Variables créées ou modifiées | Probabilité d'inclusion (sélection) à plusieurs degrés | ProbabilitéInclusion_1_ | ProbabilitéInclusion_2_ |
| | | Pondération cumulative d'échantillonnage à plusieurs degrés | Pondération Echantillon Cumulée_1_ | Pondération EchantillonCumulée_2_ |
| | Proportion d'unités échantillonnées | | | ,2 |
| Informations sur l'analyse | Hypothèses de l'estimateur | | Echantillonnage de probabilité égale sans remise | Echantillonnage de probabilité égale sans remise |
| | Probabilité d'inclusion | | Obtenu à partir de la variable ProbabilitéInclusion_1_ | Obtenu à partir de la variable ProbabilitéInclusion_2_ |

Fichier de plan : C:\property_assess.csplan
Variable de pondération : PondérationEchantillon_Final_

Le tableau récapitulatif vous permet de passer en revue votre plan d'échantillonnage et de vous assurer qu'il correspond à vos attentes.

## Récapitulatif de l'échantillonnage

Figure 13-11
*Récapitulatif*

| | Nombre d'unités échantillonnées | | Proportion d'unités échantillonnées | |
|---|---|---|---|---|
| Département | Obligatoire | Réel | Obligatoire | Réel |
| Est | 4 | 4 | 44,4% | 44,4% |
| Centre | 4 | 4 | 57,1% | 57,1% |
| Sud | 4 | 4 | 25,0% | 25,0% |
| Nord | 4 | 4 | 44,4% | 44,4% |
| Ouest | 4 | 4 | 50,0% | 50,0% |

Fichier de plan : C:\property_assess.csplan

Ce tableau récapitulatif vous permet de passer en revue la première phase de l'échantillonnage et de vérifier que celui-ci correspond au plan. Conformément à votre demande, quatre communes ont été échantillonnées dans chaque comté.

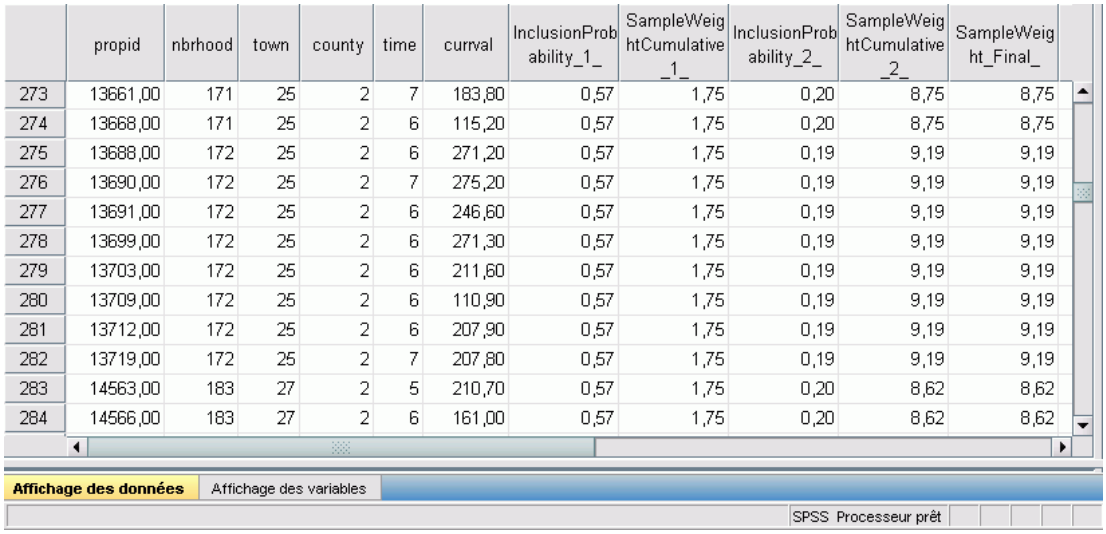
Figure 13-12
*Récapitulatif*

| | | | Nombre d'unités échantillonnées | | Proportion d'unités échantillonnées | |
|---|---|---|---|---|---|---|
| Département | Ville | Voisinnage | Obligatoire | Réel | Obligatoire | Réel |
| Est | 2 | 8 | 4 | 4 | 20,0% | 19,0% |
| | | 9 | 14 | 14 | 20,0% | 20,6% |
| | | 10 | 7 | 7 | 20,0% | 18,9% |
| | | 11 | 14 | 14 | 20,0% | 20,0% |
| | 6 | 36 | 13 | 13 | 20,0% | 20,3% |
| | | 37 | 14 | 14 | 20,0% | 20,6% |
| | | 38 | 13 | 13 | 20,0% | 20,6% |
| | 7 | 43 | 12 | 12 | 20,0% | 20,7% |
| | | 44 | 11 | 11 | 20,0% | 19,6% |
| | | 45 | 11 | 11 | 20,0% | 20,8% |
| | | 46 | 13 | 13 | 20,0% | 20,0% |
| | 9 | 57 | 13 | 13 | 20,0% | 20,6% |
| | | 58 | 5 | 5 | 20,0% | 18,5% |
| | | 59 | 11 | 11 | 20,0% | 19,3% |
| | | 60 | 13 | 13 | 20,0% | 19,4% |
| Centre | 22 | 148 | 9 | 9 | 20,0% | 19,6% |
| | | 149 | 8 | 8 | 20,0% | 20,0% |
| | | 150 | 14 | 14 | 20,0% | 20,0% |

Ce tableau récapitulatif (dont la partie supérieure apparaît ici) vous permet de passer en revue la deuxième phase de l'échantillonnage. Il permet également de vérifier que l'échantillonnage correspond au plan. Conformément à vos critères, environ 20 % des propriétés ont été échantillonnés à partir de chaque quartier de chaque commune échantillonnée au cours de la première phase.

## *Résultats de l'échantillonnage*

Figure 13-13
*Editeur de données contenant les résultats de l'échantillonnage*



| | propid | nbrhood | town | county | time | curval | InclusionProbability_1_ | SampleWeightCumulative_1_ | InclusionProbability_2_ | SampleWeightCumulative_2_ | SampleWeight_Final_ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 273 | 13661,00 | 171 | 25 | 2 | 7 | 183,80 | 0,57 | 1,75 | 0,20 | 8,75 | 8,75 |
| 274 | 13668,00 | 171 | 25 | 2 | 6 | 115,20 | 0,57 | 1,75 | 0,20 | 8,75 | 8,75 |
| 275 | 13688,00 | 172 | 25 | 2 | 6 | 271,20 | 0,57 | 1,75 | 0,19 | 9,19 | 9,19 |
| 276 | 13690,00 | 172 | 25 | 2 | 7 | 275,20 | 0,57 | 1,75 | 0,19 | 9,19 | 9,19 |
| 277 | 13691,00 | 172 | 25 | 2 | 6 | 246,60 | 0,57 | 1,75 | 0,19 | 9,19 | 9,19 |
| 278 | 13699,00 | 172 | 25 | 2 | 6 | 271,30 | 0,57 | 1,75 | 0,19 | 9,19 | 9,19 |
| 279 | 13703,00 | 172 | 25 | 2 | 6 | 211,60 | 0,57 | 1,75 | 0,19 | 9,19 | 9,19 |
| 280 | 13709,00 | 172 | 25 | 2 | 6 | 110,90 | 0,57 | 1,75 | 0,19 | 9,19 | 9,19 |
| 281 | 13712,00 | 172 | 25 | 2 | 6 | 207,90 | 0,57 | 1,75 | 0,19 | 9,19 | 9,19 |
| 282 | 13719,00 | 172 | 25 | 2 | 7 | 207,80 | 0,57 | 1,75 | 0,19 | 9,19 | 9,19 |
| 283 | 14563,00 | 183 | 27 | 2 | 5 | 210,70 | 0,57 | 1,75 | 0,20 | 8,62 | 8,62 |
| 284 | 14566,00 | 183 | 27 | 2 | 6 | 161,00 | 0,57 | 1,75 | 0,20 | 8,62 | 8,62 |



Vous pouvez visualiser les résultats de l'échantillonnage dans l'éditeur de données. Cinq nouvelles variables ont été enregistrées dans le fichier de travail. Ces variables représentent les probabilités d'insertion et les pondérations cumulatives d'échantillonnage de chaque phase, ainsi que les pondérations d'échantillonnage finales.

- Les observations contenant des valeurs pour ces variables ont été sélectionnées pour l'échantillon.
- Les observations contenant des valeurs manquantes pour les variables n'ont pas été sélectionnées.

L'agence utilisera désormais ses ressources pour recueillir les estimations en cours correspondant aux propriétés sélectionnées dans l'échantillon. Une fois que ces estimations sont disponibles, vous pouvez traiter l'échantillon avec les procédures d'analyse d'échantillons complexes, à l'aide du plan d'échantillonnage *property_assess.csplan* pour indiquer les spécifications d'échantillonnage.

# *Obtention d'un échantillon à partir d'un cadre d'échantillonnage partiel*

Une société est intéressée par la compilation et la vente d'une base de données contenant des informations d'enquête de qualité élevée. L'échantillon d'enquête doit être représentatif, mais réalisé de manière efficace, afin que les méthodes d'échantillonnage complexes soient utilisées. Le plan d'échantillonnage complet nécessite la structure suivante :

| Etape | Strates | Classes |
|---|---|---|
| 1 | Région | Province |
| 2 | Quartier | Ville |
| 3 | Sous-division | |

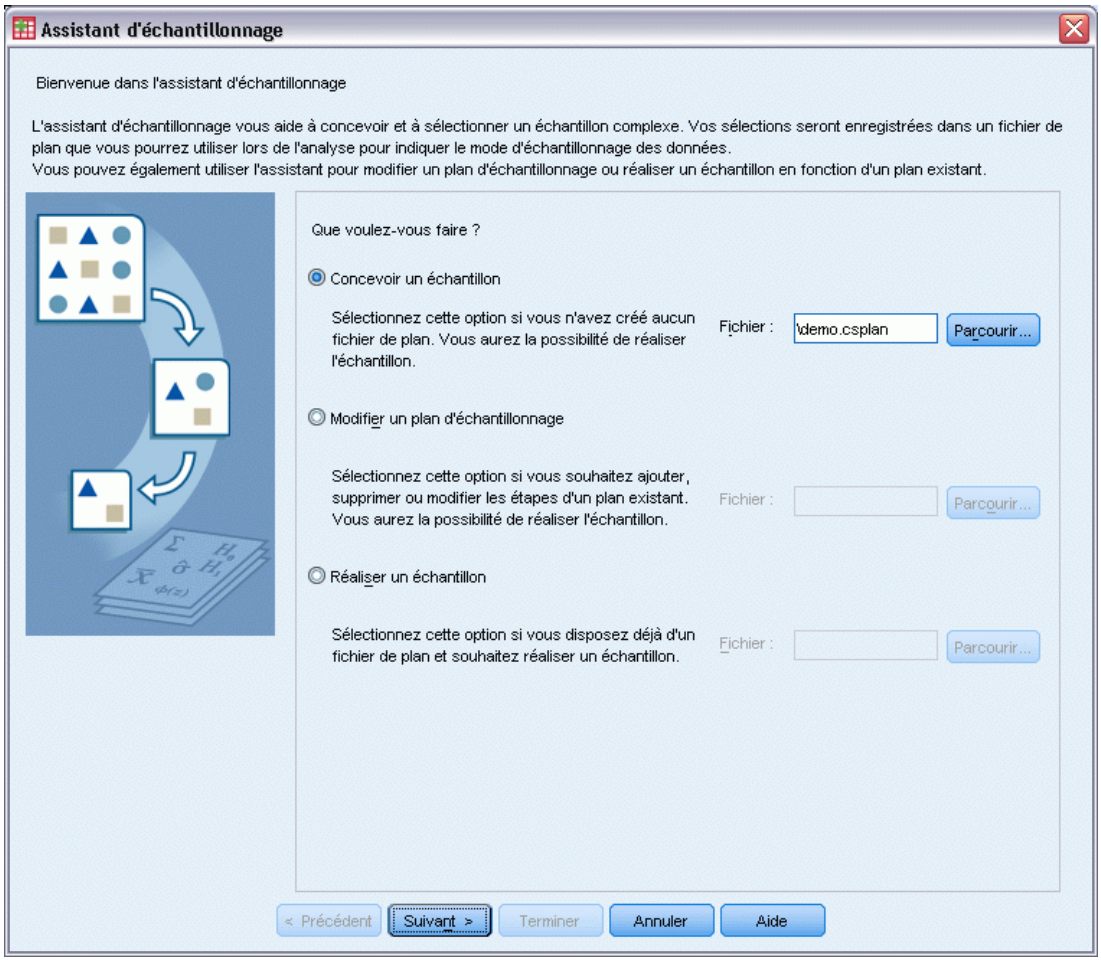
Au cours de la troisième phase, les ménages représentent l'unité d'échantillonnage principale et ceux qui sont sélectionnés feront l'objet d'une enquête. Cependant, les informations n'étant facilement disponibles qu'au niveau de la ville, la société décide d'exécuter les deux premières phases du plan, puis de rassembler les informations sur le nombre de sous-divisions et de ménages dans les villes échantillonnées. Les informations disponibles au niveau de la ville sont rassemblées dans le fichier *demo_cs_1.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*. Notez que ce fichier comporte une variable *Sous-division* qui contient tous les 1. Il s'agit d'une valeur de substitution pour la "vraie" variable dont les valeurs sont recueillies après exécution des deux premières phases du plan, et qui permet alors de spécifier le plan d'échantillonnage complet en trois phases. Utilisez l'assistant d'échantillonnage des échantillons complexes pour spécifier le plan d'échantillonnage complexe complet, puis tracez les deux premières phases.

## *Utilisation de l'assistant pour effectuer un échantillonnage à partir du premier cadre partiel*

- Pour exécuter l'assistant d'échantillonnage des échantillons complexes, sélectionnez dans le menu l'option suivante :

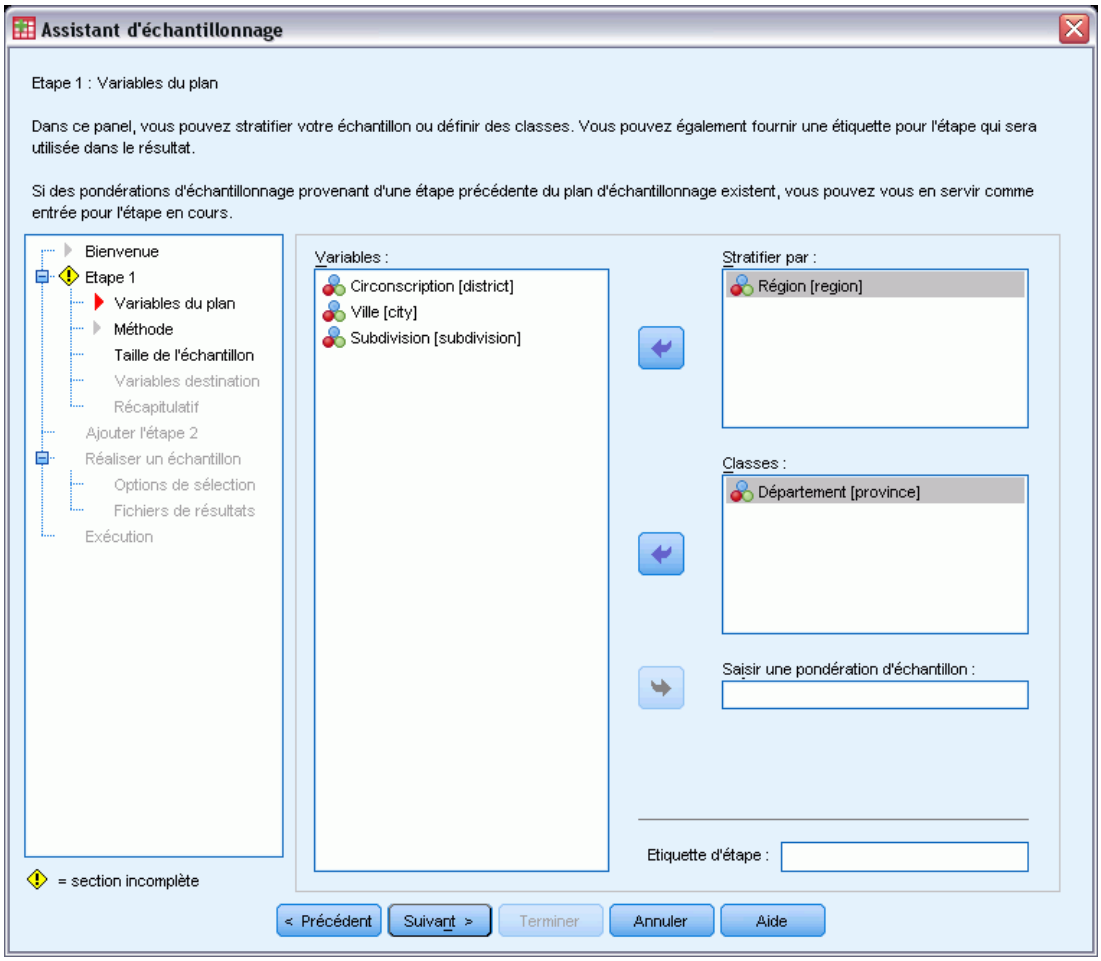
  Analyse > Echantillons complexes > Sélectionner un échantillon...

Figure 13-14
*Etape Bienvenue de l'assistant d'échantillonnage*




- Sélectionnez Concevoir un échantillon, accédez à l'emplacement auquel vous souhaitez enregistrer le fichier, et saisissez demo.csplan comme nom du fichier de plan.
- Cliquez sur Suivant.

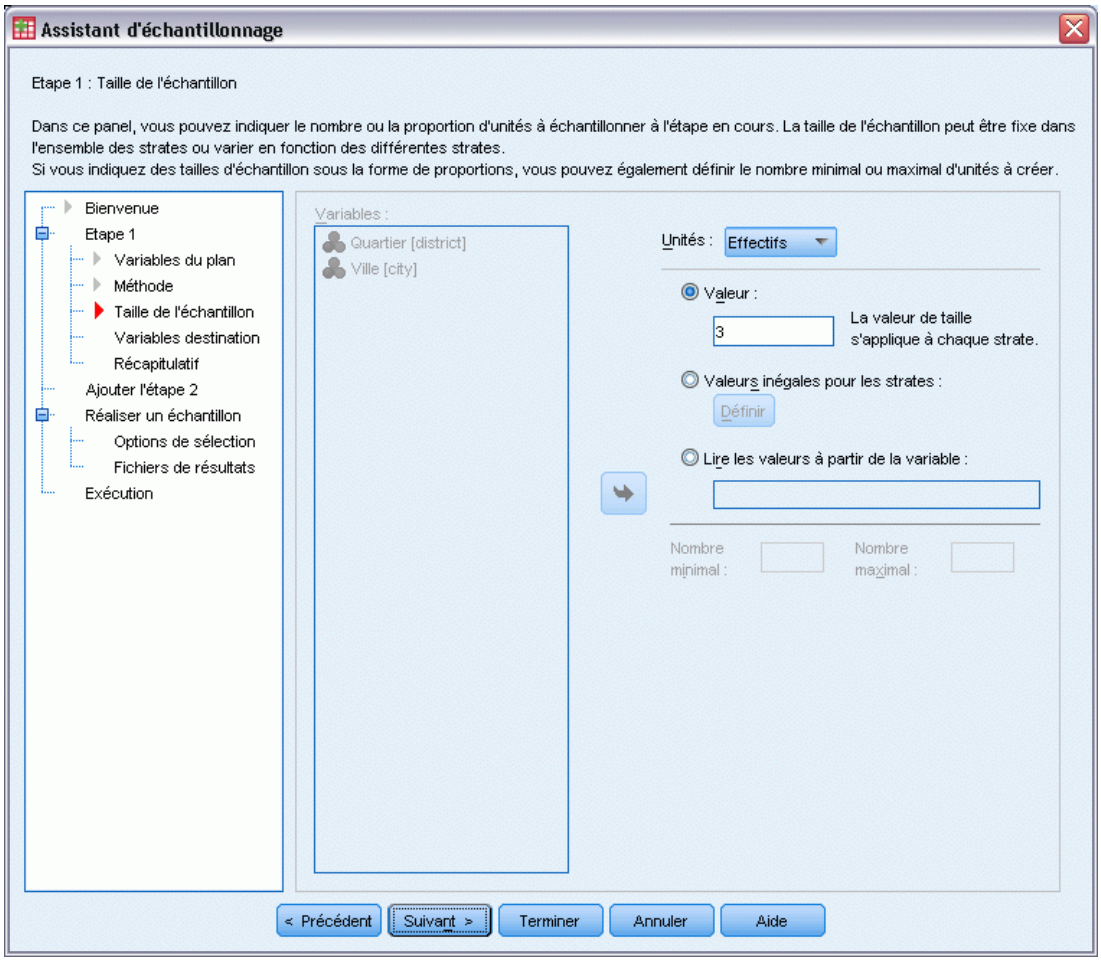
Figure 13-15
*Etape Variables de plan de l'assistant d'échantillonnage (phase 1)*




- Sélectionnez *Région* comme variable de stratification.
- Sélectionnez *Province* comme variable de grappe.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Méthode d'échantillonage.

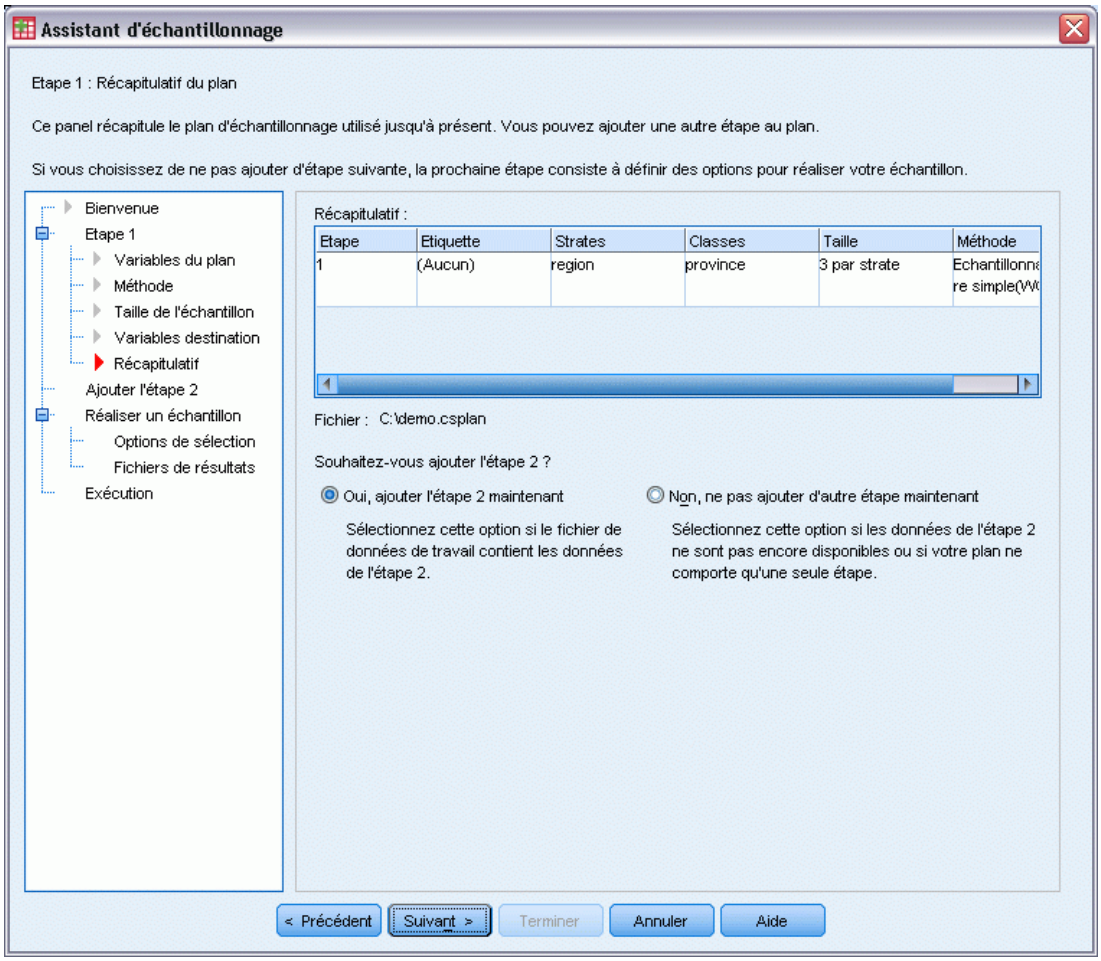
  Cette structure de plan indique que des échantillons indépendants sont réalisés pour chaque région. Au cours de cette phase, les provinces sont tracées en tant qu'unité d'échantillonnage principale à l'aide de la méthode par défaut, l'échantillonnage aléatoire simple.

Figure 13-16
*Etape Taille de l'échantillon de l'assistant d'échantillonnage (phase 1)*



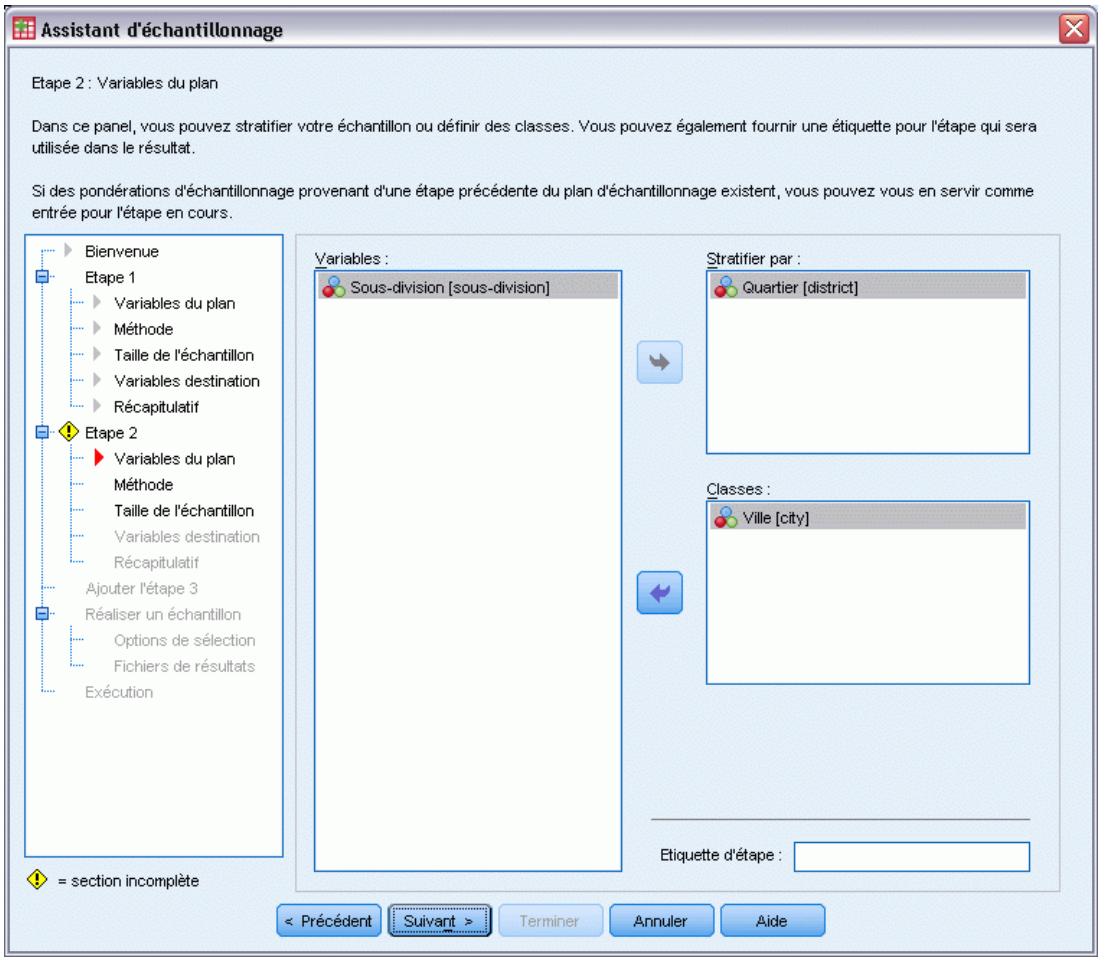

- Sélectionnez Effectifs dans la liste déroulante Unités.
- Saisissez 3 comme valeur du nombre d'unités à sélectionner au cours de cette phase.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Variables destination.

Figure 13-17
*Etape Récapitulatif du plan de l'assistant d'échantillonnage (phase 1)*




- Sélectionnez Oui, ajouter l'étape 2 maintenant.
- Cliquez sur Suivant.

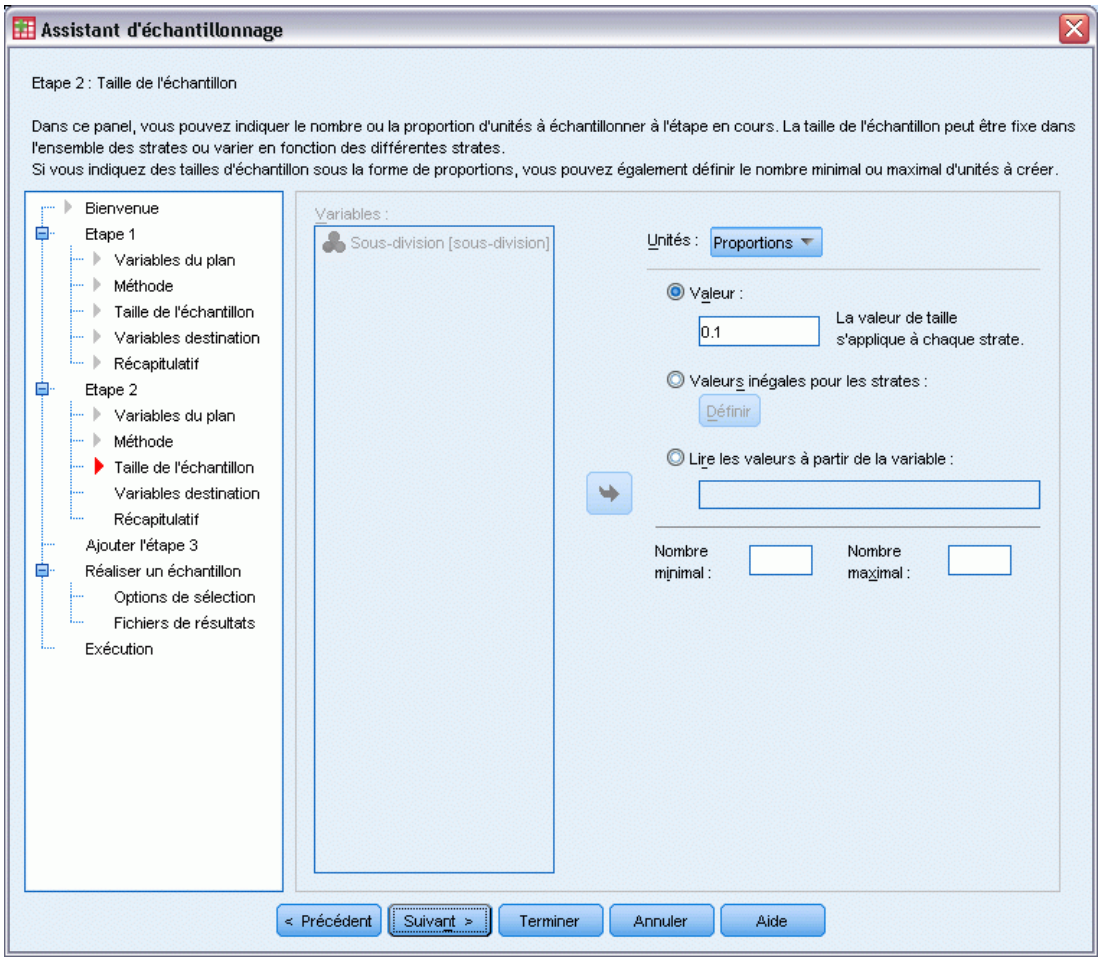
Figure 13-18
*Etape Variables de plan de l'assistant d'échantillonnage (phase 2)*




- Sélectionnez *District* comme variable de stratification.
- Sélectionnez *Ville* comme variable de grappe.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Méthode d'échantillonage.

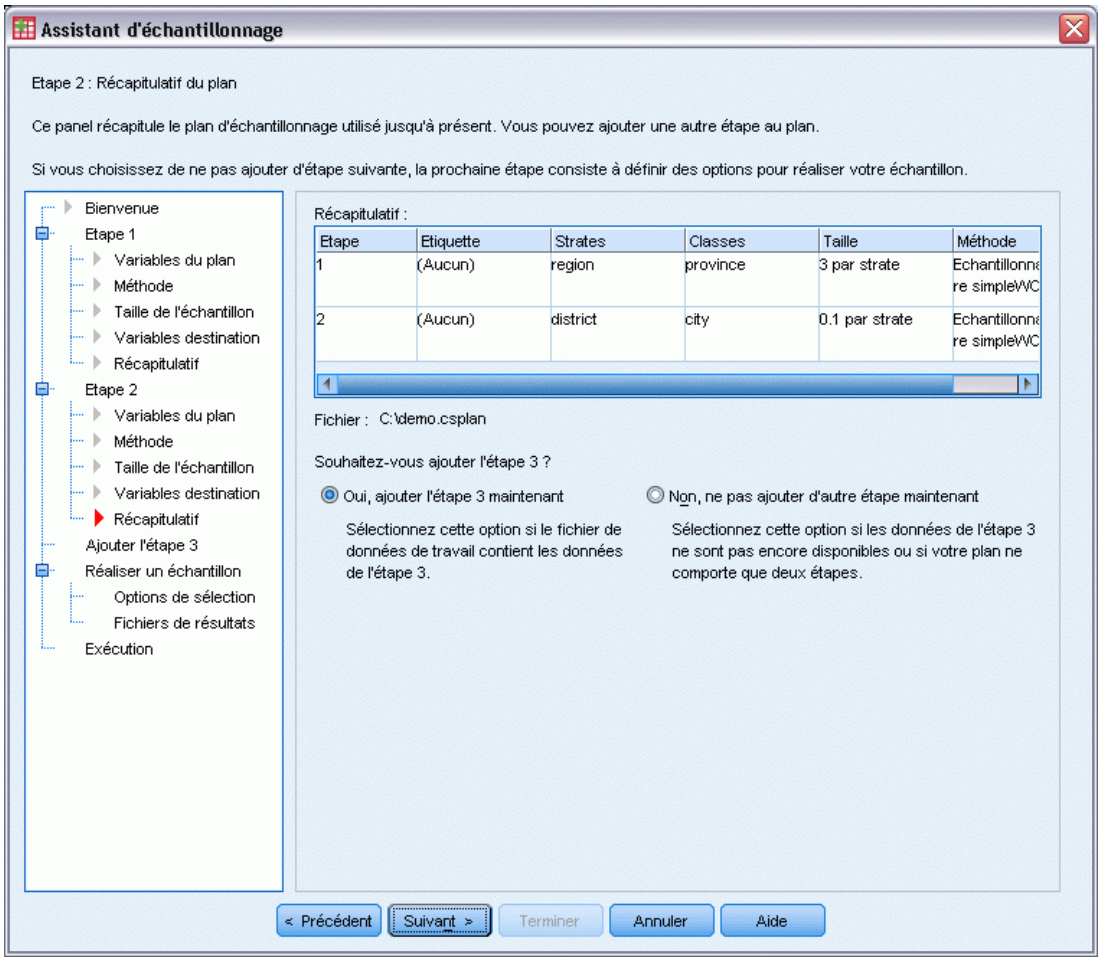
  Cette structure de plan indique que des échantillons indépendants sont réalisés pour chaque district. Au cours de cette phase, les villes sont tracées en tant qu'unité d'échantillonnage principale à l'aide de la méthode par défaut, l'échantillonnage aléatoire simple.

Figure 13-19
*Etape Taille de l'échantillon de l'assistant d'échantillonnage (phase 2)*



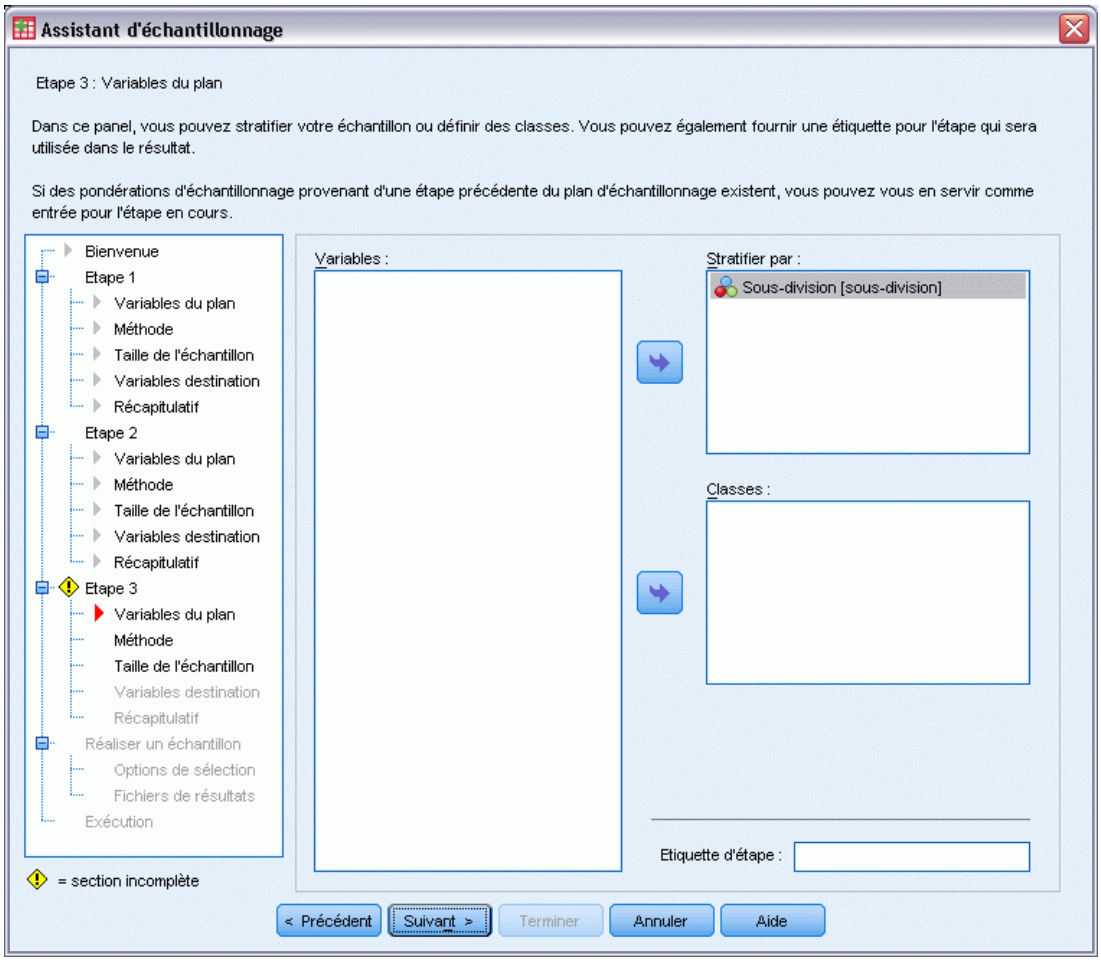

- Sélectionnez Proportions dans la liste déroulante Unités.
- Entrez 0,1 comme valeur de la proportion d'unités à échantillonner dans chaque strate.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Variables destination.

Figure 13-20
*Etape Récapitulatif du plan de l'assistant d'échantillonnage (phase 2)*




- Sélectionnez Oui, ajouter l'étape 3 maintenant.
- Cliquez sur Suivant.

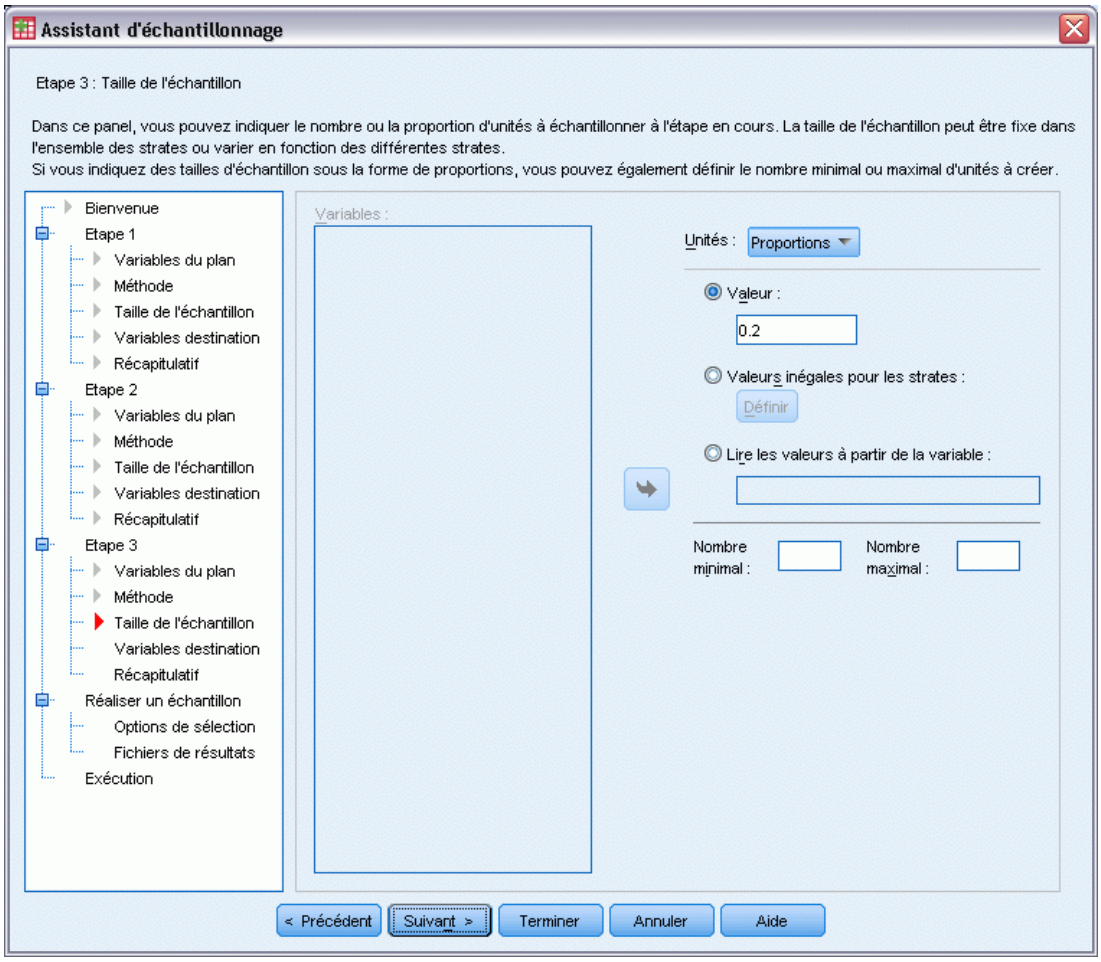
Figure 13-21
*Etape Variables de plan de l'assistant d'échantillonnage (phase 3)*




- Sélectionnez *Sous-division* comme variable de stratification.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Méthode d'échantillonage.

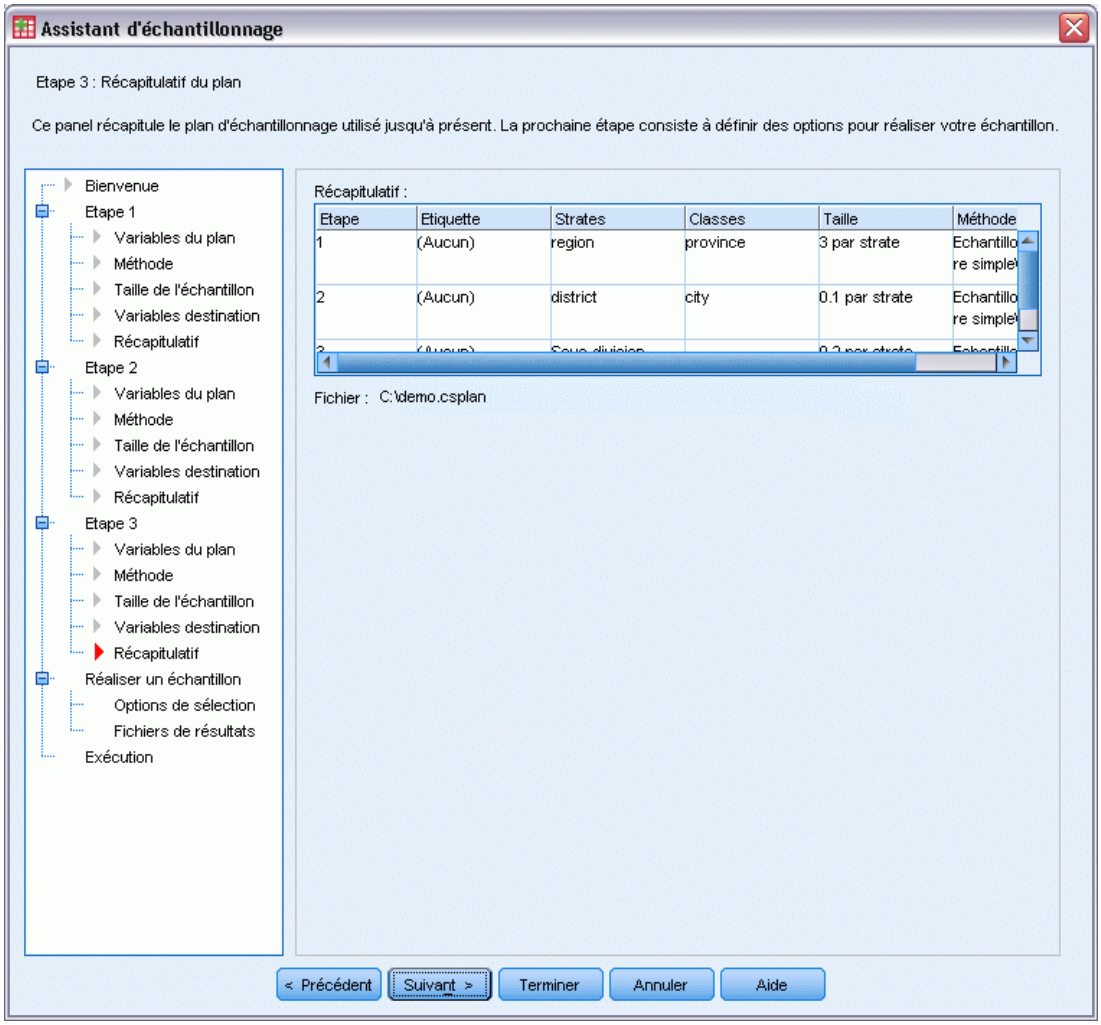
  Cette structure de plan indique que des échantillons indépendants sont réalisés pour chaque sous-division. Au cours de cette phase, les ménages sont tracés en tant qu'unité d'échantillonnage principale à l'aide de la méthode par défaut, l'échantillonnage aléatoire simple.

Figure 13-22
*Etape Taille de l'échantillon de l'assistant d'échantillonnage (phase 3)*



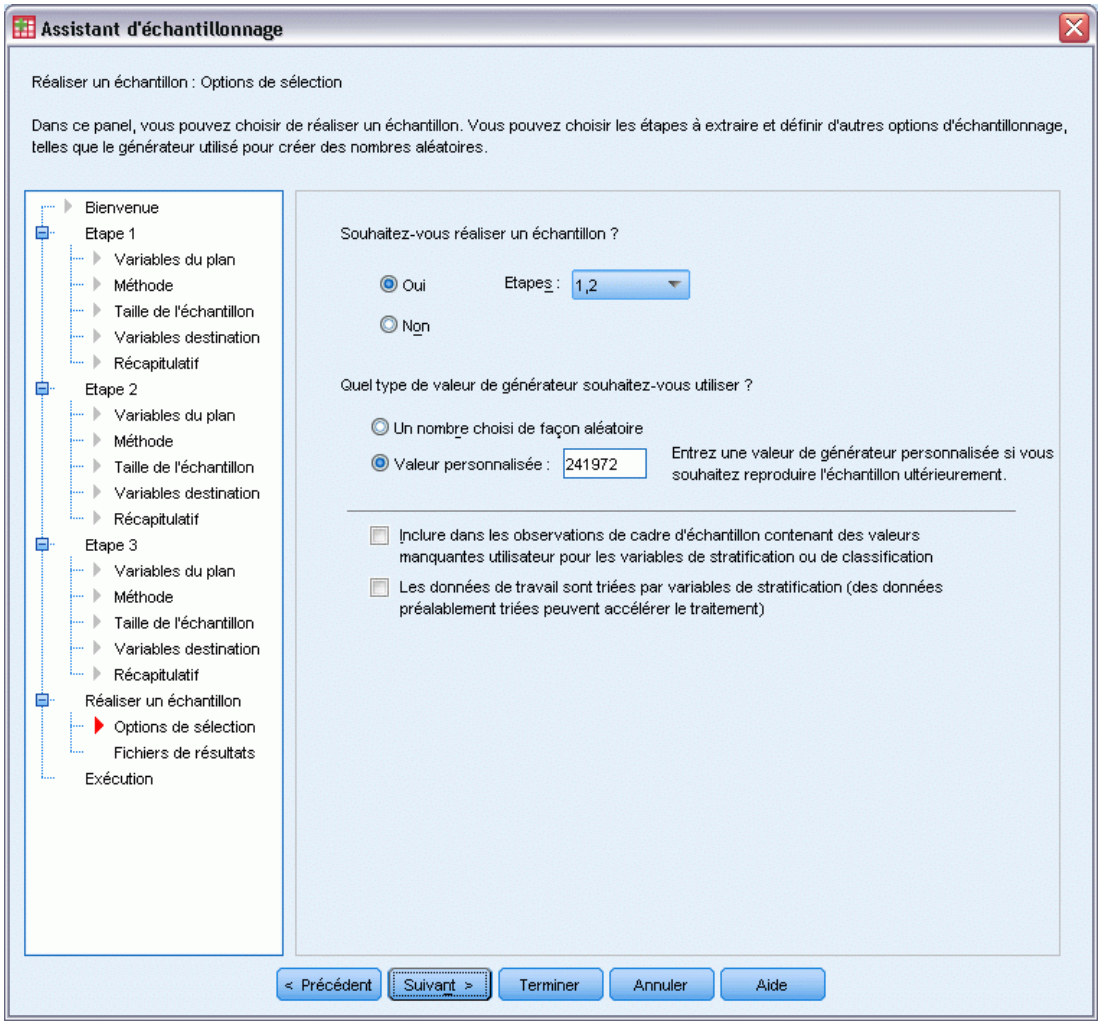

- Sélectionnez Proportions dans la liste déroulante Unités.
- Entrez la valeur 0,2 comme proportion d'unités à sélectionner au cours de cette phase.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Variables destination.

Figure 13-23
*Etape Récapitulatif du plan de l'assistant d'échantillonnage (phase 3)*




▶ Consultez le plan d'échantillonnage, puis cliquez sur Suivant.

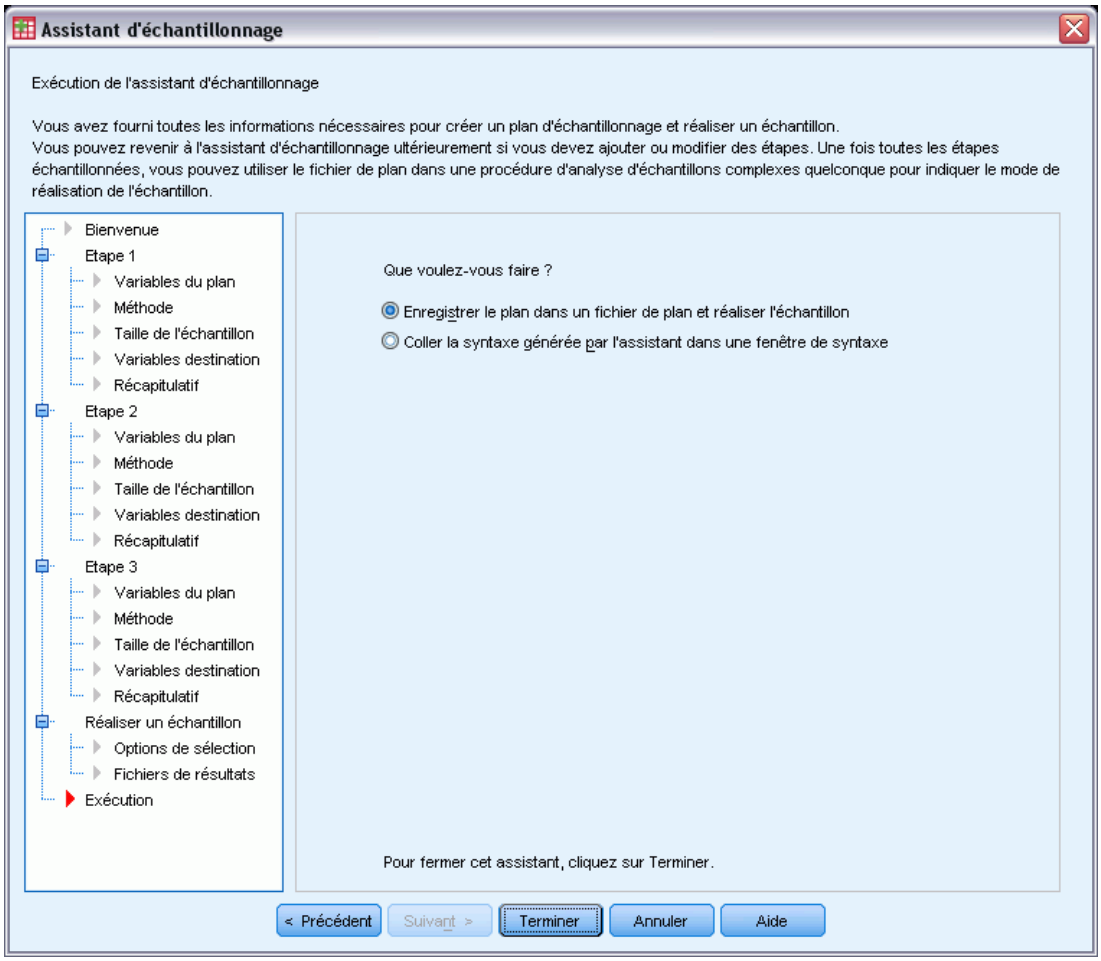
Figure 13-24
*Etape Réalisation de l'échantillon : Options de sélection de l'assistant d'échantillonnage*




- Sélectionnez 1, 2 comme phases à échantillonner maintenant.
- Sélectionnez Valeur personnalisée comme type de générateur aléatoire à utiliser, puis entrez la valeur 241972.

  L'utilisation d'une valeur personnalisée vous permet de répliquer précisément les résultats de cet exemple.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Réalisation de l'échantillon : Fichiers de résultats.

Figure 13-25
*Etape Fin de l'assistant d'échantillonnage*

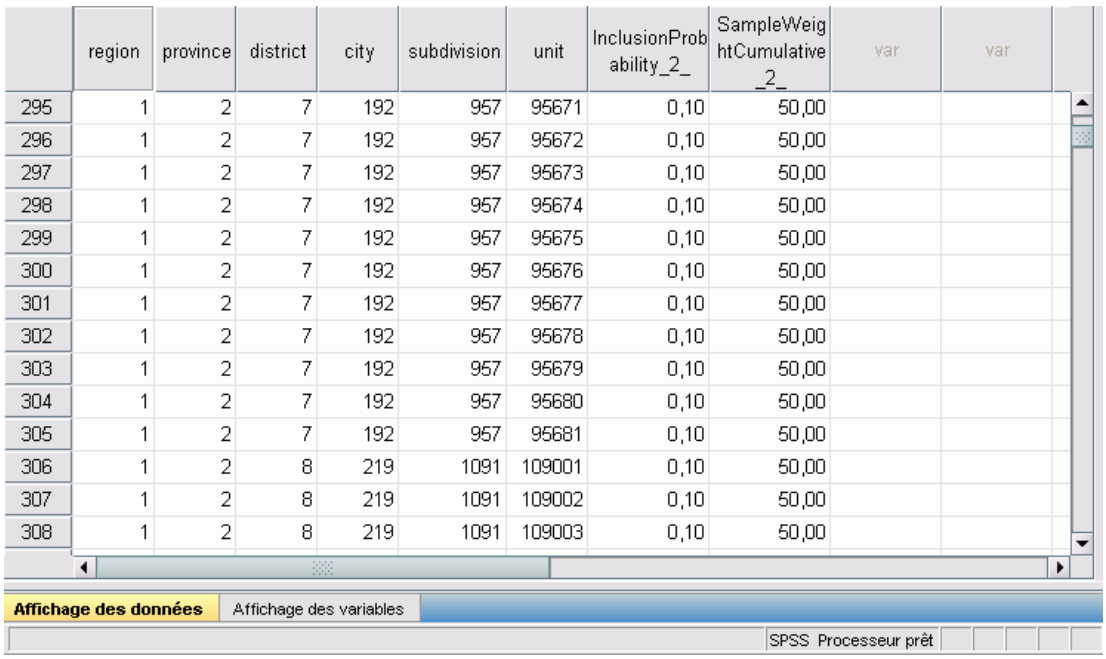



▶ Cliquez sur Terminer.

Ces sélections permettent de générer le fichier de plan d'échantillonnage *demo.csplan* et de réaliser un échantillon en fonction des deux premières phases de ce plan.

## Résultats de l'échantillonnage

Figure 13-26
*Editeur de données contenant les résultats de l'échantillonnage*



| | region | province | district | city | subdivision | unit | InclusionProbability_2_ | SampleWeightCumulative_2_ | var | var |
|---|---|---|---|---|---|---|---|---|---|---|
| 295 | 1 | 2 | 7 | 192 | 957 | 95671 | 0,10 | 50,00 | | |
| 296 | 1 | 2 | 7 | 192 | 957 | 95672 | 0,10 | 50,00 | | |
| 297 | 1 | 2 | 7 | 192 | 957 | 95673 | 0,10 | 50,00 | | |
| 298 | 1 | 2 | 7 | 192 | 957 | 95674 | 0,10 | 50,00 | | |
| 299 | 1 | 2 | 7 | 192 | 957 | 95675 | 0,10 | 50,00 | | |
| 300 | 1 | 2 | 7 | 192 | 957 | 95676 | 0,10 | 50,00 | | |
| 301 | 1 | 2 | 7 | 192 | 957 | 95677 | 0,10 | 50,00 | | |
| 302 | 1 | 2 | 7 | 192 | 957 | 95678 | 0,10 | 50,00 | | |
| 303 | 1 | 2 | 7 | 192 | 957 | 95679 | 0,10 | 50,00 | | |
| 304 | 1 | 2 | 7 | 192 | 957 | 95680 | 0,10 | 50,00 | | |
| 305 | 1 | 2 | 7 | 192 | 957 | 95681 | 0,10 | 50,00 | | |
| 306 | 1 | 2 | 8 | 219 | 1091 | 109001 | 0,10 | 50,00 | | |
| 307 | 1 | 2 | 8 | 219 | 1091 | 109002 | 0,10 | 50,00 | | |
| 308 | 1 | 2 | 8 | 219 | 1091 | 109003 | 0,10 | 50,00 | | |



Vous pouvez visualiser les résultats de l'échantillonnage dans l'éditeur de données. Cinq nouvelles variables ont été enregistrées dans le fichier de travail. Ces variables représentent les probabilités d'insertion et les pondérations cumulatives d'échantillonnage de chaque phase, ainsi que les pondérations d'échantillonnage finales des deux premières phases.

- Les villes contenant des valeurs pour ces variables ont été sélectionnées pour l'échantillon.
- Les villes contenant des valeurs manquantes pour les variables n'ont pas été sélectionnées.

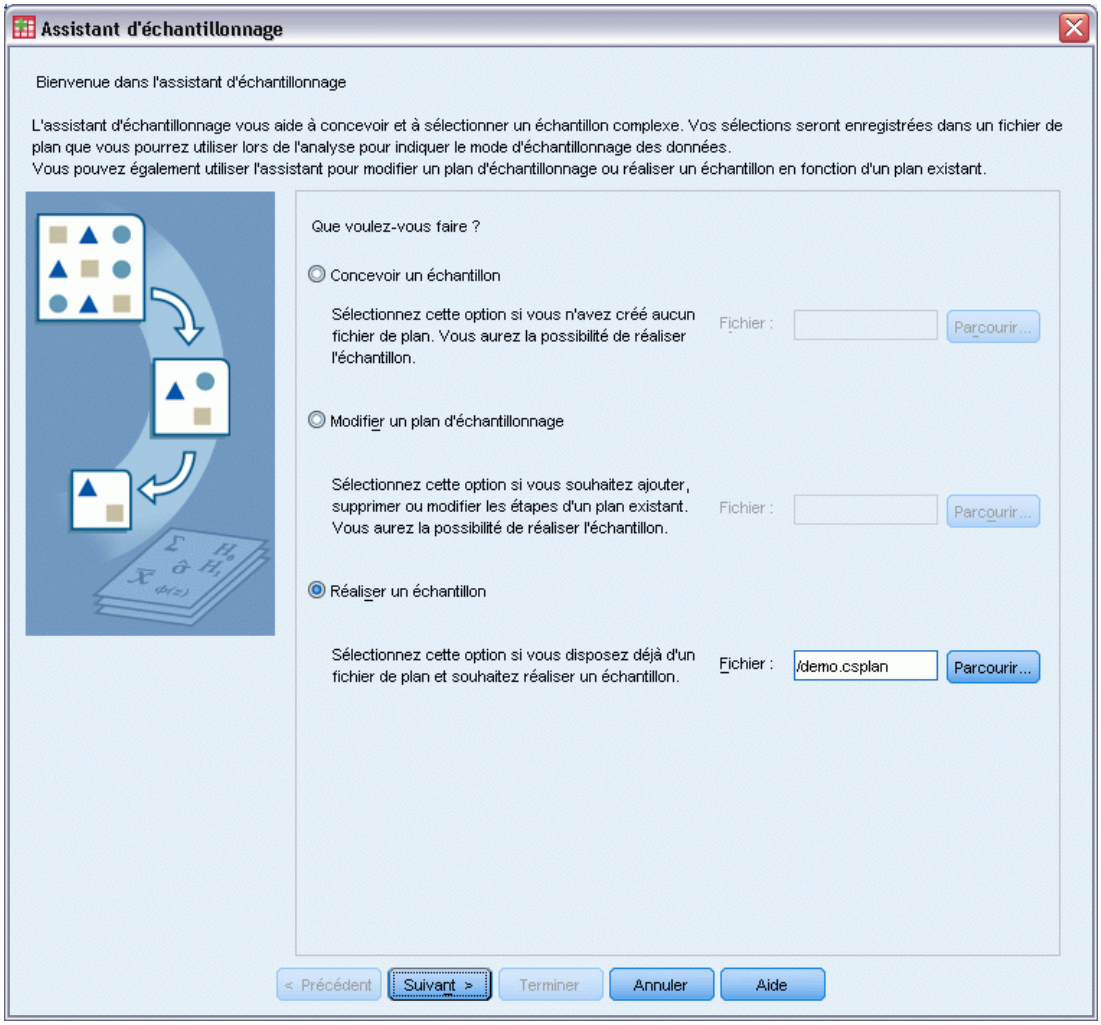
Pour chaque ville sélectionnée, la société a acquis des informations relatives aux sous-divisions et aux ménages, et les a placées dans le fichier *demo_cs_2.sav*. Utilisez ce fichier et l'assistant d'échantillonnage pour échantillonner la troisième phase de ce plan.

## Utilisation de l'assistant pour effectuer un échantillonnage à partir du deuxième cadre partiel

▶ Pour exécuter l'assistant d'échantillonnage des échantillons complexes, sélectionnez dans le menu l'option suivante :

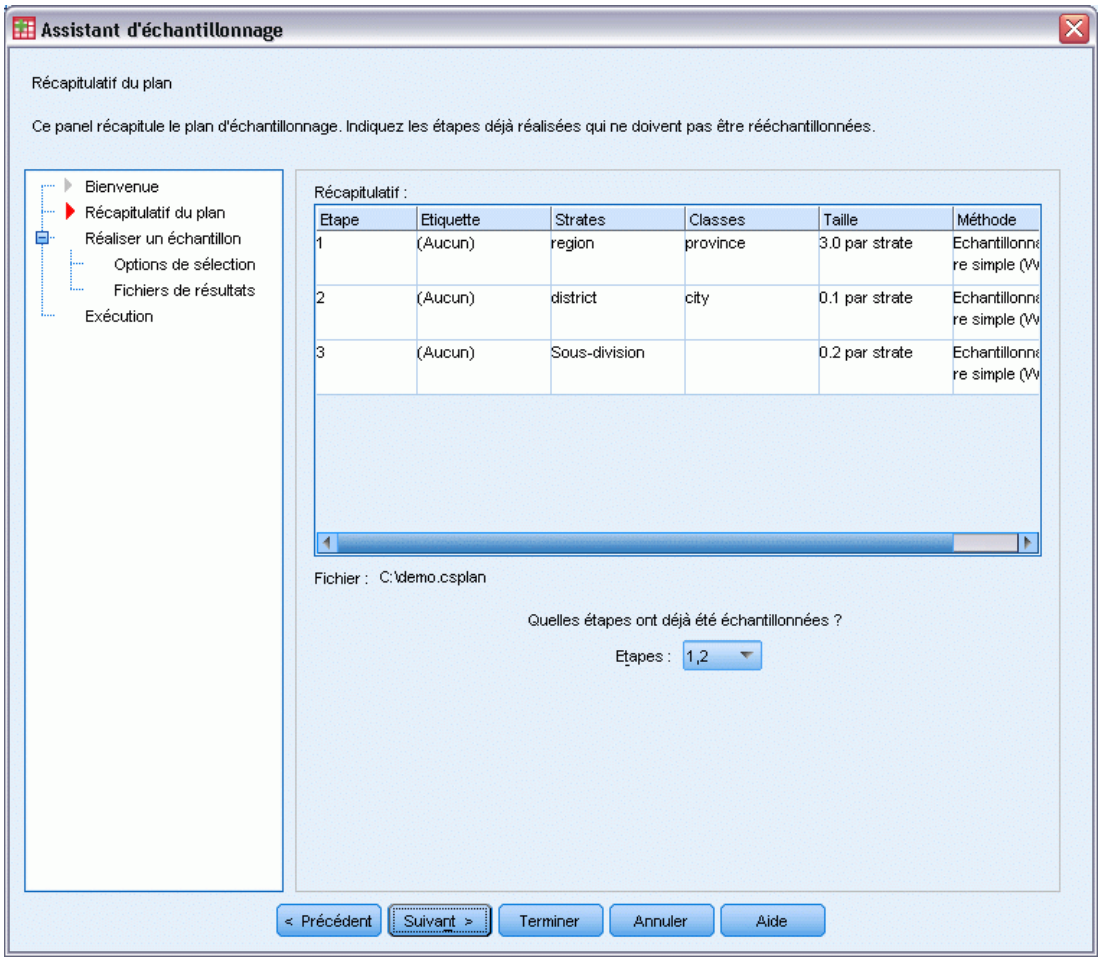
Analyse > Echantillons complexes > Sélectionner un échantillon...

Figure 13-27
*Etape Bienvenue de l'assistant d'échantillonnage*



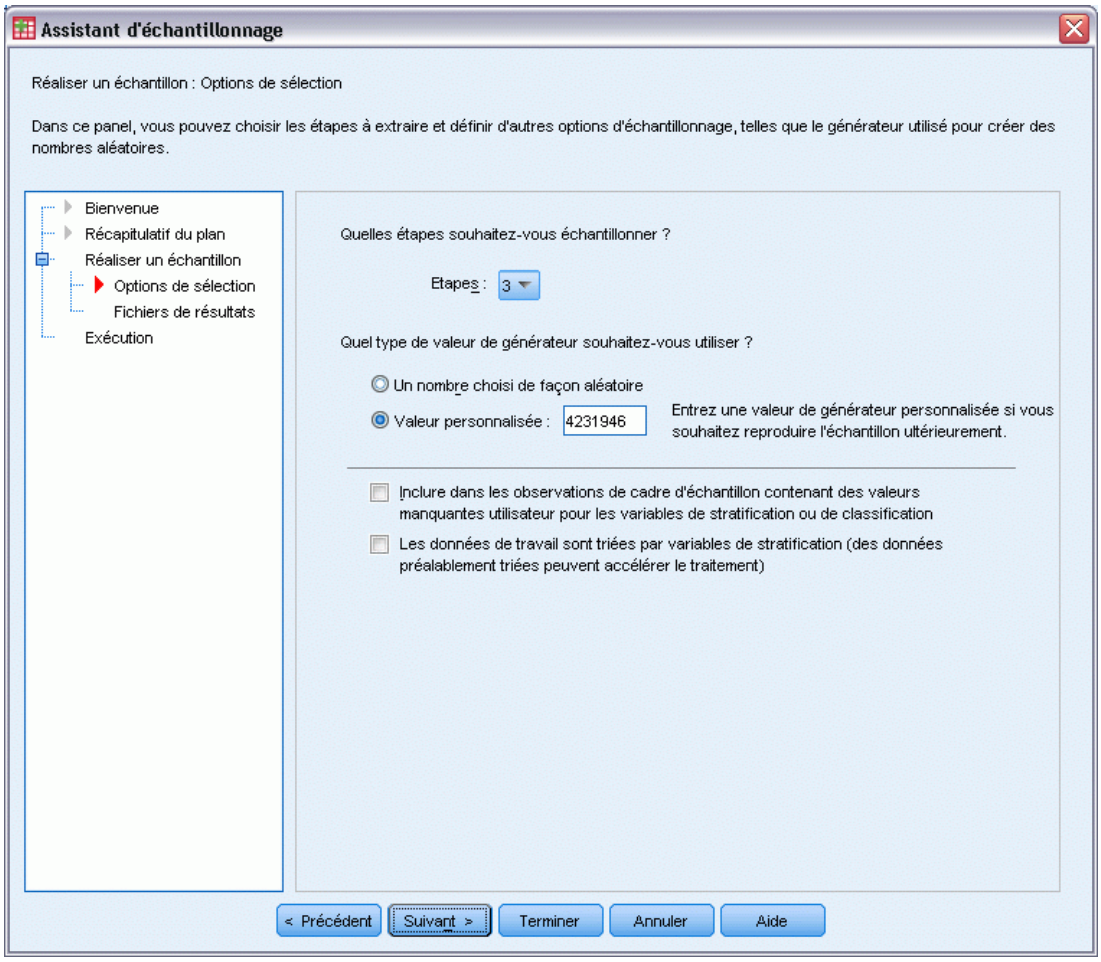

- Sélectionnez Réaliser un échantillon, accédez à l'emplacement auquel vous avez enregistré le fichier de plan, et sélectionnez le fichier de plan demo.csplan que vous avez créé.
- Cliquez sur Suivant.

Figure 13-28
*Etape Récapitulatif du plan de l'assistant d'échantillonnage (phase 3)*



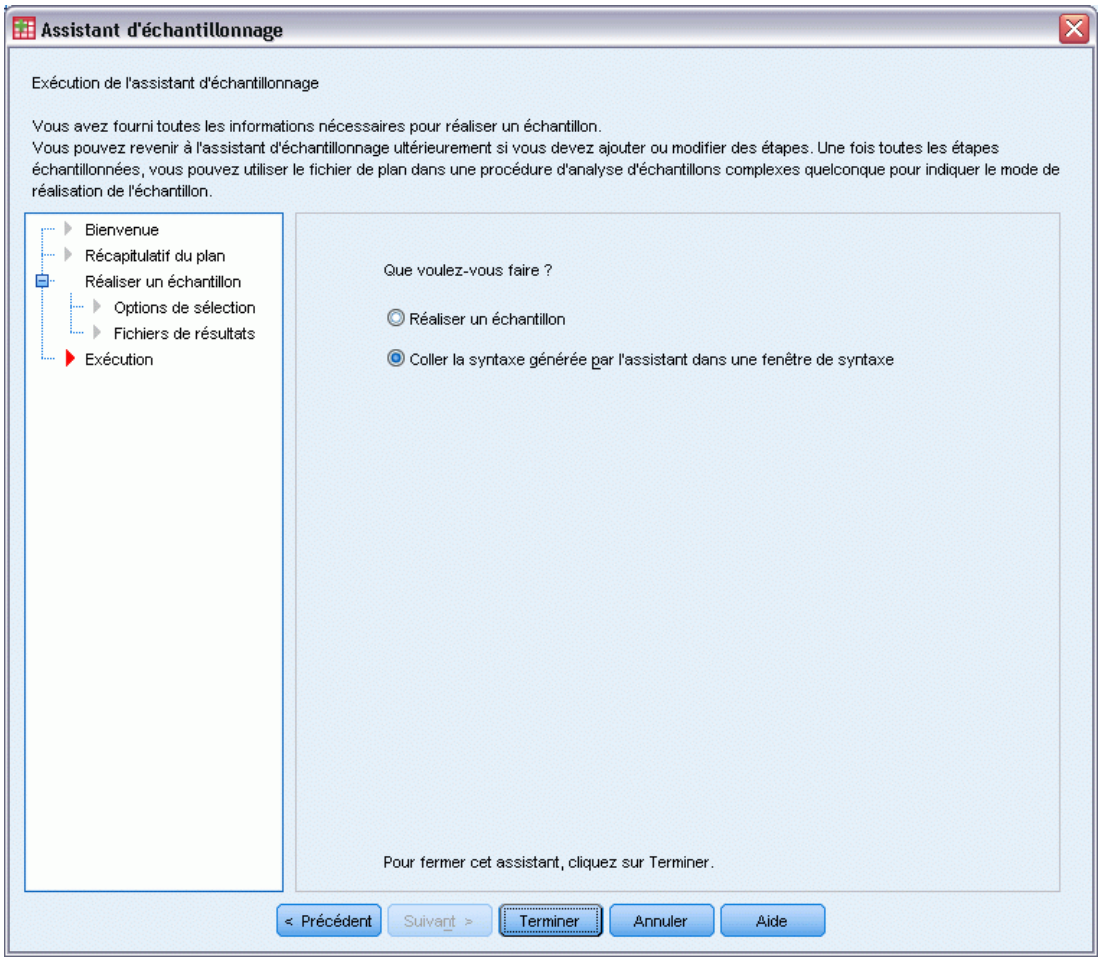

- Sélectionnez 1, 2 comme phases déjà échantillonnées.
- Cliquez sur Suivant.

Figure 13-29
*Etape Réalisation de l'échantillon : Options de sélection de l'assistant d'échantillonnage*




- Sélectionnez Valeur personnalisée comme type de générateur aléatoire à utiliser, puis entrez la valeur 4231946.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Réalisation de l'échantillon : Fichiers de résultats.

Figure 13-30
*Etape Fin de l'assistant d'échantillonnage*




- Sélectionnez Coller la syntaxe générée par l'assistant dans une fenêtre de syntaxe.
- Cliquez sur Terminer.

La syntaxe suivante est générée :

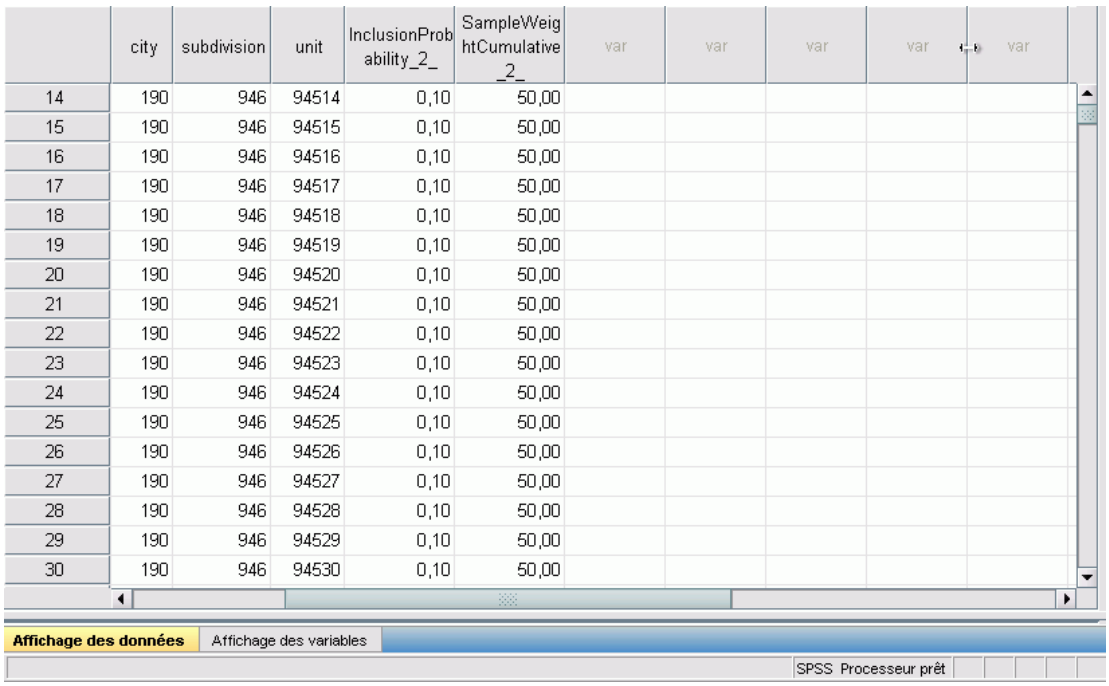
```
* Assistant d'échantillonnage.
CSSELECT
 /PLAN FILE='demo.csplan'
 /CRITERIA STAGES = 3 SEED = 4231946
 /CLASSMISSING EXCLUDE
 /DATA RENAMEVARS
 /PRINT SELECTION.
```

Dans ce cas, l'impression du récapitulatif d'échantillonnage génère un tableau encombrant qui provoque des problèmes dans l'éditeur de résultats. Pour désactiver l'affichage du récapitulatif d'échantillonnage, remplacez `SELECTION` par `CPS` dans la sous-commande `PRINT`. Exécutez ensuite la syntaxe dans la fenêtre de syntaxe.

Ces sélections réalisent un échantillon en fonction de la troisième phase du plan d'échantillonnage *demo.csplan*.

## *Résultats de l'échantillonnage*

Figure 13-31
*Editeur de données contenant les résultats de l'échantillonnage*



| | city | subdivision | unit | InclusionProbability_2_ | SampleWeightCumulative_2_ | var | var | var | var | var |
|---|---|---|---|---|---|---|---|---|---|---|
| 14 | 190 | 946 | 94514 | 0,10 | 50,00 | | | | | |
| 15 | 190 | 946 | 94515 | 0,10 | 50,00 | | | | | |
| 16 | 190 | 946 | 94516 | 0,10 | 50,00 | | | | | |
| 17 | 190 | 946 | 94517 | 0,10 | 50,00 | | | | | |
| 18 | 190 | 946 | 94518 | 0,10 | 50,00 | | | | | |
| 19 | 190 | 946 | 94519 | 0,10 | 50,00 | | | | | |
| 20 | 190 | 946 | 94520 | 0,10 | 50,00 | | | | | |
| 21 | 190 | 946 | 94521 | 0,10 | 50,00 | | | | | |
| 22 | 190 | 946 | 94522 | 0,10 | 50,00 | | | | | |
| 23 | 190 | 946 | 94523 | 0,10 | 50,00 | | | | | |
| 24 | 190 | 946 | 94524 | 0,10 | 50,00 | | | | | |
| 25 | 190 | 946 | 94525 | 0,10 | 50,00 | | | | | |
| 26 | 190 | 946 | 94526 | 0,10 | 50,00 | | | | | |
| 27 | 190 | 946 | 94527 | 0,10 | 50,00 | | | | | |
| 28 | 190 | 946 | 94528 | 0,10 | 50,00 | | | | | |
| 29 | 190 | 946 | 94529 | 0,10 | 50,00 | | | | | |
| 30 | 190 | 946 | 94530 | 0,10 | 50,00 | | | | | |



Vous pouvez visualiser les résultats de l'échantillonnage dans l'éditeur de données. Trois nouvelles variables ont été enregistrées dans le fichier de travail. Ces variables représentent les probabilités d'insertion et les pondérations cumulatives d'échantillonnage de la troisième phase, ainsi que les pondérations d'échantillonnage finales. Ces nouvelles pondérations prennent en compte les pondérations calculées au cours de l'échantillonnage des deux premières phases.

- Les unités contenant des valeurs pour ces variables ont été sélectionnées pour l'échantillon.
- Les unités contenant des valeurs manquantes par défaut pour les variables n'ont pas été sélectionnées.

La société utilisera désormais ses ressources pour obtenir des informations d'enquête pour les unités de logement sélectionnées dans l'échantillon. Une fois que ces enquêtes sont disponibles, vous pouvez traiter l'échantillon avec les procédures d'analyse d'échantillons complexes, à l'aide du plan d'échantillonnage *demo.csplan* pour indiquer les spécifications d'échantillonnage.

# Echantillonnage avec probabilité proportionnelle à la taille (PPS - Probability proportional to size)

Des élus étudiant un projet de loi devant l'assemblée législative souhaitent savoir si ce projet est populaire auprès des électeurs et déterminer le lien existant entre cette popularité et la répartition démographique des électeurs. Les enquêteurs conçoivent et mènent des entretiens en fonction d'un plan d'échantillonnage complexe.

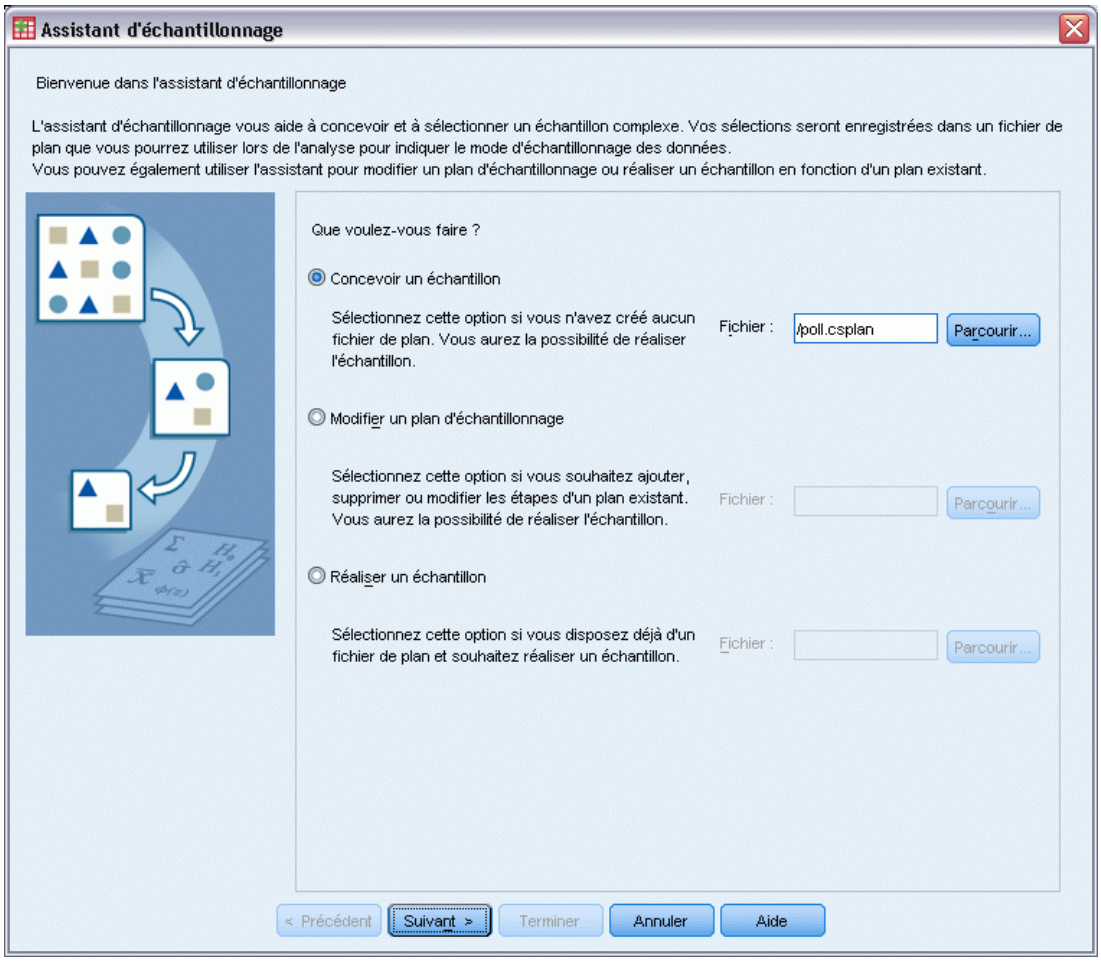
La liste des électeurs enregistrés est recueillie dans le fichier *poll_cs.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*. Utilisez l'assistant d'échantillonnage des échantillons complexes pour sélectionner un échantillon en vue d'une analyse ultérieure.

## Utilisation de l'assistant

- Pour exécuter l'assistant d'échantillonnage des échantillons complexes, sélectionnez dans le menu l'option suivante :

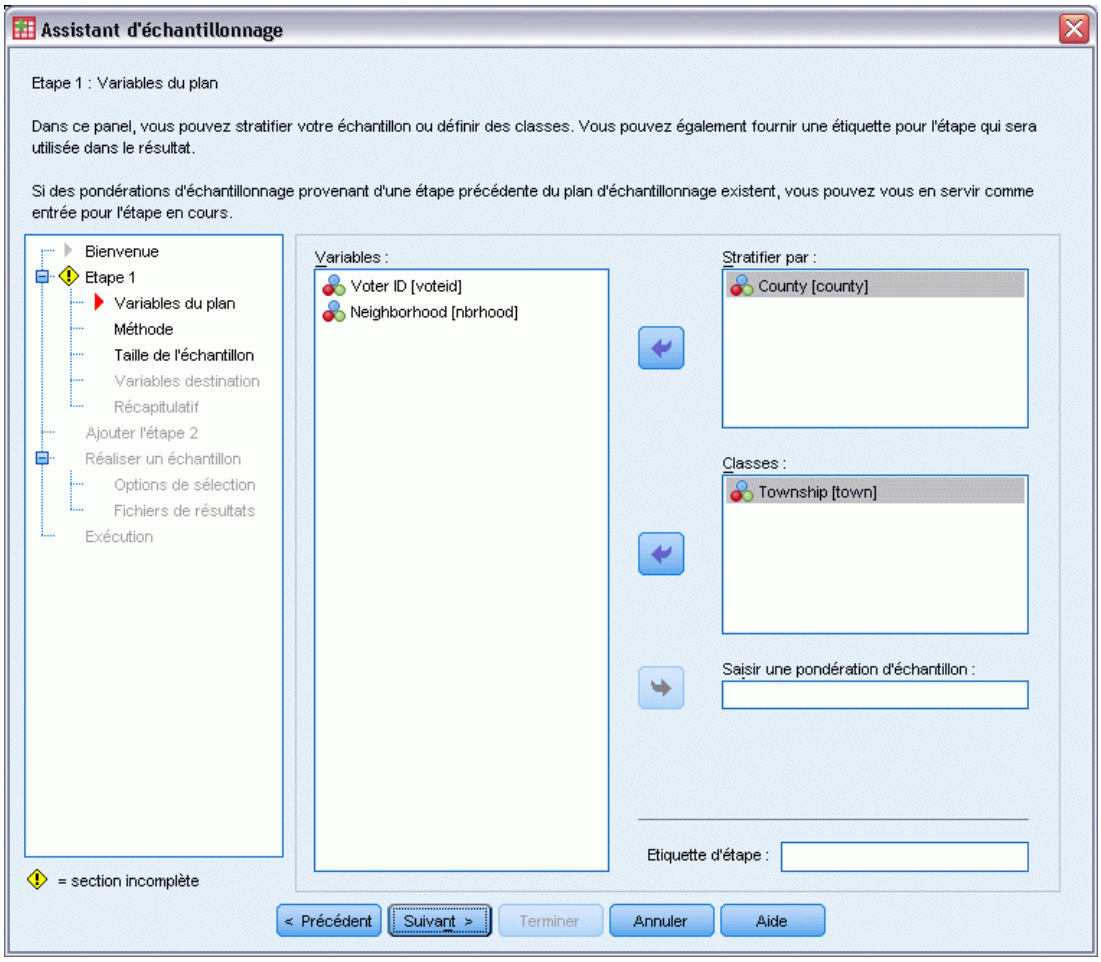
  Analyse > Echantillons complexes > Sélectionner un échantillon...

Figure 13-32
*Etape Bienvenue de l'assistant d'échantillonnage*




- Sélectionnez Concevoir un échantillon, accédez à l'emplacement auquel vous souhaitez enregistrer le fichier, et saisissez poll.csplan comme nom du fichier de plan.
- Cliquez sur Suivant.

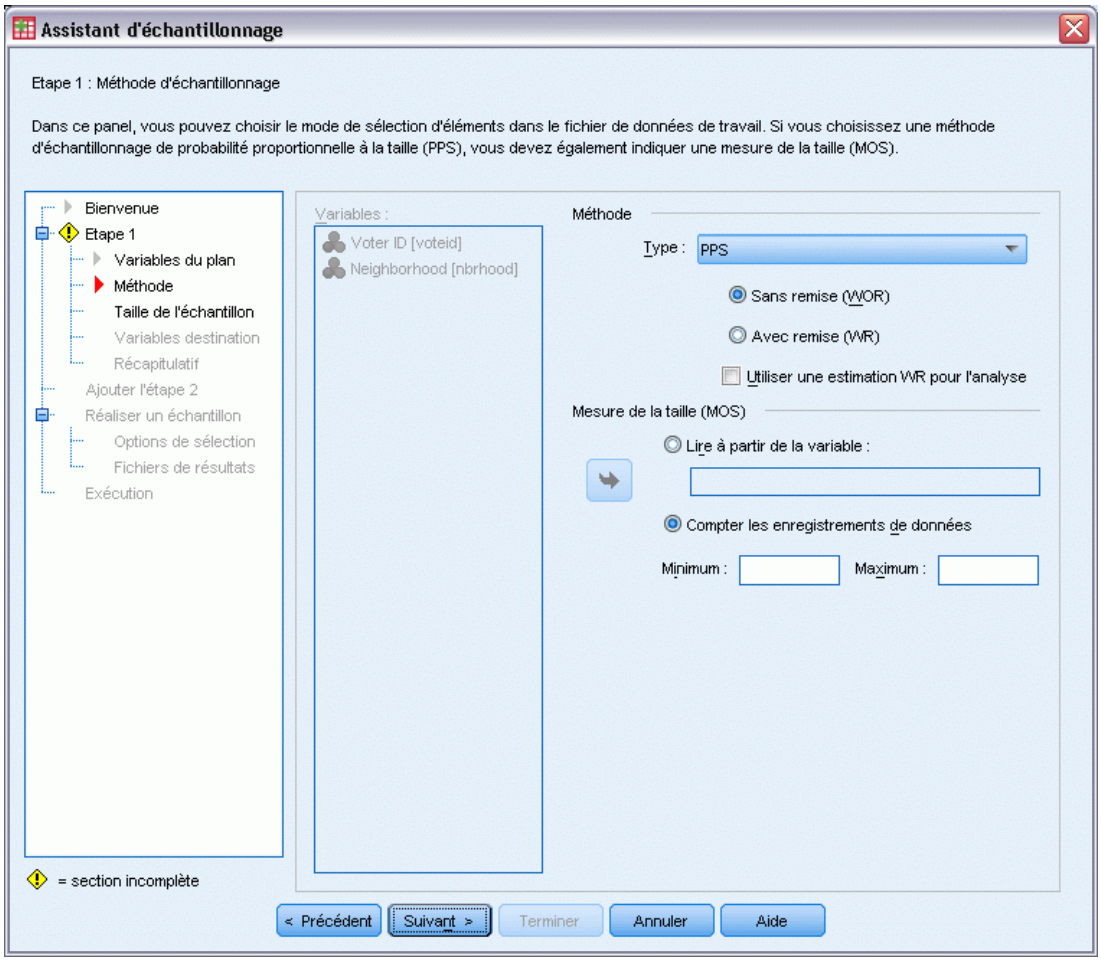
Figure 13-33
*Etape Variables de plan de l'assistant d'échantillonnage (phase 1)*




- Sélectionnez *Comté* comme variable de stratification.
- Sélectionnez *Commune* comme variable de grappe.
- Cliquez sur Suivant.

Cette structure de plan indique que des échantillons indépendants sont réalisés pour chaque comté. Au cours de cette étape, les communes sont tracées en tant qu'unité d'échantillonnage principale.

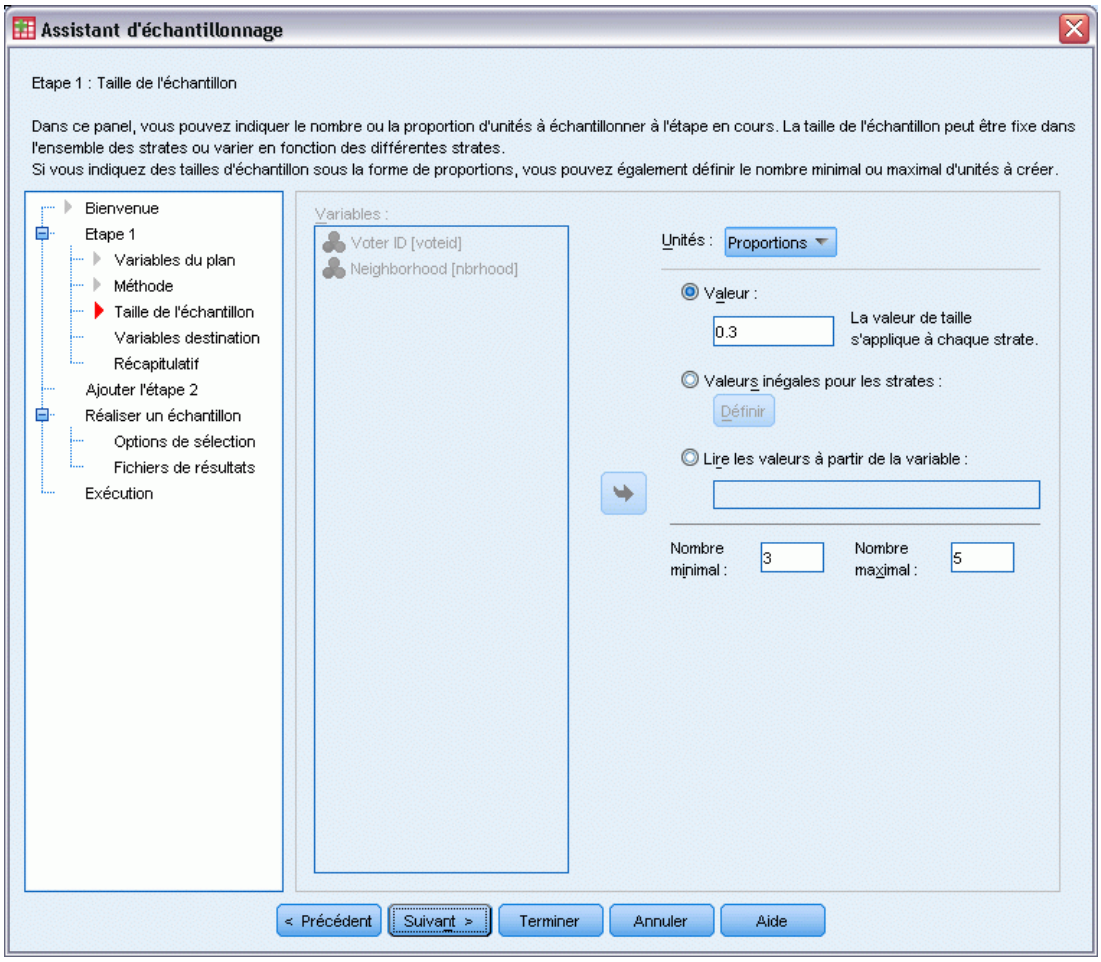
Figure 13-34
*Etape Méthode de l'assistant d'échantillonnage (phase 1)*




- Sélectionnez PPS comme méthode d'échantillonnage.
- Sélectionnez Compter les enregistrements de données comme mesure de la taille.
- Cliquez sur Suivant.

Dans chaque comté, les communes sont tracées sans remplacement avec une probabilité proportionnelle au nombre d'enregistrements pour chaque commune. Utiliser une méthode PPS génère des probabilités d'échantillonnage conjointes pour les communes. Vous indiquerez l'emplacement d'enregistrement de ces valeurs à l'étape des fichiers résultats.

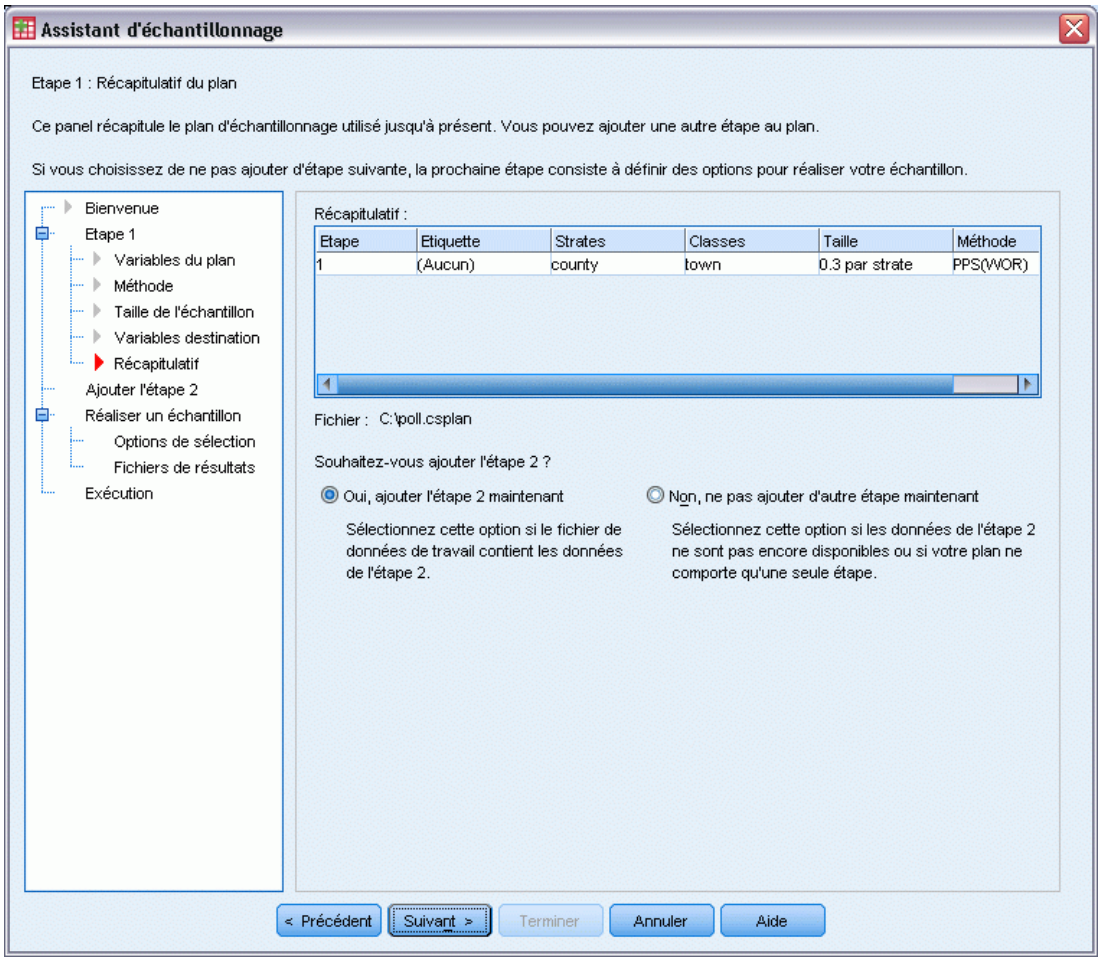
Figure 13-35
*Etape Taille de l'échantillon de l'assistant d'échantillonnage (phase 1)*




- Sélectionnez Proportions dans la liste déroulante Unités.
- Tapez 0,3 comme valeur de la proportion de communes à sélectionner par comté au cours de cette phase.

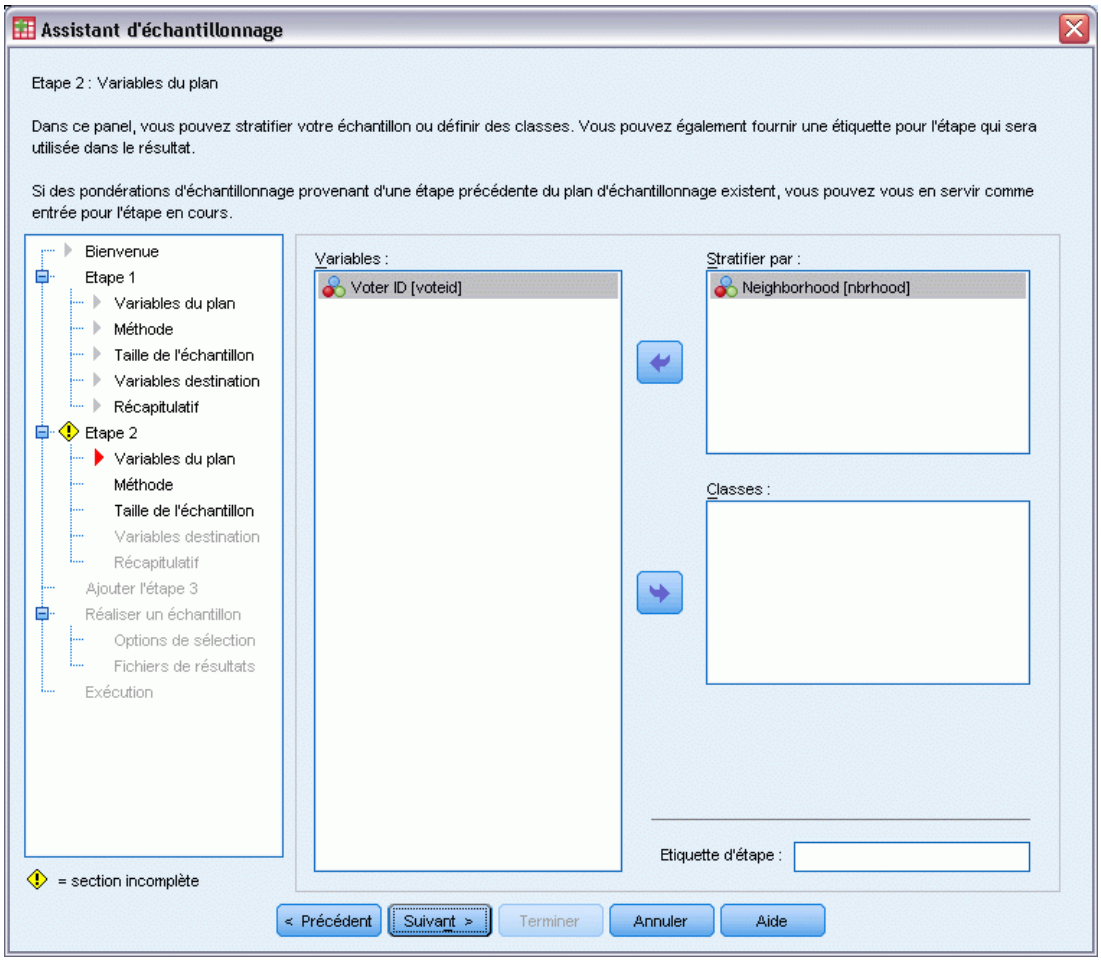
  Les législateurs du comté Ouest signalent que leur comté contient moins de communes que les autres. Pour garantir une représentation adéquate, ils souhaitent établir un minimum de trois communes échantillonnées dans chaque comté.
- Tapez 3 comme nombre minimal de communes à sélectionner et 5 comme nombre maximal.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Variables destination.

Figure 13-36
*Etape Récapitulatif du plan de l'assistant d'échantillonnage (phase 1)*




- Sélectionnez Oui, ajouter l'étape 2 maintenant.
- Cliquez sur Suivant.

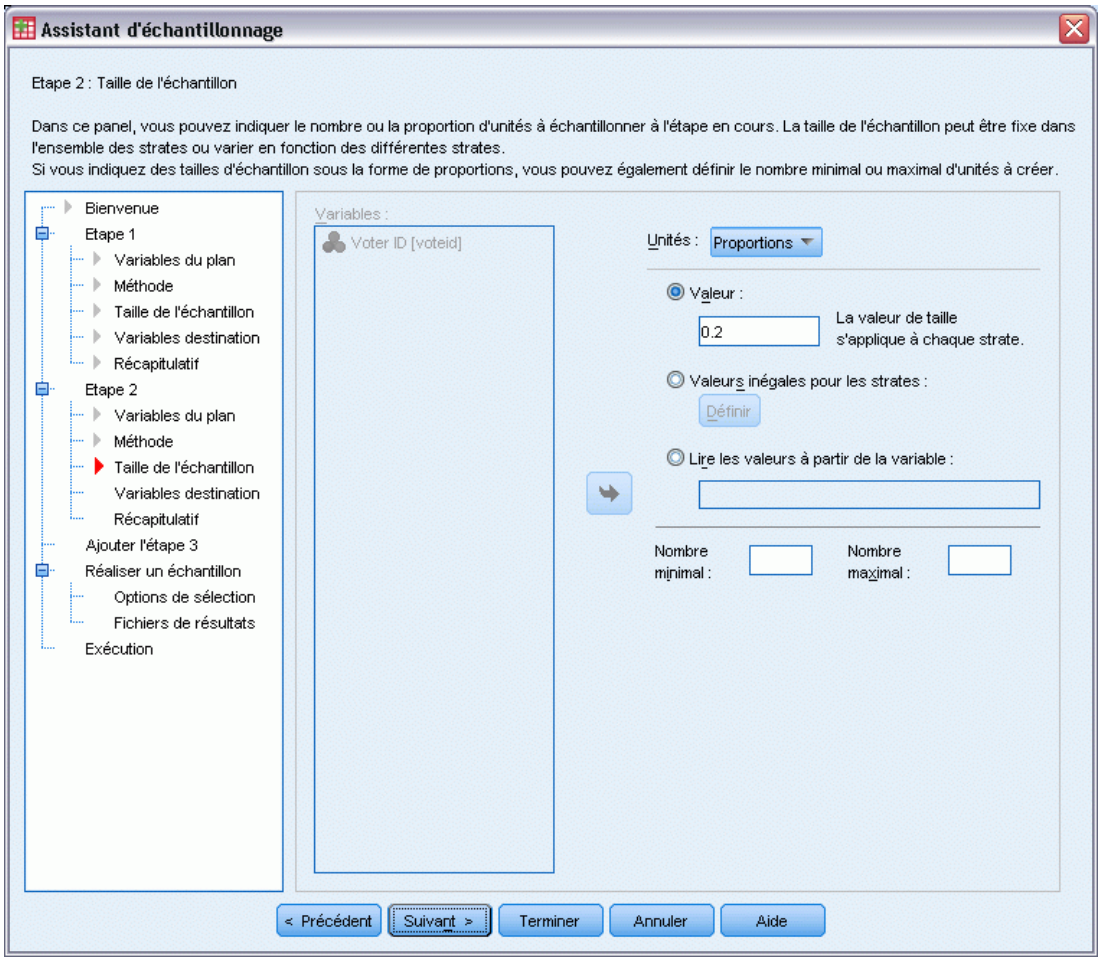
Figure 13-37
*Etape Variables de plan de l'assistant d'échantillonnage (phase 2)*




- Sélectionnez *Voisinage* comme variable de stratification.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Méthode d'échantillonage.

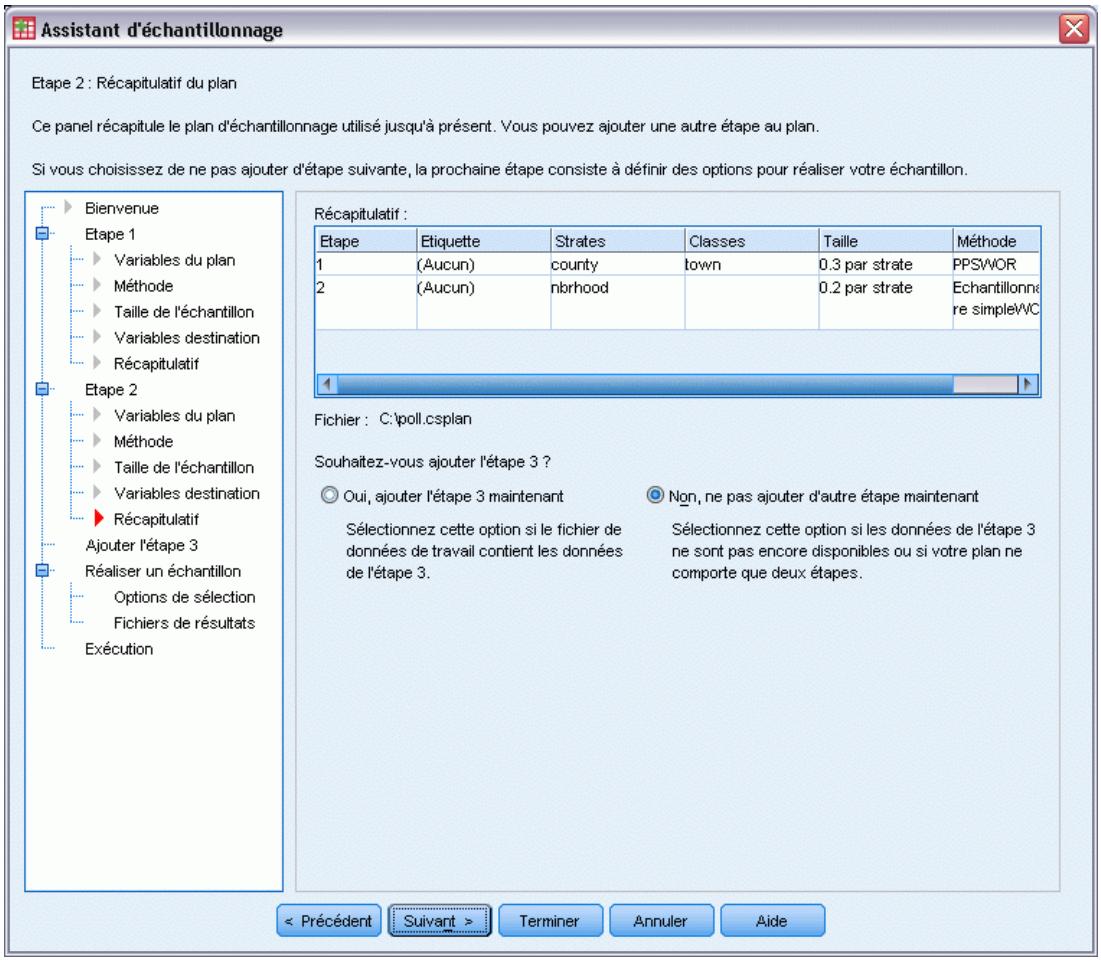
  Cette structure de plan indique que des échantillons indépendants sont réalisés pour chaque quartier des communes tracées au cours de la phase 1. Au cours de cette phase, les électeurs sont tracés en tant qu'unité d'échantillonnage principale à l'aide de la méthode Echantillonnage aléatoire simple sans remplacement.

Figure 13-38
*Etape Taille de l'échantillon de l'assistant d'échantillonnage (phase 2)*



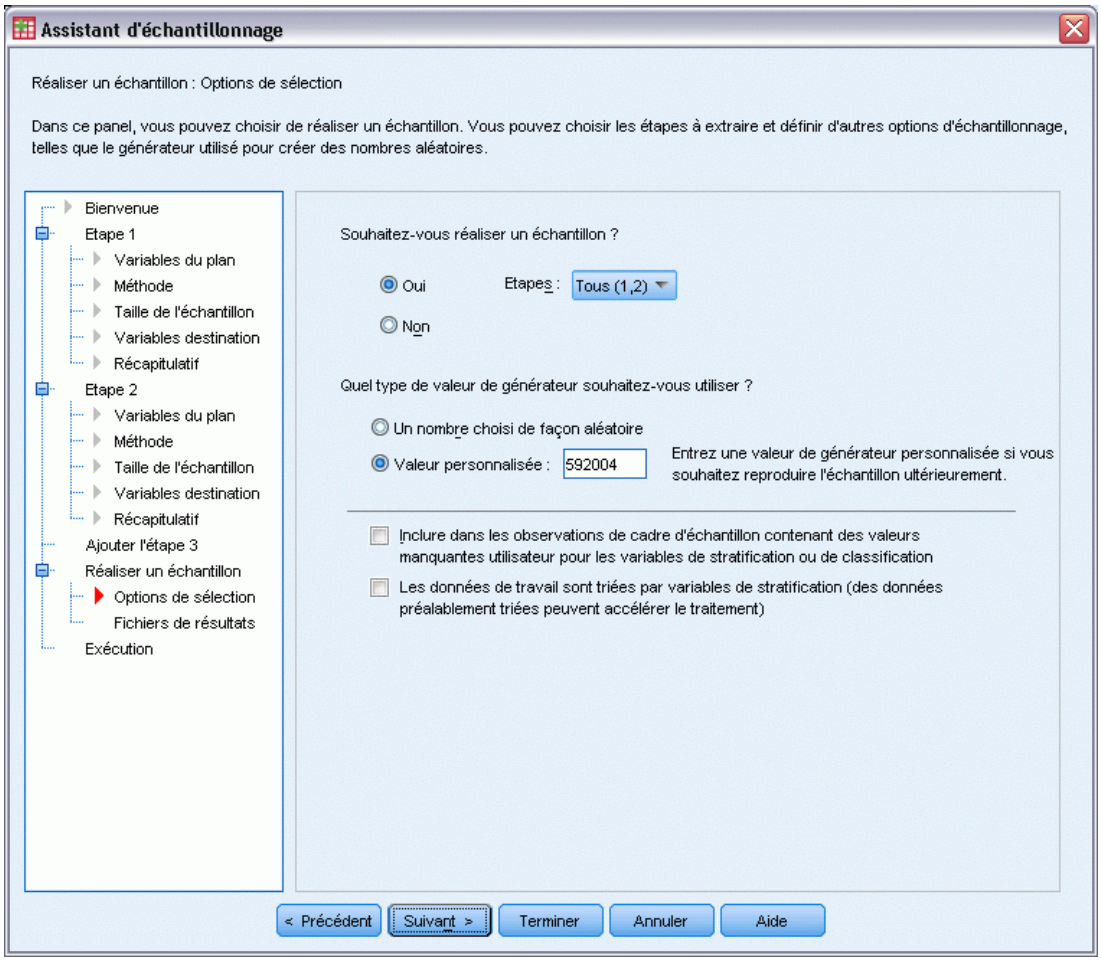

- Sélectionnez Proportions dans la liste déroulante Unités.
- Entrez 0.2 comme valeur de la proportion d'unités à échantillonner dans chaque strate.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Variables destination.

Figure 13-39
*Etape Récapitulatif du plan de l'assistant d'échantillonnage (phase 2)*




► Consultez le plan d'échantillonnage, puis cliquez sur Suivant.

Figure 13-40
*Etape Réalisation de l'échantillon : Options de sélection de l'assistant d'échantillonnage*

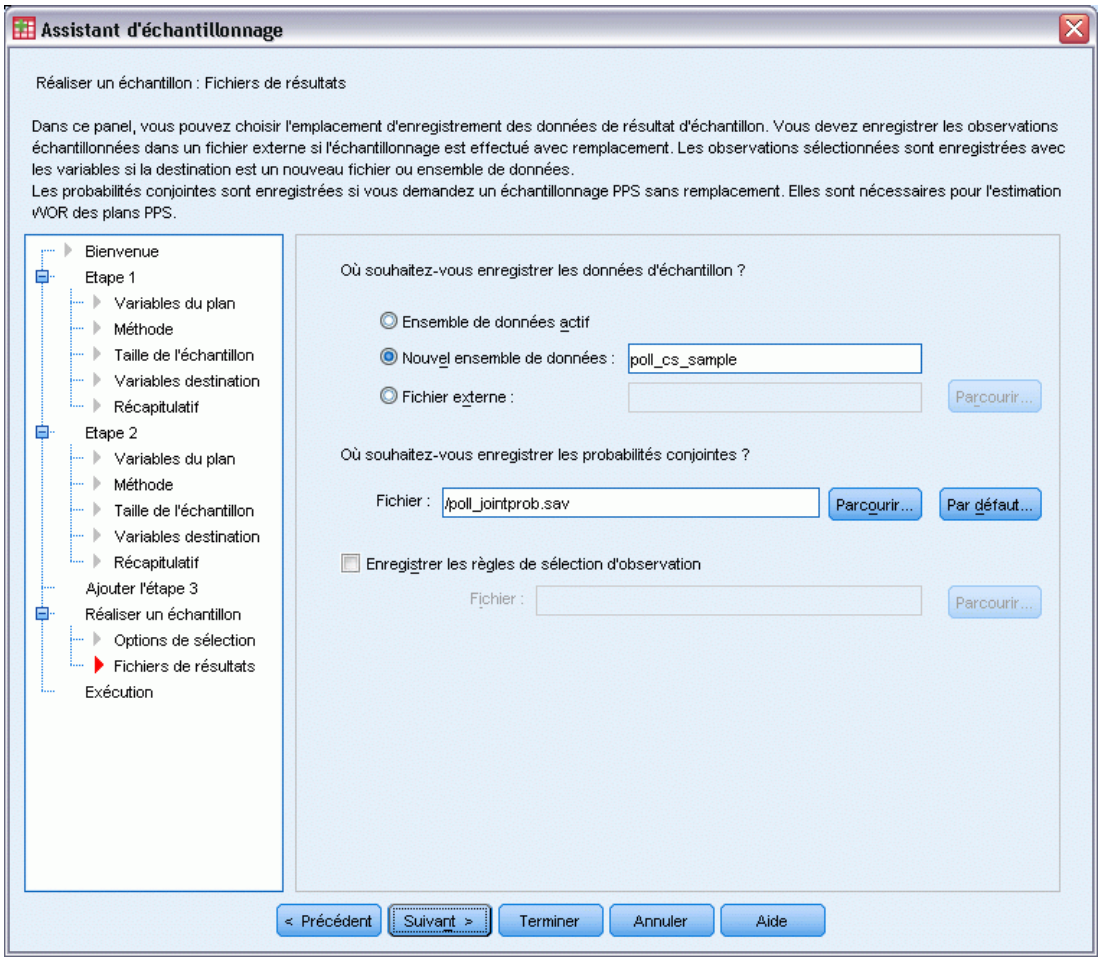



- Sélectionnez Valeur personnalisée comme type de générateur aléatoire à utiliser, puis entrez la valeur 592004.

  L'utilisation d'une valeur personnalisée vous permet de répliquer précisément les résultats de cet exemple.

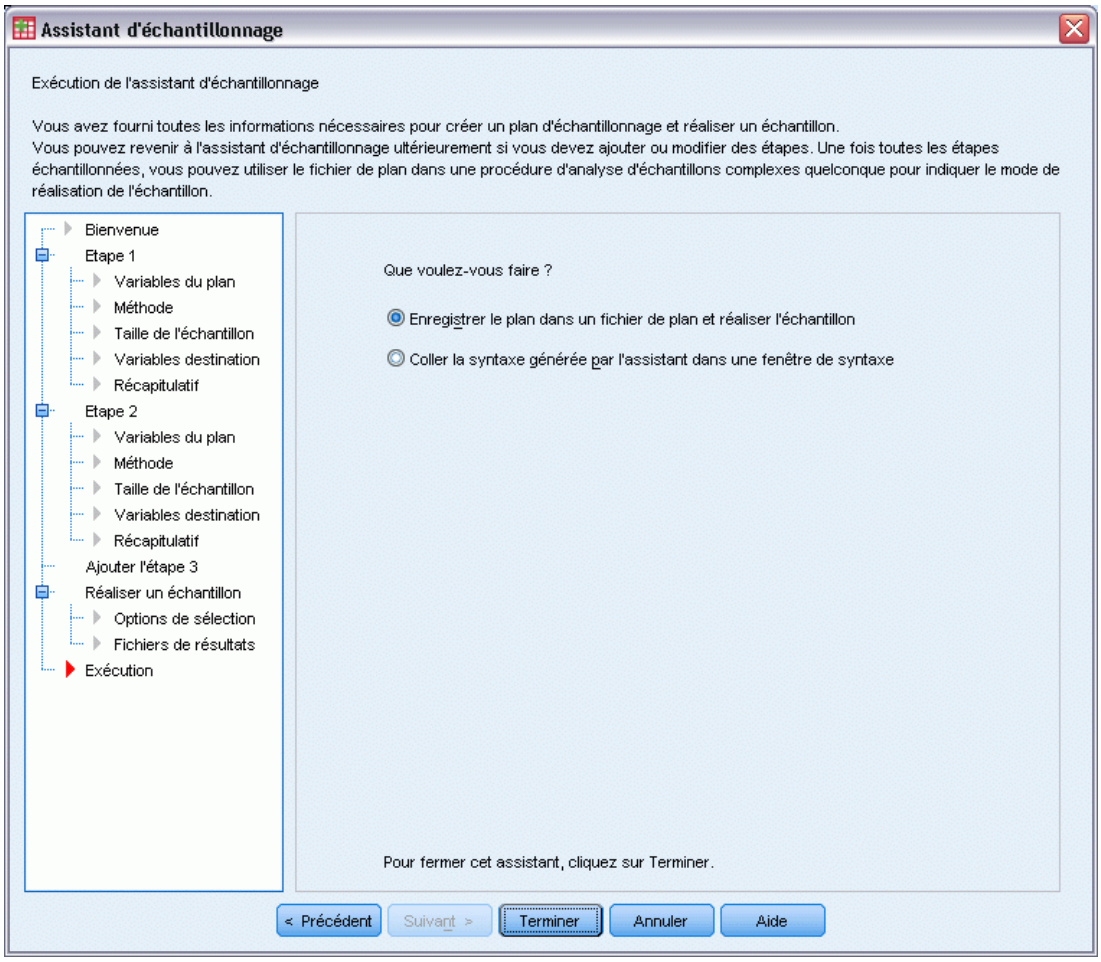
- Cliquez sur Suivant.

Figure 13-41
*Etape Réalisation de l'échantillon : Options de sélection de l'assistant d'échantillonnage*




- Choisissez d'enregistrer l'échantillon dans un nouvel ensemble de données, puis saisissez poll_cs_sample comme nom de l'ensemble de données.
- Accédez à l'emplacement auquel vous souhaitez enregistrer les probabilités conjointes, et saisissez poll_jointprob.sav comme nom du fichier de probabilités conjointes.
- Cliquez sur Suivant.

Figure 13-42
*Etape Fin de l'assistant d'échantillonnage*




▶ Cliquez sur Terminer.

Ces sélections génèrent le fichier de plan d'échantillonnage *poll.csplan* et réalisent un échantillon en fonction de ce plan, enregistrent les résultats d'échantillonnage dans le nouveau fichier de données *poll_cs_sample* et enregistrent le fichier de probabilités conjointes dans le fichier de données externe *poll_jointprob.sav*.

## *Récapitulatif du plan*

Figure 13-43
*Récapitulatif du plan*

| | | | Etape 1 | Etape 2 |
|---|---|---|---|---|
| Variables du plan | Stratification | 1 | County | Neighborhood |
| | Grappe | 1 | Township | |
| Informations sur l'échantillon | Méthode de sélection | | Echantillonnage PPS sans remise | Echantillonnage aléatoire simple sans remise |
| | Mesure de la taille | | Obtenu à partir des données | |
| | Proportion d'unités | | ,3 | ,2 |
| | Nombre minimum d'unités | | 3 | |
| | Nombre maximum d'unités | | 5 | |
| | Variables créées ou modifiées | | Probabilité Inclusion_1_ | ProbabilitéInclusion_2_ |
| | | | Pondération Echantillon Cumulée_1_ | PondérationEchantillon Cumulée_2_ |
| Informations sur l'analyse | Hypothèses de l'estimateur | | Echantillonnage de probabilité inégale sans remise (à l'aide des probabilités d'inclusion jointes) | Echantillonnage de probabilité égale sans remise |
| | Probabilité d'inclusion | | Obtenu à partir de la variable Probabilité Inclusion_1_ | Obtenu à partir de la variable ProbabilitéInclusion_2_ |

Fichier de plan : C:\poll.csplan
Variable de pondération : PondérationEchantillon_Final_

Le tableau récapitulatif vous permet de passer en revue votre plan d'échantillonnage et de vous assurer qu'il correspond à vos attentes.

## *Récapitulatif de l'échantillonnage*

Figure 13-44
*Récapitulatif*

| | Nombre d'unités échantillonnées | | Proportion d'unités échantillonnées | |
|---|---|---|---|---|
| County | Obligatoire | Réel | Obligatoire | Réel |
| Eastern | 4 | 4 | 30,0% | 30,8% |
| Central | 4 | 4 | 30,0% | 30,8% |
| Western | 3 | 3 | 30,0% | 50,0% |
| Northern | 5 | 5 | 30,0% | 33,3% |
| Southern | 3 | 3 | 30,0% | 50,0% |

Fichier de plan : C:\poll.csplan

Ce tableau récapitulatif vous permet de passer en revue la première phase de l'échantillonnage et de vérifier que celui-ci correspond au plan. Rappelez-vous que vous avez demandé un échantillon de 30 % des communes par comté. Les proportions réelles échantillonnées sont proches de 30 %, sauf dans les comtés Ouest et Sud. En effet, ces comtés ne contiennent que six communes chacun et vous avez également indiqué qu'au moins trois communes doivent être sélectionnées par comté.

Figure 13-45
*Récapitulatif*

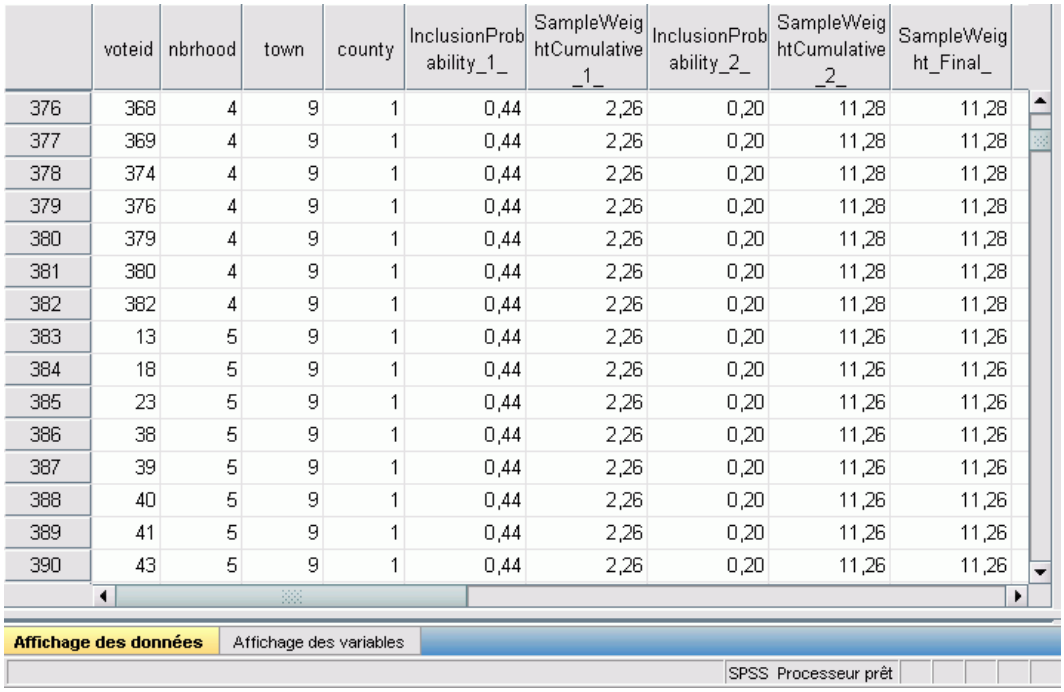
| | | | Nombre d'unités échantillonnées | | Proportion d'unités échantillonnées | |
|---|---|---|---|---|---|---|
| County | Township | Neighborhood | Obligatoire | Réel | Obligatoire | Réel |
| Eastern | 9 | 1 | 49 | 49 | 20,0% | 19,9% |
| | | 2 | 143 | 143 | 20,0% | 20,0% |
| | | 3 | 113 | 113 | 20,0% | 20,0% |
| | | 4 | 77 | 77 | 20,0% | 20,0% |
| | | 5 | 139 | 139 | 20,0% | 20,0% |
| | | 6 | 120 | 120 | 20,0% | 20,0% |
| | 10 | 1 | 149 | 149 | 20,0% | 20,1% |
| | | 2 | 117 | 117 | 20,0% | 20,0% |
| | | 3 | 116 | 116 | 20,0% | 20,0% |
| | | 4 | 69 | 69 | 20,0% | 19,9% |
| | 11 | 1 | 65 | 65 | 20,0% | 19,9% |
| | | 2 | 72 | 72 | 20,0% | 19,9% |
| | | 3 | 109 | 109 | 20,0% | 20,0% |
| | | 4 | 140 | 140 | 20,0% | 20,0% |
| | | 5 | 42 | 42 | 20,0% | 19,8% |
| | | 6 | 142 | 142 | 20,0% | 20,0% |
| | 12 | 1 | 145 | 145 | 20,0% | 20,1% |
| | | 2 | 69 | 69 | 20,0% | 20,1% |
| | | 3 | 98 | 98 | 20,0% | 20,1% |
| | | 4 | 134 | 134 | 20,0% | 20,0% |
| | | 5 | 114 | 114 | 20,0% | 20,0% |
| | | 6 | 137 | 137 | 20,0% | 19,9% |
| Central | 2 | 1 | 119 | 119 | 20,0% | 20,1% |
| | | 2 | 153 | 153 | 20,0% | 19,9% |
| | | 3 | 101 | 101 | 20,0% | 20,0% |
| | | 4 | 52 | 52 | 20,0% | 19,8% |
| | | 5 | 144 | 144 | 20,0% | 20,0% |

Fichier de plan : C:\poll.csplan

Ce tableau récapitulatif (dont la partie supérieure apparaît ici) vous permet de passer en revue la deuxième phase de l'échantillonnage. Il permet également de vérifier que l'échantillonnage correspond au plan. Conformément à vos critères, environ 20 % des électeurs ont été échantillonnés à partir de chaque quartier de chaque commune échantillonnée au cours de la première phase.

## *Résultats de l'échantillonnage*

Figure 13-46
*Editeur de données contenant les résultats de l'échantillonnage*



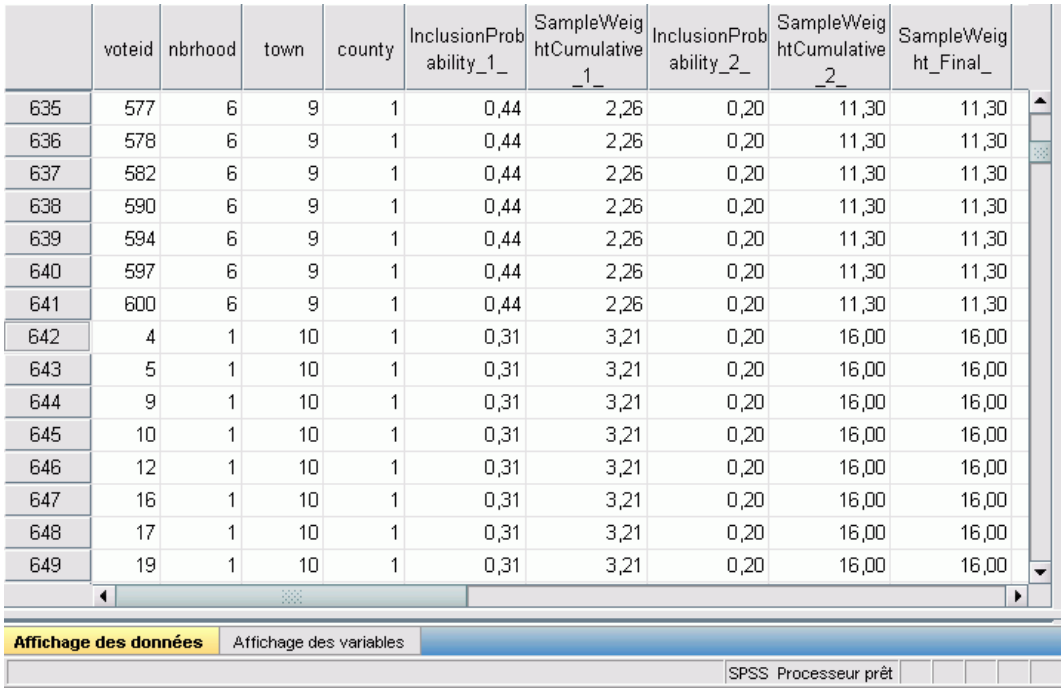
| | voteid | nbrhood | town | county | InclusionProbability_1_ | SampleWeightCumulative_1_ | InclusionProbability_2_ | SampleWeightCumulative_2_ | SampleWeight_Final_ |
|---|---|---|---|---|---|---|---|---|---|
| 376 | 368 | 4 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,28 | 11,28 |
| 377 | 369 | 4 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,28 | 11,28 |
| 378 | 374 | 4 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,28 | 11,28 |
| 379 | 376 | 4 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,28 | 11,28 |
| 380 | 379 | 4 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,28 | 11,28 |
| 381 | 380 | 4 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,28 | 11,28 |
| 382 | 382 | 4 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,28 | 11,28 |
| 383 | 13 | 5 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,26 | 11,26 |
| 384 | 18 | 5 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,26 | 11,26 |
| 385 | 23 | 5 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,26 | 11,26 |
| 386 | 38 | 5 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,26 | 11,26 |
| 387 | 39 | 5 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,26 | 11,26 |
| 388 | 40 | 5 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,26 | 11,26 |
| 389 | 41 | 5 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,26 | 11,26 |
| 390 | 43 | 5 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,26 | 11,26 |



Vous pouvez visualiser les résultats de l'échantillonnage dans le nouveau fichier de données. Cinq nouvelles variables ont été enregistrées dans le fichier de travail. Ces variables représentent les probabilités d'insertion et les pondérations cumulatives d'échantillonnage de chaque phase, ainsi que les pondérations d'échantillonnage finales. Les électeurs n'ayant pas été sélectionnés dans l'échantillon sont exclus de ce fichier de données.

Les pondérations d'échantillonnage finales sont identiques pour les électeurs situés dans le même quartier car elles sont sélectionnées en fonction d'une méthode d'échantillonnage aléatoire simple dans les quartiers. Cependant, elles sont différentes à travers les quartiers situés dans la même commune car les proportions échantillonnées ne sont pas exactement égales à 20 % dans tous les quartiers.

Figure 13-47
*Editeur de données contenant les résultats de l'échantillonnage*



| | voteid | nbrhood | town | county | InclusionProbability_1_ | SampleWeightCumulative_1_ | InclusionProbability_2_ | SampleWeightCumulative_2_ | SampleWeight_Final_ |
|---|---|---|---|---|---|---|---|---|---|
| 635 | 577 | 6 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,30 | 11,30 |
| 636 | 578 | 6 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,30 | 11,30 |
| 637 | 582 | 6 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,30 | 11,30 |
| 638 | 590 | 6 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,30 | 11,30 |
| 639 | 594 | 6 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,30 | 11,30 |
| 640 | 597 | 6 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,30 | 11,30 |
| 641 | 600 | 6 | 9 | 1 | 0,44 | 2,26 | 0,20 | 11,30 | 11,30 |
| 642 | 4 | 1 | 10 | 1 | 0,31 | 3,21 | 0,20 | 16,00 | 16,00 |
| 643 | 5 | 1 | 10 | 1 | 0,31 | 3,21 | 0,20 | 16,00 | 16,00 |
| 644 | 9 | 1 | 10 | 1 | 0,31 | 3,21 | 0,20 | 16,00 | 16,00 |
| 645 | 10 | 1 | 10 | 1 | 0,31 | 3,21 | 0,20 | 16,00 | 16,00 |
| 646 | 12 | 1 | 10 | 1 | 0,31 | 3,21 | 0,20 | 16,00 | 16,00 |
| 647 | 16 | 1 | 10 | 1 | 0,31 | 3,21 | 0,20 | 16,00 | 16,00 |
| 648 | 17 | 1 | 10 | 1 | 0,31 | 3,21 | 0,20 | 16,00 | 16,00 |
| 649 | 19 | 1 | 10 | 1 | 0,31 | 3,21 | 0,20 | 16,00 | 16,00 |



Contrairement aux électeurs de la seconde phase, les pondérations d'échantillonnage de la première phase ne sont pas identiques pour les communes au sein du même comté car elles sont sélectionnées avec probabilité proportionnelle à la taille.

Figure 13-48
*Fichier de probabilités jointes*

| | county | town | Unit_No_ | Joint_Prob_1_ | Joint_Prob_2_ | Joint_Prob_3_ | Joint_Prob_4_ | Joint_Prob_5_ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 10 | 1 | 0,31 | 0,10 | 0,11 | 0,12 | . |
| 2 | 1 | 11 | 2 | 0,10 | 0,39 | 0,15 | 0,16 | . |
| 3 | 1 | 9 | 3 | 0,11 | 0,15 | 0,44 | 0,21 | . |
| 4 | 1 | 12 | 4 | 0,12 | 0,16 | 0,21 | 0,48 | . |
| 5 | 2 | 12 | 1 | 0,22 | 0,04 | 0,07 | 0,08 | . |
| 6 | 2 | 6 | 2 | 0,04 | 0,23 | 0,07 | 0,08 | . |
| 7 | 2 | 7 | 3 | 0,07 | 0,07 | 0,41 | 0,19 | . |
| 8 | 2 | 2 | 4 | 0,08 | 0,08 | 0,19 | 0,45 | . |
| 9 | 3 | 5 | 1 | 0,58 | 0,31 | 0,32 | . | . |
| 10 | 3 | 3 | 2 | 0,31 | 0,61 | 0,36 | . | . |
| 11 | 3 | 4 | 3 | 0,32 | 0,36 | 0,63 | . | . |
| 12 | 4 | 14 | 1 | 0,26 | 0,06 | 0,06 | 0,07 | 0,09 |
| 13 | 4 | 8 | 2 | 0,06 | 0,29 | 0,07 | 0,08 | 0,10 |
| 14 | 4 | 4 | 3 | 0,06 | 0,07 | 0,29 | 0,08 | 0,10 |
| 15 | 4 | 2 | 4 | 0,07 | 0,08 | 0,08 | 0,33 | 0,12 |
| 16 | 4 | 13 | 5 | 0,09 | 0,10 | 0,10 | 0,12 | 0,43 |
| 17 | 5 | 3 | 1 | 0,74 | 0,25 | 0,27 | . | . |
| 18 | 5 | 6 | 2 | 0,25 | 0,41 | 0,13 | . | . |
| 19 | 5 | 4 | 3 | 0,27 | 0,13 | 0,43 | . | . |

Affichage des données | Affichage des variables

SPSS Processeur prêt

Le fichier *poll_jointprob.sav* contient les probabilités jointes de premier degré correspondant aux communes sélectionnées dans des comtés. *Comté* est une variable de stratification de premier degré et *Commune* est une variable de classe. Les combinaisons de ces variables identifient les unités de sondage du premier degré de manière unique. *Unité_No_* intitule les unités de sondage du premier degré de chaque strate et est utilisé pour correspondre à *Prob_Joint_1_*, *Prob_Joint_2_*, *Prob_Joint_3_*, *Prob_Joint_4_* et *Prob_Joint_5_*. Les deux premières strates ont chacune 4 unités de sondage du premier degré. C'est pourquoi les matrices de probabilités d'inclusions jointes sont de 4×4 pour ces strates et la colonne *Prob_Joint_5_* est vide pour ces lignes. De même, les dimensions des matrices de probabilités d'inclusions jointes des strates 3 et 5 sont de 3×3 et celles de la strate 4 sont de 5×5.

En parcourant les valeurs des matrices de probabilités d'inclusions jointes, vous pouvez constater qu'un fichier de probabilités jointes est nécessaire. Si vous n'utilisez pas la méthode d'échantillonnage PPS WOR, les sélections d'unités de sondage du premier degré se font indépendamment les unes des autres et leur probabilité d'inclusion jointe est simplement le produit de leurs probabilités d'inclusion. En revanche, la probabilité d'inclusion jointe des communes 9 et 10 du comté 1 est d'environ 0,11 (reportez-vous à la première observation de *Prob_Joint_3_* ou à la troisième observation de *Prob_Joint_1_*) ou inférieure au produit de leur probabilité d'inclusion (le produit de la première observation de *Prob_Joint_1_* et de la troisième observation de *Prob_Joint_3_* est 0,31×0,44 = 0,1364).

Les enquêteurs vont à présent mener des entretiens pour l'échantillon sélectionné. Une fois les résultats disponibles, vous pouvez traiter l'échantillon avec les procédures d'analyse d'échantillons complexes, à l'aide du plan d'échantillonnage *poll.csplan* pour indiquer les spécifications d'échantillonnage et du fichier *poll_jointprob.sav* pour indiquer les probabilités d'inclusion jointes nécessaires.

## Procédures apparentées

L'assistant d'échantillonnage des échantillons complexes vous permet de créer un fichier de plan d'échantillonnage et de réaliser un échantillon.

- Pour préparer un échantillon pour une analyse lorsque vous ne pouvez pas accéder au fichier de plan d'échantillonnage, utilisez l'Assistant de préparation d'analyse.

Chapitre
14

# Assistant de préparation d'analyse des échantillons complexes

L'assistant de préparation d'analyse vous guide dans la création ou la modification d'un plan d'analyse à utiliser avec les diverses procédures d'analyse des échantillons complexes. Il est plus particulièrement utile lorsque vous ne pouvez pas accéder au fichier de plan d'échantillonnage utilisé pour réaliser l'échantillon.

## Utilisation de l'assistant de préparation d'analyse des échantillons complexes pour préparer les données publiques du NHIS (National Health Interview Survey)

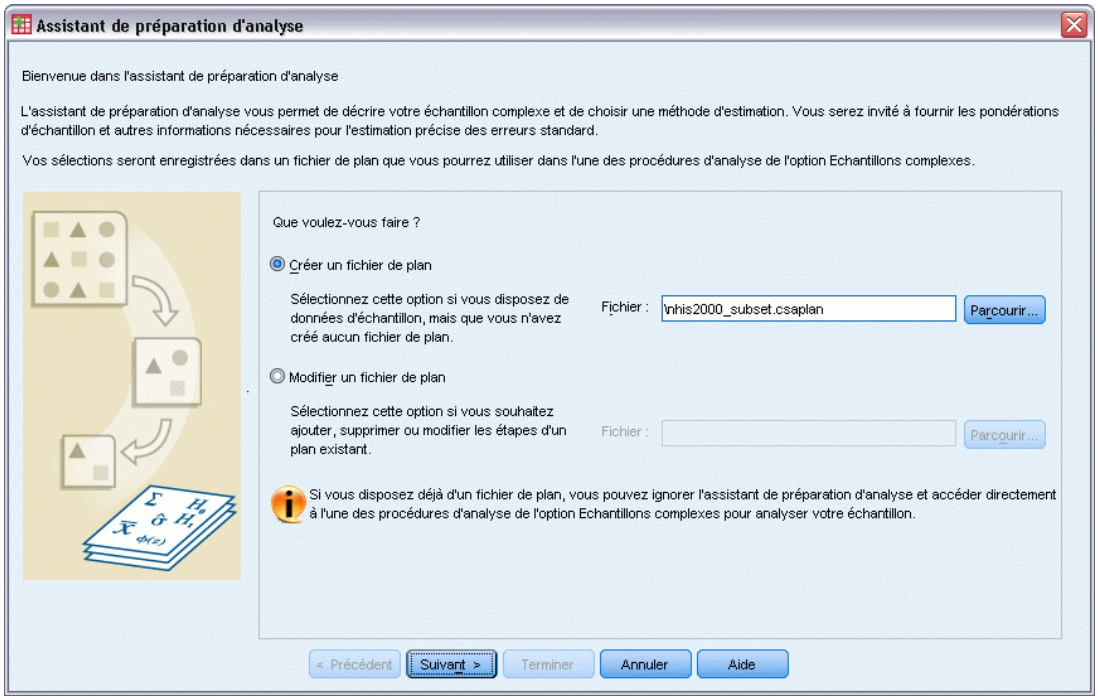
Le NHIS (National Health Interview Survey) est une enquête de grande envergure concernant la population des Etats-Unis. Des entretiens ont lieu avec un échantillon de ménages représentatifs de la population américaine. Des informations démographiques et des observations sur l'état de santé et le comportement sanitaire sont recueillies auprès des membres de chaque ménage.

Une partie de l'enquête 2000 se trouve dans le fichier *nhis2000_subset.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*. Utilisez l'assistant de préparation d'analyse des échantillons complexes pour créer un plan d'analyse pour ce fichier de données afin qu'il soit traité par les procédures d'analyse d'échantillons complexes.

### Utilisation de l'assistant

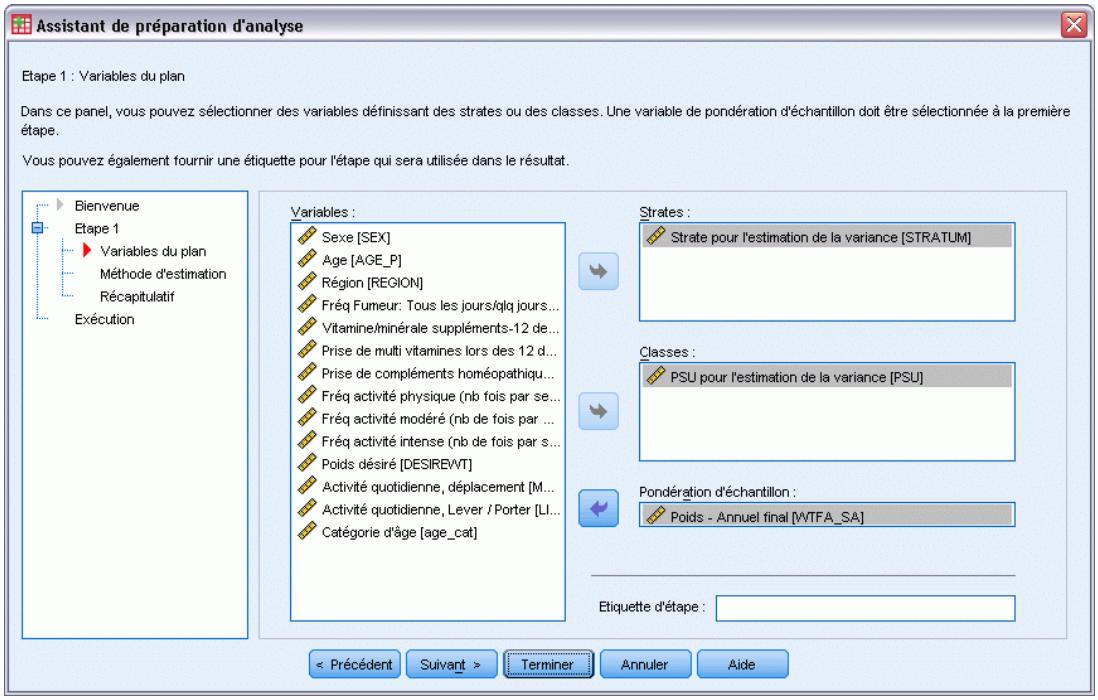
- Pour préparer un échantillon à l'aide de l'assistant de préparation d'analyse des échantillons complexes, sélectionnez dans le menu l'option suivante :
  Analyse > Echantillons complexes > Préparer pour l'analyse...

Figure 14-1
*Etape Bienvenue de l'assistant de préparation d'analyse*




- Accédez à l'emplacement auquel vous souhaitez enregistrer le fichier de plan et saisissez nhis2000_subset.csaplan comme nom du fichier de plan d'analyse.
- Cliquez sur Suivant.

Figure 14-2
*Etape Variables de plan de l'assistant de préparation d'analyse (phase 1)*




Les données sont obtenues en utilisant un échantillon complexe à plusieurs phases. Cependant, pour les utilisateurs finals, les variables de plan NHIS d'origine ont été remplacées par un ensemble simplifié de variables de plan et de pondération dont les résultats sont proches de ceux des structures de plan d'origine.

- Sélectionnez *Strate d'estimation de la variance* comme variable de strate.
- Sélectionnez *Unité de sondage du premier degré pour l'estimation de la variance* comme variable de grappe.
- Sélectionnez *Pondération – Annuelle finale* comme variable de pondération d'échantillon.
- Cliquez sur Terminer.

### *Récapitulatif*

Figure 14-3
*Récapitulatif*

| | | | Etape 1 |
|---|---|---|---|
| Variables du plan | Stratification | 1 | Strate pour 'estimation de la variance |
| | Grappe | 1 | PSU pour l'estimation de la variance |
| Informations sur l'analyse | Hypothèses de l'estimateur | | Echantillonnage avec remise |

Fichier de plan : C:\nhis2000_subset.csaplan
Variable de pondération : Poids - Annuel final
Estimateur SRS : Échantillonnage sans remplacement

Le tableau récapitulatif vous permet de passer en revue votre plan d'analyse. Le plan est composé d'une étape comprenant une variable de stratification et une variable de grappe. Une estimation avec remplacement est utilisée et le plan est enregistré dans le fichier *c:\nhis2000_subset.csaplan*. Vous pouvez utiliser ce fichier de plan pour traiter le fichier *nhis2000_subset.sav* avec les procédures d'analyse d'échantillons complexes.

## *Préparation d'une analyse lorsque les pondérations d'échantillonnage ne figurent pas dans le fichier de données*

Un responsable des prêts dispose des dossiers de ses clients, tenus selon un plan d'échantillonnage complexe. Toutefois, les pondérations d'échantillonnage ne sont pas incluses dans le fichier. Ces informations sont stockées dans le fichier *bankloan_cs_noweights.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*. En commençant par appliquer les connaissances qu'elle a du plan d'échantillonnage, la responsable souhaite utiliser l'assistant de préparation d'analyse des échantillons complexes. L'objectif est de créer un plan d'analyse pour le fichier de données, afin qu'il puisse être traité par les procédures d'analyse d'échantillons complexes.

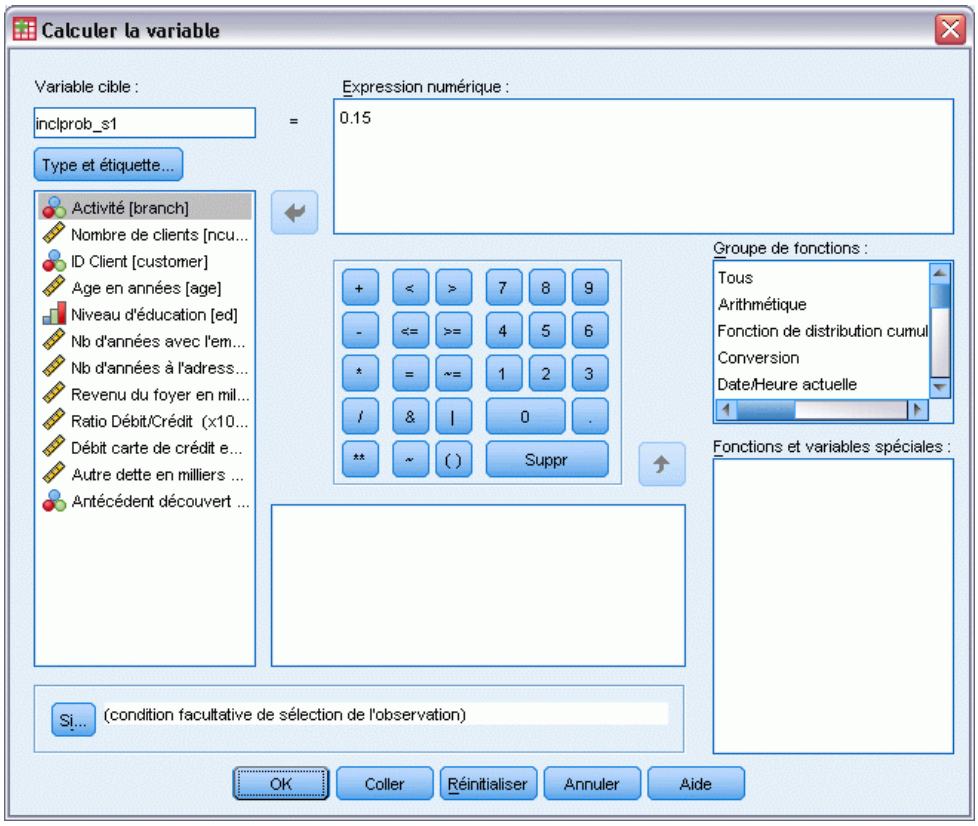
La responsable des prêts sait que les dossiers ont été sélectionnés en deux étapes. 15 succursales sur 100 sont sélectionnées à probabilité égale et sans remplacement à la première étape. Cent clients de chaque succursale ont ensuite été sélectionnés à probabilité égale et sans remplacement à la seconde étape. Les informations relatives au nombre de clients de chaque agence sont ajoutées au fichier de données. La première étape de la création d'un plan d'analyse consiste à calculer les probabilités d'inclusion à chaque étape et les pondérations d'échantillonnage finales.

### *Calcul des probabilités d'inclusion et des pondérations d'échantillonnage*

- Pour calculer les probabilités d'inclusion de la première étape, dans les différents menus, sélectionnez les éléments suivants :

  Transformer > Calculer la variable...

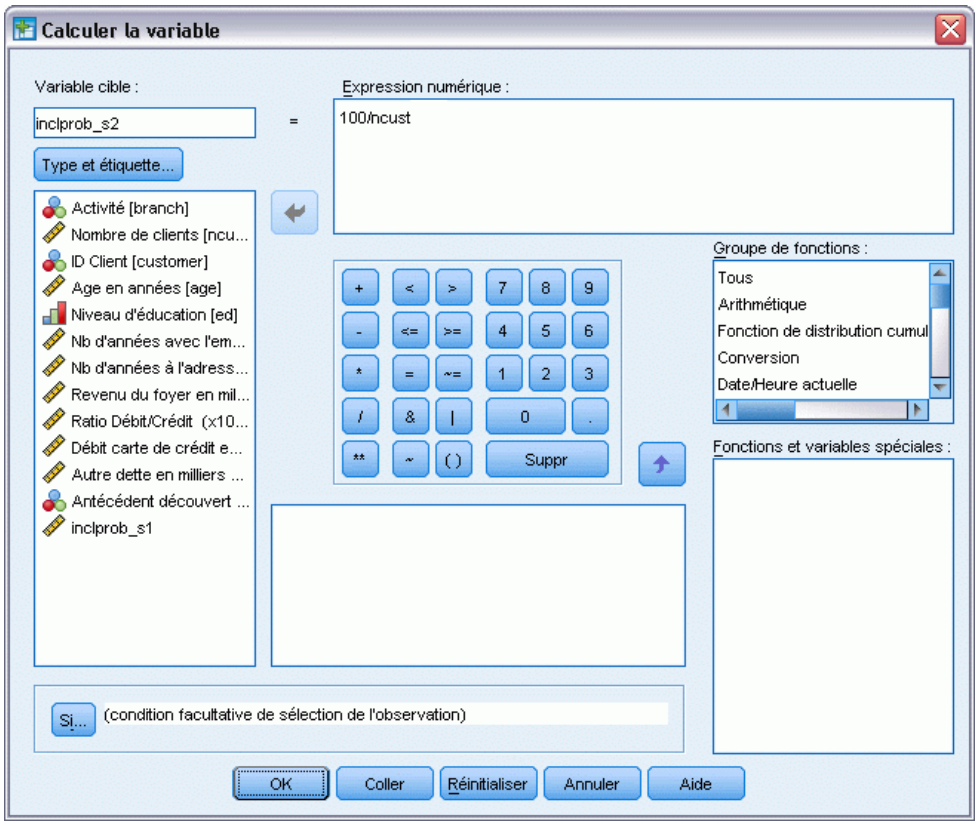
Figure 14-4
*Boîte de dialogue Calculer la variable*




Quinze succursales sur cent ont été sélectionnées sans remplacement à la première étape. Par conséquent, la probabilité de sélection d'une agence est de 15/100 = 0,15.

- Entrez la variable cible inclprob_s1.
- Entrez l'expression numérique 0,15.
- Cliquez sur OK.

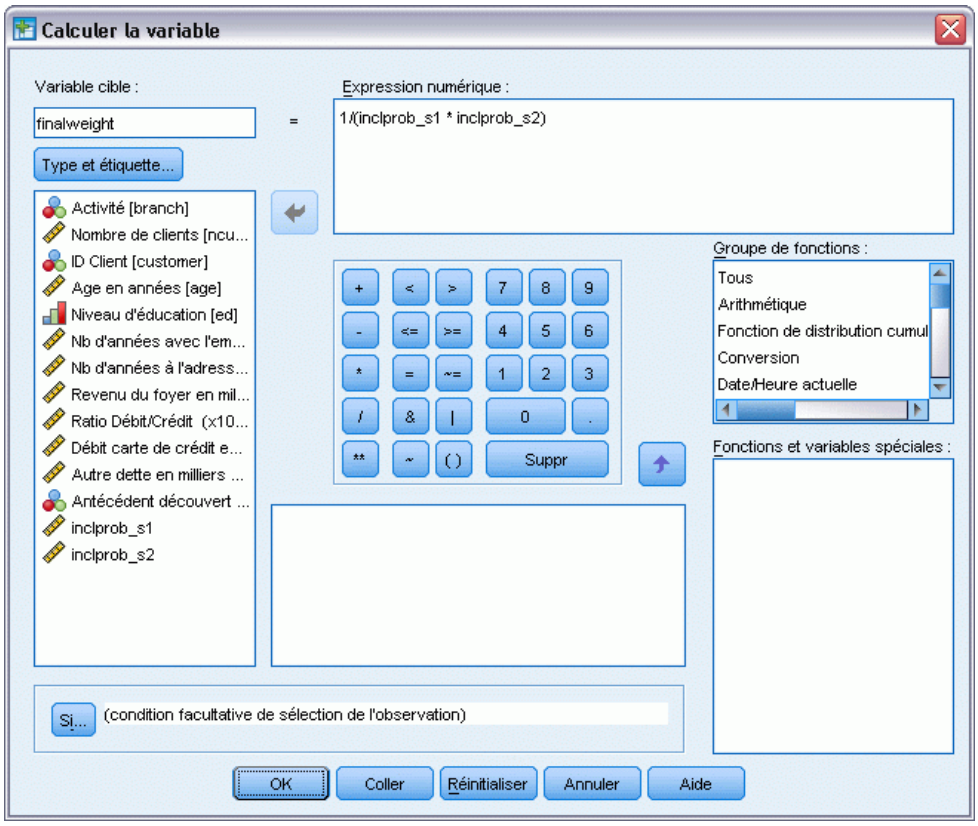
Figure 14-5
*Boîte de dialogue Calculer la variable*




Cent clients de chaque succursale ont été sélectionnés à la seconde étape. Par conséquent, la probabilité d'inclusion de cette étape pour un client d'une banque est de 100/le nombre de clients de cette banque.

- Rappelez la boîte de dialogue Calculer la variable.
- Entrez la variable cible inclprob_s2.
- Entrez l'expression numérique 100/ncust.
- Cliquez sur OK.

Figure 14-6
*Boîte de dialogue Calculer la variable*




Maintenant que vous disposez des probabilités d'inclusion de chaque étape, il est facile de calculer les pondérations d'échantillonnage finales.

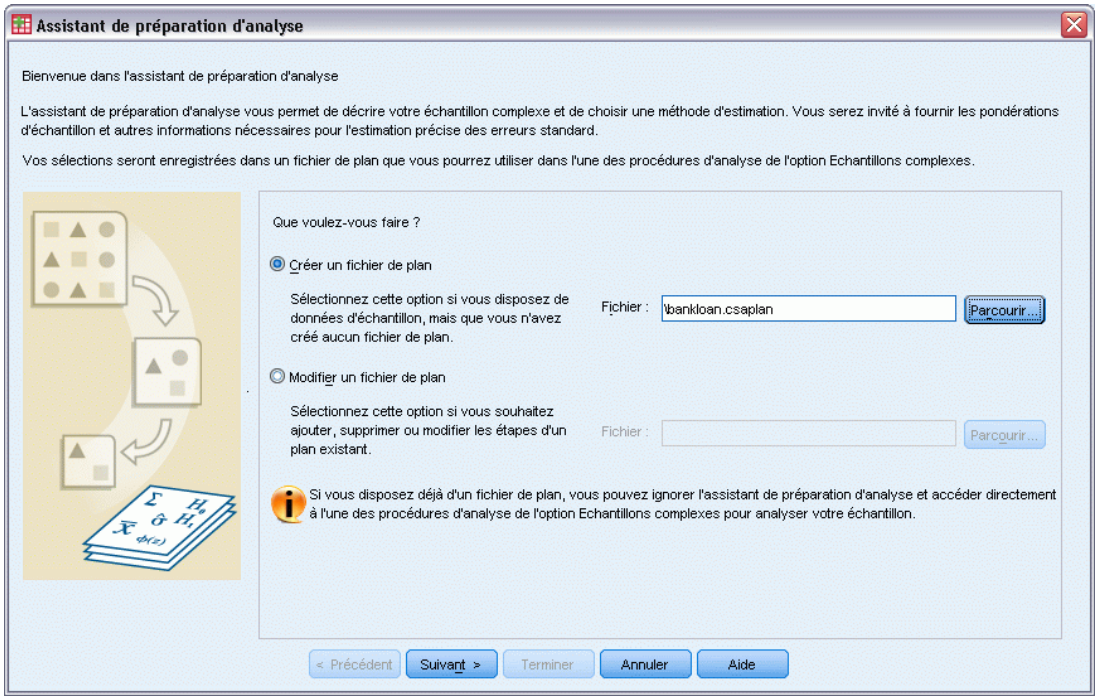
- Rappelez la boîte de dialogue Calculer la variable.
- Entrez la variable cible finalweight.
- Entrez l'expression numérique 1/(inclprob_s1 * inclprob_s2).
- Cliquez sur OK.

Vous êtes désormais prêt à créer le plan d'analyse.

## *Utilisation de l'assistant*

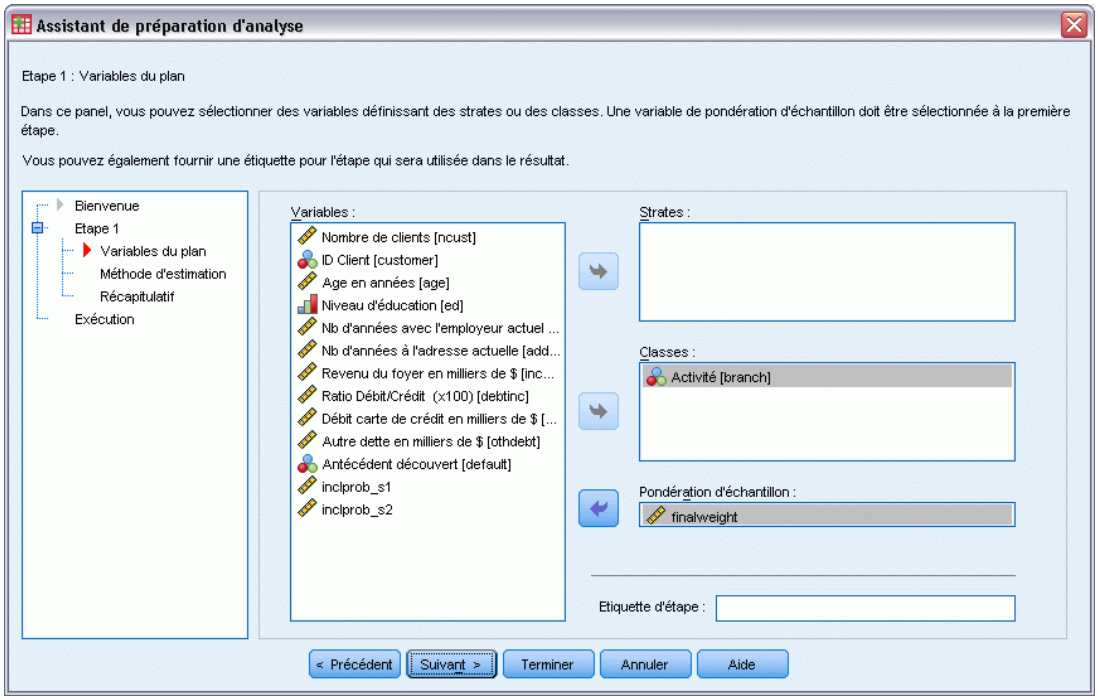
- Pour préparer un échantillon à l'aide de l'assistant de préparation d'analyse des échantillons complexes, sélectionnez dans le menu l'option suivante :
  Analyse > Echantillons complexes > Préparer pour l'analyse...

Figure 14-7
*Etape Bienvenue de l'assistant de préparation d'analyse*



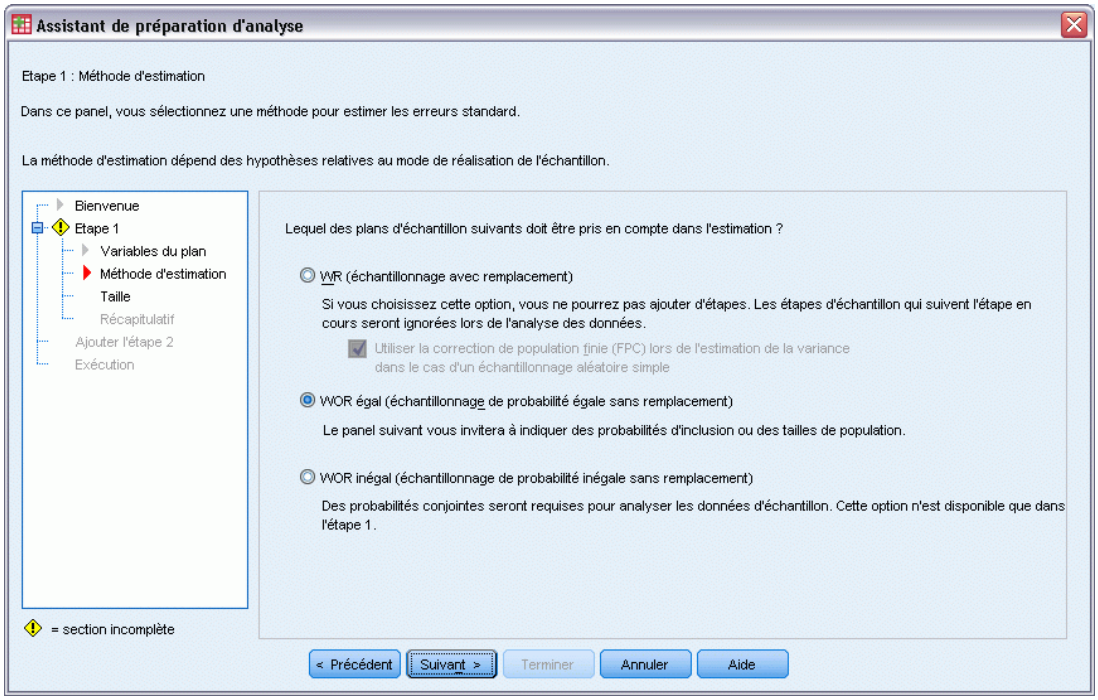

- Accédez à l'emplacement auquel vous souhaitez enregistrer le fichier de plan et saisissez bankloan.csaplan comme nom du fichier de plan d'analyse.
- Cliquez sur Suivant.

Figure 14-8
*Etape Variables de plan de l'assistant de préparation d'analyse (phase 1)*



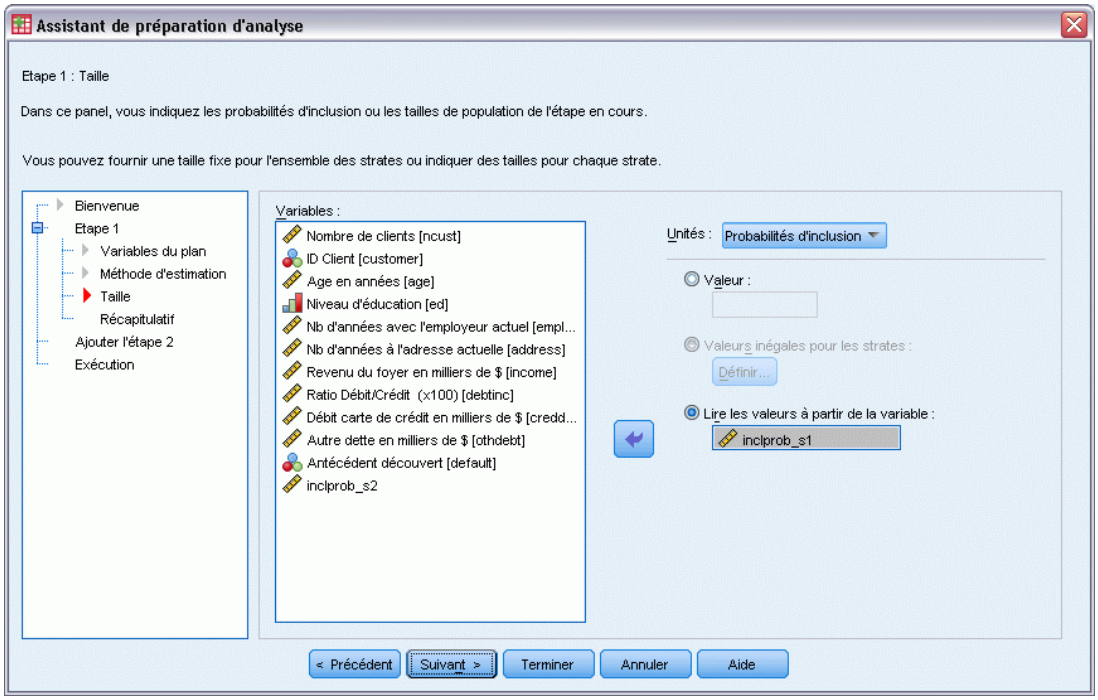

- Sélectionnez la variable de grappe *Branche*.
- Sélectionnez la variable de pondération d'échantillonnage *finalweight*.
- Cliquez sur Suivant.

Figure 14-9
*Etape Méthode d'estimation de l'assistant de préparation d'analyse (phase 1)*



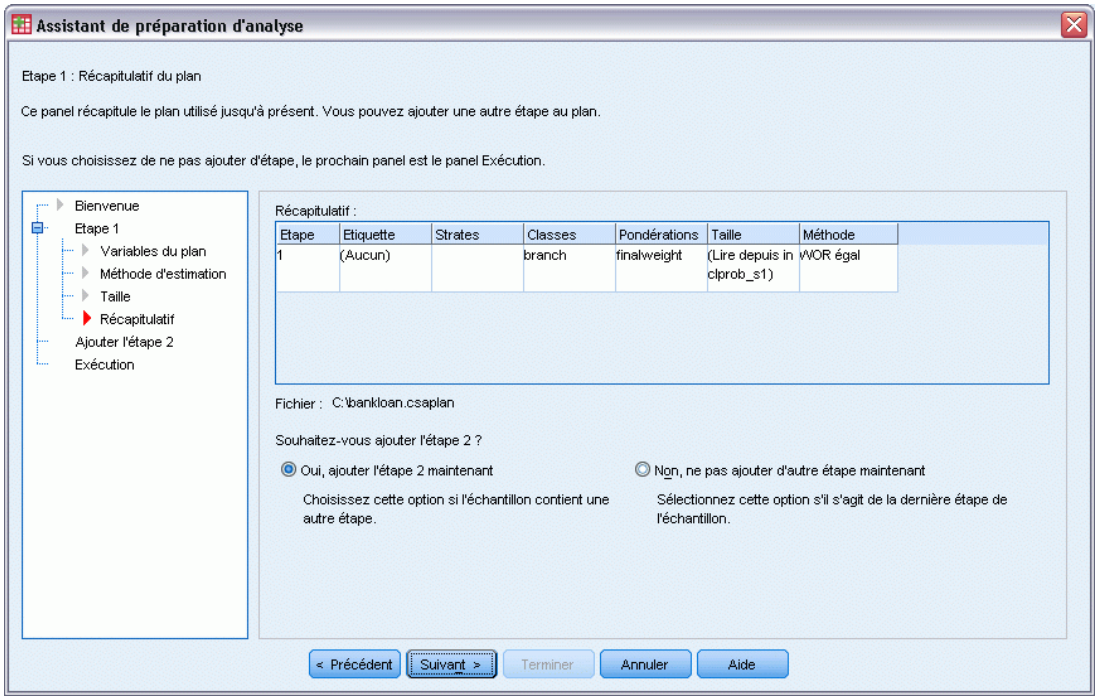

- Sélectionnez la méthode d'estimation de la première étape, WOR égal.
- Cliquez sur Suivant.

Figure 14-10
*Etape Taille de l'assistant de préparation d'analyse (phase 1)*



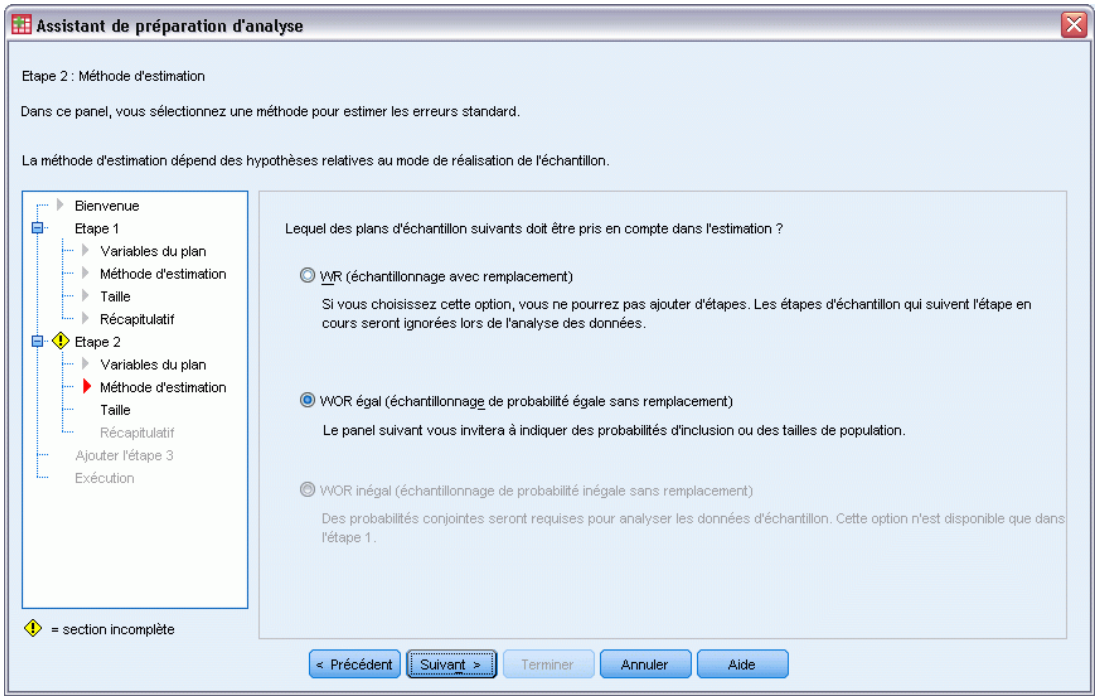

- Choisissez Lire les valeurs à partir de la variable, puis sélectionnez *inclprob_s1* en tant que variable contenant les probabilités d'inclusion de la première étape.
- Cliquez sur Suivant.

Figure 14-11
*Etape Récapitulatif du plan de l'assistant de préparation d'analyse (phase 1)*



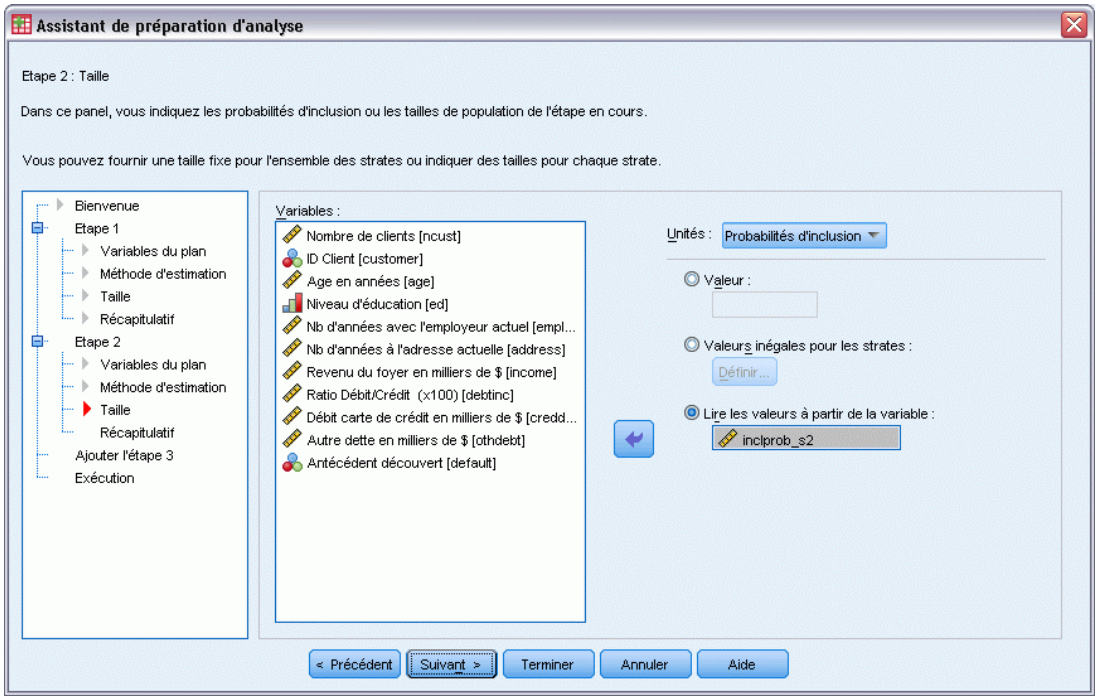

- Sélectionnez Oui, ajouter l'étape 2 maintenant.
- Cliquez sur Suivant, puis de nouveau sur Suivant à l'étape Variables de plan.

Figure 14-12
*Etape Méthode d'estimation de l'assistant de préparation d'analyse (phase 2)*




- Sélectionnez la méthode d'estimation de la seconde étape, WOR égal.
- Cliquez sur Suivant.

Figure 14-13
*Etape Taille de l'assistant de préparation d'analyse (phase 2)*




- Choisissez Lire les valeurs à partir de la variable, puis sélectionnez *inclprob_s2* en tant que variable contenant les probabilités d'inclusion de la seconde étape.
- Cliquez sur Terminer.

## *Récapitulatif*

Figure 14-14
*Récapitulatif*

| | | | Etape 1 | Etape 2 |
|---|---|---|---|---|
| Variables du plan | Grappe | 1 | Activité | |
| Informations sur l'analyse | Hypothèses de l'estimateur | | Echantillonnage de probabilité égale sans remise | Echantillonnage de probabilité égale sans remise |
| | Probabilité d'inclusion | | Obtenu à partir de la variable inclprob_s1 | Obtenu à partir de la variable inclprob_s2 |

Fichier de plan : C:\bankloan.csaplan
Variable de pondération : finalweight
Estimateur SRS : Échantillonnage sans remplacement

Le tableau récapitulatif vous permet de passer en revue votre plan d'analyse. Le plan comprend deux étapes et inclut la définition d'une variable de grappe. L'estimation à probabilité égale et sans remplacement est utilisée. Le plan est enregistré dans le fichier *c:\bankloan.csaplan*. Vous pouvez désormais utiliser ce fichier pour traiter le fichier *bankloan_noweights.sav* (qui comporte les probabilités d'inclusion et les pondérations d'échantillonnage calculées) via les procédures d'analyse d'échantillons complexes.

## *Procédures apparentées*

L'assistant de préparation d'analyse des échantillons complexes permet de préparer un échantillon pour une analyse lorsque vous ne pouvez pas accéder au fichier de plan d'échantillonnage.

- Pour créer un fichier de plan d'échantillonnage et réaliser un échantillon, utilisez l'Assistant d'échantillonnage.

Chapitre
15

# Echantillons complexes - Fréquences

La procédure Echantillons complexes - Fréquences génère les tableaux de fréquences des variables sélectionnées et affiche des statistiques univariées. Vous pouvez éventuellement classer les statistiques par sous-groupes, définis par une ou plusieurs variables qualitatives.

## Utilisation d'Echantillons complexes - Fréquences pour analyser l'utilisation des compléments nutritionnels

Un chercheur veut étudier l'utilisation des compléments nutritionnels chez les Américains à l'aide des résultats du NHIS (National Health Interview Survey) et d'un plan d'analyse existant. Pour plus d'informations, reportez-vous à la section Utilisation de l'assistant de préparation d'analyse des échantillons complexes pour préparer les données publiques du NHIS (National Health Interview Survey) dans le chapitre 14 sur p. 149.

Une partie de l'enquête 2000 se trouve dans le fichier *nhis2000_subset.sav*. Le plan d'analyse est stocké dans le fichier *nhis2000_subset.csaplan*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*. Utilisez la procédure Echantillons complexes - Fréquences pour générer les statistiques relatives à l'utilisation des compléments nutritionnels.

### Exécution de l'analyse

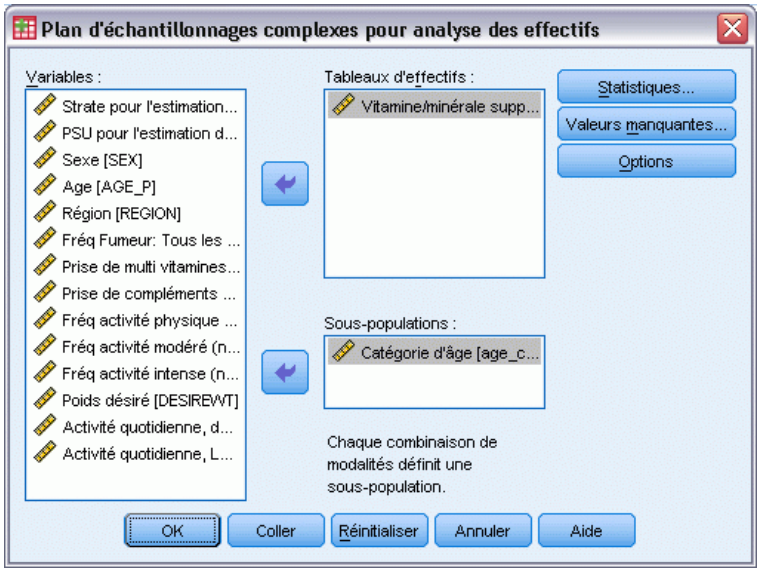
- Pour exécuter une analyse Echantillons complexes - Fréquences, sélectionnez dans le menu l'option suivante :

  Analyse > Echantillonnage > Fréquences

Figure 15-1
*Boîte de dialogue Plan d'échantillonnages complexes*

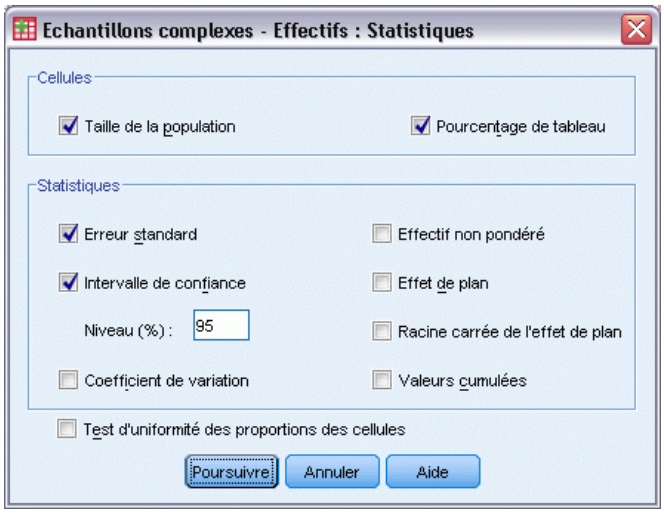

- Accédez au fichier *nhis2000_subset.csaplan* et sélectionnez-le. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*.
- Cliquez sur Poursuivre.

Figure 15-2
*Boîte de dialogue Fréquences*




- Sélectionnez *Compléments vitaminiques/minéraux – 12 derniers mois* comme variable de fréquence.
- Sélectionnez *Tranche d'âge* comme variable de sous-population.
- Cliquez sur Statistiques.

Figure 15-3
*Boîte de dialogue Fréquences : Statistiques*




- Sélectionnez Pourcentage de tableau dans le groupe Cellules.
- Sélectionnez Intervalle de confiance dans le groupe Statistiques.
- Cliquez sur Poursuivre.

▶ Cliquez sur OK dans la boîte de dialogue Fréquences.

## *Tableau des effectifs :*

Figure 15-4
*Tableau des fréquences de la variable/situation*

| | | Estimation | Erreur standard | Intervalle de confiance 95% | |
|---|---|---|---|---|---|
| | | | | Inférieur | Supérieur |
| Taille de la population | Oui | 102767095,0 | 1185127 | 100435966,8 | 105098223,19 |
| | Non | 90794234,00 | 1094402 | 88641560,18 | 92946907,816 |
| | Total | 193561329,0 | 1789099 | 190042196,1 | 197080461,94 |
| % de Total | Oui | 53,1% | ,4% | 52,4% | 53,8% |
| | Non | 46,9% | ,4% | 46,2% | 47,6% |
| | Total | 100,0% | ,0% | 100,0% | 100,0% |

Chaque statistique sélectionnée est calculée pour chaque mesure de cellule sélectionnée. La première colonne contient les estimations relatives au nombre et au pourcentage de personnes qui prennent ou ne prennent pas des compléments vitaminiques/minéraux. Les intervalles de confiance ne se chevauchant pas, vous pouvez en conclure que, dans l'ensemble, le nombre d'Américains qui prennent des compléments vitaminiques/minéraux est plus élevé que le nombre d'Américains qui n'en prennent pas.

## *Fréquence par sous-population*

Figure 15-5
*Tableau des fréquences par sous-population*

| Catégorie d'âge | | | Estimation | Erreur standard | Intervalle de confiance 95% | |
|---|---|---|---|---|---|---|
| | | | | | Inférieur | Supérieur |
| 18-24 | Taille de la population | Oui | 10018312 | 350602,4 | 9328681,921 | 10707942,079 |
| | | Non | 15472368 | 499182,4 | 14490482,996 | 16454253,004 |
| | | Total | 25490680 | 680732,8 | 24151687,776 | 26829672,224 |
| | % de Total | Oui | 39,3% | 1,0% | 37,4% | 41,2% |
| | | Non | 60,7% | 1,0% | 58,8% | 62,6% |
| | | Total | 100,0% | ,0% | 100,0% | 100,0% |
| 25-44 | Taille de la population | Oui | 39163840 | 660855,7 | 37863945,748 | 40463734,252 |
| | | Non | 39503150 | 645934,2 | 38232606,200 | 40773693,800 |
| | | Total | 78666990 | 961114,3 | 76776491,135 | 80557488,865 |
| | % de Total | Oui | 49,8% | ,6% | 48,7% | 50,9% |
| | | Non | 50,2% | ,6% | 49,1% | 51,3% |
| | | Total | 100,0% | ,0% | 100,0% | 100,0% |
| 45-64 | Taille de la population | Oui | 34154952 | 598603,7 | 32977506,572 | 35332397,428 |
| | | Non | 24005512 | 497723,8 | 23026495,959 | 24984528,041 |
| | | Total | 58160464 | 814680,4 | 56557998,654 | 59762929,346 |
| | % de Total | Oui | 58,7% | ,6% | 57,5% | 60,0% |
| | | Non | 41,3% | ,6% | 40,0% | 42,5% |
| | | Total | 100,0% | ,0% | 100,0% | 100,0% |
| 65+ | Taille de la population | Oui | 19429991 | 439459,8 | 18565579,536 | 20294402,464 |
| | | Non | 11813204 | 314238,1 | 11195101,955 | 12431306,045 |
| | | Total | 31243195 | 587623,4 | 30087347,652 | 32399042,348 |
| | % de Total | Oui | 62,2% | ,7% | 60,7% | 63,6% |
| | | Non | 37,8% | ,7% | 36,4% | 39,3% |
| | | Total | 100,0% | ,0% | 100,0% | 100,0% |

Lors du calcul des statistiques par sous-population, chaque statistique sélectionnée est calculée pour chaque mesure de cellule sélectionnée par *catégorie d'âge*. La première colonne contient les estimations relatives au nombre et au pourcentage de personnes de chaque catégorie qui prennent ou ne prennent pas des compléments vitaminiques/minéraux. Les intervalles de confiance des pourcentages du tableau ne se chevauchant pas, vous pouvez en conclure que l'utilisation de compléments vitaminiques/minéraux augmente avec l'âge.

## *Récapitulatif*

A l'aide de la procédure Echantillons complexes - Fréquences, vous avez obtenu les statistiques relatives à l'utilisation des compléments nutritionnels chez les Américains.

- Dans l'ensemble, le nombre d'Américains qui prennent des compléments vitaminiques/minéraux est plus élevé que le nombre d'Américains qui n'en prennent pas.
- L'examen des statistiques par catégorie d'âge indique que plus les Américains vieillissent, plus ils prennent des compléments vitaminiques/minéraux.

## *Procédures apparentées*

La procédure Echantillons complexes - Fréquences permet d'obtenir les statistiques descriptives univariées des variables qualitatives pour les observations obtenues via un plan d'échantillonnage complexe.

- L'assistant d'échantillonnage des échantillons complexes permet d'indiquer les spécifications du plan d'échantillonnage complexe et d'obtenir un échantillon. Le fichier de plan d'échantillonnage créé par l'assistant d'échantillonnage contient un plan d'analyse par défaut et peut être spécifié dans la boîte de dialogue Plan lors de l'analyse de l'échantillon obtenu en fonction de ce plan.
- L'assistant de préparation d'analyse des échantillons complexes est utilisé pour indiquer les spécifications d'analyse d'un échantillon complexe existant. Le fichier de plan d'analyse créé par l'assistant d'échantillonnage peut être spécifié dans la boîte de dialogue Plan lorsque vous analysez l'échantillon correspondant à ce plan.
- La procédure Tableaux croisés des échantillons complexes fournit des statistiques descriptives univariées des variables qualitatives.
- La procédure Echantillons complexes - Descriptives fournit des statistiques descriptives univariées des variables d'échelle.

Chapitre
16

# Echantillons complexes – Descriptives

La procédure Echantillons complexes – Descriptives affiche les statistiques récapitulatives univariées de plusieurs variables. Vous pouvez éventuellement classer les statistiques par sous-groupes, définis par une ou plusieurs variables qualitatives.

## Utilisation des descriptives des échantillons complexes pour analyser les niveaux d'activité

Un chercheur veut étudier les niveaux d'activité chez les Américains à l'aide des résultats du NHIS (National Health Interview Survey) et d'un plan d'analyse existant. Pour plus d'informations, reportez-vous à la section Utilisation de l'assistant de préparation d'analyse des échantillons complexes pour préparer les données publiques du NHIS (National Health Interview Survey) dans le chapitre 14 sur p. 149.

Une partie de l'enquête 2000 se trouve dans le fichier *nhis2000_subset.sav*. Le plan d'analyse est stocké dans le fichier *nhis2000_subset.csaplan*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*. Utilisez la procédure Echantillons complexes – Descriptives pour générer les statistiques descriptives univariées relatives aux niveaux d'activité.

### Exécution de l'analyse

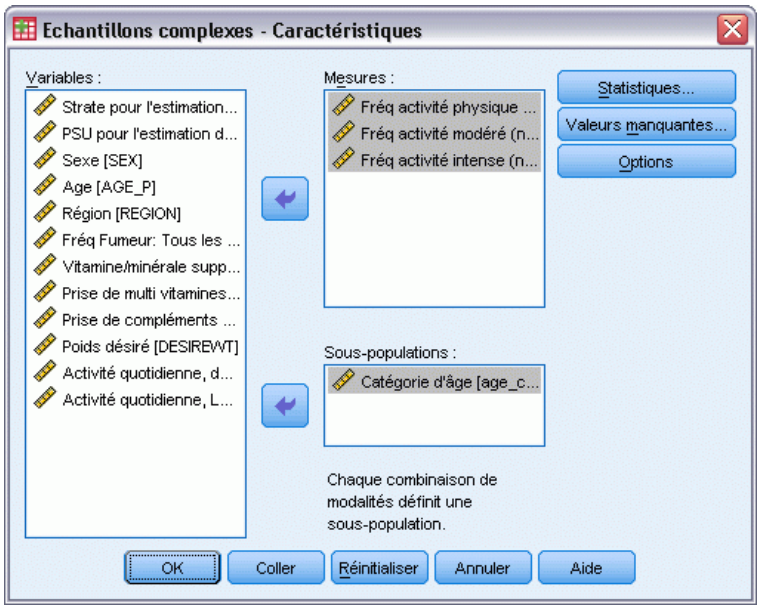
- Pour exécuter une analyse des descriptives des échantillons complexes, sélectionnez les options suivantes dans le menu :

  Analyse > Echantillonnage > Descriptives

Figure 16-1
*Boîte de dialogue Plan d'échantillonnages complexes*

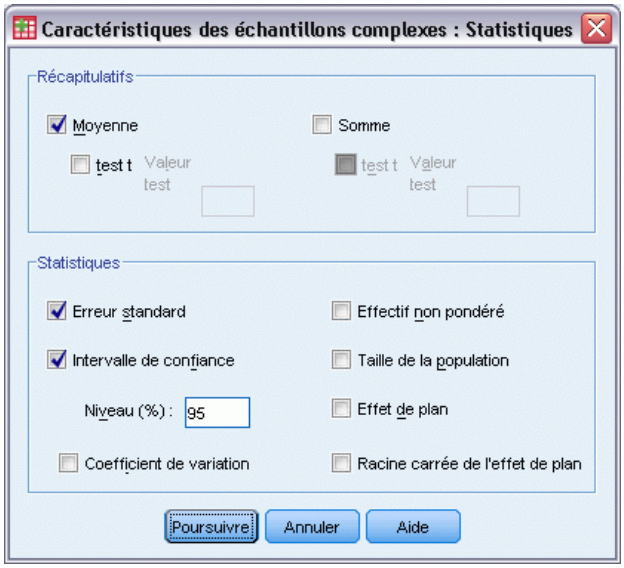

- Accédez au fichier *nhis2000_subset.csaplan* et sélectionnez-le. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*.
- Cliquez sur Poursuivre.

Figure 16-2
*Boîte de dialogue Descriptives*




- Sélectionnez les variables de mesure *Fréq activité modérée (nbre de fois/sem)* à *Fréq activité intense (nbre de fois/sem)*.
- Sélectionnez *Tranche d'âge* comme variable de sous-population.
- Cliquez sur Statistiques.

Figure 16-3
*Boîte de dialogue Statistiques Descriptives*




- Sélectionnez Intervalle de confiance dans le groupe Statistiques.

- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Echantillons complexes – Descriptives.

## *Statistiques univariées*

Figure 16-4
*Statistiques univariées.*

| | | Estimation | Erreur standard | Intervalle de confiance 95% | |
|---|---|---|---|---|---|
| | | | | Inférieur | Supérieur |
| Moyenne | Fréq activité physique (nb fois par semaine) | 3,73 | ,033 | 3,66 | 3,79 |
| | Fréq activité modéré (nb de fois par semaine) | 4,90 | ,041 | 4,82 | 4,98 |
| | Fréq activité intense (nb de fois par semaine) | 3,52 | ,042 | 3,43 | 3,60 |

Chaque statistique sélectionnée est calculée pour chaque mesure de variable. La première colonne contient les estimations du nombre moyen de fois par semaine où une personne pratique un type d'activité donné. Les intervalles de confiance des moyennes ne se chevauchent pas. Vous pouvez en conclure que, globalement, les Américains pratiquent une activité intense moins souvent qu'une activité plus douce, et une activité plus douce moins souvent qu'une activité modérée.

## *Statistiques univariées par sous-population*

Figure 16-5
*Statistiques univariées par sous-population*

| Catégorie d'âge | | | Estimation | Erreur standard | Intervalle de confiance 95% | |
|---|---|---|---|---|---|---|
| | | | | | Inférieur | Supérieur |
| 18-24 | Moyenne | Fréq activité physique (nb fois par semaine) | 3,92 | ,087 | 3,75 | 4,09 |
| | | Fréq activité modéré (nb de fois par semaine) | 5,18 | ,137 | 4,91 | 5,45 |
| | | Fréq activité intense (nb de fois par semaine) | 3,45 | ,085 | 3,28 | 3,62 |
| 25-44 | Moyenne | Fréq activité physique (nb fois par semaine) | 3,55 | ,048 | 3,46 | 3,65 |
| | | Fréq activité modéré (nb de fois par semaine) | 4,73 | ,056 | 4,62 | 4,84 |
| | | Fréq activité intense (nb de fois par semaine) | 3,28 | ,052 | 3,18 | 3,38 |
| 45-64 | Moyenne | Fréq activité physique (nb fois par semaine) | 3,79 | ,063 | 3,66 | 3,91 |
| | | Fréq activité modéré (nb de fois par semaine) | 4,88 | ,070 | 4,74 | 5,02 |
| | | Fréq activité intense (nb de fois par semaine) | 3,65 | ,092 | 3,47 | 3,84 |
| 65+ | Moyenne | Fréq activité physique (nb fois par semaine) | 4,18 | ,111 | 3,96 | 4,39 |
| | | Fréq activité modéré (nb de fois par semaine) | 5,22 | ,084 | 5,06 | 5,39 |
| | | Fréq activité intense (nb de fois par semaine) | 4,66 | ,155 | 4,36 | 4,97 |

Chaque statistique sélectionnée est calculée pour chaque mesure de variable en fonction des valeurs de la *tranche d'âge*. La première colonne contient les estimations du nombre moyen de fois par semaine où les personnes de chaque tranche d'âge pratiquent un type d'activité donné. Les intervalles de confiance des moyennes vous permettent de tirer des conclusions intéressantes.

- En ce qui concerne les activités dynamiques et modérées, les 25–44 ans sont moins actifs que les 18–24 ans et les 45–64, et les 45–64 ans sont moins actifs que les personnes âgées de plus de 65 ans.
- Pour ce qui est des activités intenses, les 25–44 ans sont moins actifs que les 45–64 ans, et les 18–24 ans et les 45–64 ans sont moins actifs que les personnes âgées de plus de 65 ans.

### *Récapitulatif*

A l'aide de la procédure Echantillons complexes – Descriptives, vous avez obtenu les statistiques relatives aux niveaux d'activité chez les Américains.

- En règle générale, les Américains consacrent plus ou moins de temps à différents types d'activité.
- Répertoriés par tranches d'âge, les résultats montrent globalement que, après leurs études secondaires, les Américains sont moins actifs qu'auparavant, puis qu'ils s'adonnent de plus en plus à la pratique d'une activité sportive en vieillissant.

## *Procédures apparentées*

La procédure Echantillons complexes – Descriptives permet d'obtenir les statistiques descriptives univariées des mesures d'échelle pour les observations obtenues via un plan d'échantillonnage complexe.

- L'assistant d'échantillonnage des échantillons complexes permet d'indiquer les spécifications du plan d'échantillonnage complexe et d'obtenir un échantillon. Le fichier de plan d'échantillonnage créé par l'assistant d'échantillonnage contient un plan d'analyse par défaut et peut être spécifié dans la boîte de dialogue Plan lors de l'analyse de l'échantillon obtenu en fonction de ce plan.
- L'assistant de préparation d'analyse des échantillons complexes est utilisé pour indiquer les spécifications d'analyse d'un échantillon complexe existant. Le fichier de plan d'analyse créé par l'assistant d'échantillonnage peut être spécifié dans la boîte de dialogue Plan lorsque vous analysez l'échantillon correspondant à ce plan.
- La procédure Echantillons complexes Rapports indique les statistiques descriptives des ratios des mesures d'échelle
- La procédure Echantillons complexes - Fréquences indique les statistiques descriptives univariées des variables qualitatives.

Chapitre
17

# Tableaux croisés des échantillons complexes

La procédure Echantillons complexes - Tableaux croisés génère les tableaux croisés des paires de variables sélectionnées et affiche des statistiques à deux entrées. Vous pouvez éventuellement classer les statistiques par sous-groupes, définis par une ou plusieurs variables qualitatives.

## Utilisation de tableaux croisés des échantillons complexes pour mesurer le risque relatif d'un événement

Une société de vente d'abonnements à des magazines envoie généralement chaque mois des courriers aux personnes dont le nom est recensé dans une base de données qu'elle a acquise. Le taux de réponse est habituellement faible. Vous devez donc trouver un moyen de mieux cibler les clients potentiels. En partant de l'hypothèse que les personnes abonnées à un journal sont plus susceptibles de s'abonner à un magazine, il est proposé d'adresser le publipostage à ces personnes.

Utilisez la procédure Echantillons complexes - Tableaux croisés pour vérifier cette théorie en construisant un tableau 2*2 liant l'*abonnement à un journal* et la *réponse obtenue*, et en calculant le risque relatif qu'une personne abonnée à un journal réponde au publipostage. Ces informations sont rassemblées dans le fichier *demo_cs.sav* et doivent être analysées à l'aide du fichier de plan d'échantillonnage *demo.csplan*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*.

### Exécution de l'analyse

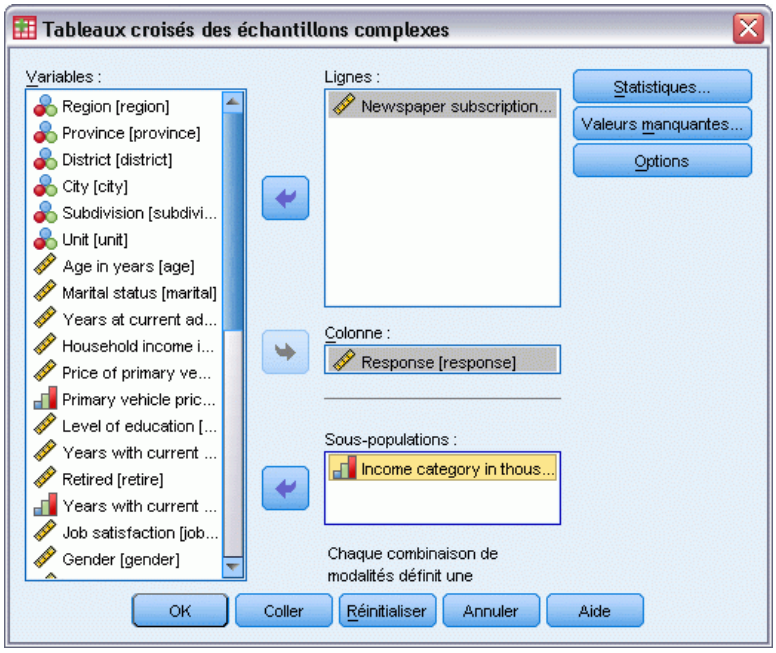
- Pour exécuter une analyse des tableaux croisés des échantillons complexes, sélectionnez les options suivantes dans le menu :

  Analyse > Echantillonnage > Tableaux croisés

Figure 17-1
*Boîte de dialogue Plan d'échantillonnages complexes*

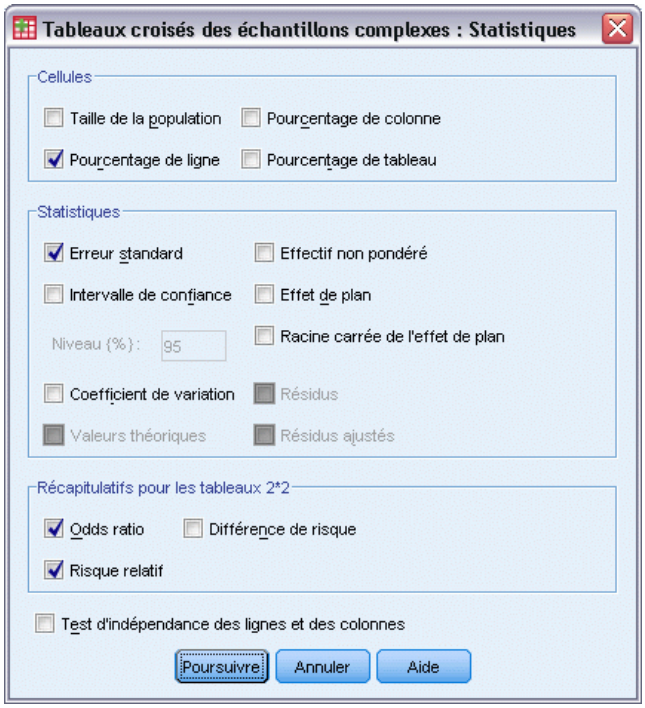

- Recherchez et sélectionnez *demo.csplan*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*.
- Cliquez sur Poursuivre.

Figure 17-2
*Boîte de dialogue Tableaux croisés*




- Sélectionnez la variable en ligne *Abonnement à un journal*.
- Sélectionnez *Réponse* comme variable en colonne.
- Il est également intéressant de voir les résultats répertoriés par catégorie de revenu. Sélectionnez par conséquent la variable de sous-population *Catégorie de revenu en milliers*.
- Cliquez sur Statistiques.

Figure 17-3
*Boîte de dialogue Tableaux croisé : Statistiques*




- Dans le groupe Cellules, désélectionnez Taille de la population et sélectionnez Pourcentage de ligne.
- Dans le groupe Récapitulatifs pour les tableaux 2*2, sélectionnez Odds ratio et Risque relatif.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Echantillons complexes - Tableaux croisés.

  Ces sélections permettent de générer un tableau croisé et d'obtenir une estimation du risque liant l'*abonnement à un journal* et la *réponse obtenue*. Des tableaux distincts répertoriant les résultats en fonction de la *catégorie de revenu en milliers* sont également créés.

## *Tableau croisé*

Figure 17-4
*Tableau croisé liant l'abonnement à un journal et la réponse obtenue*

| Abonnement à un magasine | | | Réponse | | |
|---|---|---|---|---|---|
| | | | Oui | Non | Total |
| Oui | % dans Abonnement à un magasine | Estimation | 17.2% | 82.8% | 100,0% |
| | | Erreur standard | 1.0% | 1.0% | ,0% |
| Non | % dans Abonnement à un magasine | Estimation | 10.3% | 89.7% | 100,0% |
| | | Erreur standard | ,9% | ,9% | ,0% |
| Total | % dans Abonnement à un magasine | Estimation | 12.8% | 87.2% | 100,0% |
| | | Erreur standard | ,8% | ,8% | ,0% |

Le tableau croisé indique que, globalement, peu de personnes ont répondu au publipostage. Toutefois, la proportion est plus élevée parmi les personnes abonnées à un journal.

## Estimation du risque

Figure 17-5
*Estimation du risque relatif à l'abonnement à un journal et à la réponse obtenue*

| | | | Estimation |
|---|---|---|---|
| Abonnement à un magasine * Réponse | Odds Ratios | | 1,812 |
| | Risque relatif | Pour la cohorte Réponse = Oui | 1,673 |
| | | Pour la cohorte Réponse = Non | ,923 |

Les statistiques ne sont calculées que pour les tableaux 2 par 2 où toutes les cellules sont observées.

Le risque relatif est le ratio des probabilités d'un événement. Le risque relatif d'une réponse au publipostage est le ratio de la probabilité qu'une personne abonnée à un journal réponde par rapport à la probabilité qu'une personne non abonnée réponde. Par conséquent, l'estimation du risque relatif correspond simplement au calcul suivant : 17,2%/10,3% = 1,673. De même, le risque relatif de non-réponse est le ratio de la probabilité qu'un abonné ne réponde pas par rapport à la probabilité qu'une personne non abonnée ne réponde pas. L'estimation de ce risque relatif est de 0,923. Au vu de ces résultats, vous pouvez estimer qu'une personne abonnée à un journal est 1,673 fois plus susceptible de répondre au publipostage qu'une personne qui n'est abonnée à aucun journal ou a 0,923 fois autant de chances de ne pas répondre.

L'odds ratio est le ratio des chances de l'événement. Les chances d'un événement correspondent au ratio de la probabilité que l'événement se produise par rapport à la probabilité qu'il ne se produise pas. Par conséquent, l'estimation des chances qu'un abonné à un journal réponde au publipostage est de 17,2 %/82,8 % = 0,208. De même, l'estimation des chances qu'une personne non abonnée réponde est de 10,3 %/89,7 % = 0,115. L'estimation de l'odds ratio est donc de 0,208/0,115 = 1,812 (à une erreur d'arrondi près dans les étapes impliquées). L'odds ratio est également le ratio du risque relatif de réponse par rapport au risque relatif de non-réponse, soit 1,673/0,923 = 1,812.

### Odds ratio et risque relatif

Ratio de ratios, l'odds ratio est très difficile à interpréter. En revanche, l'interprétation du risque relatif est plus facile. Seul, l'odds ratio n'est donc pas très utile. Dans certains cas cependant, l'estimation du risque relatif est rarement concluante. L'odds ratio peut alors servir à obtenir une approximation du risque relatif de l'événement étudié. L'odds ratio doit être utilisé comme

approximation du risque relatif de l'événement étudié lorsque les deux conditions suivantes sont remplies :

- La probabilité de l'événement étudié est faible (<0,1). Cette condition garantit que l'odds ratio fournira une bonne approximation par rapport au risque relatif. Dans cet exemple, l'événement étudié est la réponse au publipostage.
- L'étude est dotée d'un plan de contrôle d'observation. Cette condition indique que l'estimation classique du risque relatif ne sera certainement pas concluante. Les études de contrôle d'observation sont rétrospectives. Elles sont la plupart du temps utilisées lorsque l'événement étudié est improbable, ou que le plan d'une éventuelle expérience est irréaliste ou immoral.

Aucune de ces conditions n'est remplie dans le présent exemple, étant donné que la proportion globale des répondants est de 12,8 % et que l'étude n'est pas dotée d'un plan de contrôle d'observation. Il est donc plus sûr d'indiquer 1,673 comme risque relatif plutôt que la valeur de l'odds ratio.

## *Estimation du risque par sous-population*

Figure 17-6
*Estimation du risque relatif à l'abonnement à un journal et la réponse obtenue, en contrôlant la catégorie de revenu*

| Catégories de | | | | Estimation |
|---|---|---|---|---|
| Inf à $25 | Abonnement à un magasine * Réponse | Odds Ratios | | 2,712 |
| | | Risque relatif | Pour la cohorte Réponse = Oui | 2,241 |
| | | | Pour la cohorte Réponse = Non | ,826 |
| $25 - $49 | Abonnement à un magasine * Réponse | Odds Ratios | | 1,794 |
| | | Risque relatif | Pour la cohorte Réponse = Oui | 1,645 |
| | | | Pour la cohorte Réponse = Non | ,917 |
| $50 - $74 | Abonnement à un magasine * Réponse | Odds Ratios | | 1,168 |
| | | Risque relatif | Pour la cohorte Réponse = Oui | 1,152 |
| | | | Pour la cohorte Réponse = Non | ,986 |
| $75 + | Abonnement à un magasine * Réponse | Odds Ratios | | 1,242 |
| | | Risque relatif | Pour la cohorte Réponse = Oui | 1,227 |
| | | | Pour la cohorte Réponse = Non | ,988 |

Les statistiques ne sont calculées que pour les tableaux 2 par 2 où toutes les cellules sont observées.

Les estimations de risque relatif sont calculées séparément pour chaque catégorie de revenu. Notez que le risque relatif d'une réponse positive de la part des abonnés à un journal semble décroître au fur et à mesure que le revenu augmente, ce qui indique qu'il doit être possible de cibler les destinataires des publipostages plus précisément.

### *Récapitulatif*

A l'aide des estimations de risque des tableaux croisés des échantillons complexes, vous avez découvert que vous pouvez accroître le taux de réponse aux publipostages directs en ciblant les personnes abonnées à des journaux. Vous avez en outre constaté que les estimations de risque ne sont pas forcément constantes d'une *catégorie de revenu* à l'autre. Vous pouvez donc augmenter encore plus le taux de réponse en vous adressant aux abonnés aux revenus les plus faibles.

## *Procédures apparentées*

La procédure Echantillons complexes - Tableaux croisés permet d'obtenir les statistiques descriptives du tableau croisé des variables qualitatives pour les observations obtenues via un plan d'échantillonnage complexe.

- L'assistant d'échantillonnage des échantillons complexes permet d'indiquer les spécifications du plan d'échantillonnage complexe et d'obtenir un échantillon. Le fichier de plan d'échantillonnage créé par l'assistant d'échantillonnage contient un plan d'analyse par défaut et peut être spécifié dans la boîte de dialogue Plan lors de l'analyse de l'échantillon obtenu en fonction de ce plan.
- L'assistant de préparation d'analyse des échantillons complexes est utilisé pour indiquer les spécifications d'analyse d'un échantillon complexe existant. Le fichier de plan d'analyse créé par l'assistant d'échantillonnage peut être spécifié dans la boîte de dialogue Plan lorsque vous analysez l'échantillon correspondant à ce plan.
- La procédure Echantillons complexes - Fréquences indique les statistiques descriptives univariées des variables qualitatives.

Chapitre
18

# Echantillons complexes – Rapports

La procédure Echantillons complexes – Rapports affiche les statistiques récapitulatives univariées des rapports de variables. Vous pouvez éventuellement classer les statistiques par sous-groupes, définis par une ou plusieurs variables qualitatives.

## Utilisation des rapports d'échantillons complexes pour évaluer la valeur d'une propriété

Une agence d'état est chargée d'assurer l'application d'impôts fonciers équitables dans chaque comté. Les impôts étant calculés à partir de la valeur estimée de la propriété, l'agence souhaite identifier les valeurs des propriétés dans chaque comté afin de s'assurer que les registres des comtés sont mis à jour uniformément. Les ressources permettant de réaliser les évaluations étant limitées, l'agence a choisi d'utiliser la méthode de l'échantillonnage complexe pour sélectionner des propriétés.

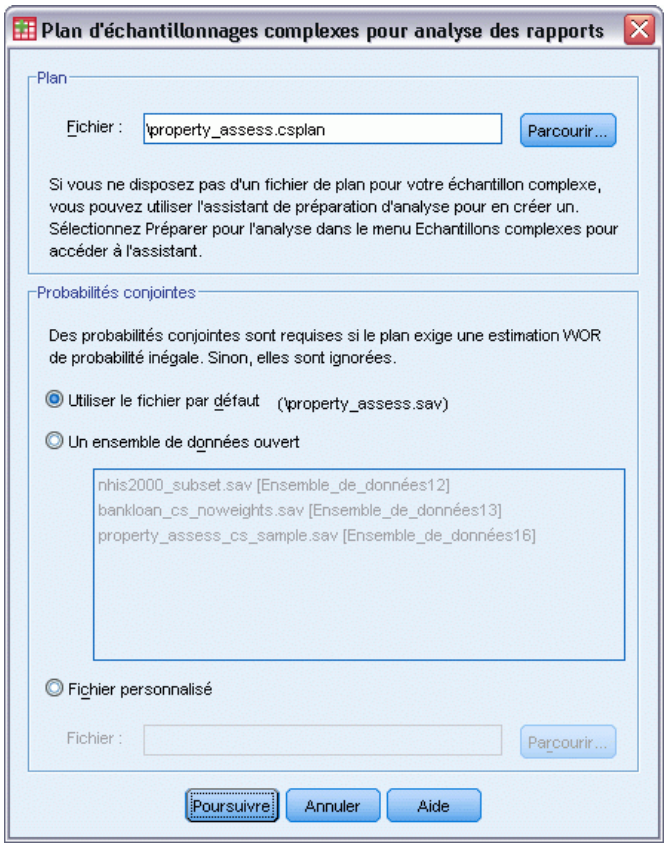
L'échantillon de propriétés sélectionné et les informations de l'évaluation réalisée sont regroupés dans *property_assess_cs_sample.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*. Utilisez les rapports d'échantillons complexes pour connaître l'évolution de la valeur des propriétés dans les cinq comtés, depuis la dernière évaluation.

### Exécution de l'analyse

- Pour exécuter une analyse des rapports d'échantillons complexes, sélectionnez les options suivantes dans le menu :

  Analyse > Echantillonnage > Ratios...

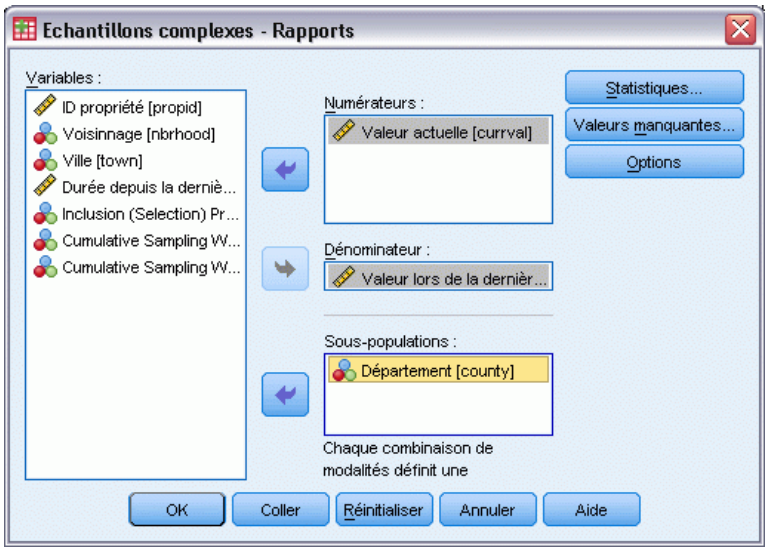
Figure 18-1
*Boîte de dialogue Plan d'échantillonnages complexes*




- Accédez au fichier *property_assess.csplan* et sélectionnez-le. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20.*

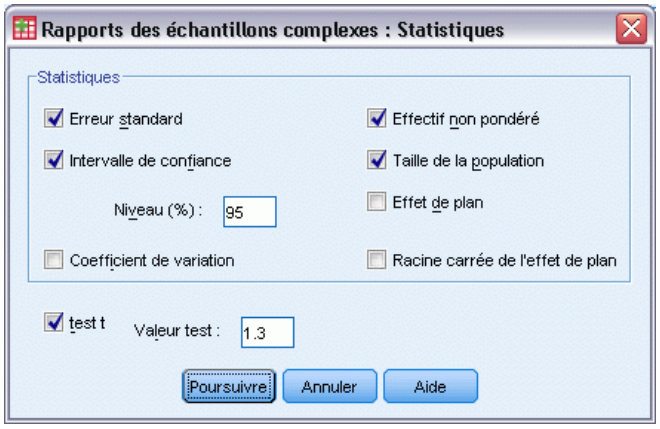
- Cliquez sur Poursuivre.

Figure 18-2
*Boîte de dialogue Rapports*




- Sélectionnez la variable de numérateur *Valeur courante*.
- Sélectionnez la variable de dénominateur *Valeur lors de la dernière évaluation*.
- Sélectionnez la variable de sous-population *Comté*.
- Cliquez sur Statistiques.

Figure 18-3
*Boîte de dialogue Rapports : Statistiques*




- Dans le groupe Statistiques, sélectionnez Intervalle de confiance, Effectif non pondéré et Taille de la population.
- Sélectionnez test t et entrez la valeur de test 1,3.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Echantillons complexes – Rapports.

## *Ratios*

Figure 18-4
*Tableau des rapports*

| Département | Numérateur | Dénominateur | Estimation du rapport | Erreur standard | Intervalle de confiance 95% | | V |
|---|---|---|---|---|---|---|---|
| | | | | | Inférieur | Supérieur | |
| Est | Valeur actuelle | Durée depuis la dernière évaluation | 1,381 | ,068 | 1,236 | 1,525 | |
| Centre | Valeur actuelle | Durée depuis la dernière évaluation | 1,364 | ,064 | 1,227 | 1,502 | |
| Oest | Valeur actuelle | Durée depuis la dernière évaluation | 1,524 | ,053 | 1,410 | 1,638 | |
| Nord | Valeur actuelle | Durée depuis la dernière évaluation | 1,277 | ,032 | 1,208 | 1,346 | |
| Sud | Valeur actuelle | Durée depuis la dernière évaluation | 1,195 | ,029 | 1,134 | 1,256 | |

L'affichage par défaut du tableau est très large : vous devez donc le faire pivoter pour mieux le visualiser.

### *Tableau des rapports pivotant*

- ▶ Double-cliquez sur le tableau pour l'activer.
- ▶ A partir des menus du Viewer, sélectionnez :
  Tableau pivotant > Structure pivotante
- ▶ Faites glisser le *numérateur*, puis le *dénominateur* de la ligne à la strate.
- ▶ Faites glisser le *comté* de la ligne à la colonne.
- ▶ Faites glisser les *statistiques* de la colonne à la ligne.
- ▶ Fermez la fenêtre Structure pivotante.

## *Tableau des rapports pivoté*

Figure 18-5
*Tableau des rapports pivoté*

Numérateur: Valeur actuelle
Dénominateur: Inclusion (Selection) Probability for Stage 1

| | | Département | | | | |
|---|---|---|---|---|---|---|
| | | Est | Centre | Ouest | Nord | Sud |
| Estimation du rapport | | 1.381 | 1.364 | 1.524 | 1.277 | 1.195 |
| Erreur standard | | .068 | .064 | .053 | .032 | .029 |
| Intervalle de confiance 95% | Inférieur | 1.236 | 1.227 | 1.410 | 1.208 | 1.134 |
| | Supérieur | 1.525 | 1.502 | 1.638 | 1.346 | 1.256 |
| Test d'hypothèse | Valeur de test | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 |
| | t | 1.191 | .997 | 4.201 | -.702 | -3.646 |
| | ddl | 15 | 15 | 15 | 15 | 15 |
| | Sig. | .252 | .334 | .001 | .493 | .002 |
| Effectif non pondéré | | 168 | 179 | 202 | 205 | 220 |

Maintenant que vous avez fait pivoter le tableau des rapports, vous pouvez comparer les statistiques des comtés plus aisément.

- Les estimations de rapport vont de 1,195 dans le comté du sud à 1,524 dans le comté de l'ouest.
- Les erreurs standard connaissent en outre une certaine variabilité avec un minimum de 0,029 pour le comté du sud et un maximum de 0,068 pour le comté de l'est.
- Certains des intervalles de confiance ne se chevauchent pas ; vous pouvez en conclure que les rapports du comté de l'ouest sont plus élevés que ceux des comtés du nord et du sud.
- Enfin, notez que les valeurs de signification des tests *t* des comtés de l'ouest et du sud sont inférieures à 0,05. Il s'agit d'une mesure plus objective. Vous pouvez en déduire que le rapport du comté de l'ouest est supérieur à 1,3 et celui du comté du sud inférieur à 1,3.

## *Récapitulatif*

La procédure Echantillons complexes – Rapports vous a permis d'obtenir diverses statistiques sur les rapports entre les variables *Valeur courante* et *Valeur lors de la dernière évaluation*. D'après les résultats, il existe probablement des inégalités dans l'évaluation des impôts fonciers des différents comtés, à savoir :

- Les rapports du comté de l'ouest sont élevés, ce qui indique que ses registres sont moins à jour que ceux des autres comtés en ce qui concerne l'appréciation des valeurs des propriétés. Les impôts fonciers sont probablement trop faibles dans ce comté.
- Les rapports du comté du sud sont faibles, ce qui indique que ses registres sont plus à jour que ceux des autres comtés en ce qui concerne l'appréciation des valeurs des propriétés. Les impôts fonciers sont probablement trop élevés dans ce comté.
- Les rapports du comté du sud sont moins élevés que ceux du comté de l'ouest, mais restent au niveau de la valeur visée (1,3).

Les ressources permettant de connaître les valeurs des propriétés dans le comté du sud seront réattribuées au comté de l'ouest afin que les rapports de ces comtés s'alignent sur ceux des autres comtés et sur l'objectif de 1,3.

## *Procédures apparentées*

La procédure Echantillons complexes – Rapports permet d'obtenir les statistiques descriptives univariées du ratio des mesures d'échelle pour les observations obtenues via un plan d'échantillonnage complexe.

- L'assistant d'échantillonnage des échantillons complexes permet d'indiquer les spécifications du plan d'échantillonnage complexe et d'obtenir un échantillon. Le fichier de plan d'échantillonnage créé par l'assistant d'échantillonnage contient un plan d'analyse par défaut et peut être spécifié dans la boîte de dialogue Plan lors de l'analyse de l'échantillon obtenu en fonction de ce plan.
- L'assistant de préparation d'analyse des échantillons complexes est utilisé pour indiquer les spécifications d'analyse d'un échantillon complexe existant. Le fichier de plan d'analyse créé par l'assistant d'échantillonnage peut être spécifié dans la boîte de dialogue Plan lorsque vous analysez l'échantillon correspondant à ce plan.
- La procédure Echantillons complexes - Descriptives fournit des statistiques descriptives des variables d'échelle.

Chapitre
19

# Modèle linéaire général des échantillons complexes

La procédure relative au modèle linéaire général des échantillons complexes effectue une analyse de régression linéaire, ainsi qu'une analyse de la variance et de la covariance, pour les échantillons réalisés à l'aide de méthodes d'échantillonnage complexes. Vous pouvez également demander une analyse pour une sous-population.

## Utilisation d'Echantillons complexes - Modèle linéaire général pour ajuster ANOVA à deux facteurs

Une chaîne d'épiceries a interrogé, en fonction d'un plan complexe, un groupe de clients au sujet de leurs habitudes de consommation. Compte tenu des résultats de l'enquête et de la somme dépensée par les clients au cours du mois précédent, l'enseigne souhaite voir si la fréquence des achats est liée aux dépenses mensuelles, et ce en prenant en compte le sexe du client et en intégrant le plan d'échantillonnage.

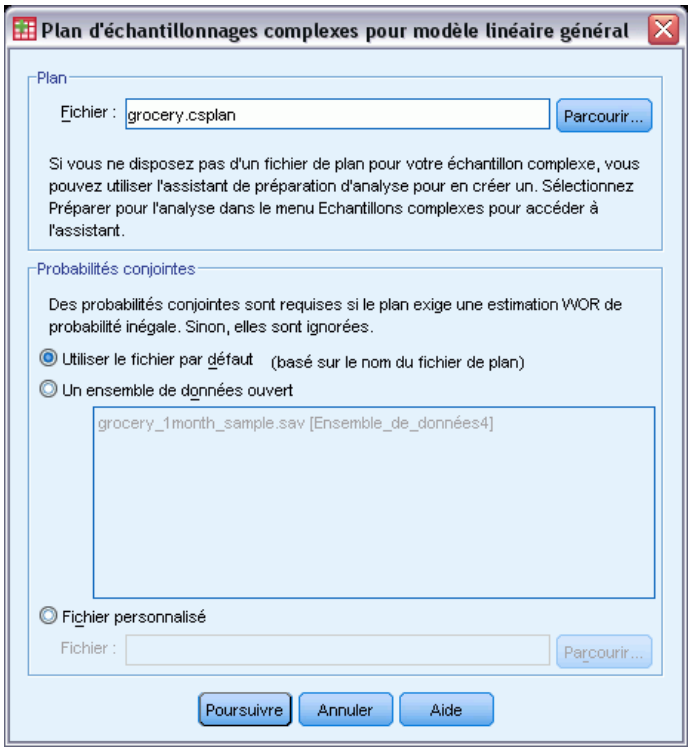
Ces informations sont regroupées dans le fichier *grocery_1month_sample.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*. Utilisez la procédure relative au modèle linéaire général des échantillons complexes pour exécuter ANOVA à 2 facteurs (ou d'ordre 2) sur les budgets consacrés.

### Exécution de l'analyse

- Pour exécuter une analyse relative au modèle linéaire général des échantillons complexes, sélectionnez les options suivantes dans le menu :

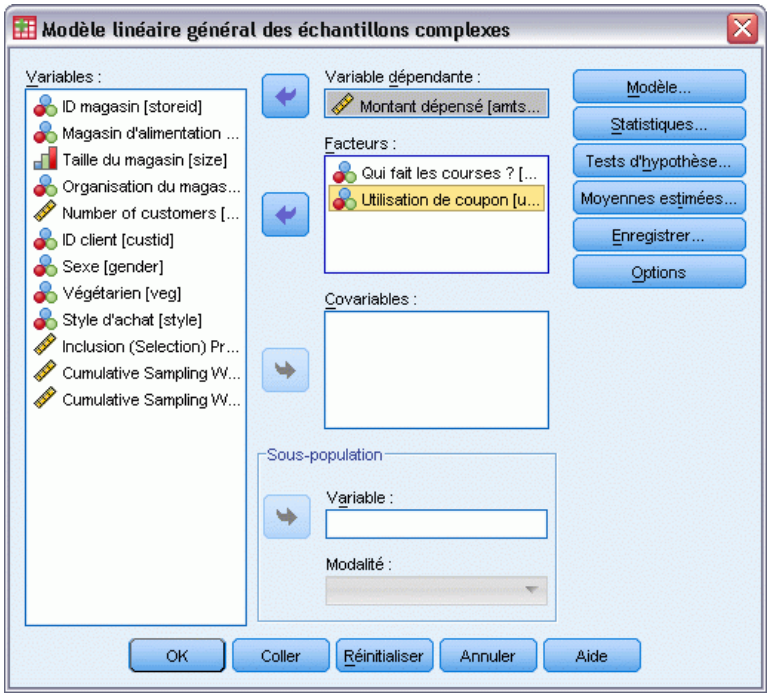
  Analyse > Echantillonnage > Modèle linéaire général...

Figure 19-1
*Boîte de dialogue Plan d'échantillonnages complexes*



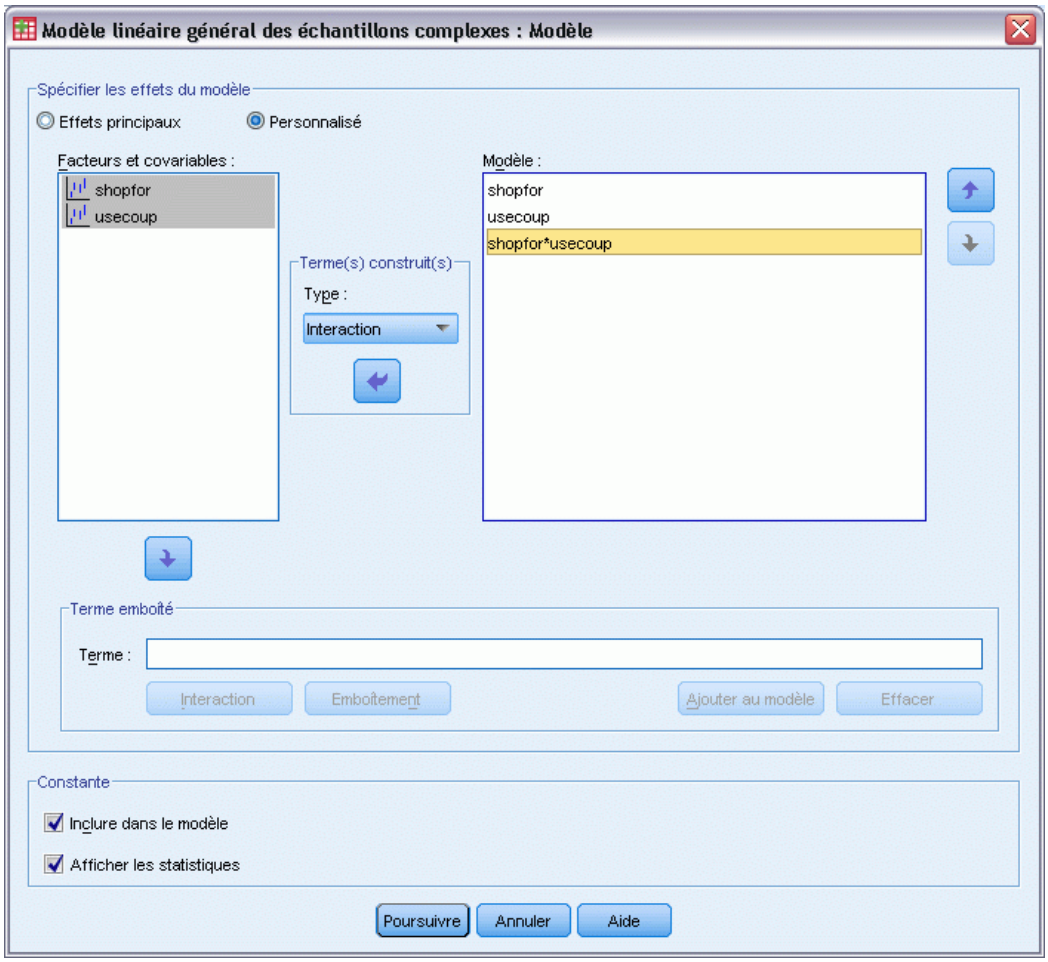

- Accédez au fichier *grocery.csplan* et sélectionnez-le. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*.
- Cliquez sur Poursuivre.

Figure 19-2
*Boîte de dialogue Modèle linéaire général*



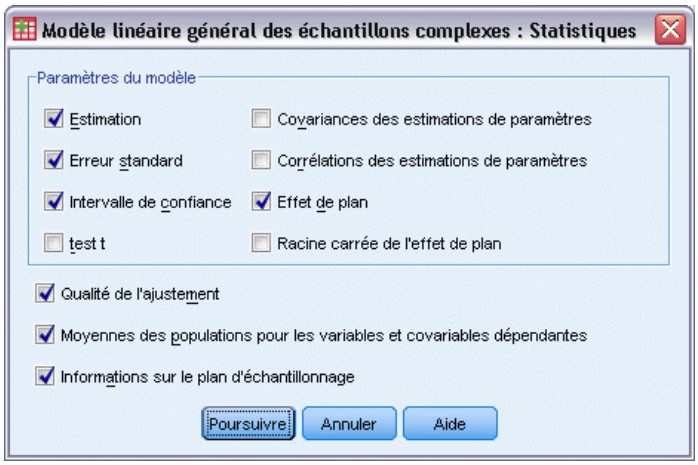

- Sélectionnez l'option *Budget consacré* comme variable dépendante.
- Sélectionnez *Fait les courses pour* et *Utilise des coupons* comme facteurs.
- Cliquez sur Modèle.

Figure 19-3
*Boîte de dialogue Modèle*



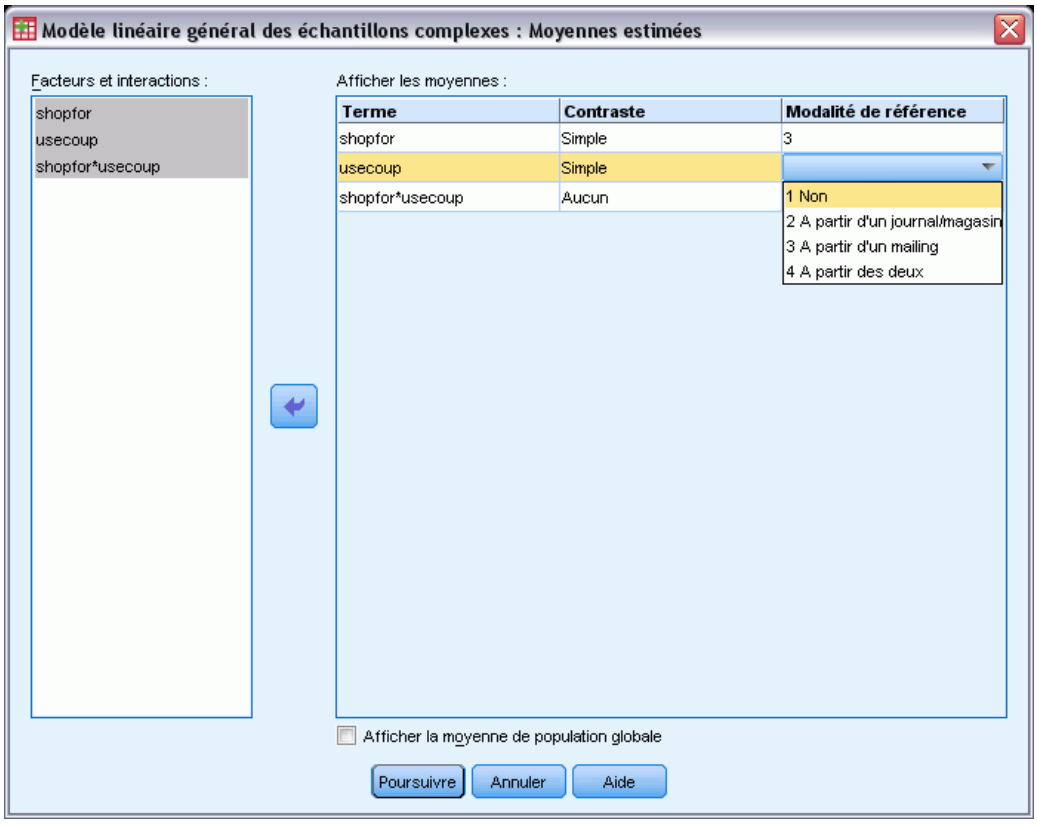

- Choisissez de générer un modèle personnalisé.
- Sélectionnez Effets principaux comme type de terme pour la génération et sélectionnez *CoursesPour* et *UtilisCoupons* comme termes du modèle.
- Sélectionnez Interaction comme type de terme pour la génération et ajoutez l'interaction *CoursesPour*UtilisCoupons* comme terme du modèle.
- Cliquez sur Poursuivre.
- Cliquez sur Statistiques dans la boîte de dialogue Modèle linéaire général.

Figure 19-4
*Boîte de dialogue Statistiques du modèle linéaire général*




- Sélectionnez Estimer, Erreur standard, Intervalle de confiance et Effet de plan dans le groupe de paramètres du modèle.
- Cliquez sur Poursuivre.
- Cliquez sur Moyennes estimées dans la boîte de dialogue Modèle linéaire général.

Figure 19-5
*Boîte de dialogue Moyennes estimées du modèle linéaire général*




- Choisissez d'afficher les moyennes pour *CoursesPour*, *UtilisCoupons* et l'interaction *CoursesPour*UtilisCoupons*.
- Sélectionnez un contraste simple et 3 Soi-même et famille comme modalité de référence pour *UtilisCoupons*. Notez que, une fois sélectionnée, la modalité apparaît comme « 3 » dans la boîte de dialogue.
- Sélectionnez un contraste simple et 1 Non comme modalité de référence pour *UtilisCoupons*.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Modèle linéaire général.

## Récapitulatif des modèles

Figure 19-6
*Statistique R-deux*

| | |
|---|---|
| R-deux | ,601 |

a. Modèle : Montant dépensé = (Ordonnée à l'origine) + shopfor + usecoup + shopfor * usecoup

*R*-deux, le coefficient de détermination est une mesure de la force de l'ajustement du modèle. Il indique qu'environ 60 % de la variation dans *Budget consacré* est expliquée par le modèle, ce qui vous permet d'expliquer certaines choses. Vous voulez peut-être ajouter d'autres variables indépendantes au modèle pour en améliorer l'ajustement.

## Tests des effets de modèle

Figure 19-7
*Tests des effets inter-sujets*

| Source | df1 | df2 | Wald F | Sig. |
|---|---|---|---|---|
| (Modèle corrigé) | 11,000 | 3,000 | 127,231 | ,001 |
| (Ordonnée à l'origine) | 1,000 | 13,000 | 6321,597 | ,000 |
| shopfor | 2,000 | 12,000 | 643.,593 | ,000 |
| usecoup | 3,000 | 11,000 | 87,453 | ,000 |
| shopfor * usecoup | 6,000 | 8,000 | 10,688 | ,020 |

a. Modèle : Montant dépensé = (Ordonnée à l'origine) + shopfor + usecoup + shopfor * usecoup

Chaque terme du modèle, ainsi que le modèle dans son ensemble, est testé pour savoir si son effet est égal à 0. Les termes avec des valeurs de signification inférieures à 0,05 ont des effets perceptibles. Tous les termes contribuent donc au modèle.

## Estimations de paramètre

Figure 19-8
*Estimations des paramètres*

| Paramètre | Estimation | Erreur std. | Intervalle de confiance 95% | | Effet du plan |
|---|---|---|---|---|---|
| | | | Inférieur | Supérieur | |
| (Intercept) | 518,249 | 11,731 | 492,905 | 543,592 | 1,387 |
| [shopfor=1] | -174,757 | 10,762 | -198,0 | -151,51 | ,950 |
| [shopfor=2] | -129,443 | 11,455 | -154,2 | -104,70 | ,925 |
| [shopfor=3] | ,000[a] | . | . | . | . |
| [usecoup=1] | -140,838 | 10,180 | -162,8 | -118,85 | ,649 |
| [usecoup=2] | -63,026 | 13,195 | -91,531 | -34,520 | ,940 |
| [usecoup=3] | -31,375 | 9,726 | -52,387 | -10,363 | ,564 |
| [usecoup=4] | ,000[a] | . | . | . | . |
| [shopfor=1] * [usecoup=1] | 41,693 | 11,170 | 17,562 | 65,824 | ,606 |
| [shopfor=1] * [usecoup=2] | 44,505 | 18,068 | 5,471 | 83,539 | 1,413 |
| [shopfor=1] * [usecoup=3] | 9,204 | 11,057 | -14,684 | 33,092 | ,594 |
| [shopfor=1] * [usecoup=4] | ,000[a] | . | . | . | . |
| [shopfor=2] * [usecoup=1] | 89,211 | 10,967 | 65,518 | 112,903 | ,533 |
| [shopfor=2] * [usecoup=2] | 54,267 | 14,949 | 21,972 | 86,562 | ,836 |
| [shopfor=2] * [usecoup=3] | 17,884 | 13,753 | -11,828 | 47,595 | ,797 |
| [shopfor=2] * [usecoup=4] | ,000[a] | . | . | . | . |
| [shopfor=3] * [usecoup=1] | ,000[a] | . | . | . | . |
| [shopfor=3] * [usecoup=2] | ,000[a] | . | . | . | . |
| [shopfor=3] * [usecoup=3] | ,000[a] | . | . | . | . |
| [shopfor=3] * [usecoup=4] | ,000[a] | . | . | . | . |

a. Paramétré sur zéro car ce paramètre est redondant.
b. Modèle : Montant dépensé = (Ordonnée à l'origine) + shopfor + + usecoup + shopfor * usecoup

Les estimations des paramètres indiquent l'effet de chaque variable indépendante sur le *montant consacré*. La valeur 421,784 de la constante indique que la chaîne d'épicerie peut s'attendre à ce qu'un client avec une famille et utilisant des coupons de journal et de publipostage ciblé dépense 421,78 € en moyenne. Vous pouvez voir que la constante est associée à ces niveaux de facteurs car il s'agit des niveaux de facteurs dont les paramètres sont redondants.

- Les coefficients *coursespour* suggèrent que, parmi les clients qui utilisent à la fois les coupons de journal et de publipostage, ceux qui n'ont pas de famille tendent à dépenser moins que ceux mariés, qui eux-mêmes tendent à dépenser moins que ceux avec des personnes à charge à leur domicile. Puisque les tests des effets de modèle montrent que ce terme contribue au modèle, ces différences ne sont pas dues au hasard.
- Les coefficients *UtilisCoupons* suggèrent que les montants dépensés par les clients avec des personnes à charge à leur domicile baissent avec une utilisation de coupons décroissante. Il existe un certain niveau d'incertitude dans les estimations mais les intervalles de confiance n'incluent pas la valeur 0.
- Les coefficients d'interaction suggèrent que les clients qui n'utilisent pas de coupons ou uniquement des coupons de journaux et n'ont pas de personne à charge tendent à dépenser plus que ce que à quoi vous pouvez vous attendre. Si une portion de paramètre d'interaction est redondante, le paramètre d'interaction est redondant.
- L'écart-type des valeurs des effets de plan indique que certaines erreurs standard calculées pour ces estimations de paramètres sont plus élevées que celles que vous auriez obtenues si vous aviez supposé que ces observations venaient d'un échantillon aléatoire simple, alors que d'autres sont plus petites. Il est extrêmement important d'incorporer les informations de plan d'échantillonnage à votre analyse car vous risquez sinon de conclure que le coefficient *Utilise des coupons=3* n'est pas différent de 0 !

Les estimations de paramètres sont utiles pour quantifier l'effet de chaque terme de modèle, mais les tableaux de moyennes marginales estimées peuvent faciliter l'interprétation des résultats des modèles.

## Moyennes marginales estimées

Figure 19-9
*Moyennes marginales estimées par niveaux de Fait les courses pour*

| Qui fait les courses ? | Moyenne | Erreur std. | Intervalle de confiance 95% | |
|---|---|---|---|---|
| | | | Inférieur | Supérieur |
| Lui/Elle même | 308,5326 | 3,94286 | 300,0145 | 317,0506 |
| Lui/Elle même ou conjoint(e) | 370,3361 | 4,87908 | 359,7955 | 380,8767 |
| Lui/Elle même ou famille | 459,4392 | 7,19769 | 443,8895 | 474,9888 |

Ce tableau affiche les moyennes marginales estimées du modèle et les erreurs standard du *montant consacré* aux niveaux de facteurs *Fait les courses pour*. Ce tableau est utile pour explorer les différences entre les niveaux de ce facteur. Dans cet exemple, on attend d'un client qui fait les courses pour lui-même qu'il dépense environ 246,09 € alors qu'on attend d'un client marié qu'il dépense 295,39 € et qu'on attend d'un client avec personnes à charge qu'il dépense 366,44 €. Pour savoir si ceci est dû à une réelle différence ou à une variation aléatoire, consultez les résultats du test.

Figure 19-10
*Résultats de test individuels pour les moyennes marginales estimées de sexe*

| Contraste simple Qui fait les courses ?[a] | Estimation de contraste | Valeur hypothétique | Différence (valeur estimée - valeur hypothétique) | Erreur std. | df1 | df2 | Wald F | Sig. |
|---|---|---|---|---|---|---|---|---|
| Niveau Lui/Elle même et niveau Lui/Elle même ou famille | -150,907 | ,000 | -150,907 | 4,903 | 1,000 | 13,00 | 947,41 | ,000 |
| Niveau Lui/Elle même ou conjoint(e) et niveau Lui/Elle même ou famille | -89,103 | ,000 | -87,103 | 5,903 | 1,000 | 13,00 | 227,84 | ,000 |

a. Modalité de référence = Lui/Elle même ou famille

Le tableau des tests individuels affiche deux contrastes simples dans les dépenses.

- L'estimation de contraste est la différence entre les dépenses pour les niveaux répertoriés de *Fait les courses pour*.
- La valeur hypothétique de 0,00 confirme qu'il n'y a aucune différence dans les dépenses.
- La statistique de Wald *F*, avec les degrés de liberté affichés est utilisée pour tester si la différence entre une estimation de contraste et une valeur hypothétique est due à une variation aléatoire.
- Puisque les valeurs de signification sont inférieures à 0,05, vous pouvez en conclure qu'il existe des différences dans les dépenses.

Les valeurs des estimations de contraste sont différentes des estimations de paramètres. Ceci est dû à l'existence d'un terme d'interaction contenant l'effet *Fait les courses pour*. En conséquence, l'estimation de paramètre pour *CoursesPour=1* est un contraste simple entre les niveaux *Soi-même* et *Soi-même et famille* au niveau *Des deux* de la variable *Utilise des coupons*. L'estimation de contraste dans ce tableau est considéré par sa moyenne des niveaux de *Utilise des coupons*.

Figure 19-11
*Résultats de tests généraux pour les moyennes marginales estimées de sexe*

| df1 | df2 | Wald F | Sig. |
|---|---|---|---|
| 2,000 | 12,000 | 643,593 | ,000 |

Le tableau de test général rapporte les résultats d'un test de tous les contrastes du tableau de test individuel. Sa valeur de signification inférieure à 0,05 confirme qu'il existe une différence dans les dépenses entre les niveaux de *Fait les courses pour*.

Figure 19-12
*Moyennes marginales estimées par niveaux du comportement lors des courses*

| Utilisation de coupon | Moyenne | Erreur std. | Intervalle de confiance 95% | |
|---|---|---|---|---|
| | | | Inférieur | Supérieur |
| Non | 319,6455 | 6,51429 | 305,5722 | 333,7188 |
| A partir d'un journal/magasine | 386,7469 | 4,32295 | 377,4077 | 396,0861 |
| A partir d'un mailing | 394,5028 | 5,54218 | 382,5297 | 406,4760 |
| A partir des deux | 416,8486 | 6,51260 | 402,7790 | 430,9182 |

Ce tableau affiche les moyennes marginales estimées du modèle et les erreurs standard du *montant consacré* aux niveaux de facteurs *Utilise des coupons*. Ce tableau est utile pour explorer les différences entre les niveaux de ce facteur. Dans cet exemple, on attend d'un client qui n'utilise

pas de coupons qu'il dépense environ 255 €, alors qu'on attend d'un client utilisant des coupons qu'il dépense beaucoup plus.

Figure 19-13
*Résultats de test individuels pour les moyennes marginales estimées du comportement lors des courses*

| Contraste simple Utilisation de coupon[a] | Estimation de contraste | Valeur hypothétique | Différence (valeur estimée - valeur hypothétique) | Erreur std. | df1 | df2 | Wald F | Sig. |
|---|---|---|---|---|---|---|---|---|
| Niveau A partir d'un journal/magasine et niveau Non | 67,101 | ,000 | 67,101 | 6,537 | 1,00 | 13,0 | 105,35 | ,000 |
| Niveau A partir d'un mailing et niveau Non | 74,857 | ,000 | 74,857 | 5,875 | 1,00 | 13,0 | 162,33 | ,000 |
| Niveau A partir des deux et niveau Non | 97,203 | ,000 | 97,203 | 5,603 | 1,00 | 13,0 | 300,92 | ,000 |

a. Modalité de référence = Non

Le tableau des tests individuels affiche trois contrastes simples, comparant les dépenses des clients qui n'utilisent pas de coupons à celles de ceux qui en utilisent.

Puisque les valeurs de signification des tests sont inférieures à 0,05, vous pouvez conclure que les clients qui utilisent des coupons tendent à dépenser plus que ceux qui n'en utilisent pas.

Figure 19-14
*Résultats de test généraux pour les moyennes marginales estimées du comportement lors des courses*

| df1 | df2 | Wald F | Sig. |
|---|---|---|---|
| 3,000 | 11,000 | 87,453 | ,000 |

Le tableau de résultats de test généraux rapporte les résultats d'un test de tous les contrastes du tableau de test individuel. Sa valeur de signification inférieure à 0,05 confirme qu'il existe une différence dans les dépenses entre les niveaux d'*Utilise des coupons*. Notez que les tests généraux pour *Utilise des coupons* et *Fait les courses pour* sont équivalents aux tests des effets du modèle car les valeurs de contrastes hypothétiques sont égales à 0.

Figure 19-15
*Moyennes marginales estimées par niveaux de sexe par comportement lors des courses*

| Qui fait les courses ? | Utilisation de coupon | Moyenne | Erreur std. | Intervalle de confiance 95% | |
|---|---|---|---|---|---|
| | | | | Inférieur | Supérieur |
| Lui/Elle même | Non | 244,3471 | 6,00949 | 231,3664 | 257,3298 |
| | A partir d'un journal/magasine | 324,9708 | 5,94134 | 312,1353 | 337,8063 |
| | A partir d'un mailing | 321,3207 | 4,11028 | 312,4410 | 330,2005 |
| | A partir des deux | 343,4916 | 6,57845 | 329,2729 | 357,7034 |
| Lui/Elle même ou conjoint(e) | Non | 337,1783 | 7,12181 | 321,7925 | 352,5640 |
| | A partir d'un journal/magasine | 380,0468 | 7,91038 | 362,9574 | 397,1361 |
| | A partir d'un mailing | 375,3141 | 6,22468 | 361,8665 | 388,7617 |
| | A partir des deux | 388,8054 | 7,12101 | 373,4214 | 404,1894 |
| Lui/Elle même ou famille | Non | 377,4111 | 11,58215 | 352,3894 | 402,4328 |
| | A partir d'un journal/magasine | 455,2232 | 6,14420 | 441,9494 | 468,4969 |
| | A partir d'un mailing | 486,8736 | 10,76529 | 463,6166 | 510,1306 |
| | A partir des deux | 518,2488 | 11,73120 | 492,9050 | 543,5925 |

Ce tableau affiche les moyennes marginales estimées du modèle, les erreurs standard et les intervalles de confiance du *montant consacré* pour les combinaisons de facteurs *Fait les courses pour* et *Utilise des coupons*. Ce tableau est utile pour explorer l'effet d'interaction entre les deux facteurs trouvé dans les tests des effets du modèle.

### *Récapitulatif*

Dans cet exemple, les moyennes marginales estimées ont révélé les différences de dépenses entre les clients à différents niveaux de *Fait les courses pour* et *Utilise des coupons*. Les tests des effets du modèle l'ont confirmé, de même que l'existence d'un effet d'interaction *Fait les courses pour***Utilise des coupons*. Le tableau récapitulatif des modèles a révélé que le modèle présent explique relativement plus de la moitié de la variation des données et pourrait probablement être amélioré en ajoutant plus de variables indépendantes.

## *Procédures apparentées*

La procédure relative au modèle linéaire général des échantillons complexes est un outil utile pour modéliser une variable d'échelle lorsque les observations ont été conçues en fonction d'un schéma d'échantillonnage complexe.

- L'assistant d'échantillonnage des échantillons complexes permet d'indiquer les spécifications du plan d'échantillonnage complexe et d'obtenir un échantillon. Le fichier de plan d'échantillonnage créé par l'assistant d'échantillonnage contient un plan d'analyse par défaut et peut être spécifié dans la boîte de dialogue Plan lors de l'analyse de l'échantillon obtenu en fonction de ce plan.

- L'assistant de préparation d'analyse des échantillons complexes est utilisé pour indiquer les spécifications d'analyse d'un échantillon complexe existant. Le fichier de plan d'analyse créé par l'assistant d'échantillonnage peut être spécifié dans la boîte de dialogue Plan lorsque vous analysez l'échantillon correspondant à ce plan.
- La procédure de régression logistique des échantillons complexes vous permet de modéliser une réponse qualitative.
- La procédure de régression ordinale des échantillons complexes vous permet de modéliser une réponse quantitative.

Chapitre
20

# Régression logistique des échantillons complexes

La procédure de régression logistique des échantillons complexes effectue une analyse de régression logistique sur une variable dépendante binaire ou multinomiale, pour les échantillons réalisés à l'aide de méthodes d'échantillonnage complexes. Vous pouvez également demander une analyse pour une sous-population.

## Utilisation de la régression logistique des échantillons complexes pour évaluer le risque de crédit

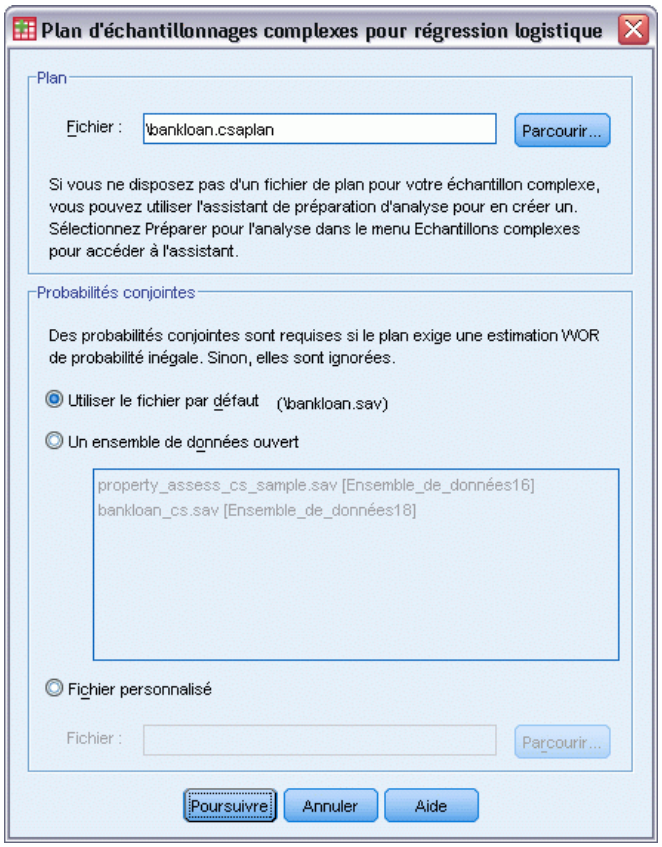
Si vous êtes responsable des prêts dans une banque, vous souhaitez certainement être capable d'identifier les caractéristiques qui indiquent les personnes susceptibles de manquer à leurs engagements, afin de les utiliser pour identifier les bons et les mauvais risques de crédit.

Supposons qu'un responsable des prêts ait recueilli l'historique des prêts octroyés aux clients à différents guichets, en fonction d'un plan complexe. Ces informations sont contenues dans le fichier *bankloan_cs.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*. Ce responsable souhaite voir si la probabilité de défaut de paiement est liée à l'âge, au parcours professionnel et au montant de la dette, et ce, en intégrant le plan d'échantillonnage.

### Exécution de l'analyse

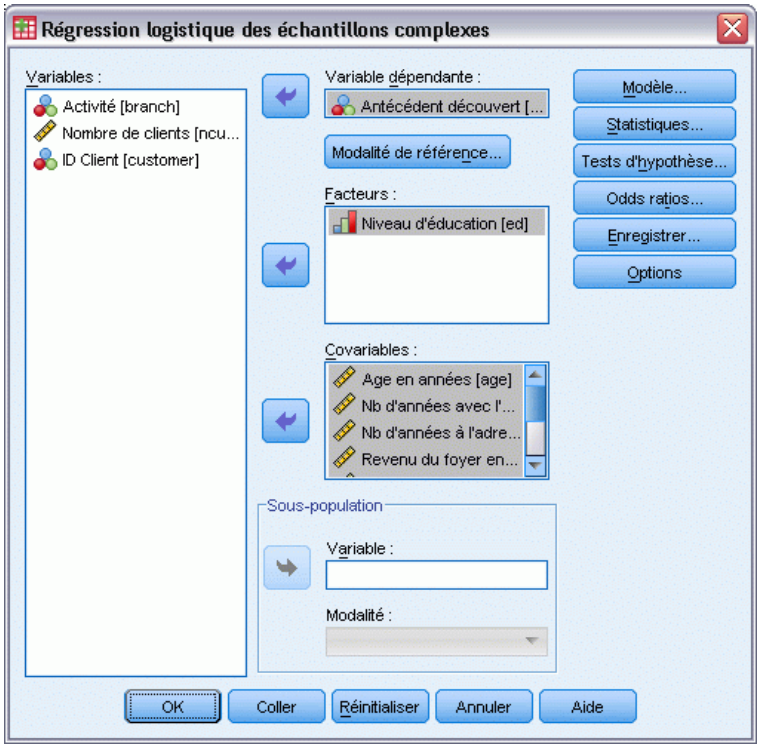
- Pour créer le modèle de régression logistique, sélectionnez, à partir du menu :
  Analyse > Echantillonnage > Régression logistique...

Figure 20-1
*Boîte de dialogue Plan d'échantillonnages complexes*



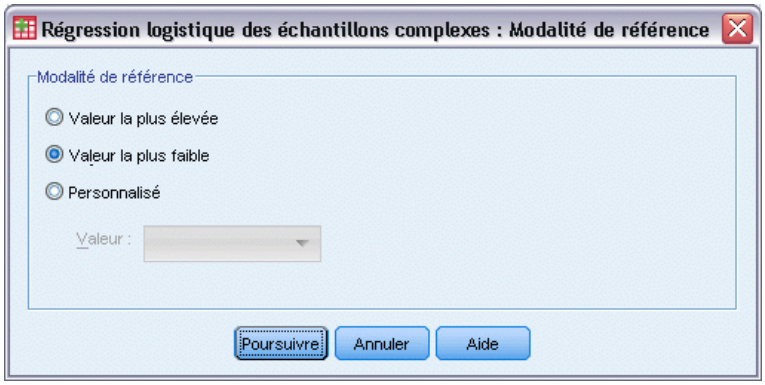

- Recherchez et sélectionnez *bankloan.csaplan*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*.
- Cliquez sur Poursuivre.

Figure 20-2
*Boîte de dialogue Régression logistique*




- Sélectionnez *Manquement précédent* comme variable dépendante.
- Sélectionnez *Niveau d'éducation* comme facteur.
- Sélectionnez les options de *Age en années* à *Autres dettes en milliers* comme covariables.
- Sélectionnez *Manquement précédent* et cliquez sur Modalité de référence.

Figure 20-3
*Boîte de dialogue Modalité de référence de la régression logistique*



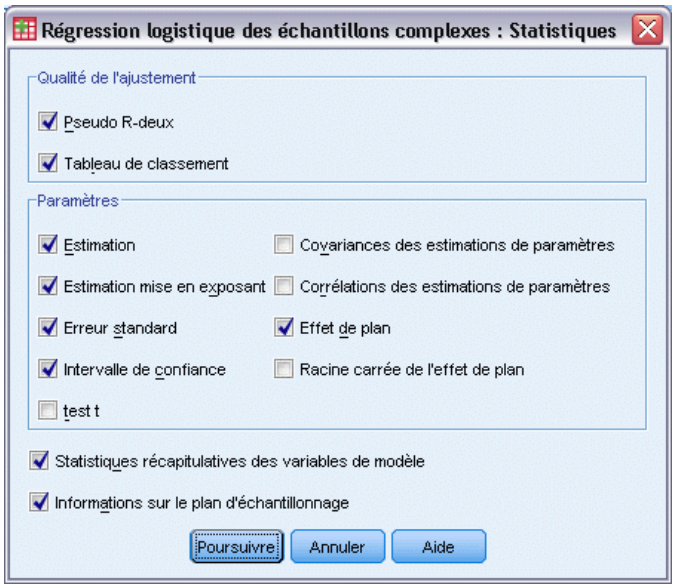

- Sélectionnez Valeur la moins élevée comme modalité de référence.

  Vous définissez ainsi la modalité « N'a pas manqué » comme modalité de référence ; ainsi les odds ratios rapportés dans les résultats indiquent que la probabilité de manquement augmente avec des odds ratios croissants.

- Cliquez sur Poursuivre.

- Cliquez sur Statistiques dans la boîte de dialogue Régression logistique.

Figure 20-4
*Boîte de dialogue Statistiques de régression logistique*



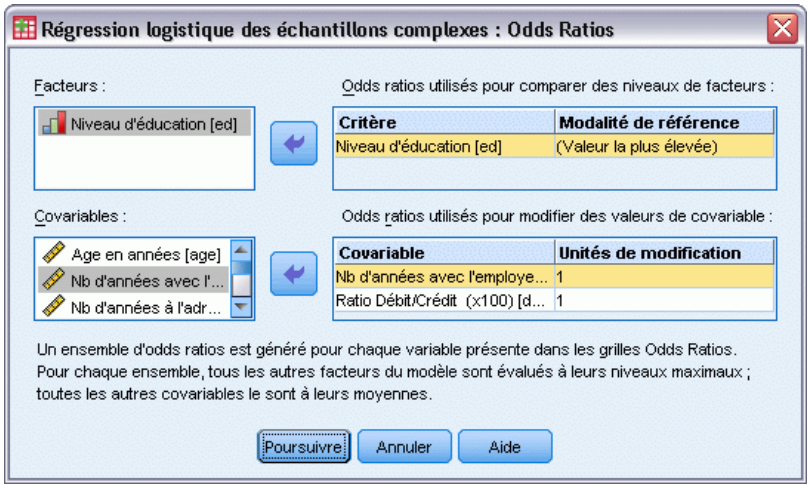

- Sélectionnez Tableau de classement dans le groupe Ajustement du modèle

- Sélectionnez Estimer, Estimation mise en exposant, Erreur standard, Intervalle de confiance et Effet de plan dans le groupe de paramètres.

- Cliquez sur Poursuivre.

▶ Cliquez sur Odds Ratios dans la boîte de dialogue Régression logistique.

Figure 20-5
*Boîte de dialogue Odds Ratios de la régression logistique*




▶ Choisissez de créer des odds ratios pour le facteur *ed* et les covariables *emploi* et *dettrev*.

▶ Cliquez sur Poursuivre.

▶ Cliquez sur OK dans la boîte de dialogue Régression logistique.

## *Pseudo R-deux*

Figure 20-6
*Statistiques de pseudo R-deux*

| | |
|---|---|
| Cox et Snell | ,330 |
| Nagelkerke | ,451 |
| McFadden | ,304 |

Variable dépendante : Antécédent découvert (modalité de référence = Non) Modèle : (Ordonnée à l'origine), ed, age, employ, address, income, debtinc, creddebt, othdebt

Dans un modèle de régression linéaire, le coefficient de détermination $R^2$ récapitule la proportion de la variance dans la variable dépendante associée aux prédicteurs (variables indépendantes). Les valeurs de $R^2$ les plus élevées indiquent que le modèle explique davantage de variation, jusqu'à un maximum de 1. Pour les modèles de régression comportant une variable dépendante catégorielle, il n'est pas possible de calculer une seule statistique $R^2$ regroupant toutes les caractéristiques de $R^2$ du modèle de régression linéaire ; c'est pourquoi ces approximations sont calculées. Les méthodes suivantes sont utilisées pour estimer le coefficient de détermination.

- Le $R^2$(Cox et Snell, 1989) de Cox & Snell est basé sur le log de vraisemblance du modèle comparé au log de vraisemblance du modèle de la ligne de base. Cependant, avec des résultats catégoriels, la valeur théorique maximale est inférieure à 1, y compris dans un modèle « parfait ».

- Le $R^2$(Nagelkerke, 1991) de Nagelkerke est une version ajustée du *R*-carré de Cox & Snell qui ajuste l'échelle de la statistique pour couvrir la totalité de l'intervalle compris entre 0 et 1.
- Le $R^2$(McFadden, 1974) de McFadden est une autre version, basée sur les noyaux de log-vraisemblance, pour le modèle à constante seulement et le modèle entièrement estimé.

Les raisons pour lesquelles une valeur de $R^2$ peut être considérée comme "bonne" varient selon les champs d'application. Si ces statistiques peuvent être subjectives en soi, elles sont particulièrement utiles pour comparer des modèles en compétition pour les mêmes données. Le modèle dont la statistique $R^2$ est la plus élevée est le "meilleur" selon cette mesure.

## *Classification*

Figure 20-7
*Tableau de classement*

| | Prévisions | | |
|---|---|---|---|
| Observations | Non | Oui | Pourcentage correct |
| Non | 188289,7 | 31871,267 | 85,5% |
| Oui | 49970,600 | 77675,133 | 60,9% |
| Pourcentage global | 68,5% | 31,5% | 76,5% |

Variable dépendante : Antécédent découvert (modalité de référence = Non) Modèle
: (Ordonnée à l'origine), ed, age, employ, address, income, debtinc, creddebt, othdebt

Le tableau de classement affiche les résultats pratiques de l'utilisation du modèle de régression logistique. Pour chaque observation, la réponse prévue est *Oui* si le logit prévu de ce modèle est supérieur à 0. Les observations sont pondérées par *finalweight*, afin que le tableau de classement rapporte les performances attendues du modèle dans la population.

- Les cellules de la diagonale sont des prévisions correctes.
- Les cellules hors de la diagonale sont des prévisions incorrectes.

En vous basant sur les observations utilisées pour créer le modèle, vous pouvez envisager de classer correctement 85,5 % des personnes ne manquant pas à leurs engagements dans la population utilisant ce modèle. De même, vous pouvez envisager de classer correctement 60,9 % des personnes manquant à leurs engagements. Au total, vous pouvez envisager de classer 76,5 % des observations correctement ; cependant, du fait que ce tableau a été construit sur des observations utilisées pour créer le modèle, ces estimations sont probablement un peu trop optimistes.

## *Tests des effets de modèle*

Figure 20-8
*Tests des effets inter-sujets*

| Source | df1 | df2 | Wald F | Sig. |
|---|---|---|---|---|
| (Modèle corrigé) | 11,000 | 4,000 | 14,669 | ,010 |
| (Ordonnée à l'origine) | 1,000 | 14,000 | 5,777 | ,031 |
| ed | 4,000 | 11,000 | 1,683 | ,224 |
| age | 1,000 | 14,000 | 5,352 | ,036 |
| employ | 1,000 | 14,000 | 88,244 | ,000 |
| address | 1,000 | 14,000 | 1,123 | ,307 |
| income | 1,000 | 14,000 | ,007 | ,932 |
| debtinc | 1,000 | 14,000 | 27,632 | ,000 |
| creddebt | 1,000 | 14,000 | 33,402 | ,000 |
| othdebt | 1,000 | 14,000 | ,709 | ,414 |

Variable dépendante : Antécédent découvert (modalité de référence = Non)
Modèle
: (Ordonnée à l'origine), ed, age, employ, address, income, debtinc, creddebt, othdebt

Chaque terme du modèle, ainsi que le modèle dans son ensemble, est testé pour savoir si son effet est égal à 0. Les termes avec des valeurs de signification inférieures à 0,05 ont des effets perceptibles. De cette manière, *âge*, *emploi*, *dettrev* et *dettcred* contribuent au modèle, contrairement aux autres effets principaux. Dans une analyse approfondie des données, vous allez probablement supprimer *éd*, *adresse*, *revenu* et *autrdettes* de la considération du modèle.

## *Estimations de paramètre*

Figure 20-9
*Estimations des paramètres*

| Antécédent découvert | Paramètre | B | Erreur std. | Intervalle de confiance 95% Inférieur | Intervalle de confiance 95% Supérieur | Effet du plan | Exp (B) | Intervalle de confiance à 95 % pour Exp(B) Inférieur | Intervalle de confiance à 95 % pour Exp(B) Supérieur |
|---|---|---|---|---|---|---|---|---|---|
| Oui | (Ordonnée à l'origine) | -1,140 | ,399 | -1,995 | -,284 | ,665 | ,320 | ,136 | ,753 |
| | [ed=1] | ,720 | ,340 | -,010 | 1,449 | ,862 | 2,054 | ,990 | 4,259 |
| | [ed=2] | ,684 | ,371 | -,112 | 1,481 | 1,247 | 1,983 | ,894 | 4,397 |
| | [ed=3] | ,518 | ,307 | -,140 | 1,177 | ,813 | 1,679 | ,869 | 3,244 |
| | [ed=4] | ,789 | ,302 | ,142 | 1,437 | ,817 | 2,202 | 1,152 | 4,208 |
| | [ed=5] | ,000[a] | . | . | . | . | 1,000 | . | . |
| | age | -,023 | ,010 | -,043 | -,002 | ,418 | ,978 | ,958 | ,998 |
| | employ | -,225 | ,024 | -,277 | -,174 | 1,200 | ,798 | ,758 | ,840 |
| | address | -,028 | ,026 | -,085 | ,029 | ,651 | ,972 | ,919 | 1,029 |
| | income | ,000 | ,003 | -,007 | ,006 | 1,410 | 1,000 | ,993 | 1,006 |
| | debtinc | ,095 | ,018 | ,056 | ,134 | 1,222 | 1,100 | 1,058 | 1,143 |
| | creddebt | ,493 | ,085 | ,310 | ,676 | 1,373 | 1,637 | 1,363 | 1,966 |
| | othdebt | ,026 | ,031 | -,041 | ,094 | 1,219 | 1,027 | ,960 | 1,098 |

Variable dépendante : Antécédent découvert (modalité de référence = Non) Modèle
: (Ordonnée à l'origine), ed, age, employ, address, income, debtinc, creddebt, othdebt

a. Paramétré sur zéro car ce paramètre est redondant.

Le tableau des estimations de paramètres récapitule l'effet de chaque variable indépendante. Notez que les valeurs du paramètre affectent la probabilité de la modalité « A manqué » relative à la modalité « N'a pas manqué ». Cependant, les paramètres avec des coefficients positifs

augmentent la probabilité de manquement aux engagements, alors que les paramètres avec des coefficients négatifs diminuent la probabilité de manquement aux engagements.

La signification d'un coefficient de régression logistique n'est pas aussi évidente que celle d'un coefficient de régression linéaire. Bien que *B* soit pratique pour tester les effets du modèle, *Exp(B)* est plus facile à interpréter. *Exp(B)* représente la modification du rapport dans la cote de l'événement étudié attribuable à l'augmentation d'une unité de la variable indépendante pour les variables indépendantes qui ne font pas partie de termes d'interaction. Par exemple, *Exp(B)* pour *emploi* est égal à 0,798, ce qui signifie que la cote de manquement des personnes qui ont le même employeur depuis deux ans est 0,798 fois la cote des manquements de ceux qui ont le même employeur depuis un an, tout le reste étant identique.

Les effets de plan indiquent que certaines erreurs standard calculées pour ces estimations de paramètres sont plus élevées que celles que vous auriez obtenues si vous aviez supposé que ces observations provenaient d'un échantillon aléatoire simple, alors que d'autres sont plus petites. Il est extrêmement important d'incorporer les informations de plan d'échantillonnage à votre analyse car vous risquez sinon de conclure, par exemple, que le coefficient âge n'est pas différent de 0 !

## *Odds Ratios*

Figure 20-10
*Odds ratios pour le niveau d'éducation*

| | | Antécédent découvert | Odds Ratios | Intervalle de confiance 95% | |
|---|---|---|---|---|---|
| | | | | Inférieur | Supérieur |
| Niveau d'éducation | Niveau bac et Bac+5 et plus | Oui | 2,054 | ,990 | 4,259 |
| | Bac+2 et Bac+5 et plus | Oui | 1,983 | ,894 | 4,397 |
| | Bac+3 et Bac+5 et plus | Oui | 1,679 | ,869 | 3,244 |
| | Bac+4 et Bac+5 et plus | Oui | 2,202 | 1,152 | 4,208 |

Variable dépendante : Antécédent découvert (modalité de référence = Non) Modèle : (Ordonnée à l'origine), ed, age, employ, address, income, debtinc, creddebt, othdebt

a. Les valeurs ci-après sont fixées pour les covariables et les facteurs utilisés pour le calcul : Niveau d'éducation=Bac+5 et plus; Age en années=34,19; Nb d'années avec l'employeur actuel=6,99; Nb d'années à l'adresse actuelle=6,32; Revenu du foyer en milliers de $=60,16; Ratio Débit/Crédit (x100)=9,9341; Débit carte de crédit en milliers de $=1,9764; Autre dette en milliers de $=3,9164

Ce tableau affiche les odds ratios de *Manquement précédent* aux niveaux des facteurs du *niveau d'éducation*. Les valeurs rapportées sont les rapports des cotes de manquement allant de *N'a pas terminé le lycée* à *Diplôme de premier cycle*, par rapport à la cote de manquement pour *Diplôme de second cycle*. Ainsi, un odds ratio de 2,054 dans la première ligne du tableau signifie que la cote de manquement pour une personne qui n'a pas terminé le lycée est 2,054 fois la cote de manquement d'une personne diplômée du second cycle.

Figure 20-11
*Odds ratios pour le nombre d'années travaillées chez l'employeur actuel*

| Unités de modification | | Antécédent découvert | Odds Ratios | Intervalle de confiance 95% | |
|---|---|---|---|---|---|
| | | | | Inférieur | Supérieur |
| Nb d'années avec l'employeur actuel | 1,000 | Oui | ,798 | ,758 | ,840 |

Variable dépendante : Antécédent découvert (modalité de référence = Non) Modèle : (Ordonnée à l'origine), ed, age, employ, address, income, debtinc, creddebt, othdebt

a. Les valeurs ci-après sont fixées pour les covariables et les facteurs utilisés pour le calcul : Niveau d'éducation=Bac+5 et plus; Age en années=34,19; Nb d'années avec l'employeur actuel=6,99; Nb d'années à l'adresse actuelle=6,32; Revenu du foyer en milliers de $=60,16; Ratio Débit/Crédit (x100)=9,9341; Débit carte de crédit en milliers de $=1,9764; Autre dette en milliers de $=3,9164

Ce tableau affiche les odds ratios de *Manquement précédent* pour une modification d'unité de la covariable *Nb d'années avec l'employeur actuel*. La valeur reportée est le rapport de la cote de manquement pour une personne avec 7,99 années d'ancienneté comparé à la cote de manquement d'une personne avec 6,99 années (la moyenne).

Figure 20-12
*Odds ratios pour les dettes en fonction du taux de revenu*

| Unités de modification | | Antécédent découvert | Odds Ratios | Intervalle de confiance 95% | |
|---|---|---|---|---|---|
| | | | | Inférieur | Supérieur |
| Ratio Débit/Crédit (x100) | 1,000 | Oui | 1,100 | 1,058 | 1,143 |

Variable dépendante : Antécédent découvert (modalité de référence = Non) Modèle : (Ordonnée à l'origine), ed, age, employ, address, income, debtinc, creddebt, othdebt

a. Les valeurs ci-après sont fixées pour les covariables et les facteurs utilisés pour le calcul : Niveau d'éducation=Bac+5 et plus; Age en années=34,19; Nb d'années avec l'employeur actuel=6,99; Nb d'années à l'adresse actuelle=6,32; Revenu du foyer en milliers de $=60,16; Ratio Débit/Crédit (x100)=9,9341; Débit carte de crédit en milliers de $=1,9764; Autre dette en milliers de $=3,9164

Ce tableau affiche les odds ratios de *Manquement précédent* pour une modification d'unité de la covariable *Rapport dette/revenu*. La valeur reportée est le rapport de la cote de manquement pour une personne avec un taux de dette par revenu de 10,9341 comparé à la cote de manquement d'une personne avec 9,9341 (la moyenne).

Notez que comme aucune de ces variables indépendantes ne fait partie des termes d'interaction, les valeurs des odds ratios rapportées dans ces tableaux sont égales aux valeurs des estimations de paramètres exponentielles. Lorsqu'une variable indépendante fait partie d'un terme d'interaction, son odds ratio tel qu'il est rapporté dans ces tableaux dépend également des valeurs des autres variables indépendantes qui constituent l'interaction.

## *Récapitulatif*

A l'aide de la procédure Régression logistique des échantillons complexes, vous avez construit un modèle pour prévoir la probabilité qu'un client donné manque à son prêt.

Un problème important pour les responsables des prêts est le coût des erreurs de type I et de type II. En fait, quel est le coût du classement d'une personne manquant à ses engagements dans la catégorie des personnes ne manquant pas à leurs engagements (type I) ? Quel est le coût du classement d'une personne ne manquant pas à ses engagements dans la catégorie des personnes manquant à leurs engagements (type II) ? Si les mauvaises dettes sont votre préoccupation principale, alors minimisez votre erreur de type I et maximisez votre **sensibilité**.

Si le développement de votre base client est la priorité, abaissez alors votre erreur de type II et maximisez votre **spécificité**. Habituellement, les deux sont des préoccupations majeures, vous devez donc choisir une règle de décision optimisant à la fois la sensibilité et la spécificité pour classer les clients.

## Procédures apparentées

La procédure de régression logistique des échantillons complexes est un outil utile pour modéliser une variable qualitative lorsque les observations ont été conçues en fonction d'un schéma d'échantillonnage complexe.

- L'assistant d'échantillonnage des échantillons complexes permet d'indiquer les spécifications du plan d'échantillonnage complexe et d'obtenir un échantillon. Le fichier de plan d'échantillonnage créé par l'assistant d'échantillonnage contient un plan d'analyse par défaut et peut être spécifié dans la boîte de dialogue Plan lors de l'analyse de l'échantillon obtenu en fonction de ce plan.
- L'assistant de préparation d'analyse des échantillons complexes est utilisé pour indiquer les spécifications d'analyse d'un échantillon complexe existant. Le fichier de plan d'analyse créé par l'assistant d'échantillonnage peut être spécifié dans la boîte de dialogue Plan lorsque vous analysez l'échantillon correspondant à ce plan.
- La procédure Echantillons complexes - Modèle linéaire général vous permet de modéliser une réponse d'échelle.
- La procédure de régression ordinale des échantillons complexes vous permet de modéliser une réponse quantitative.

Chapitre
21

# Régression ordinale des échantillons complexes

La procédure de régression ordinale des échantillons complexes crée un modèle de prévision pour une variable dépendante ordinale, pour les échantillons réalisés à l'aide des méthodes d'échantillonnage complexe. Vous pouvez également demander une analyse pour une sous-population.

## Utilisation de la procédure de régression ordinale des échantillons complexes pour analyser des résultats d'enquête

Des élus étudiant un projet de loi devant l'assemblée législative souhaitent savoir si ce projet est populaire auprès des électeurs et déterminer le lien existant entre cette popularité et la répartition démographique des électeurs. Les enquêteurs conçoivent et mènent des entretiens en fonction d'un plan d'échantillonnage complexe.

Les résultats de l'enquête sont recueillis dans le fichier *poll_cs_sample.sav*. Le plan d'échantillonnage utilisé par les enquêteurs se trouve dans le fichier *poll.csplan*. Ce plan faisant appel à une méthode de probabilité proportionnelle à la taille (PPS - Probability proportional to size), il existe également un fichier contenant les probabilités de sélection conjointes (*poll_jointprob.sav*). Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*.Utilisez la procédure de régression ordinale des échantillons complexes afin d'ajuster un modèle concernant la cote de popularité du projet de loi en fonction de la répartition démographique des électeurs.

### Exécution de l'analyse

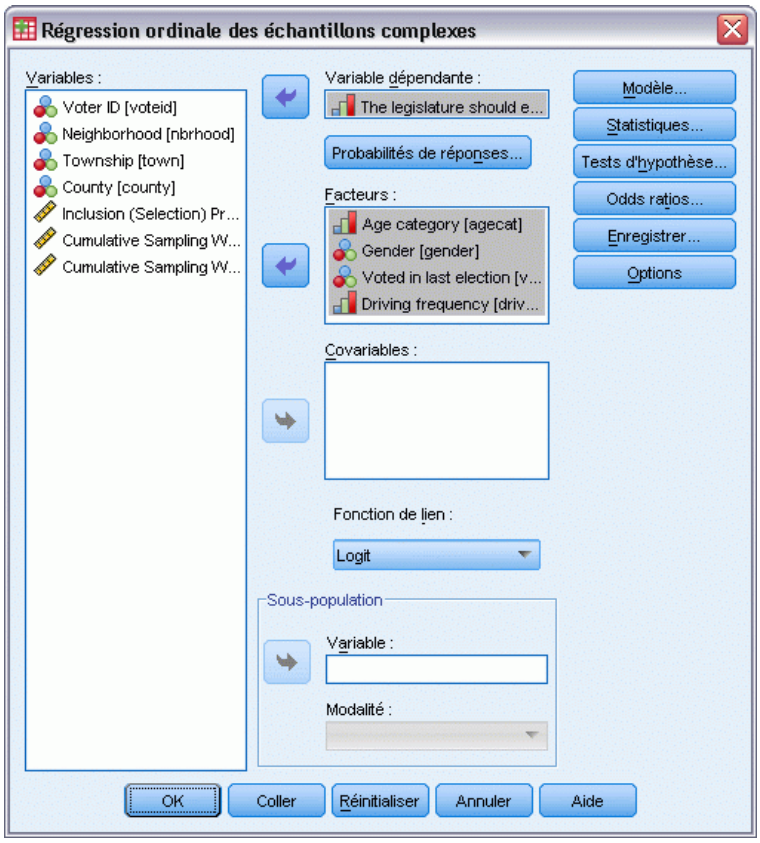
- Pour exécuter une analyse Echantillons complexes - Régression ordinale, sélectionnez les options suivantes dans le menu :

  Analyse > Echantillonnage > Régression ordinale…

Figure 21-1
*Boîte de dialogue Plan d'échantillonnages complexes*

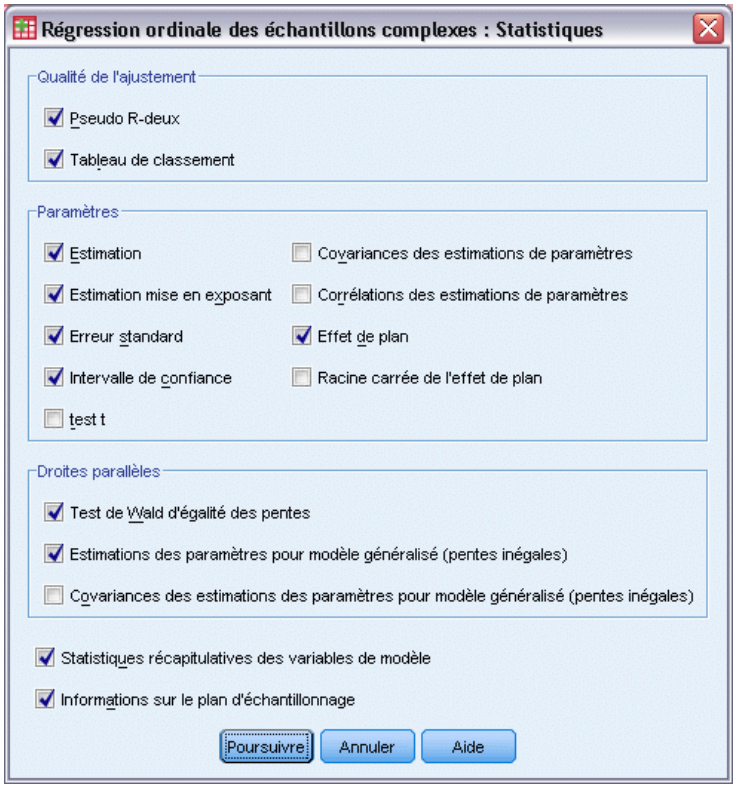

- Accédez au fichier *poll.csplan*, puis sélectionnez-le comme fichier de plan. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*.
- Sélectionnez le fichier de probabilités conjointes *poll_jointprob.sav*.
- Cliquez sur Poursuivre.

Figure 21-2
*Boîte de dialogue Régression ordinale*



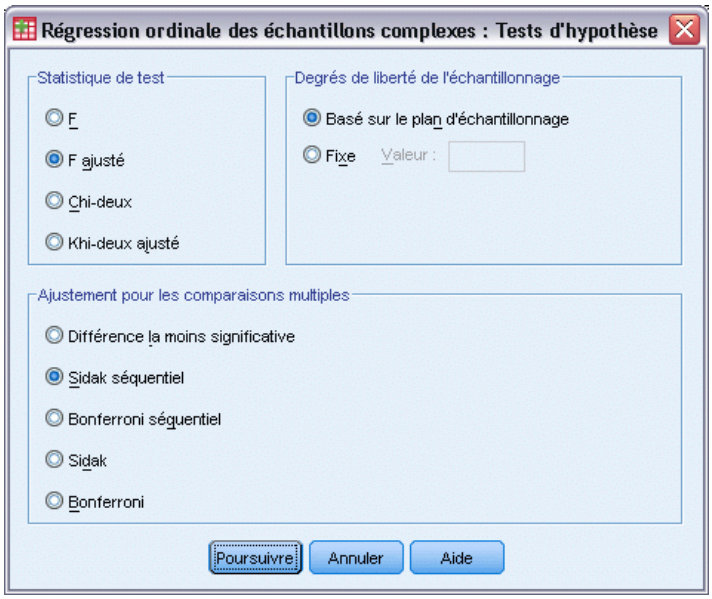

- Sélectionnez la variable *L'assemblée législative doit promulguer une taxe sur le gaz* comme variable dépendante.
- Sélectionnez *Tranche d'âge* à *Effectif* comme facteurs.
- Cliquez sur Statistiques.

Figure 21-3
*Boîte de dialogue Statistiques de régression ordinale*




- Sélectionnez Tableau de classement dans le groupe Ajustement du modèle.
- Sélectionnez Estimer, Estimation mise en exposant, Erreur standard, Intervalle de confiance et Effet de plan dans le groupe de paramètres.
- Sélectionnez Test de Wald d'égalité des pentes et Estimation des paramètres pour modèle généralisé (pentes inégales).
- Cliquez sur Poursuivre.
- Cliquez sur Tests d'hypothèse dans la boîte de dialogue Echantillons complexes - Régression ordinale.

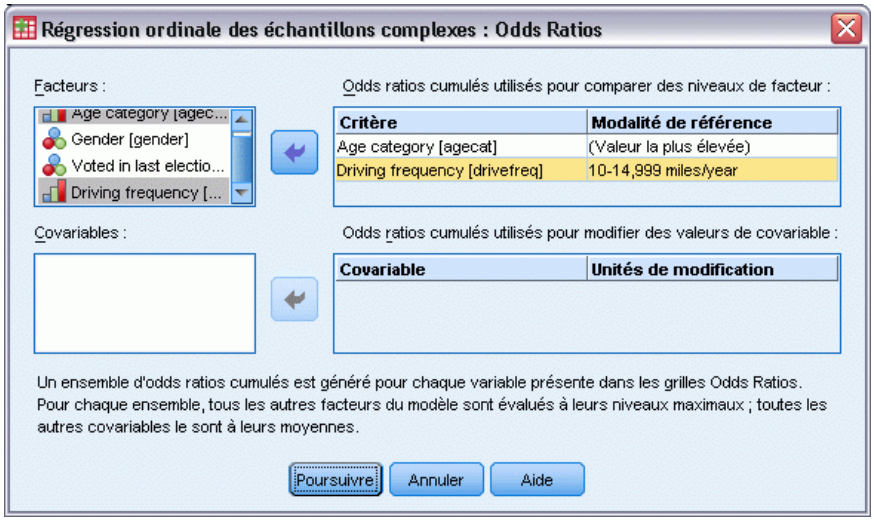
Figure 21-4
*Boîte de dialogue Tests d'hypothèse*




Même pour un nombre moyen de variables indépendantes et de modalités de réponses, la statistique du test *F* de Wald peut s'avérer incalculable pour le test des droites parallèles.

- Sélectionnez F ajusté dans le groupe Statistique de test.
- Sélectionnez Sidak séquentiel comme méthode d'ajustement des comparaisons multiples.
- Cliquez sur Poursuivre.
- Cliquez sur Odds Ratios dans la boîte de dialogue Echantillons complexes - Régression ordinale.

Figure 21-5
*Boîte de dialogue Odds ratios de régression ordinale*




- Choisissez de produire des odds ratios cumulés pour *Tranche d'âge* et *Effectif*.
- Sélectionnez 10-14 999 miles/an, un trajet annuel plus « typique » que le maximum, comme modalité de référence pour *Effectif*.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Echantillons complexes - Régression ordinale.

## *Pseudo R-deux*

Figure 21-6
*Pseudo R-deux*

| | |
|---|---|
| Cox et Snell | ,179 |
| Nagelkerke | ,191 |
| McFadden | ,071 |

Variable dépendante : The legislature should enact a gas tax (Croissant)
Modèle : (Seuil), agecat, gender, votelast, drivefreq
Fonction de lien : Logit

Dans un modèle de régression linéaire, le coefficient de détermination $R^2$ récapitule la proportion de la variance dans la variable dépendante associée aux prédicteurs (variables indépendantes). Les valeurs de $R^2$ les plus élevées indiquent que le modèle explique davantage de variation, jusqu'à un maximum de 1. Pour les modèles de régression comportant une variable dépendante catégorielle, il n'est pas possible de calculer une seule statistique $R^2$ regroupant toutes les caractéristiques de $R^2$ du modèle de régression linéaire ; c'est pourquoi ces approximations sont calculées. Les méthodes suivantes sont utilisées pour estimer le coefficient de détermination.

- Le $R^2$(Cox et Snell, 1989) de Cox & Snell est basé sur le log de vraisemblance du modèle comparé au log de vraisemblance du modèle de la ligne de base. Cependant, avec des résultats catégoriels, la valeur théorique maximale est inférieure à 1, y compris dans un modèle « parfait ».
- Le $R^2$(Nagelkerke, 1991) de Nagelkerke est une version ajustée du *R*-carré de Cox & Snell qui ajuste l'échelle de la statistique pour couvrir la totalité de l'intervalle compris entre 0 et 1.
- Le $R^2$(McFadden, 1974) de McFadden est une autre version, basée sur les noyaux de log-vraisemblance, pour le modèle à constante seulement et le modèle entièrement estimé.

Les raisons pour lesquelles une valeur de $R^2$ peut être considérée comme "bonne" varient selon les champs d'application. Si ces statistiques peuvent être subjectives en soi, elles sont particulièrement utiles pour comparer des modèles en compétition pour les mêmes données. Le modèle dont la statistique $R^2$ est la plus élevée est le "meilleur" selon cette mesure.

## Tests des effets de modèle

Figure 21-7
*Tests des effets de modèle*

| Source | df1 | df2 | Wald F ajusté | Sig. | Sidak Sig. séquentiel |
|---|---|---|---|---|---|
| agecat | 2,283 | 31,966 | 6,215 | ,004 | ,003 |
| gender | 1,000 | 14,000 | ,046 | ,834 | ,834 |
| votelast | 1,000 | 14,000 | ,076 | ,787 | ,787 |
| drivefreq | 3,785 | 52,987 | 228,015 | ,000 | ,000 |

Variable dépendante : The legislature should enact a gas tax (Croissant)
Modèle : (Seuil), agecat, gender, votelast, drivefreq
Fonction de lien : Logit

Chaque terme du modèle est testé pour savoir si son effet est égal à 0. Les termes avec des valeurs de signification inférieures à 0,05 ont des effets perceptibles. Par conséquent, les variables *agecat* et *drivefreq* contribuent au modèle, contrairement aux autres effets principaux. Dans une analyse approfondie des données, vous pourriez envisager de supprimer les variables *gender* et *votelast* du modèle.

## Estimations de paramètre

Le tableau des estimations de paramètres récapitule l'effet de chaque variable indépendante. Bien que l'interprétation des coefficients de ce modèle soit difficile du fait de la nature de la fonction de lien, les signes des coefficients des covariables et ceux des valeurs relatives des coefficients des niveaux de facteurs peuvent donner d'importantes indications concernant les effets des variables indépendantes du modèle.

- Pour les covariables, des coefficients positifs (négatifs) indiquent des relations positives (inverses) entre les variables indépendantes et les résultats. Une valeur de covariable croissante avec un coefficient positif correspond à une probabilité croissante de se situer dans l'une des modalités de résultats cumulés les plus « élevées ».

- Pour les facteurs, un niveau de facteur avec un coefficient supérieur indique une plus grande probabilité de se situer dans l'une des modalités de résultats cumulés les plus « élevées ». Le signe d'un coefficient d'un niveau de facteur dépend de l'effet de ce niveau de facteur par rapport à la modalité de référence.

Figure 21-8
*Estimations des paramètres*

| Paramètre \: | | B | Erreur std. | Intervalle de confiance 95% Inférieur | Intervalle de confiance 95% Supérieur | Effet du plan | Exp (B) \: | Intervalle de confiance à 95 % pour Exp(B) Inférieur | Intervalle de confiance à 95 % pour Exp(B) Supérieur |
|---|---|---|---|---|---|---|---|---|---|
| Seuil | [opinion_gastax=1] | -3,343 | ,104 | -3,566 | -3,120 | 1,132 | ,035 | ,028 | ,044 |
| | [opinion_gastax=2] | -1,910 | ,098 | -2,120 | -1,700 | 1,058 | ,148 | ,120 | ,183 |
| | [opinion_gastax=3] | -,674 | ,090 | -,866 | -,482 | ,915 | ,510 | ,421 | ,618 |
| Régression | [agecat=1] | -,324 | ,079 | -,494 | -,154 | 1,793 | ,723 | ,610 | ,858 |
| | [agecat=2] | -,138 | ,054 | -,255 | -,022 | 1,158 | ,871 | ,775 | ,978 |
| | [agecat=3] | -,095 | ,076 | -,257 | ,068 | 2,206 | ,909 | ,773 | 1,070 |
| | [agecat=4] | ,000[a] | . | . | . | . | 1,00 | . | . |
| | [gender=0] | -,008 | ,035 | -,084 | ,068 | ,949 | ,992 | ,920 | 1,071 |
| | [gender=1] | ,000[a] | . | . | . | . | 1,00 | . | . |
| | [votelast=0] | -,011 | ,039 | -,095 | ,073 | 1,103 | ,989 | ,909 | 1,076 |
| | [votelast=1] | ,000[a] | . | . | . | . | 1,00 | . | . |
| | [drivefreq=1] | -3,751 | ,153 | -4,079 | -3,423 | 1,117 | ,023 | ,017 | ,033 |
| | [drivefreq=2] | -3,003 | ,116 | -3,251 | -2,755 | 1,226 | ,050 | ,039 | ,064 |
| | [drivefreq=3] | -2,295 | ,114 | -2,540 | -2,050 | 1,585 | ,101 | ,079 | ,129 |
| | [drivefreq=4] | -1,570 | ,092 | -1,769 | -1,372 | 1,078 | ,208 | ,171 | ,254 |
| | [drivefreq=5] | -,812 | ,089 | -1,003 | -,621 | ,941 | ,444 | ,367 | ,537 |
| | [drivefreq=6] | ,000[a] | . | . | . | . | 1,00 | . | . |

Variable dépendante : The legislature should enact a gas tax (Croissant)
Modèle : (Seuil), agecat, gender, votelast, drivefreq
Fonction de lien : Logit

a. Paramétré sur zéro car ce paramètre est redondant.

Vous pouvez tirer les conclusions suivantes en fonction des estimations des paramètres :

- Les personnes des tranches d'âge inférieures affichent un plus grand soutien pour le projet de loi que les personnes de la tranche d'âge la plus élevée.
- Les personnes qui conduisent moins souvent affichent un plus grand soutien pour le projet de loi que celles qui conduisent plus souvent.
- Les coefficients des variables *gender* et *votelast*, en plus de ne pas être statistiquement significatifs, semblent faibles par rapport aux autres coefficients.

Les effets de plan indiquent que certaines erreurs standard calculées pour ces estimations de paramètres sont plus élevées que celles que vous auriez obtenues à l'aide d'un échantillon aléatoire simple, alors que d'autres sont plus faibles. Il est capital d'incorporer les informations du plan d'échantillonnage à votre analyse, sinon vous risquez de conclure, par exemple, que le coefficient du troisième niveau de la *tranche d'âge*, *[agecat=3]*, diffère significativement de 0 !

## *Classification*

Figure 21-9
*Informations sur les variables catégorielles*

| | | Effectif pondéré | Pourcentage pondéré |
|---|---|---|---|
| The legislature should enact a gas tax[a] | Strongly agree | 25132,955 | 21,3% |
| | Agree | 32261,425 | 27,3% |
| | Disagree | 29477,417 | 24,9% |
| | Strongly disagree | 31314,203 | 26,5% |
| Age category | 18-30 | 20509,504 | 17,4% |
| | 31-45 | 35380,506 | 29,9% |
| | 46-60 | 34865,792 | 29,5% |
| | >60 | 27430,198 | 23,2% |
| Gender | Male | 61424,547 | 52,0% |
| | Female | 56761,453 | 48,0% |
| Voted in last election | No | 70607,216 | 59,7% |
| | Yes | 47578,784 | 40,3% |
| Driving frequency | Do not own car | 3437,137 | 2,9% |
| | <10,000 miles/year | 10816,349 | 9,2% |
| | 10-14,999 miles/year | 32539,364 | 27,5% |
| | 15-19,999 miles/year | 39179,814 | 33,2% |
| | 20-29,999 miles/year | 25617,804 | 21,7% |
| | >=30,000 miles/year | 6595,532 | 5,6% |
| Taille de la population | | 118186,0 | 100,0% |

a. Les valeurs des variables dépendantes sont triées dans l'ordre croissant.

Etant donné les données observées, le modèle « nul » (qui est un modèle sans variable indépendante) classerait tous les clients dans le groupe modal *D'accord*. Ainsi, le modèle nul serait correct pour 27,3 % des observations.

Figure 21-10
*Tableau de classement*

| | Prévisions | | | | |
|---|---|---|---|---|---|
| Observations | Strongly agree | Agree | Disagree | Strongly disagree | Pourcentage correct |
| Strongly agree | 7067,567 | 12130,814 | 3875,825 | 2058,750 | 28,1% |
| Agree | 4271,234 | 14464,286 | 7320,767 | 6205,137 | 44,8% |
| Disagree | 2024,816 | 11703,368 | 7108,487 | 8640,746 | 24,1% |
| Strongly disagree | 889,869 | 8169,109 | 6946,522 | 15308,703 | 48,9% |
| Pourcentage global | 12,1% | 39,3% | 21,4% | 27,3% | 37,2% |

Variable dépendante : The legislature should enact a gas tax (Croissant)
Modèle : (Seuil), agecat, gender, votelast, drivefreq
Fonction de lien : Logit

Le tableau de classement affiche les résultats pratiques de l'utilisation du modèle. Pour chaque observation, la réponse prévue est la modalité de réponse dotée de la probabilité de prédiction la plus élevée du modèle. Les observations sont pondérées en fonction de la valeur *Pondération d'échantillonnage finale*, de sorte que le tableau de classement reporte les performances attendues du modèle dans la population.

- Les cellules de la diagonale sont des prévisions correctes.
- Les cellules hors de la diagonale sont des prévisions incorrectes.

Le modèle classe correctement 9,9 % de plus ou 37,2 % des observations. Plus particulièrement, le modèle est considérablement plus efficace pour classer les personnes qui sont *d'accord* ou *pas du tout d'accord*, et un peu moins efficace pour classer celles qui sont *plutôt pas d'accord*.

## *Odds Ratios*

Les **odds ratios cumulés** sont définis comme le rapport entre la probabilité que la variable dépendante adopte une valeur inférieure ou égale à une modalité de réponse donnée et la probabilité qu'elle adopte une valeur supérieure à la modalité de réponse. L'**odds ratio cumulé** est le rapport des odds cumulés pour différentes valeurs de variable indépendante. Il est étroitement lié aux estimations de paramètre exponentielles. Fait intéressant, l'odds ratio cumulé lui-même ne dépend pas de la modalité de réponse.

Figure 21-11
*Odds ratios cumulés de la tranche d'âge*

| | | Odds ratio cumulé | Intervalle de confiance 95% | | Effet du plan | Effet du plan (racine carrée) |
|---|---|---|---|---|---|---|
| | | | Inférieur | Supérieur | | |
| Age category | 18-30 et >60 | 1,383 | 1,166 | 1,639 | 1,793 | 1,339 |
| | 31-45 et >60 | 1,148 | 1,022 | 1,290 | 1,158 | 1,076 |
| | 46-60 et >60 | 1,100 | ,935 | 1,294 | 2,206 | 1,485 |

Variable dépendante : The legislature should enact a gas tax (Croissant)
Modèle : (Seuil), agecat, gender, votelast, drivefreq
Fonction de lien : Logit

a. Les valeurs ci-après sont fixées pour les covariables et les facteurs utilisés pour le calcul : Age category=>60; Gender=Female; Voted in last election=Yes; Driving frequency=>=30,000 miles/year

Ce tableau affiche les odds ratios cumulés des niveaux de facteurs de la *tranche d'âge*. Les valeurs rapportées sont les ratios des odds cumulés des tranches *18–30* à *46–60*, par rapport aux odds cumulés de la tranche > *60*. Par conséquent, l'odds ratio de 1,383 sur la première ligne du tableau signifie que les odds cumulés d'une personne âgée de 18–30 ans équivalent à 1,383 fois les odds cumulés d'une personne de plus de 60 ans. La *tranche d'âge* n'intervenant dans aucun terme d'interaction, les odds ratios sont simplement les rapports des estimations de paramètres exponentielles. Par exemple, l'odds ratio cumulé de la tranche d'âge *18–30* par rapport à la tranche >*60* est de 1,00/0,723 = 1,383.

Figure 21-12
*Odds ratios de Effectif*

| | | Odds ratio cumulé | Intervalle de confiance 95% | | Effet du plan | Effet du plan (racine carrée) |
|---|---|---|---|---|---|---|
| | | | Inférieur | Supérieur | | |
| Driving frequency | Do not own car et 10-14,999 miles/year | 4,288 | 2,878 | 6,390 | 2,345 | 1,531 |
| | <10,000 miles/year et 10-14,999 miles/year | 2,030 | 1,656 | 2,488 | 1,838 | 1,356 |
| | 15-19,999 miles/year et 10-14,999 miles/year | ,484 | ,430 | ,546 | 1,450 | 1,204 |
| | 20-29,999 miles/year et 10-14,999 miles/year | ,227 | ,193 | ,267 | 2,095 | 1,448 |
| | >=30,000 miles/year et 10-14,999 miles/year | ,101 | ,079 | ,129 | 1,585 | 1,259 |

Variable dépendante : The legislature should enact a gas tax (Croissant)
Modèle : (Seuil), agecat, gender, votelast, drivefreq
Fonction de lien : Logit

a. Les valeurs ci-après sont fixées pour les covariables et les facteurs utilisés pour le calcul : Age category=>60; Gender=Female; Voted in last election=Yes; Driving frequency=>=30,000 miles/year

Ce tableau affiche les odds ratios cumulés des niveaux de facteurs de la variable *Effectif* en utilisant *10–14 999 miles/an* comme modalité de référence. La variable *Effectif* n'intervenant dans aucun terme d'interaction, les odds ratios sont simplement les rapports des estimations de paramètres exponentielles. Par exemple, l'odds ratio cumulé de *20–29 999 miles/an* par rapport à *10–14 999 miles/an* est de 0,101/0,444 = 0,227.

## Modèle cumulé généralisé

Figure 21-13
*Test de droites parallèles*

| df1 | df2 | Wald F ajusté | Sig. | Sidak Sig. séquentiel |
|---|---|---|---|---|
| 8,769 | 122,767 | 1,894 | ,061 | ,392 |

Variable dépendante : The legislature should enact a gas tax (Croissant)
Modèle : (Seuil), agecat, gender, votelast, drivefreq
Fonction de lien : Logit

Le test des droites parallèles permet de déterminer si l'hypothèse selon laquelle les paramètres sont identiques pour toutes les modalités de réponses est fondée. Ce test compare le modèle estimé comportant un ensemble de coefficients pour toutes les modalités à un modèle généralisé comportant un ensemble distinct de coefficients pour chaque modalité.

Le test *F* de Wald est un test composite de la matrice de contraste pour l'hypothèse des droites parallèles qui fournit des valeurs de *p* asymptotiquement correctes ; pour les échantillons de taille petite à moyenne, la statistique *F* de Wald ajustée donne de bons résultats. La valeur de signification est proche de 0,05, ce qui suggère que le modèle généralisé peut améliorer l'ajustement du modèle. Cependant, le test ajusté Procédure de Sidak séquentielle rapporte une valeur de signification assez élevée (0,392) qui, dans l'ensemble, indique qu'aucun élément ne montre clairement la nécessité de rejeter l'hypothèse des droites parallèles. Le test Procédure de Sidak séquentielle commence par des tests Wald de contraste individuels pour fournir une valeur de *p* globale. Ces résultats doivent être comparables au résultat du test Wald composite. Le fait

qu'ils soient si différents dans cet exemple est assez surprenant, mais peut être lié à l'existence de nombreux contrastes dans le test et aux degrés de liberté d'un plan relativement petit.

Figure 21-14
*Estimations des paramètres pour le modèle cumulé généralisé (affichage partiel)*

| The legislature should enact a gas tax | Paramètre \: | B | Erreur std. | Intervalle de confiance 95% | |
|---|---|---|---|---|---|
| | | | | Inférieur | Supérieur |
| Strongly agree | (Seuil) | -3,681 | ,221 | -4,155 | -3,207 |
| | [agecat=1] | -,320 | ,096 | -,525 | -,115 |
| | [agecat=2] | -,075 | ,071 | -,227 | ,077 |
| | [agecat=3] | -,022 | ,073 | -,180 | ,135 |
| | [agecat=4] | ,000[a] | . | | |
| | [gender=0] | -,082 | ,054 | -,197 | ,033 |
| | [gender=1] | ,000[a] | . | | |
| | [votelast=0] | ,008 | ,052 | -,104 | ,120 |
| | [votelast=1] | ,000[a] | . | | |
| | [drivefreq=1] | -4,096 | ,267 | -4,669 | -3,523 |
| | [drivefreq=2] | -3,367 | ,237 | -3,876 | -2,857 |
| | [drivefreq=3] | -2,678 | ,224 | -3,158 | -2,199 |
| | [drivefreq=4] | -1,928 | ,213 | -2,384 | -1,471 |
| | [drivefreq=5] | -1,015 | ,252 | -1,555 | -,476 |
| | [drivefreq=6] | ,000[a] | . | | |
| Agree | (Seuil) | -1,963 | ,153 | -2,291 | -1,635 |
| | [agecat=1] | -,385 | ,095 | -,587 | -,182 |
| | [agecat=2] | -,130 | ,069 | -,279 | ,018 |
| | [agecat=3] | -,139 | ,101 | -,356 | ,077 |
| | [agecat=4] | ,000[a] | . | | |
| | [gender=0] | -,004 | ,040 | -,090 | ,082 |
| | [gender=1] | ,000[a] | . | | |
| | [votelast=0] | ,009 | ,059 | -,117 | ,135 |
| | [votelast=1] | ,000[a] | . | | |
| | [drivefreq=1] | -3,867 | ,318 | -4,549 | -3,185 |
| | [drivefreq=2] | -3,005 | ,175 | -3,380 | -2,630 |
| | [drivefreq=3] | -2,290 | ,187 | -2,691 | -1,888 |
| | [drivefreq=4] | -1,633 | ,166 | -1,988 | -1,278 |
| | [drivefreq=5] | -,909 | ,137 | -1,204 | -,615 |
| | [drivefreq=6] | ,000[a] | . | | |

De plus, les valeurs estimées des coefficients du modèle généralisé ne diffèrent pas beaucoup des estimations correspondant à l'hypothèse des droites parallèles.

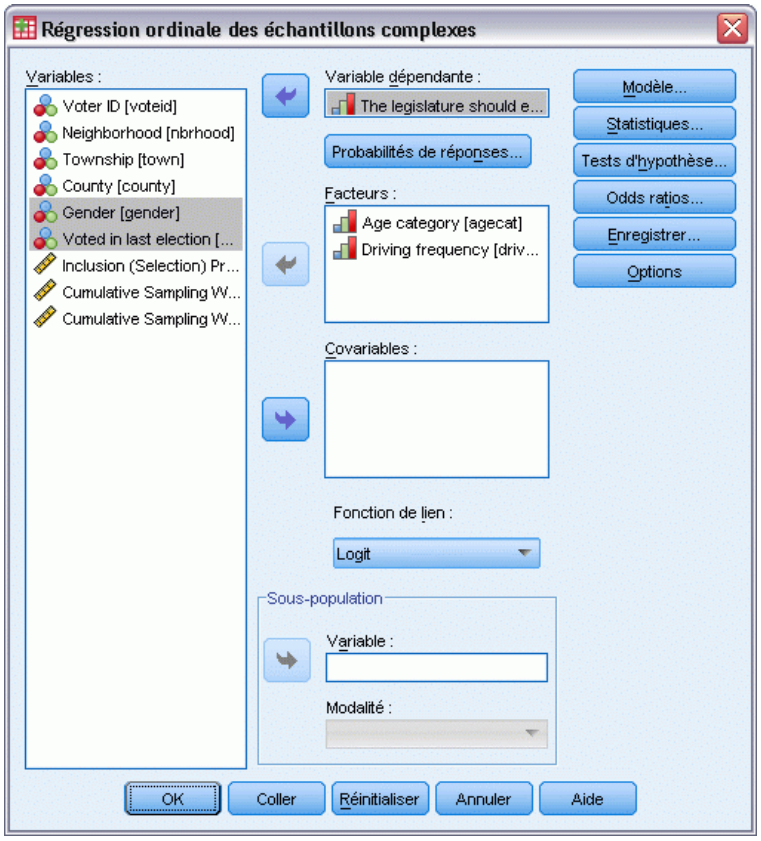
## *Suppression des variables indépendantes non significatives*

Les tests des effets de modèle indiquent que les coefficients de modèle des variables *Sexe* et *Résultats des dernières élections* ne sont pas, d'un point de vue statistique, significativement différents de 0.

► Pour générer un modèle réduit, rouvrez la boîte de dialogue Echantillons complexes - Régression ordinale.

▶ Cliquez sur Poursuivre dans la boîte de dialogue Plan.

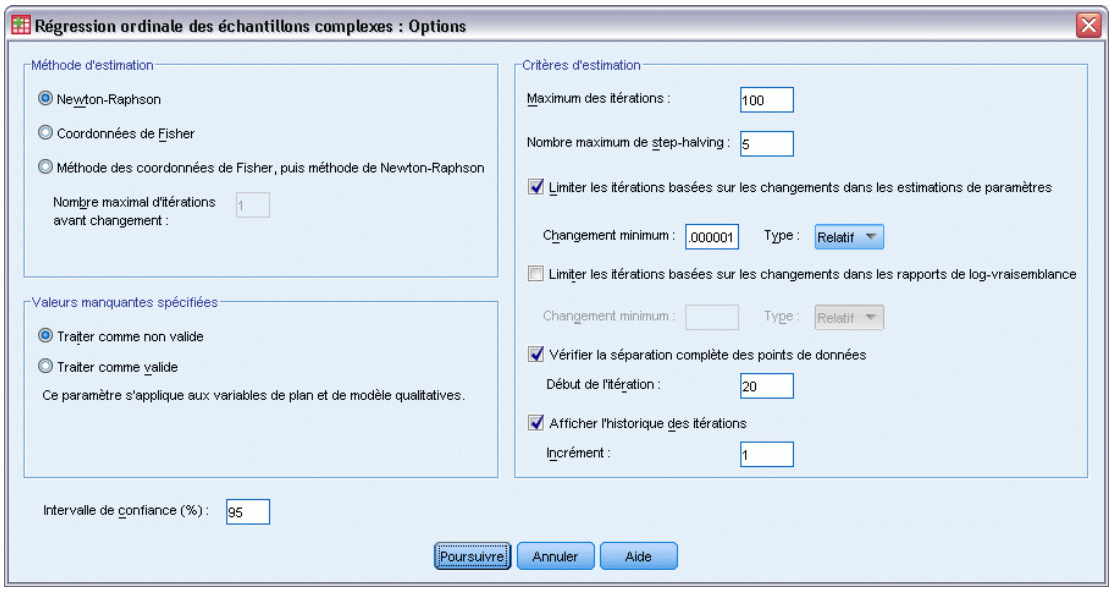
Figure 21-15
*Boîte de dialogue Régression ordinale*




▶ Désélectionnez *Sexe* et *Résultats des dernières élections* comme facteurs.

▶ Cliquez sur Options.

Figure 21-16
*Boîte de dialogue Régression ordinale : Options*




▶ Sélectionnez Afficher l'historique des itérations.

L'historique des itérations est utile pour diagnostiquer les problèmes rencontrés par l'algorithme d'estimation.

▶ Cliquez sur Poursuivre.

▶ Cliquez sur OK dans la boîte de dialogue Echantillons complexes - Régression ordinale.

## *Avertissements*

Figure 21-17
*Avertissements relatifs au modèle réduit*

La valeur de log-vraisemblance ne peut pas être augmentée au-delà du nombre maximal de step-halving.
La procédure CSORDINAL continue malgré les avertissements ci-dessus. Le reste des résultats affichés est basé sur la dernière itération. La validité de l'ajustement au modèle est incertaine.
Le message suivant s'applique au modèle cumulé généralisé.
La valeur de log-vraisemblance ne peut pas être augmentée au-delà du nombre maximal de step-halving.

Les avertissements indiquent que l'estimation du modèle réduit s'est terminée avant que les estimations des paramètres n'atteignent la convergence car la log-vraisemblance n'a pu être augmentée avec aucun changement, ou « étape », des valeurs courantes des estimations des paramètres.

Figure 21-18
*Avertissements relatifs au modèle réduit*

| Numéro de l'itération[b] | N dichotomie | Pseudo -2 Log vraisemblance | Seuil | | | Régression | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | [opinion_gastax=1] | [opinion_gastax=2] | [opinion_gastax=3] | [agecat=1] | [agecat=2] | [agecat=3] | [drivefreq=1] | [drivefreq=2] | [drivefreq=3] | [drivefreq=4] | [drivefreq=5] |
| 0 | 0 | 326640,3 | -1,309 | -,058 | 1,020 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 |
| 1 | 0 | 303567,5 | -3,242 | -1,881 | -,704 | -,323 | -,137 | -,094 | -3,841 | -2,970 | -2,248 | -1,563 | -,835 |
| 2 | 0 | 303336,3 | -3,327 | -1,897 | -,664 | -,325 | -,139 | -,095 | -3,740 | -2,998 | -2,291 | -1,568 | -,811 |
| 3 | 0 | 303335,9 | -3,333 | -1,900 | -,664 | -,326 | -,139 | -,096 | -3,750 | -3,003 | -2,295 | -1,570 | -,812 |
| 4 | 0 | 303335,9 | -3,333 | -1,900 | -,664 | -,326 | -,139 | -,096 | -3,750 | -3,003 | -2,295 | -1,570 | -,812 |
| 5[a] | 5 | 303335,9 | -3,333 | -1,900 | -,664 | -,326 | -,139 | -,096 | -3,750 | -3,003 | -2,295 | -1,570 | -,812 |

Les paramètres redondants ne sont pas affichés. Leurs valeurs sont toujours de zéro dans toutes les itérations.
Variable dépendante: The legislature should enact a gas tax (Croissant)
Modèle : (Seuil), agecat, drivefreq
Fonction de lien : Logit

a. La valeur de log-vraisemblance ne peut pas être augmentée au-delà du nombre maximal de step-halving.

b. La méthode de Newton-Raphson a été utilisée pour estimer les paramètres.

D'après l'historique des itérations, les changements survenus dans les estimations des paramètres au cours des quelques dernières itérations sont relativement mineurs pour que le message d'avertissement ne vous préoccupe pas de façon excessive.

## *Comparaison de modèles*

Figure 21-19
*Pseudo R-deux du modèle réduit*

| | |
|---|---|
| Cox et Snell | ,179 |
| Nagelkerke | ,191 |
| McFadden | ,071 |

Variable dépendante : The legislature should enact a gas tax (Croissant)
Modèle : (Seuil), agecat, gender, votelast, drivefreq
Fonction de lien : Logit

Les valeurs $R^2$ du modèle réduit sont identiques à celles du modèle d'origine. Ceci plaide en faveur du modèle réduit.

Figure 21-20
*Tableau de classement du modèle réduit*

| Observations | Prévisions | | | | |
|---|---|---|---|---|---|
| | Strongly agree | Agree | Disagree | Strongly disagree | Pourcentage correct |
| Strongly agree | 7067,567 | 12823,258 | 3183,380 | 2058,750 | 28,1% |
| Agree | 4271,234 | 15684,090 | 6100,963 | 6205,137 | 48,6% |
| Disagree | 2024,816 | 13157,809 | 5654,047 | 8640,746 | 19,2% |
| Strongly disagree | 889,869 | 9226,578 | 5889,053 | 15308,703 | 48,9% |
| Pourcentage global | 12,1% | 43,1% | 17,6% | 27,3% | 37,0% |

Variable dépendante : The legislature should enact a gas tax (Croissant)
Modèle : (Seuil), agecat, drivefreq
Fonction de lien : Logit

Le tableau de classement complique un peu les choses. Le taux de classement global de 37 % du modèle réduit est comparable au modèle d'origine, ce qui plaide en faveur du modèle réduit. Cependant, le modèle réduit change la réponse prévue de 3,8 % des électeurs de *Pas d'accord* à *D'accord* et parmi ces électeurs, plus de la moitié a répondu *Pas d'accord* ou *Pas du tout d'accord*. Il s'agit d'une distinction très importante à considérer avec soin avant de choisir le modèle réduit.

### *Récapitulatif*

Avec la procédure de régression ordinale des échantillons complexes, vous avez construit des modèles en compétition pour le niveau de soutien du projet de loi en fonction de la répartition démographique des électeurs. Le test des droites parallèles indique qu'un modèle cumulé généralisé n'est pas nécessaire. Les tests des effets de modèle suggèrent que les variables *Sexe* et *Résultats des dernières élections* peuvent être exclues du modèle, et que la valeur pseudo $R^2$ et le taux de classement global du modèle réduit fonctionnent bien par rapport au modèle d'origine. Cependant, le modèle réduit classe de manière incorrecte davantage d'électeurs dans la scission *D'accord*/*Pas d'accord*, si bien que les législateurs préfèrent pour l'instant conserver le modèle d'origine.

## *Procédures apparentées*

La procédure de régression ordinale des échantillons complexes est un outil utile pour modéliser une variable ordinale lorsque les observations ont été conçues en fonction d'un schéma d'échantillonnage complexe.

- L'assistant d'échantillonnage des échantillons complexes permet d'indiquer les spécifications du plan d'échantillonnage complexe et d'obtenir un échantillon. Le fichier de plan d'échantillonnage créé par l'assistant d'échantillonnage contient un plan d'analyse par défaut et peut être spécifié dans la boîte de dialogue Plan lors de l'analyse de l'échantillon obtenu en fonction de ce plan.
- L'assistant de préparation d'analyse des échantillons complexes est utilisé pour indiquer les spécifications d'analyse d'un échantillon complexe existant. Le fichier de plan d'analyse créé par l'assistant d'échantillonnage peut être spécifié dans la boîte de dialogue Plan lorsque vous analysez l'échantillon correspondant à ce plan.

- La procédure Echantillons complexes - Modèle linéaire général vous permet de modéliser une réponse d'échelle.
- La procédure de régression logistique des échantillons complexes vous permet de modéliser une réponse qualitative.

Chapitre
22

# Régression de Cox des échantillons complexes

La procédure de la régression de Cox des échantillons complexes effectue une analyse de survie pour les échantillons réalisés à l'aide de méthodes d'échantillonnage complexes.

## Utilisation d'une variable indépendante chronologique dans la régression de Cox des échantillons complexes

Une administration chargée de l'application de la loi s'inquiète des taux de récidive dans sa juridiction. L'une des mesures de récidive est le temps qui s'écoule avant la deuxième arrestation des délinquants. L'agence souhaite modéliser le temps s'écoulant jusqu'à la deuxième arrestation à l'aide de la régression de Cox sur un échantillon réalisé au moyen de méthodes d'échantillonnage complexes, mais elle craint que l'hypothèse des hasards proportionnels ne soit pas valide sur l'ensemble des tranches d'âge.

Les personnes libérées suite à leur première arrestation au mois de juin 2003 ont été sélectionnées à partir de départements échantillonnés, et leurs antécédents judiciaires ont été étudiés jusqu'à la fin du mois de juin 2006. L'échantillon se trouve dans *recidivism_cs_sample.sav*. Le plan d'échantillonnage utilisé se trouve dans le fichier *recidivism_cs.csplan*. Ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS), il existe également un fichier contenant les probabilités de sélection conjointes (*recidivism_cs_jointprob.sav*). Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20*. Utilisez la régression de Cox des échantillons complexes pour évaluer la validité de l'hypothèse des hasards proportionnels et ajuster un modèle avec des variables indépendantes chronologiques, le cas échéant.

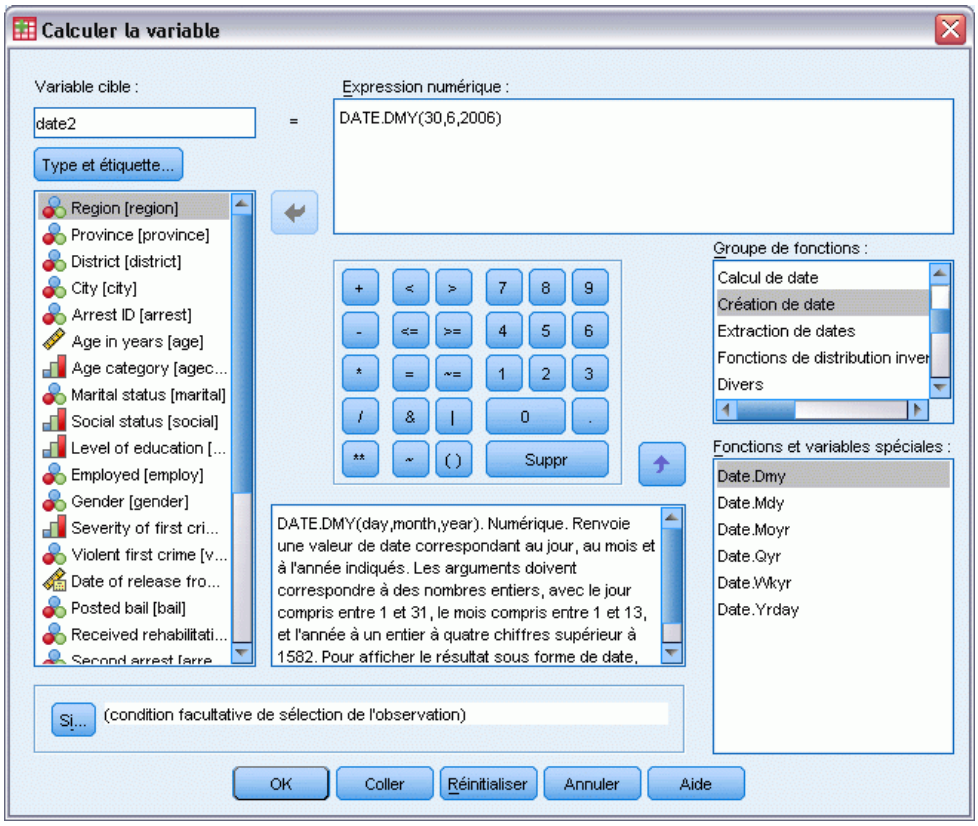
### Préparation des données

L'ensemble de données contient les dates de libération suite à la première et à la seconde arrestation. Comme la régression de Cox analyse les temps de survie, vous devez calculer le temps écoulé entre ces deux dates.

Cependant, *Date de la deuxième arrestation [date2]* contient des observations avec la valeur 03/10/1582 qui est une valeur manquante pour les variables de date. Il s'agit de personnes n'ayant pas récidivé et nous tenons à les inclure dans le modèle en tant qu'observations censurées à droite. La période de suivi est arrivée à son terme le 30 juin 2006. Nous allons donc recoder 03/10/1582 en 30/06/2006.

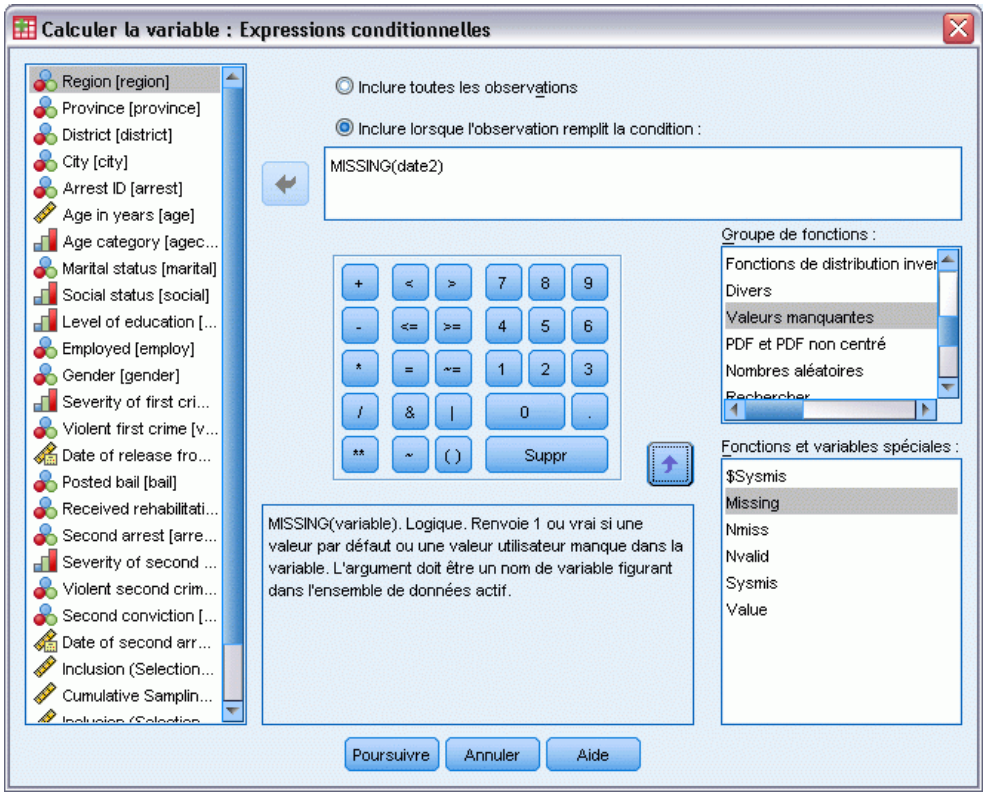
- Pour recoder ces valeurs, à partir des menus, sélectionnez :
  Transformer > Calculer la variable...

Figure 22-1
*Boîte de dialogue Calculer la variable*




- Entrez la variable cible date2.
- Entrez l'expression DATE.DMY(30,6,2006).
- Cliquez sur Si.

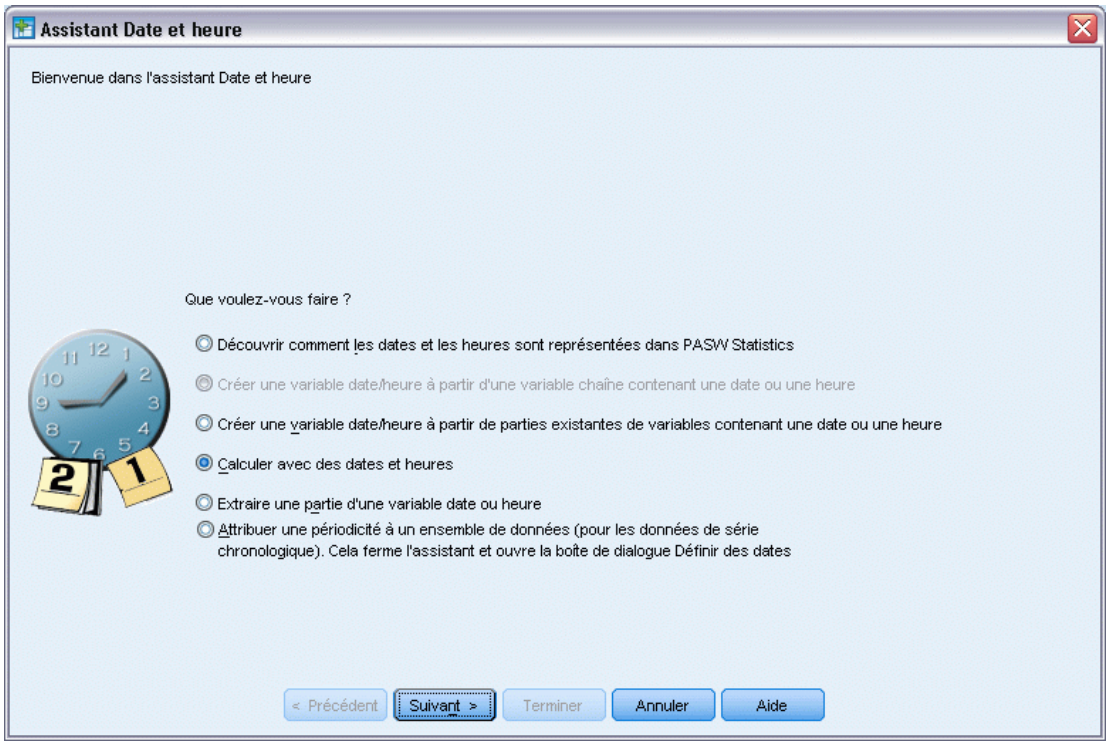
Figure 22-2
*Boîte de dialogue Expression conditionnelle Calculer la variable*




- Sélectionnez Inclure si l'observation remplit la condition :
- Entrez l'expression MISSING(date2).
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Calculer la variable.
- Puis, pour calculer la durée entre la première et la seconde arrestation, à partir des menus, sélectionnez :

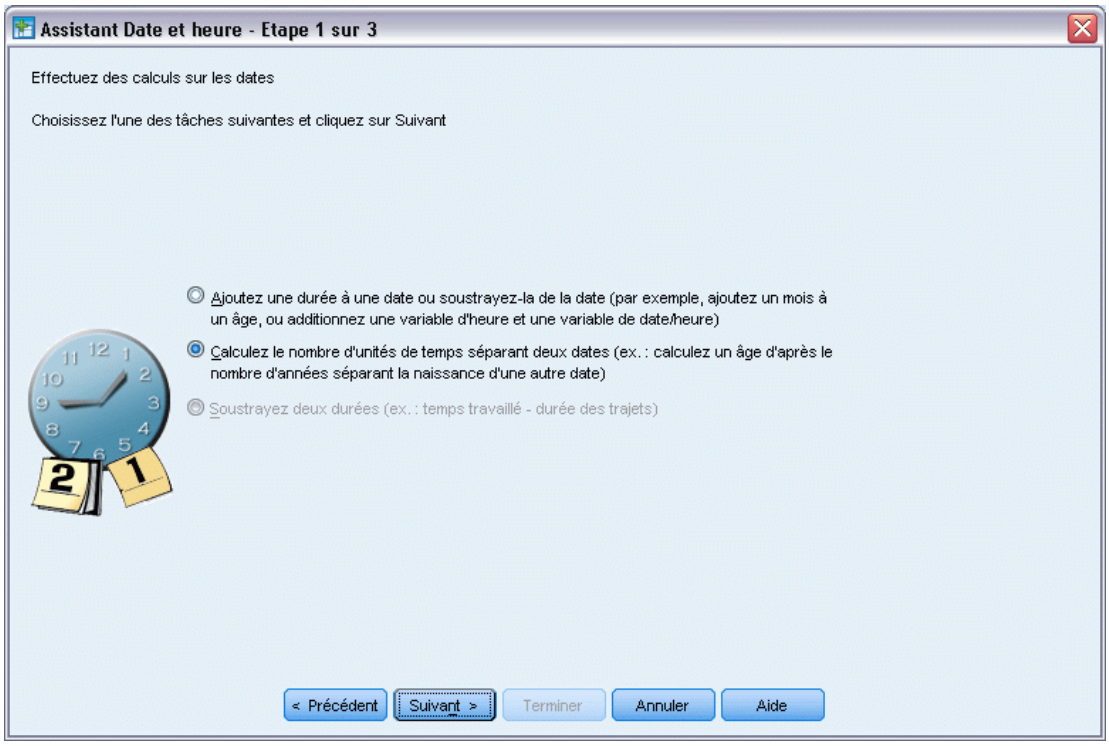
  Transformer > Assistant Date et heure...

Figure 22-3
*Etape Bienvenue de l'assistant Date et heure*



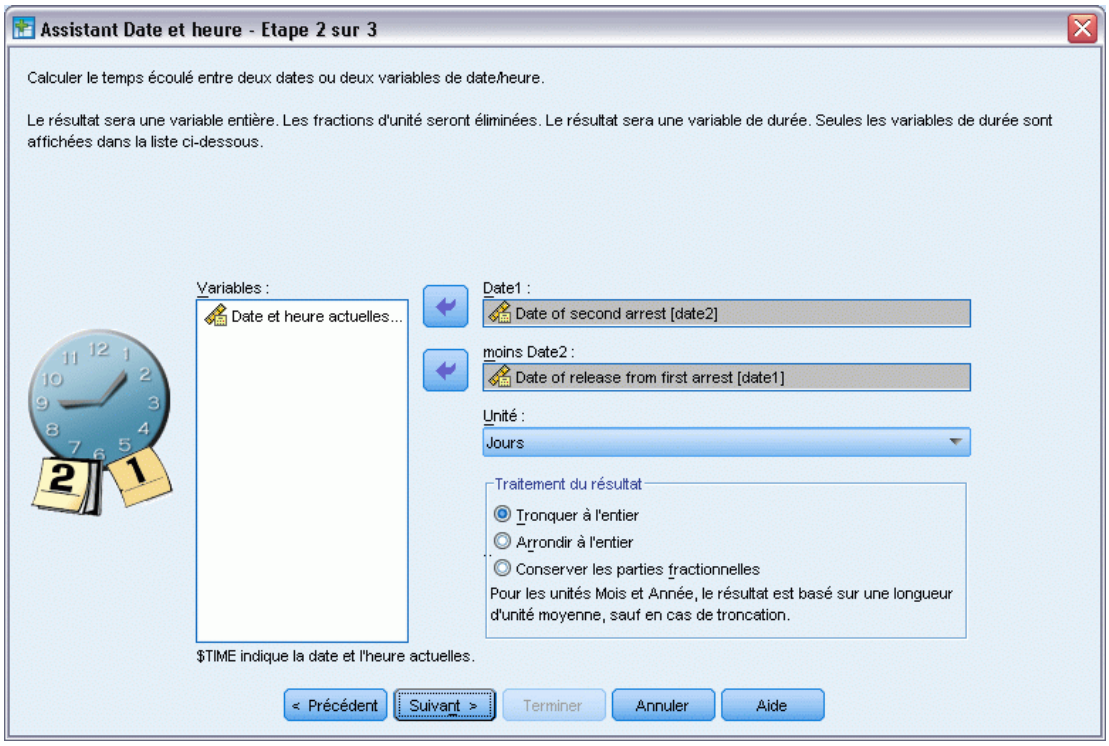

- Sélectionnez Calculer avec des dates et heures.
- Cliquez sur Suivant.

Figure 22-4
*Etape de calculs sur les dates de l'assistant Date et heure*



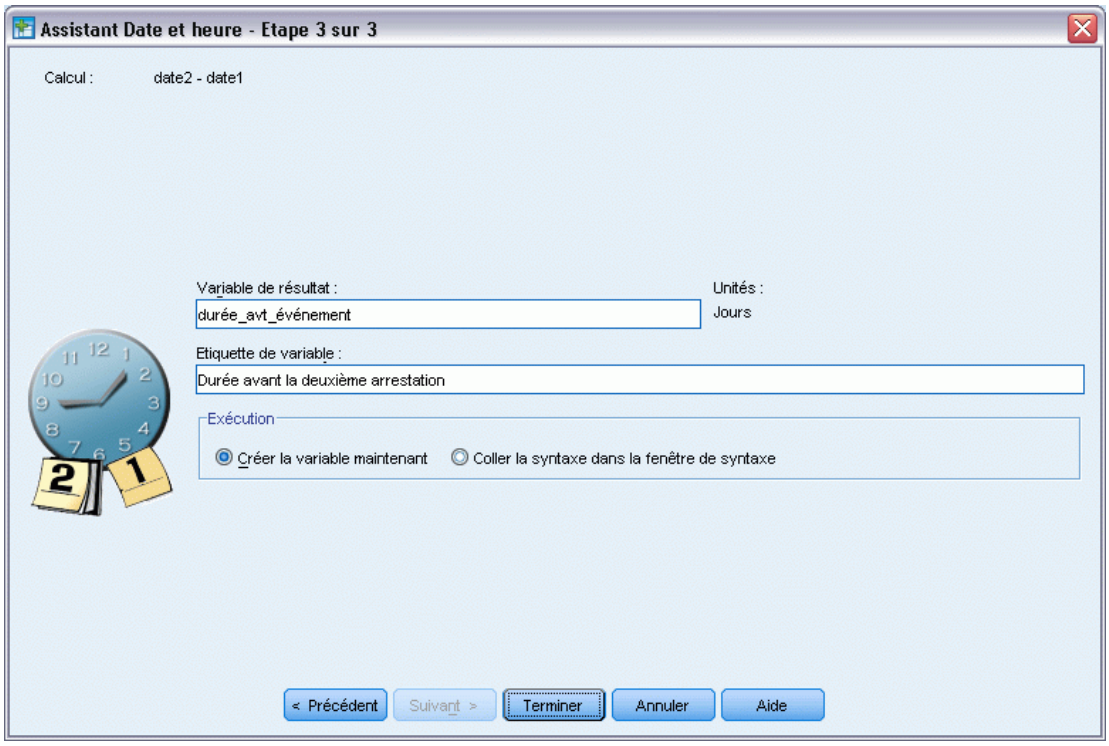

- Sélectionnez Calculez le nombre d'unités de temps séparant deux dates.
- Cliquez sur Suivant.

Figure 22-5
*Etape de calcul du nombre d'unités de temps séparant deux dates de l'assistant Date et heure*




- Sélectionnez *Date de la deuxième arrestation [date2]* comme première date.
- Sélectionnez *Date de libération après la première arrestation [date1]* comme date à soustraire de la première date.
- Sélectionnez Jours comme unité.
- Cliquez sur Suivant.

Figure 22-6
*Etape de calcul de l'assistant Date et heure*




- Tapez *durée_avt_événement* pour le nom de la variable représentant le temps écoulé entre les deux dates.
- Tapez *Durée avant la deuxième arrestation* comme étiquette de variable.
- Cliquez sur Terminer.

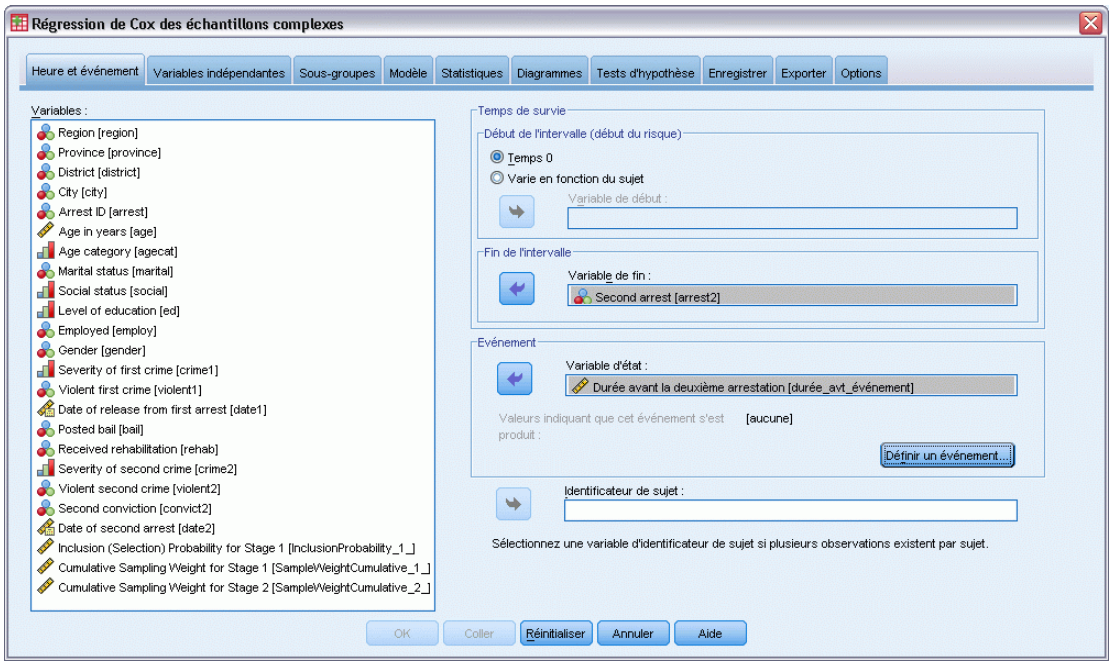
## *Exécution de l'analyse*

- Pour exécuter une analyse de régression de Cox des échantillons complexes, sélectionnez les options suivantes dans le menu :
  Analyse > Echantillonnage > Modèle de Cox

Figure 22-7
*Boîte de dialogue Plan d'échantillonnages complexes pour régression de Cox*

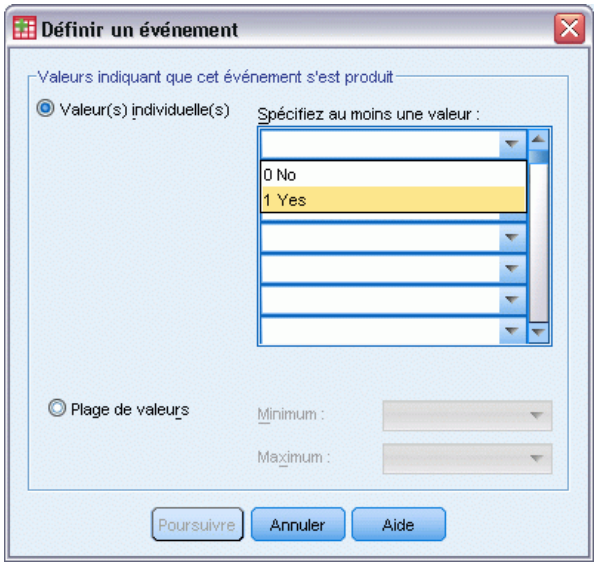

- Accédez au répertoire des fichiers d'exemple et sélectionnez *recidivism_cs.csplan* comme fichier de plan.
- Sélectionnez Fichier personnalisé dans le groupe des probabilités conjointes, accédez au répertoire des fichiers d'exemple et sélectionnez *recidivism_cs_jointprob.sav*.
- Cliquez sur Poursuivre.

Figure 22-8
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Heure et événement*



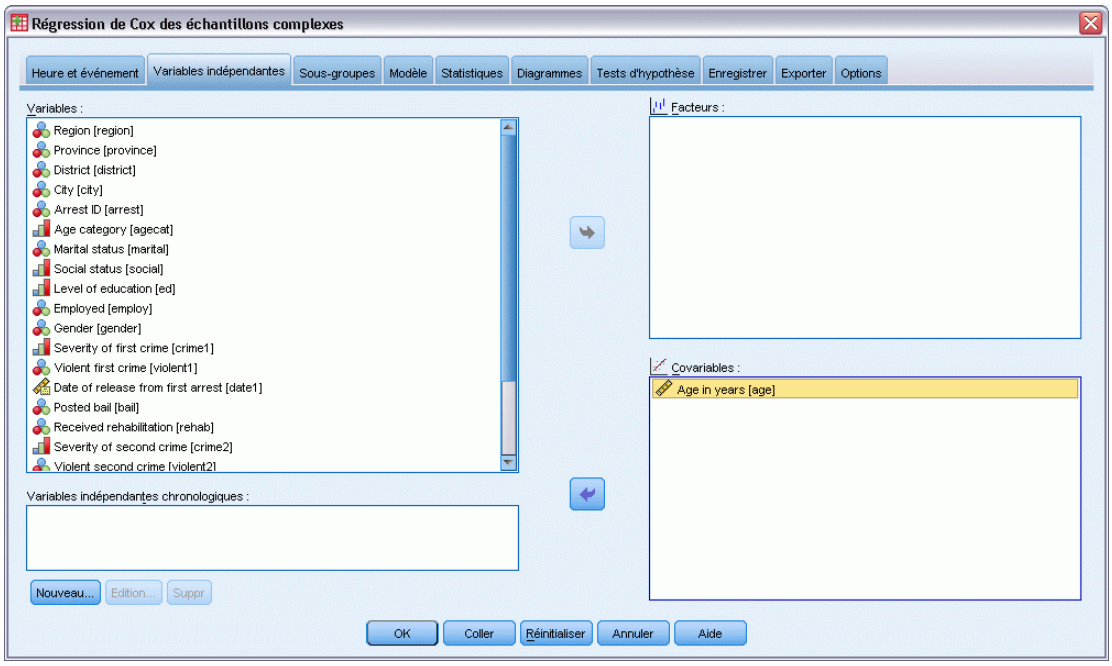

- Sélectionnez *Durée avant la deuxième arrestation [durée_avt_événement]* comme variable définissant la fin de l'intervalle.
- Sélectionnez *Deuxième arrestation [arrest2]* comme variable définissant si l'événement s'est produit.
- Cliquez sur Définir un événement.

Figure 22-9
*Boîte de dialogue Définir un événement*



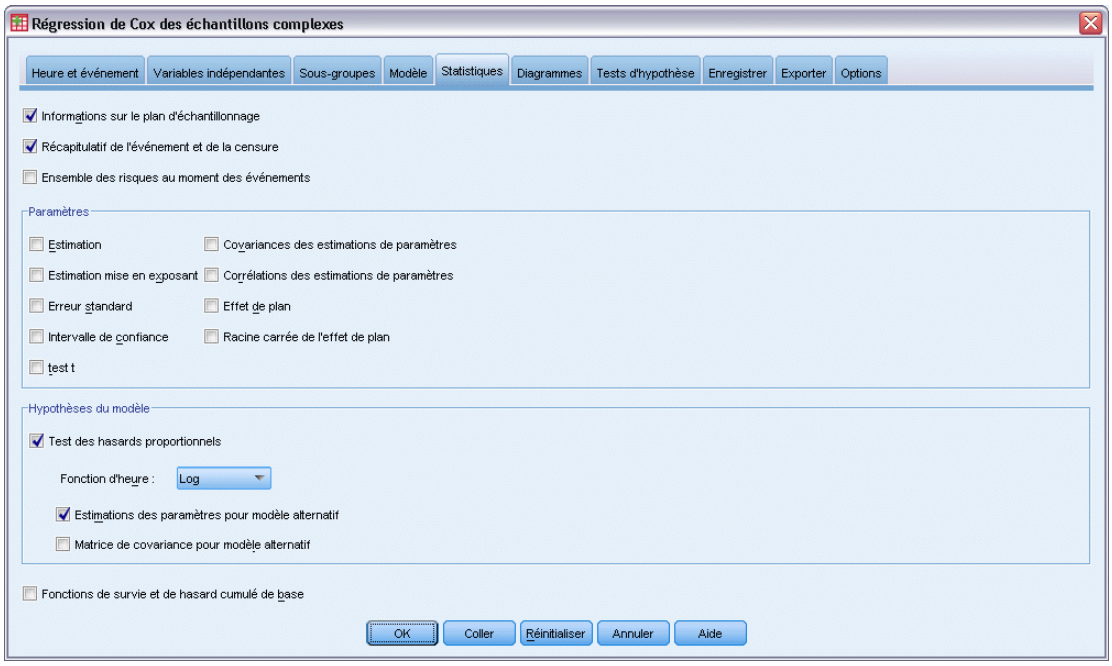

- Sélectionnez 1 oui comme valeur indiquant que l'événement étudié (nouvelle arrestation) s'est produit.
- Cliquez sur Poursuivre.
- Cliquez sur l'onglet Variables indépendantes.

Figure 22-10
*Boîte de dialogue Régression de Cox, onglet Variables indépendantes*




- Sélectionnez *Age en années [âge]* comme covariable.
- Cliquez sur l'onglet Statistiques.

Figure 22-11
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Statistiques*




- Sélectionnez Test des hasards proportionnels, puis Log comme fonction heure dans le groupe Hypothèses du modèle.
- Sélectionnez Estimations des paramètres pour modèle alternatif
- Cliquez sur OK.

## *Informations sur le plan d'échantillonnage*

Figure 22-12
*Informations sur le plan d'échantillonnage*

| | | | N |
|---|---|---|---|
| Effectifs non pondérés | Valide | Sujets | 1674 |
| | | Observations | 1674 |
| | | Observations non valides | 4013 |
| | | Observations totales | 5687 |
| Valide | | Taille de sujet de population | 89984,214 |
| Etape 1 | Valide | Strates | 4 |
| | | Unités | 20 |
| Valide | | Degrés de liberté de plan d'échantillonnage | 16 |

Ce tableau contient des informations sur le plan d'échantillonnage se rapportant à l'estimation du modèle.

- Il y a une observation par sujet et les 5 687 observations sont toutes utilisées dans l'analyse.

- L'échantillon représente moins de 2 % de l'ensemble de la population estimée.
- Le plan demandait 4 strates et 5 unités par strate pour un total de 20 unités dans la première étape du plan. Les degrés de liberté du plan d'échantillonnage sont estimés par 20−4=16.

## *Tests des effets de modèle*

Figure 22-13
*Tests des effets de modèle*

| So... | df1 | df2 | Wald F | Sig. |
|---|---|---|---|---|
| age | 1,000 | 16,000 | ,017 | ,897 |

Variable Durée de survie\ : Time to second arrest
Variable Etat événement: Second arrest = 1
Modèle: age

Dans le modèle des hasards proportionnels, la valeur de signification pour la variable indépendante *âge* est inférieure à 0,05 et semble, de ce fait, contribuer au modèle.

## *Test des hasards proportionnels*

Figure 22-14
*Test général des hasards proportionnels*

| df1 | df2 | Wald F | Sig. |
|---|---|---|---|
| 1,000 | 16,000 | 11,437 | ,004 |

Variable Durée de survie\ : Time to second arrest
Variable Etat événement: Second arrest = 1
Modèle: age, age*_TF

Figure 22-15
*Estimations des paramètres pour modèle alternatif*

| Paramètre | B | Erreur std. | Intervalle de confiance 95% | |
|---|---|---|---|---|
| | | | Inférieur | Supérieur |
| age | ,045 | ,014 | ,016 | ,073 |
| age*_TF[a] | -,008 | ,002 | -,012 | -,003 |

Variable Durée de survie\ : Time to second arrest
Variable Etat événement: Second arrest = 1
Modèle: age, age*_TF

a. Fonction Heure : Log

b. Méthode Ex aequo\ : Efron

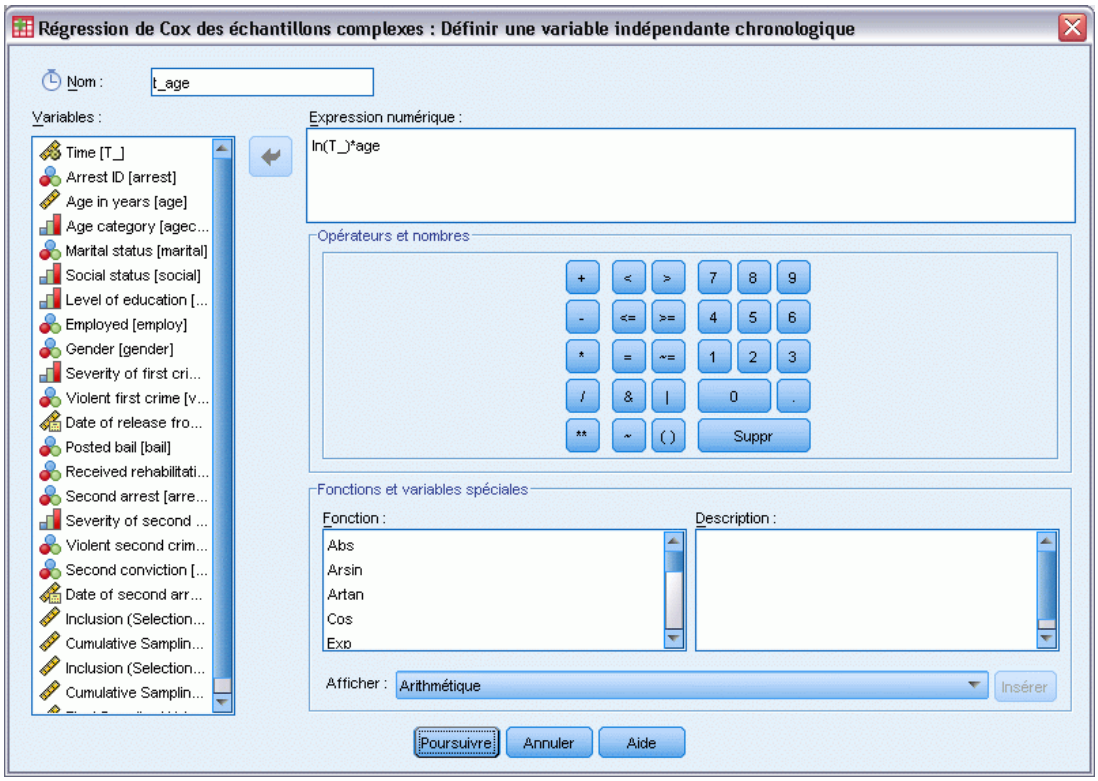
La valeur de signification pour le test général des hasards proportionnels est inférieure à 0,05, ce qui indique que l'hypothèse des hasards proportionnels n'est pas respectée. La fonction log heure est utilisée pour le modèle alternatif. Il sera donc facile de reproduire cette variable indépendante chronologique.

## *Ajout d'une variable indépendante chronologique*

▶ Affichez de nouveau la boîte de dialogue Régression de Cox des échantillons complexes et cliquez sur l'onglet Variables indépendantes.

▶ Cliquez sur Nouveau.

Figure 22-16
*Boîte de dialogue Régression de Cox - Définir une variable indépendante chronologique*

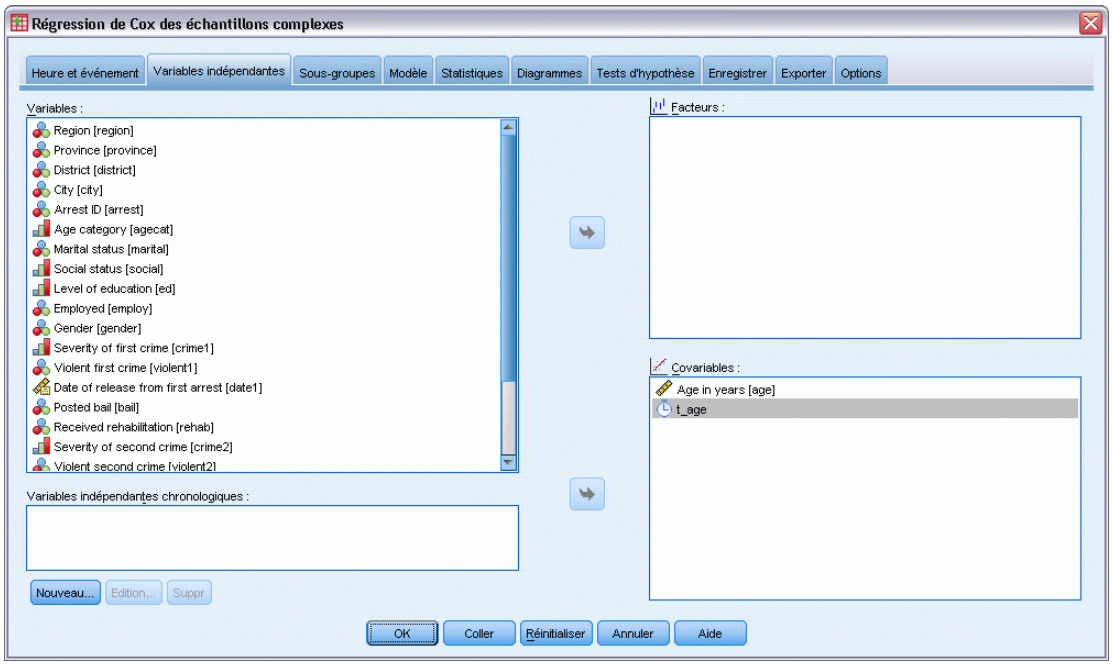



▶ Tapez t_âge pour le nom de la variable indépendante chronologique que vous souhaitez définir.

▶ Entrez l'expression numérique ln(T_)*age.

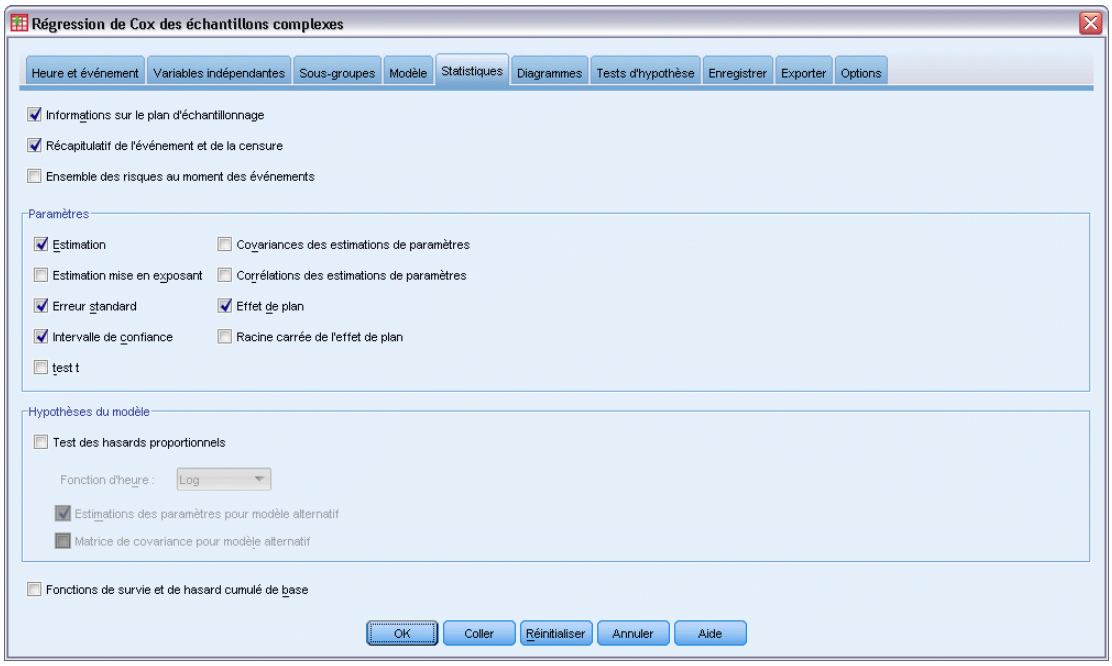
▶ Cliquez sur Poursuivre.

Figure 22-17
*Boîte de dialogue Régression de Cox, onglet Variables indépendantes*




- Sélectionnez *t_âge* comme covariable.
- Cliquez sur l'onglet Statistiques.

Figure 22-18
*Boîte de dialogue Régression de Cox, onglet Variables indépendantes*




- Sélectionnez Estimation, Erreur standard, Intervalle de confiance et Effet de plan dans le groupe de paramètres du modèle.
- Désélectionnez Test des hasards proportionnels et Estimations des paramètres pour modèle alternatif dans le groupe Hypothèses du modèle.
- Cliquez sur OK.

### *Tests des effets de modèle*

Figure 22-19
*Tests des effets de modèle*

| Source | ddl1 | ddl2 | F de Wald | Sig. |
|---|---|---|---|---|
| âge | 1,000 | 16,000 | ,015 | 0,91 |
| t_âge | 1,000 | 16,000 | 29,924 | 5,136E-5 |

Variable de temps de survie : Durée avant la deuxième arrestation
Variable d'état d'événement : Deuxième arrestation = 1
Modèle : âge

Avec l'ajout de la variable indépendante chronologique, la valeur de signification pour *âge* est 0,91, ce qui indique que sa contribution au modèle est remplacée par celle de *t_âge*.

### *Estimations de paramètre*

Figure 22-20
*Estimations des paramètres*

| Paramètre | B | Erreur type | Intervalle de confiance à 95 % | | Effet de plan |
|---|---|---|---|---|---|
| | | | Inférieur | Supérieur | |
| âge | -,002 | 0,01 | -,030 | ,027 | ,702 |
| t_âge | -,012 | ,002 | -,017 | -,008 | ,666 |

Variable de temps de survie : Durée avant la deuxième arrestation
Variable d'état d'événement : Deuxième arrestation = 1
Modèle : âge, t_âge

Si vous observez les estimations des paramètres et les erreurs standard, vous pouvez vous rendre compte du fait que vous avez reproduit le modèle alternatif à partir du test des hasards proportionnels. En spécifiant explicitement le modèle, vous pouvez demander des statistiques des paramètres et des diagrammes supplémentaires. Dans le cas présent, nous avons demandé l'effet de plan. La valeur de *t_âge* qui est inférieure à 1 indique que l'erreur standard pour *t_âge* est inférieure à ce que vous pourriez obtenir si vous supposiez que l'ensemble de données était un échantillon aléatoire simple. Dans ce cas, l'effet de *t_âge* serait toujours significatif d'un point de vue statistique, mais les intervalles de confiance seraient plus larges.

## *Observations multiples par sujet dans la régression de Cox des échantillons complexes*

Des chercheurs étudiant les temps de survie de patients qui quittent un programme de rééducation suite à un accident ischémique doivent faire face à un certain nombre de problèmes.

**Observations multiples par sujet.** Des variables représentant les antécédents médicaux des patients devraient être utiles en tant que variables indépendantes. Au fil du temps, les patients peuvent vivre des événements médicaux majeurs modifiant leurs antécédents médicaux. Dans cet ensemble de données, l'occurrence d'infarctus du myocarde, d'accidents ischémiques ou hémorragiques est signalée, et le moment de l'événement enregistré. Vous pouvez créer des prédicteurs chronologiques calculables dans la procédure pour inclure ces informations dans le modèle, mais il serait plus pratique d'utiliser plusieurs observations par sujet. A l'origine, les variables ont été codées pour que les antécédents du patient soient enregistrés sur l'ensemble des variables. Vous devrez donc restructurer l'ensemble de données.

**Troncation à gauche.** Le début du risque commence au moment de l'accident ischémique. Cependant, si l'échantillon ne comprend que des patients ayant survécu au programme de rééducation, il est tronqué à gauche car les temps de survie observés sont « augmentés » par la durée de la rééducation. Pour représenter cette information, indiquez l'heure à laquelle ils ont quitté le programme de rééducation comme heure d'entrée dans l'étude.

**Pas de plan d'échantillonnage.** L'ensemble de données n'a pas été collecté par le biais d'un plan d'échantillonnage complexe et est considéré comme un échantillon aléatoire simple. Vous devrez créer un plan d'analyse pour utiliser la régression de Cox des échantillons complexes.

L'ensemble de données se trouve dans *stroke_survival.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Complex Samples 20.* Utilisez l'Assistant de restructuration des données pour préparer les données pour l'analyse,

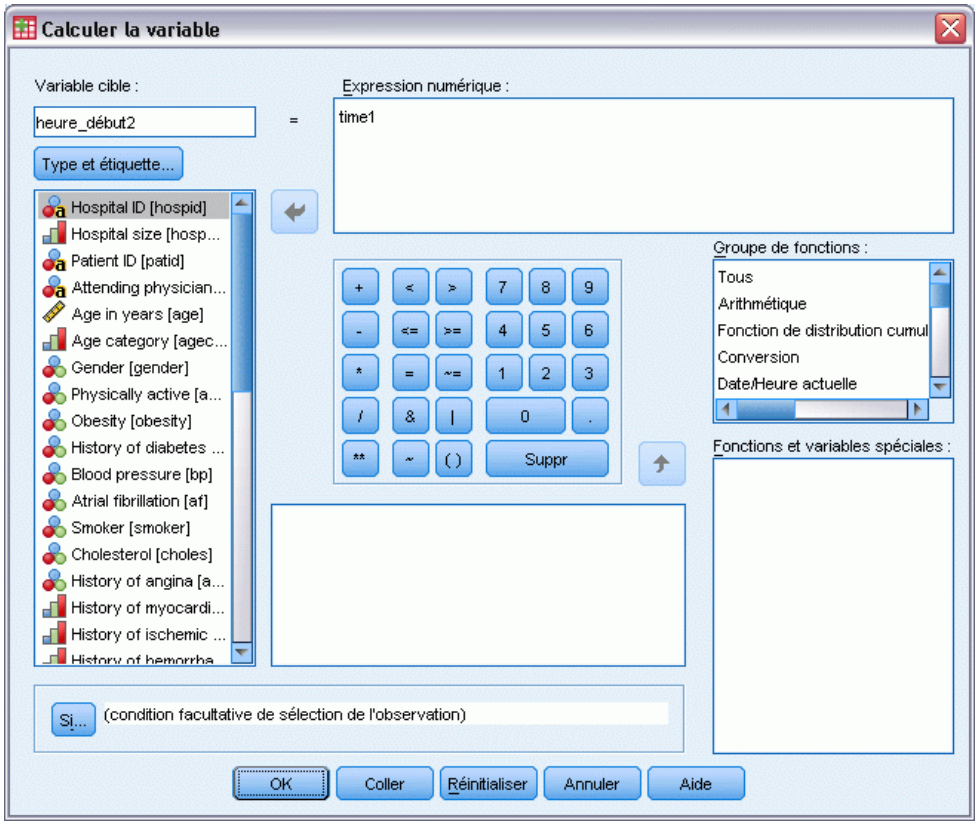
puis l'Assistant de préparation d'analyse pour créer un plan d'échantillonnage aléatoire simple, et enfin la régression de Cox des échantillons complexes pour construire un modèle pour les durées de survie.

## Préparation des données pour l'analyse

Avant la restructuration des données, vous devrez créer deux variables secondaires pour permettre la restructuration.

- Pour calculer une nouvelle variable, à partir des menus sélectionnez :
  Transformer > Calculer la variable...

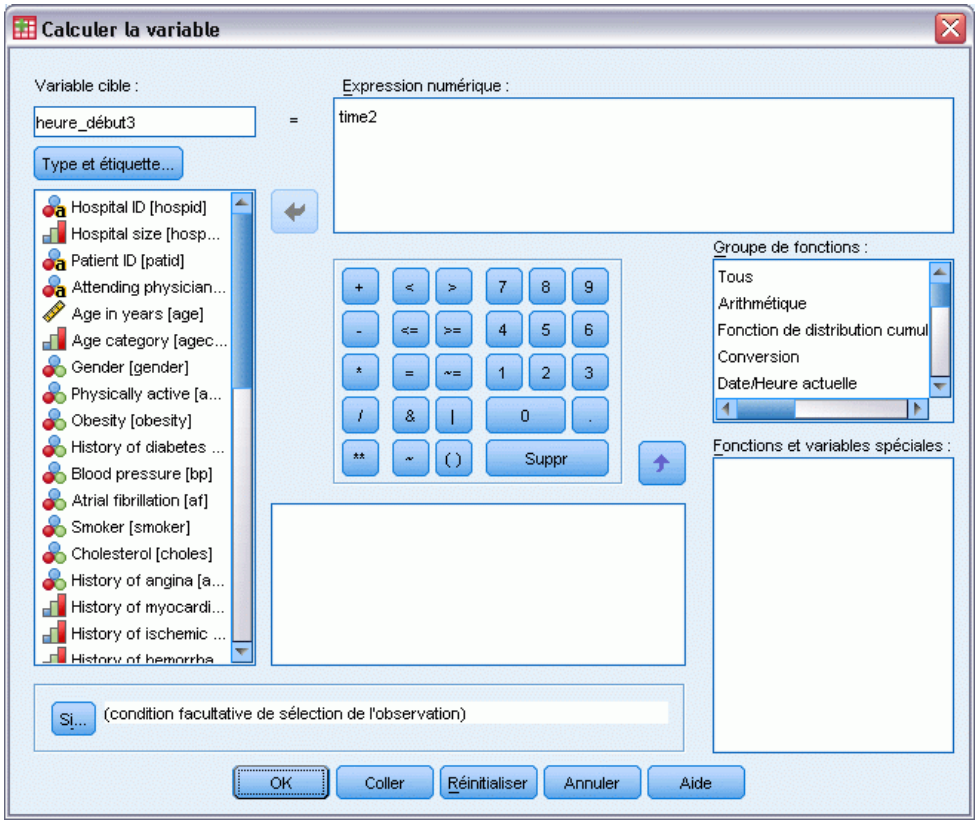
Figure 22-21
*Boîte de dialogue Calculer la variable*




- Tapez *heure_début2* comme variable cible.
- Entrez l'expression numérique *time1*.
- Cliquez sur OK.

▶ Rappelez la boîte de dialogue Calculer la variable.

Figure 22-22
*Boîte de dialogue Calculer la variable*



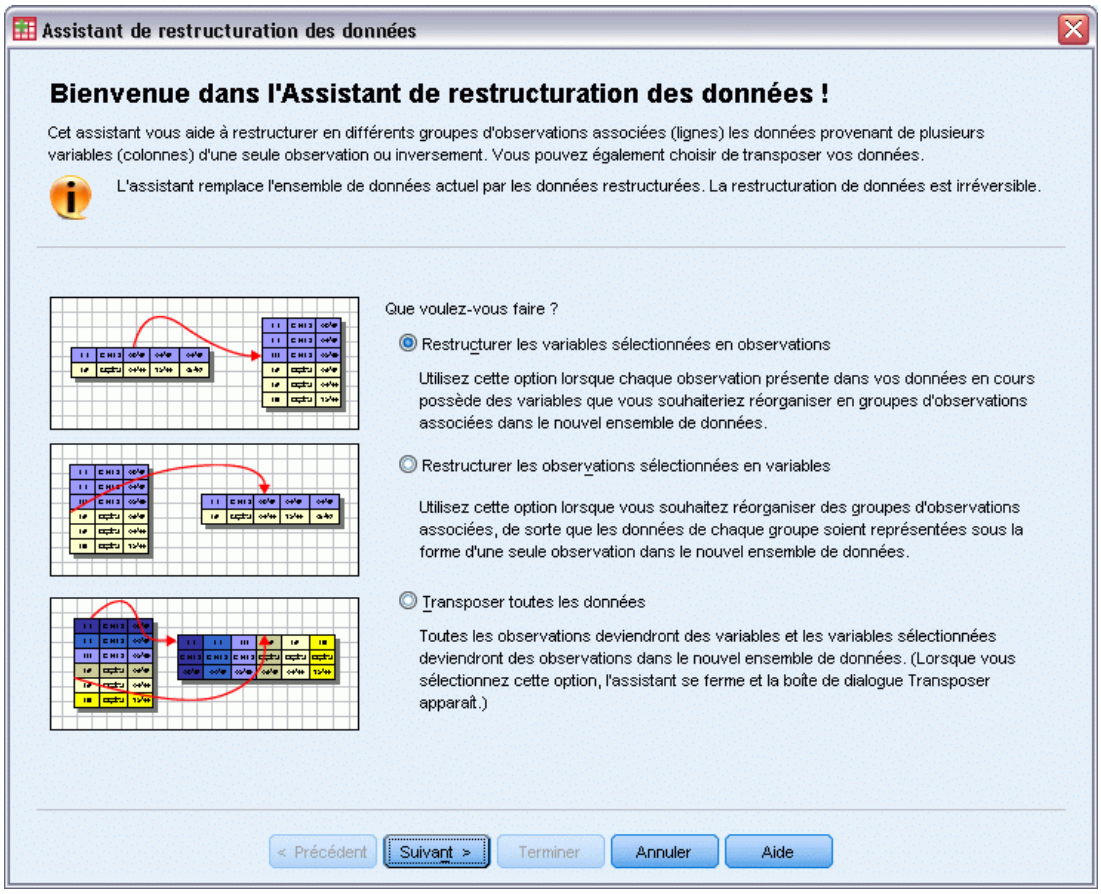

▶ Tapez *heure_début3* comme variable cible.

▶ Entrez l'expression numérique *time2*.

▶ Cliquez sur OK.

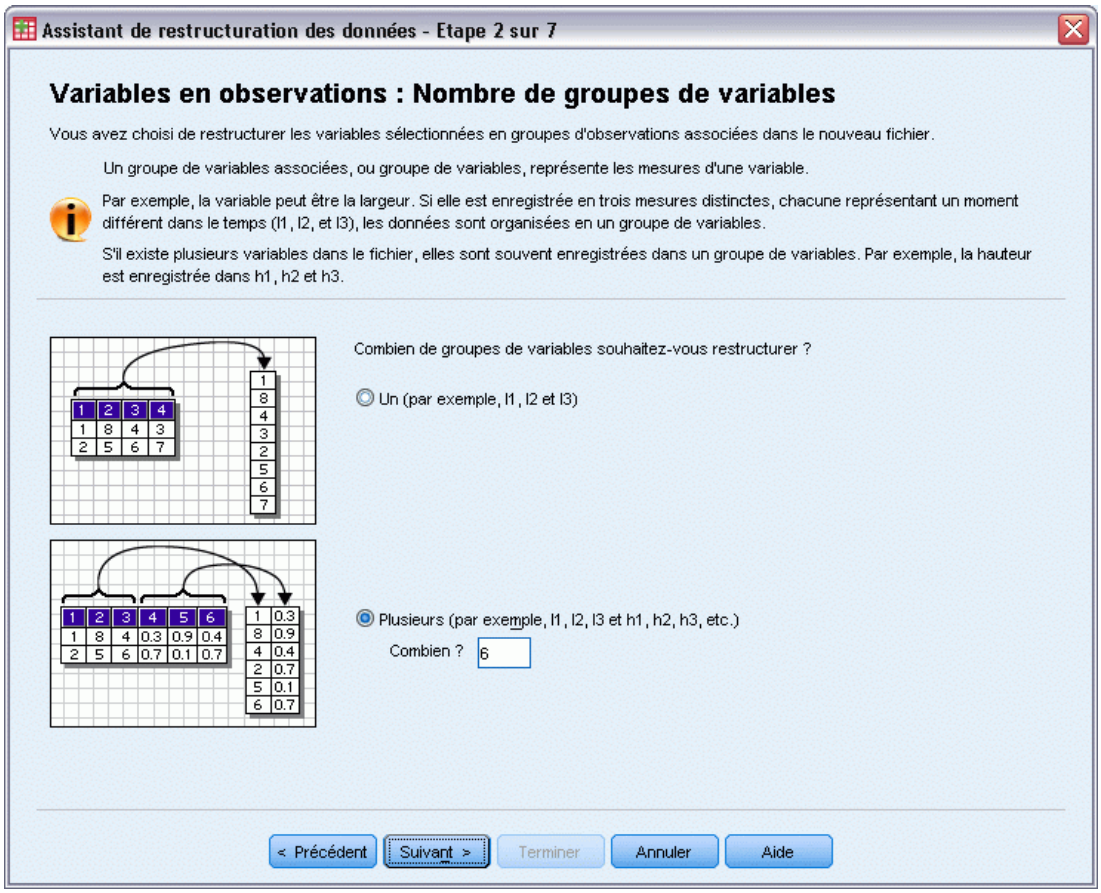
▶ Pour restructurer les données de variables en observations, à partir des menus, sélectionnez : Données > Restructurer...

Figure 22-23
*Etape Bienvenue de l'Assistant de restructuration des données*



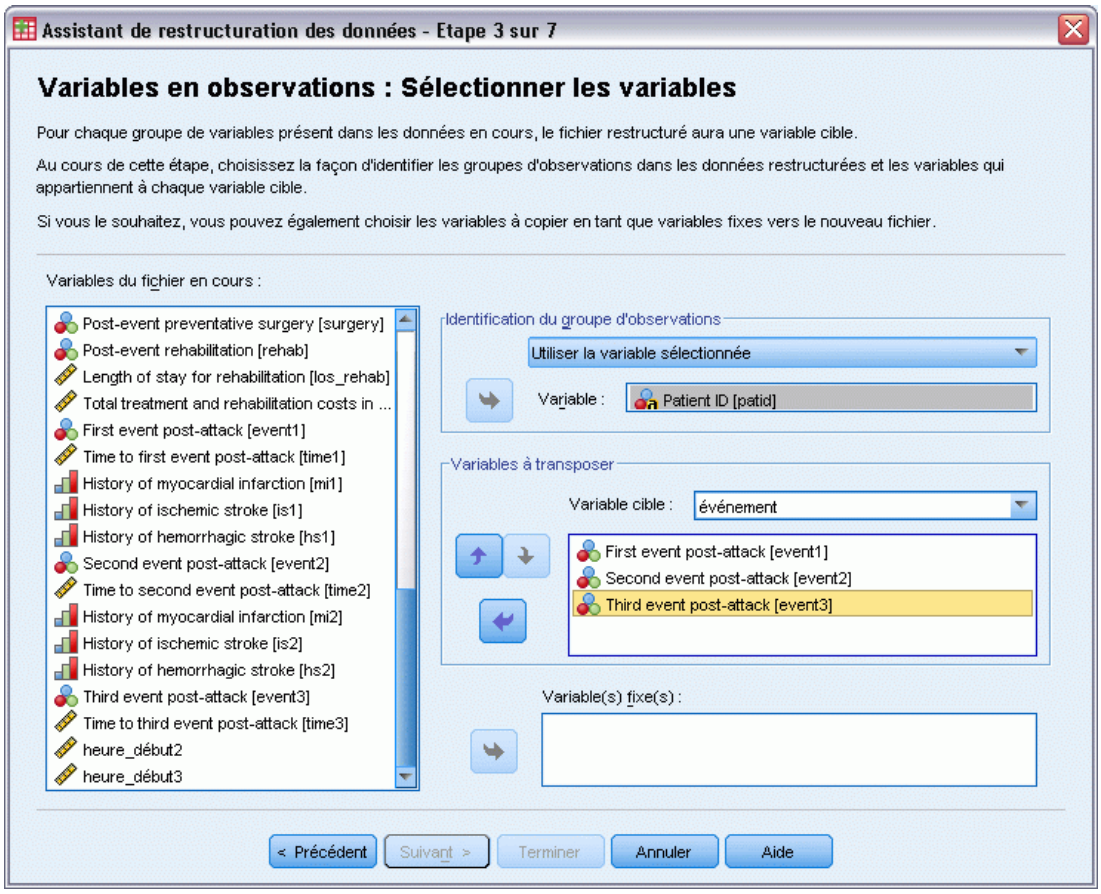

- Assurez-vous que l'option Restructurer les variables sélectionnées en observations est sélectionnée.
- Cliquez sur Suivant.

Figure 22-24
*Etape Variables en observations – Nombre de groupes de variables de l'Assistant de restructuration des données*



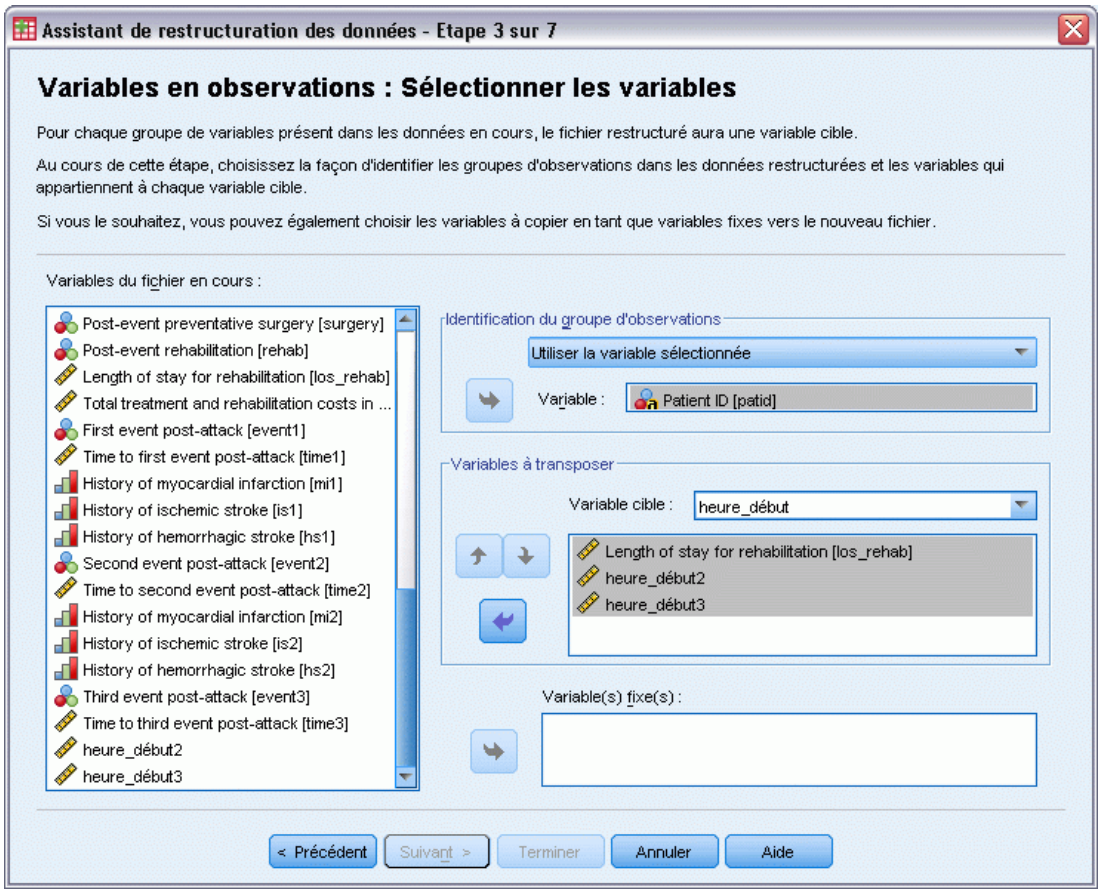

- Sélectionnez Plusieurs groupes de variables pour restructurer les données.
- Entrez 6 comme nombre de groupes.
- Cliquez sur Suivant.

Figure 22-25
*Etape Variables en observations - Sélectionner les variables de l'Assistant de restructuration des données*



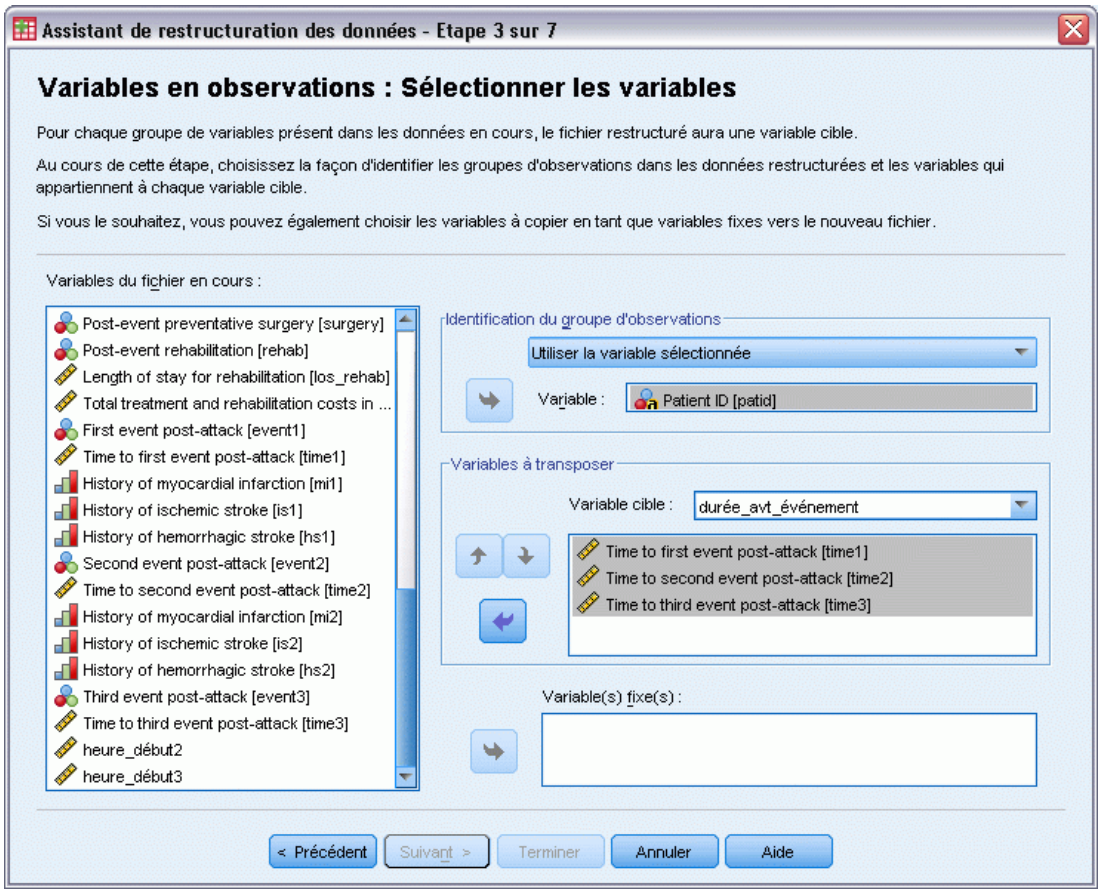

- Dans le groupe d'identification des groupes d'observations, sélectionnez Utiliser la variable sélectionnée, puis *ID patient [idpat]* comme identificateur de sujet.
- Tapez événement comme première variable cible.
- Sélectionnez *Premier événement après l'attaque [événement1]*, *Deuxième événement après l'attaque [événement2]*, et *Troisième événement après l'attaque [événement3]* comme variables à transposer.
- Sélectionnez *trans2* à partir de la liste des variables cible.

Figure 22-26
*Etape Variables en observations - Sélectionner les variables de l'Assistant de restructuration des données*



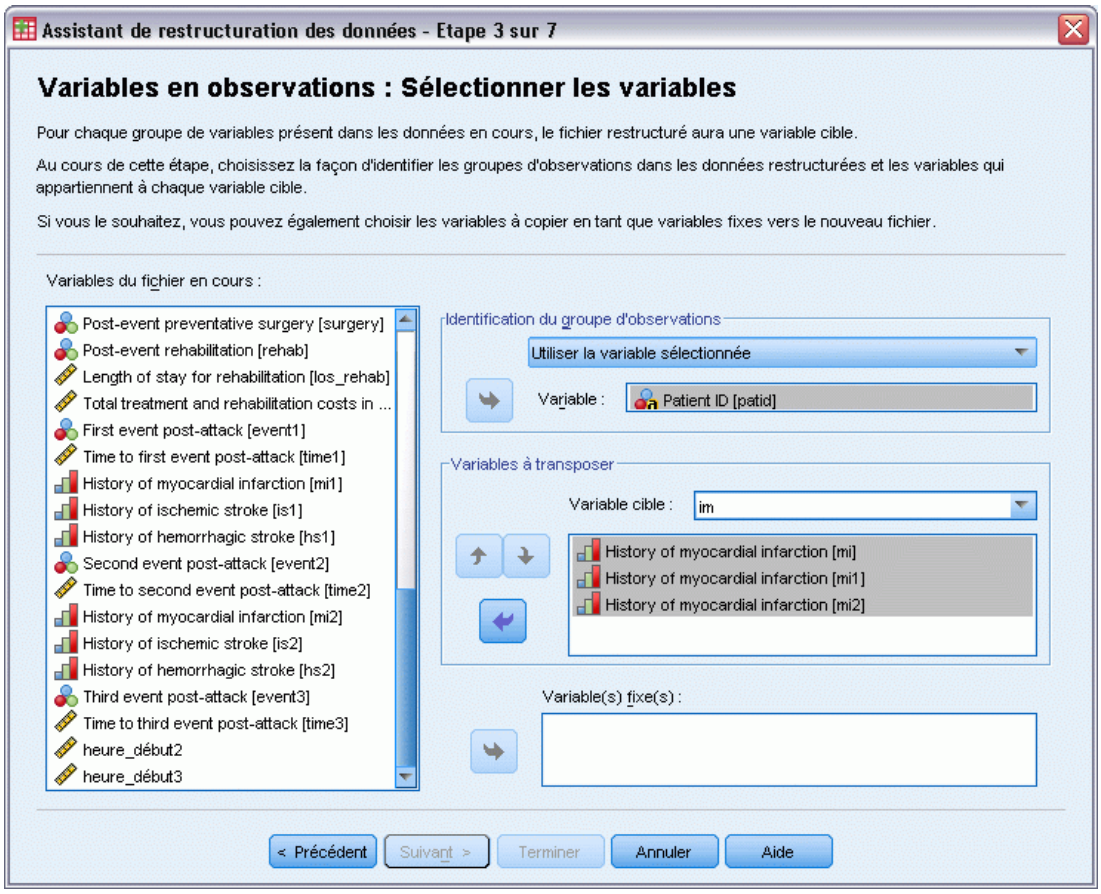

- Tapez heure_début comme variable cible.
- Sélectionnez *Durée du séjour nécessaire à la rééducation [dds_rééduc]*, *heure_début2*, et *heure_début3* comme variables à transposer. Les variables *Durée avant premier événement après attaque [durée1]* et *Durée avant deuxième événement après attaque [durée2]* seront utilisées pour créer les heures de fin, et chaque variable ne peut apparaître que dans une seule liste de variables à transposer. Par conséquent, *heure_début2* et *heure_début3* étaient nécessaires.
- Sélectionnez *trans3* à partir de la liste des variables cible.

Figure 22-27
*Etape Variables en observations - Sélectionner les variables de l'Assistant de restructuration des données*



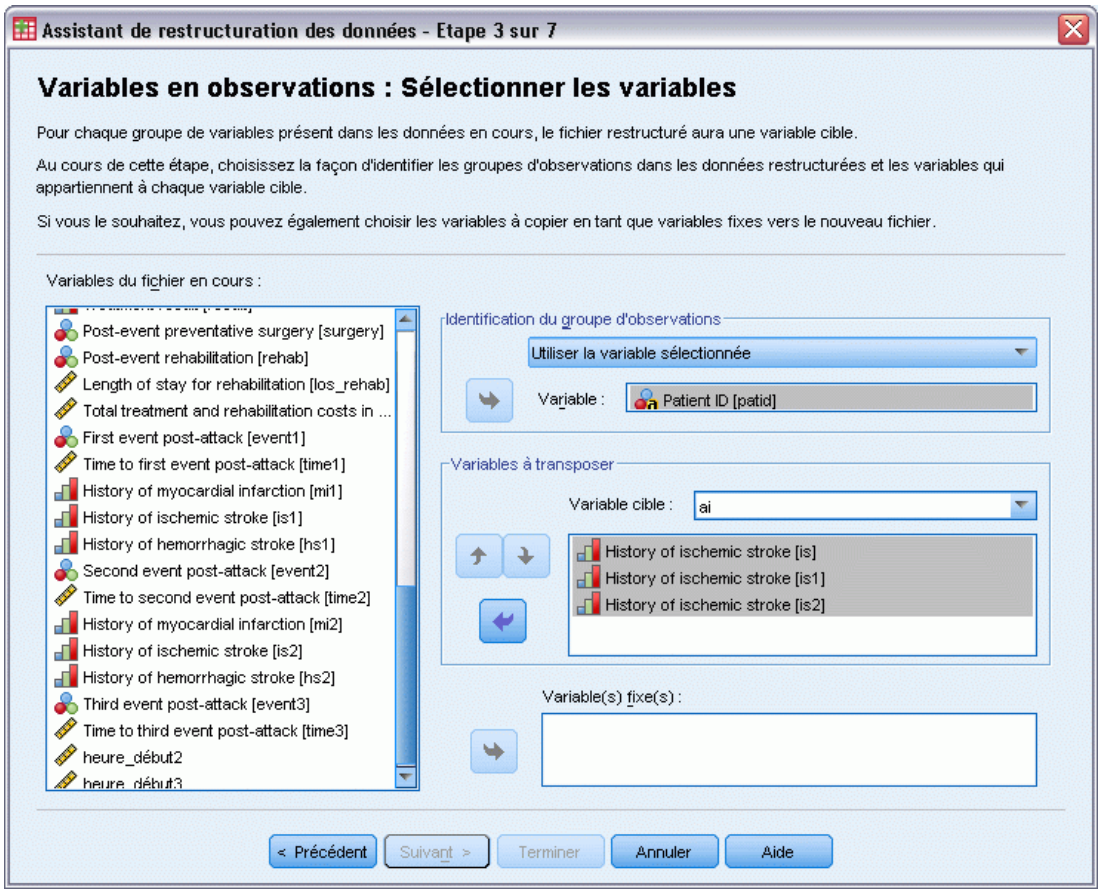

- Tapez durée_avt_événement comme valeur cible.
- Sélectionnez *Durée avant premier événement après attaque [durée1]*, *Durée avant deuxième événement après attaque [durée2]*, et *Durée avant troisième événement après attaque [durée3]* comme variables à transposer.
- Sélectionnez *trans4* à partir de la liste des variables cible.

Figure 22-28
*Etape Variables en observations - Sélectionner les variables de l'Assistant de restructuration des données*



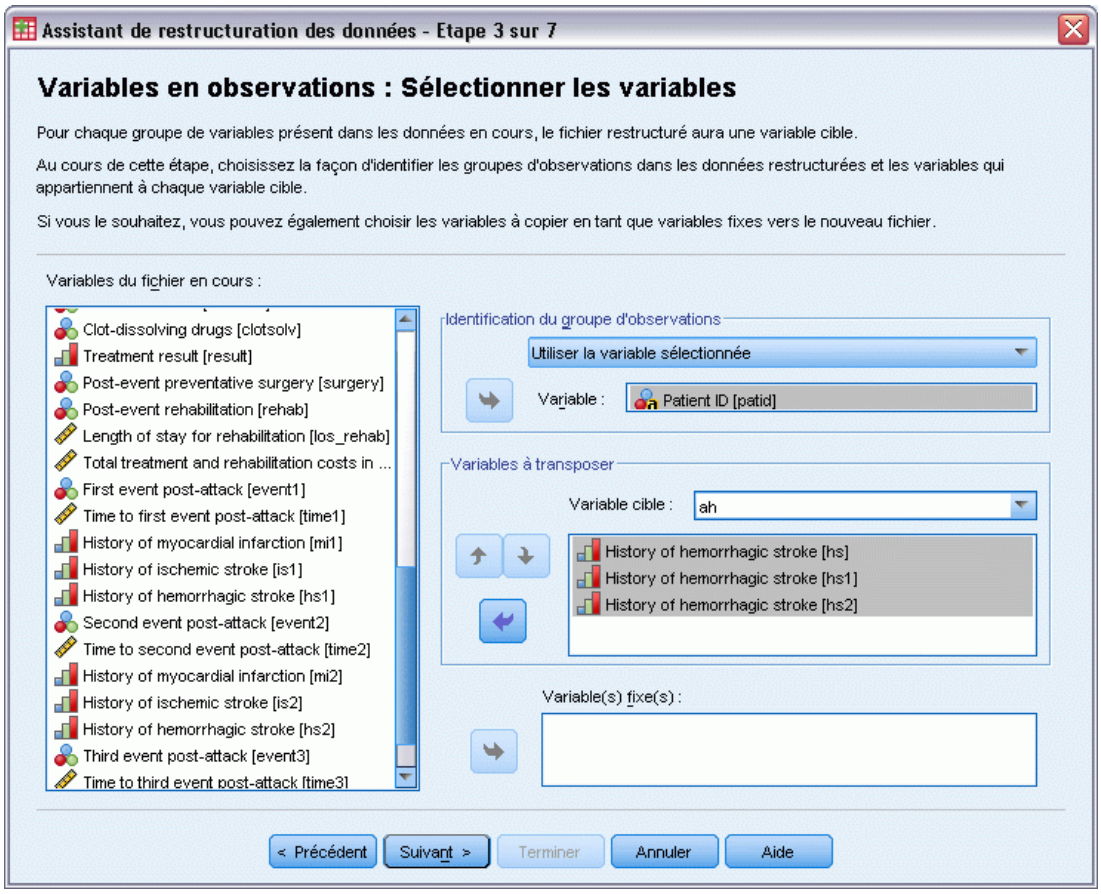

- Tapez im comme variable cible.
- Sélectionnez *Antécédents d'infarctus du myocarde [im]*, *Antécédents d'infarctus du myocarde [im1]*, et *Antécédents d'infarctus du myocarde [im2]* comme variables à transposer.
- Sélectionnez *trans5* à partir de la liste des variables cible.

Figure 22-29
*Etape Variables en observations - Sélectionner les variables de l'Assistant de restructuration des données*



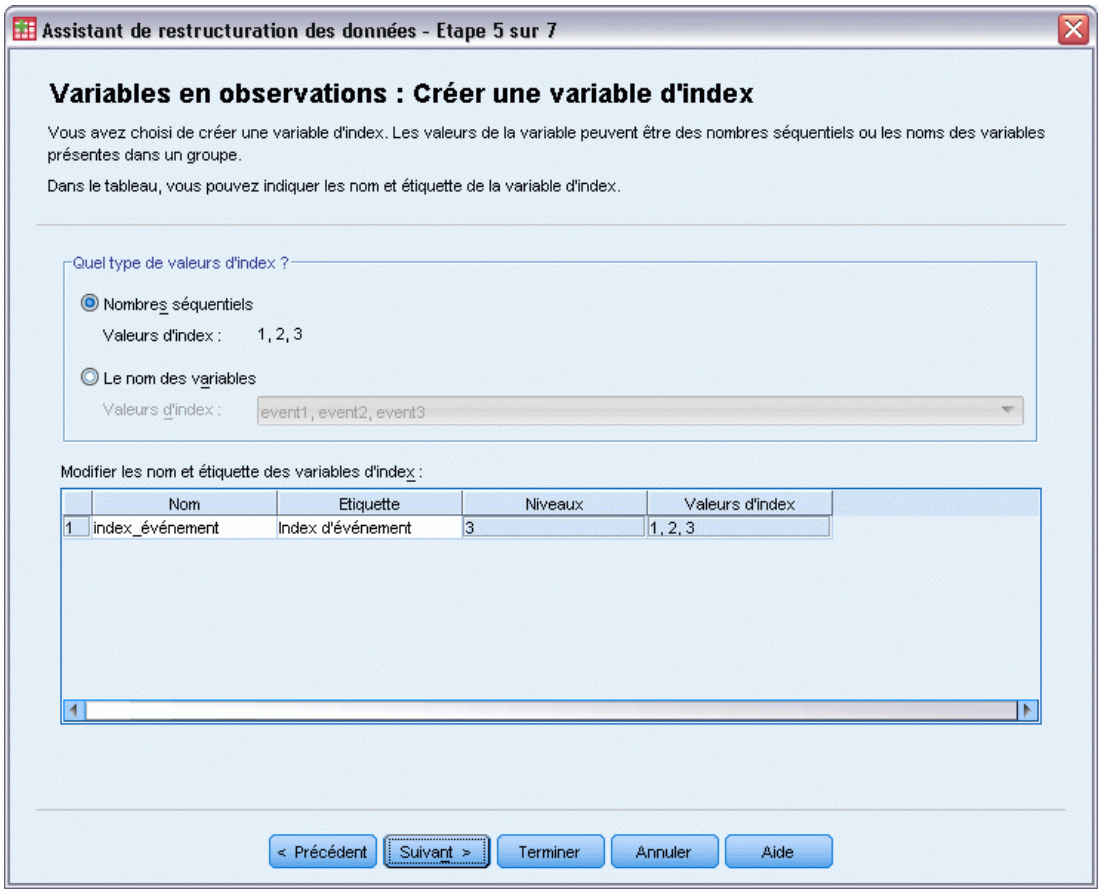

- Tapez ai comme variable cible.
- Sélectionnez *Antécédents d'accidents ischémiques [ai]*, *Antécédents d'accidents ischémiques [ai1]*, et *Antécédents d'accidents ischémiques [ai2]* comme variables à transposer.
- Sélectionnez *trans6* à partir de la liste des variables cible.

Figure 22-30
*Etape Variables en observations - Sélectionner les variables de l'Assistant de restructuration des données*



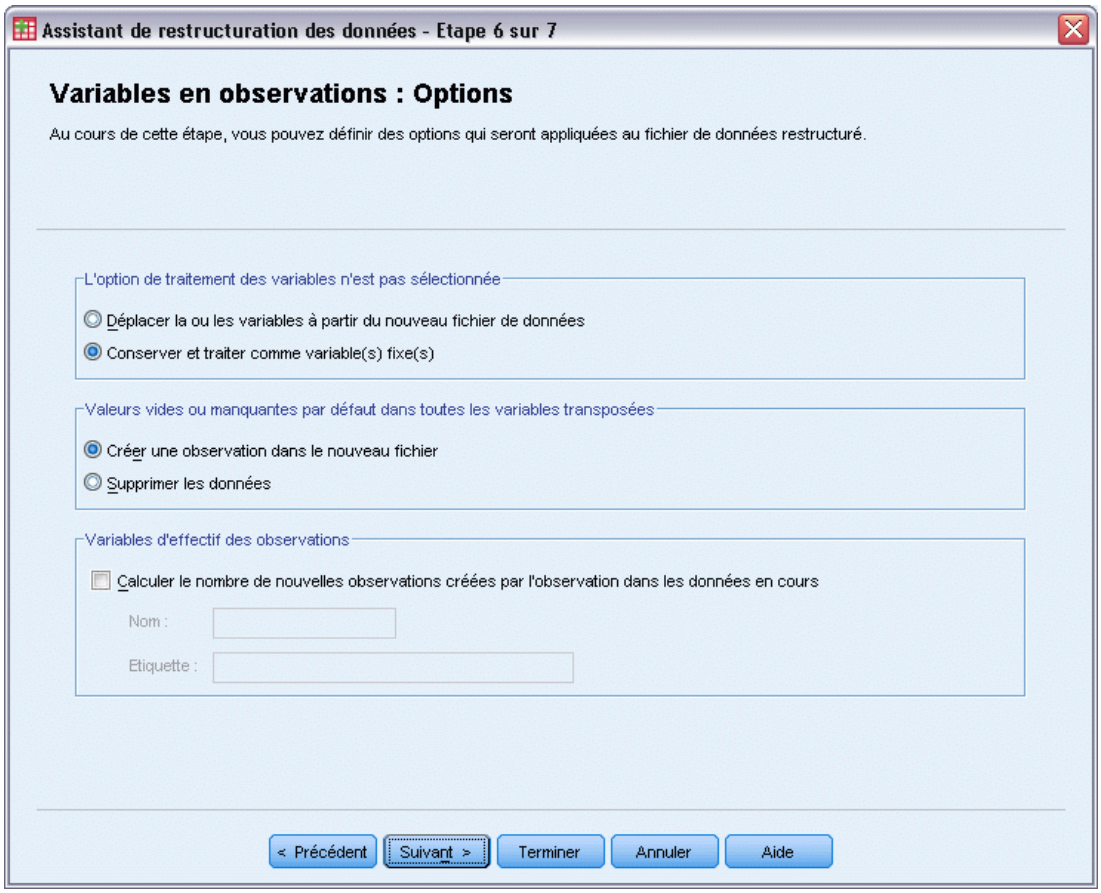

- Tapez ah comme variable cible.
- Sélectionnez *Antécédents d'accidents hémorragiques [ah]*, *Antécédents d'accidents hémorragiques [ah1]*, et *Antécédents d'accidents hémorragiques [ah2]* comme variables à transposer.
- Cliquez sur Suivant, puis sur Suivant à l'étape Créer des variables d'index.

Figure 22-31
*Etape Variables en observations - Créer une variable d'index de l'Assistant de restructuration des données*




- Tapez index_événement comme nom de la variable d'index et tapez Index d'événement comme étiquette de variable.
- Cliquez sur Suivant.

Figure 22-32
*Etape Variables en observations - Créer une variable d'index de l'Assistant de restructuration des données*




- Assurez-vous que l'option Conserver et traiter comme variable(s) fixe(s) est sélectionnée.
- Cliquez sur Terminer.

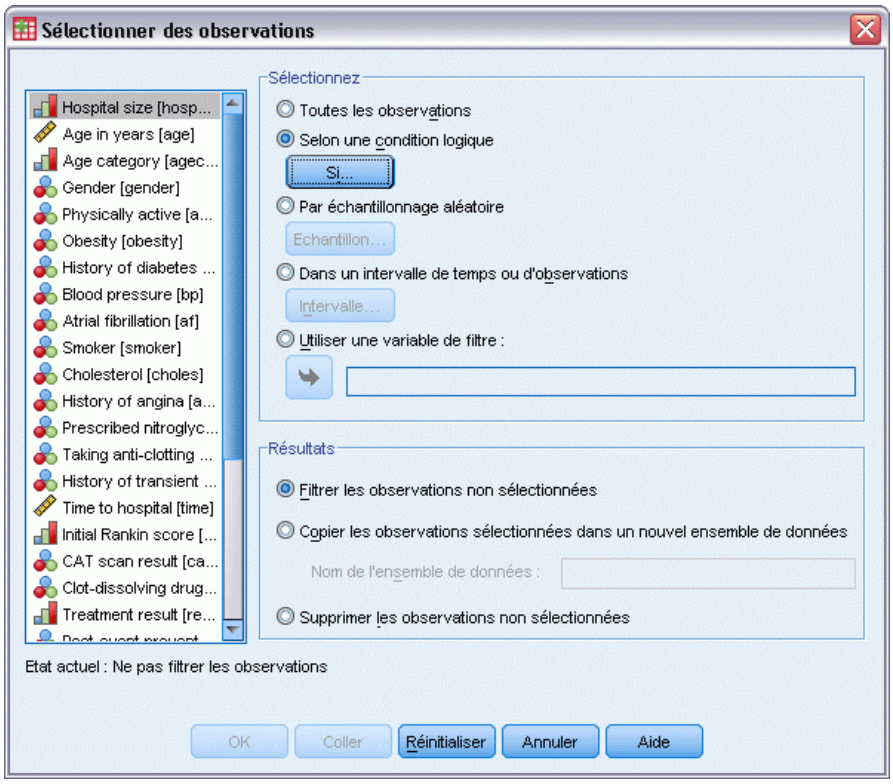
Figure 22-33
*Restructuration des données*

| event_index | event | start_time | time_to_event | mi | is | hs |
|---|---|---|---|---|---|---|
| 1 | 0 | 3 | 1500 | 0 | 1 | 0 |
| 2 | -4 | 1500 | -4 | -4 | -4 | -4 |
| 3 | -4 | . | -4 | -4 | -4 | -4 |
| 1 | 1 | 33 | 1311 | 0 | 1 | 0 |
| 2 | 4 | 1311 | 1325 | 1 | 1 | 0 |
| 3 | -3 | 1325 | -3 | -3 | -3 | -3 |
| 1 | 4 | 12 | 1098 | 1 | 1 | 0 |
| 2 | -3 | 1098 | -3 | -3 | -3 | -3 |
| 3 | -3 | . | -3 | -3 | -3 | -3 |
| 1 | 4 | 4 | 1356 | 0 | 1 | 0 |
| 2 | -3 | 1356 | -3 | -3 | -3 | -3 |
| 3 | -3 | . | -3 | -3 | -3 | -3 |

Les données restructurées contiennent trois observations pour chaque patient. Cependant, de nombreux patients ont connu moins de trois événements. Il existe donc de nombreuses observations avec des valeurs (manquantes) négatives pour *événement*. Il vous suffit de les filtrer à partir de l'ensemble de données.

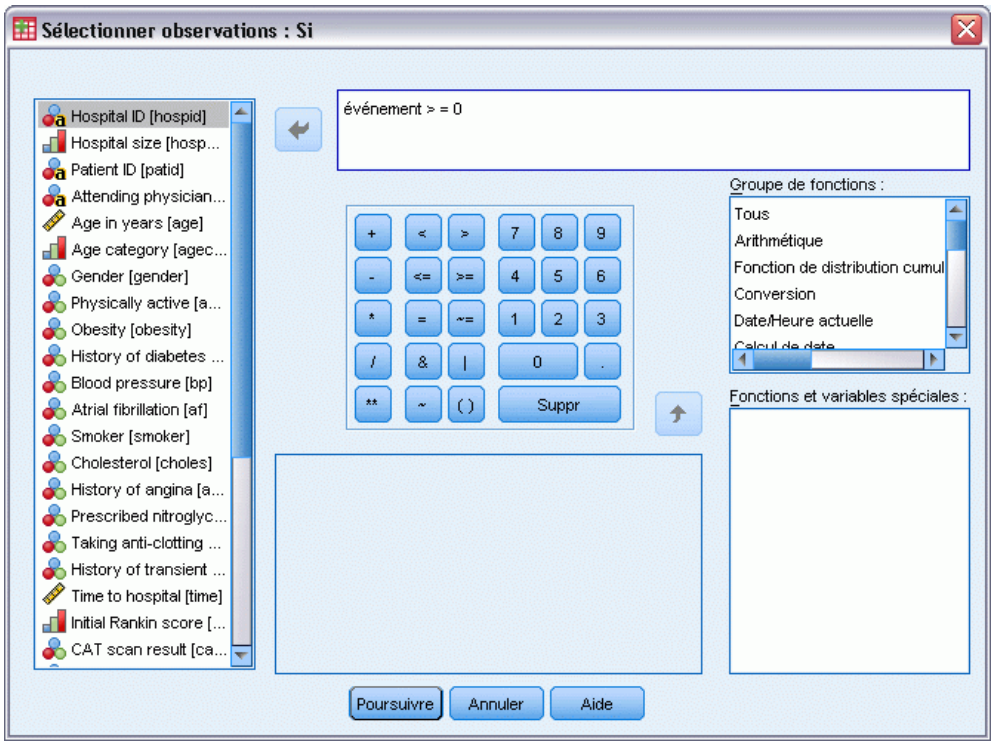
- Pour filtrer les observations, à partir des menus, choisissez :
  Données > Sélectionner des observations

Figure 22-34
*Boîte de dialogue Sélectionner des observations*



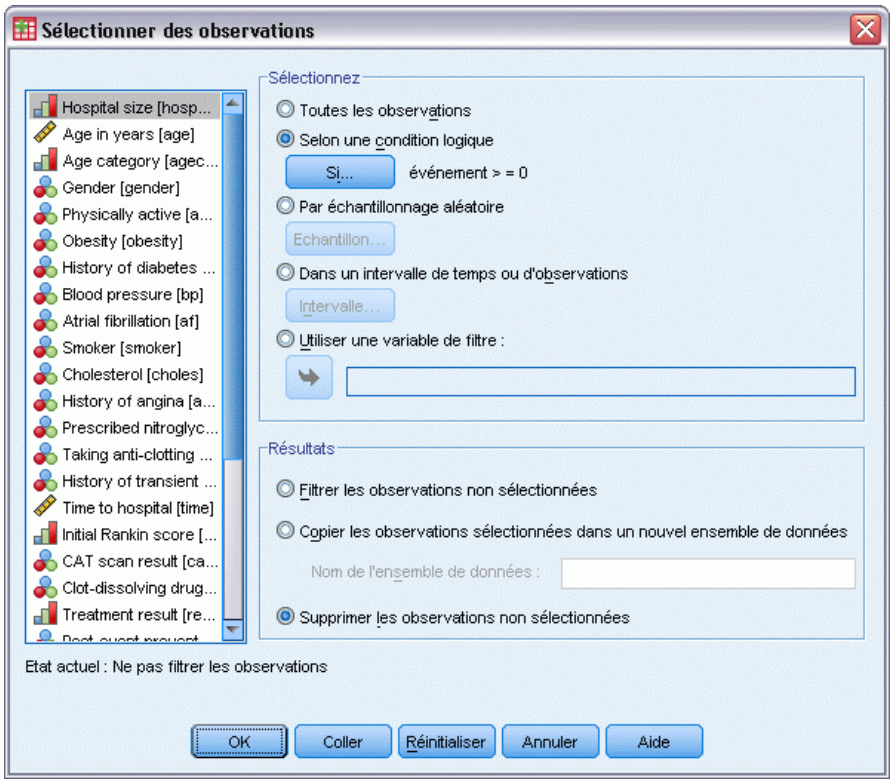

- Sélectionnez Selon une condition logique.
- Cliquez sur Si.

Figure 22-35
*Boîte de dialogue Sélectionner des observations*




- Tapez l'expression conditionnelle événement > = 0.
- Cliquez sur Poursuivre.

Figure 22-36
*Boîte de dialogue Sélectionner des observations*



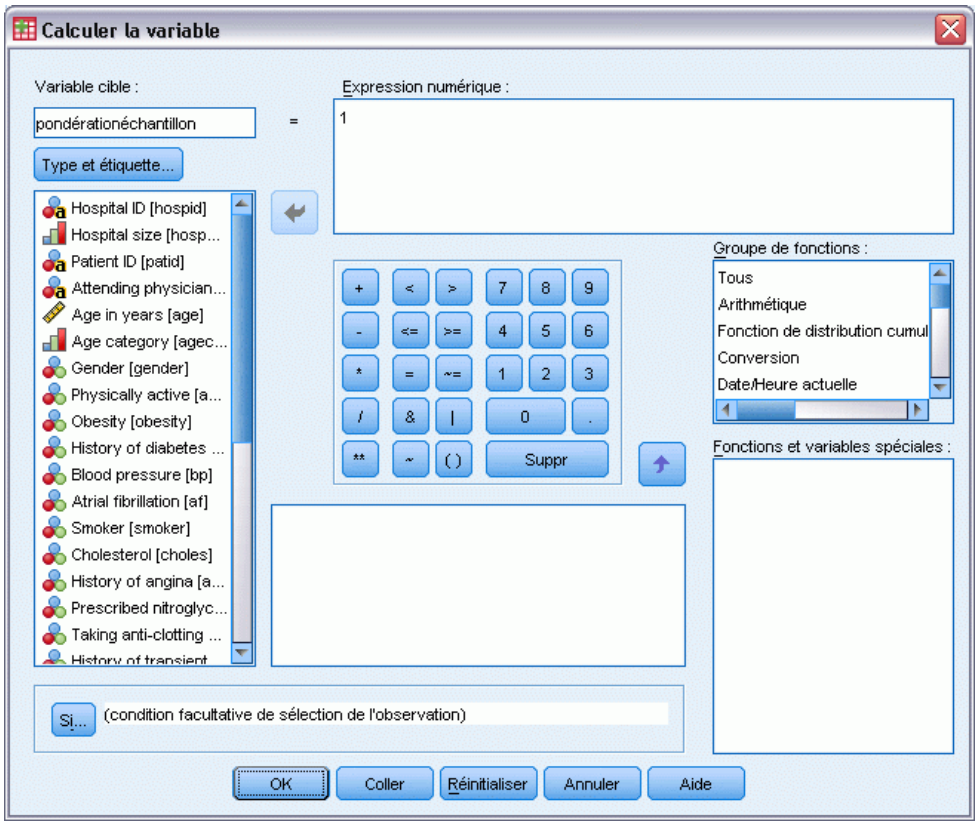

- Sélectionnez Supprimer les observations non sélectionnées.
- Cliquez sur OK.

## *Création d'un plan d'analyse d'échantillonnage aléatoire simple*

Vous êtes désormais prêt à créer le plan d'analyse d'échantillonnage aléatoire simple.

- Tout d'abord, vous devez créer une variable de pondération d'échantillonnage. A partir des menus, sélectionnez :
  Transformer > Calculer la variable...

Figure 22-37
*Boîte de dialogue principale Régression de Cox*



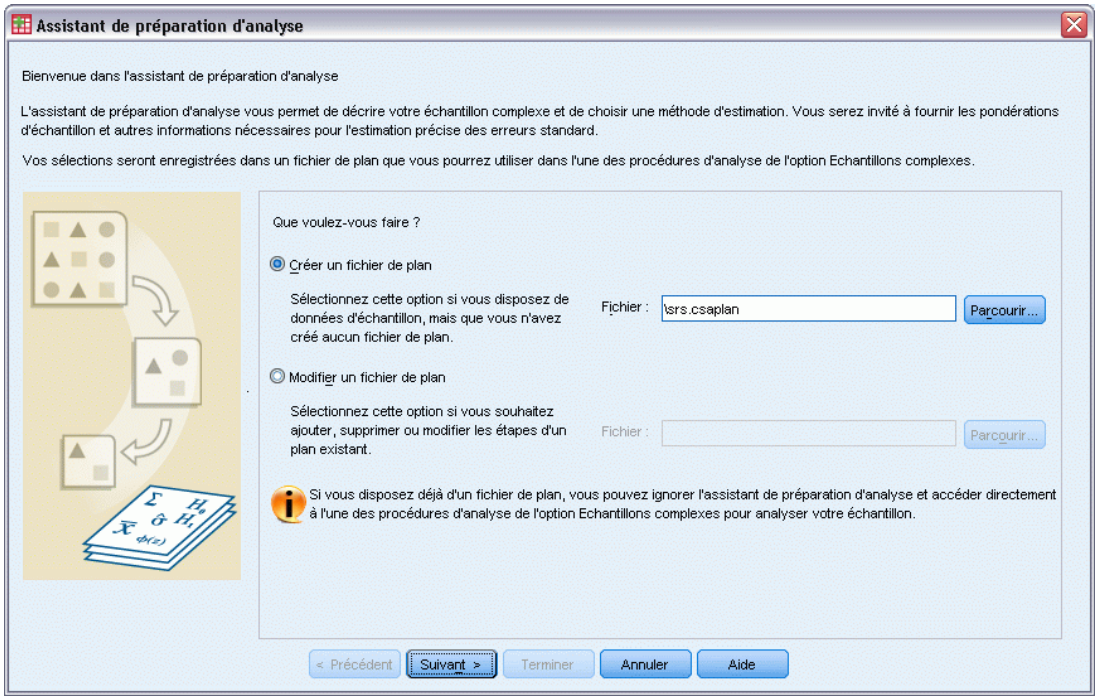

- Tapez pondérationéchantillon comme variable cible.
- Entrez l'expression numérique 1.
- Cliquez sur OK.

  Vous êtes désormais prêt à créer le plan d'analyse.

  *Remarque :* Il existe un fichier de plan, *srs.csaplan*, dans le répertoire des fichiers d'exemple que vous pouvez utiliser si vous voulez ignorer les instructions suivantes et passer à l'analyse des données.

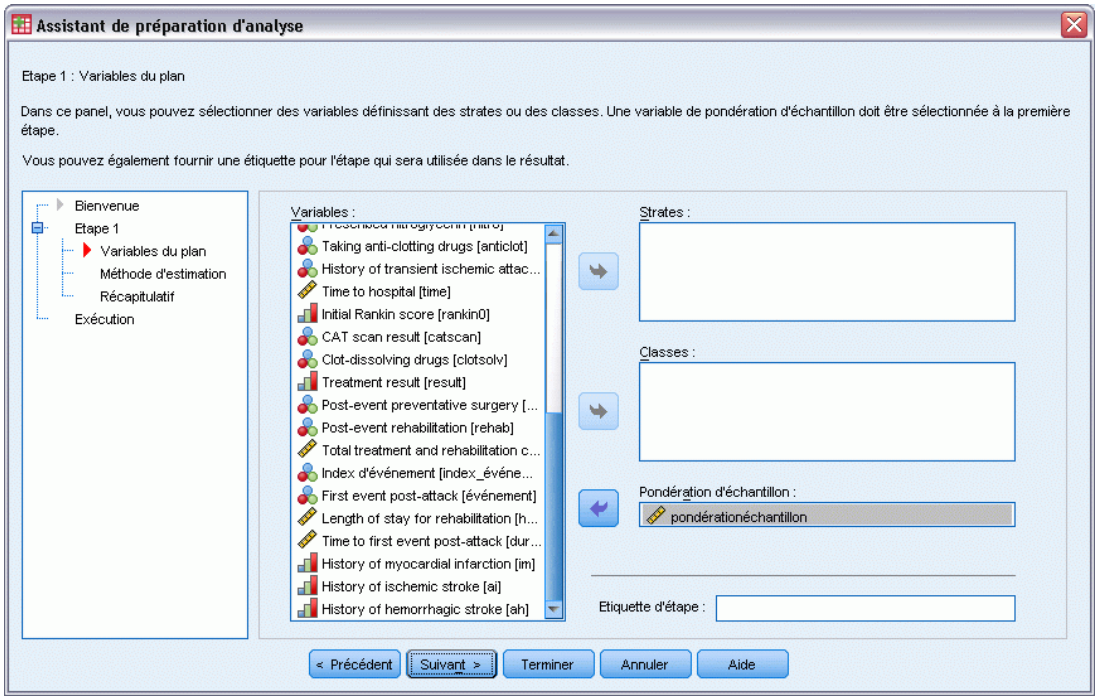
- Pour créer le plan d'analyse, à partir des menus, sélectionnez :
  Analyse > Echantillonnage > Préparer pour l'analyse...

Figure 22-38
*Etape Bienvenue de l'assistant de préparation d'analyse*



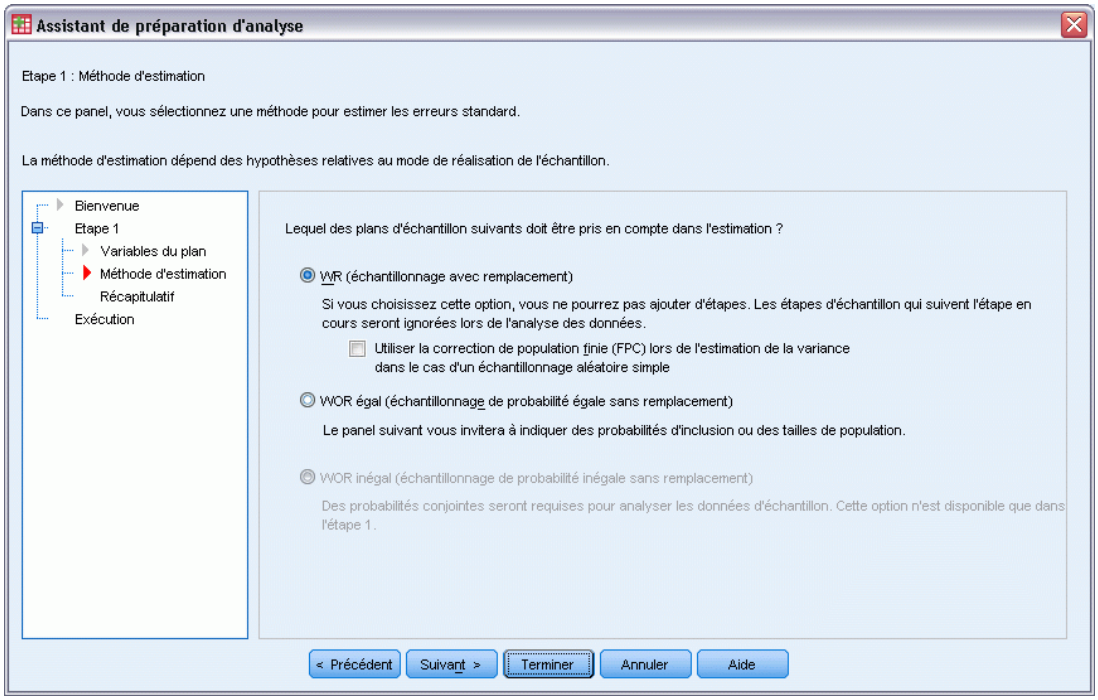

- Sélectionnez Créer un fichier de plan et entrez srs.csaplan comme nom du fichier. Vous pouvez également accéder à l'emplacement où vous souhaitez l'enregistrer.
- Cliquez sur Suivant.

Figure 22-39
*Assistant de préparation d'analyse - Variables du plan*




- Sélectionnez la variable de pondération d'échantillonnage *pondérationéchantillon*.
- Cliquez sur Suivant.

Figure 22-40
*Assistant de préparation d'analyse - Méthode d'estimation*




- Désélectionnez Utiliser la correction de population finie.
- Cliquez sur Terminer.

Vous êtes désormais prêt à exécuter l'analyse.

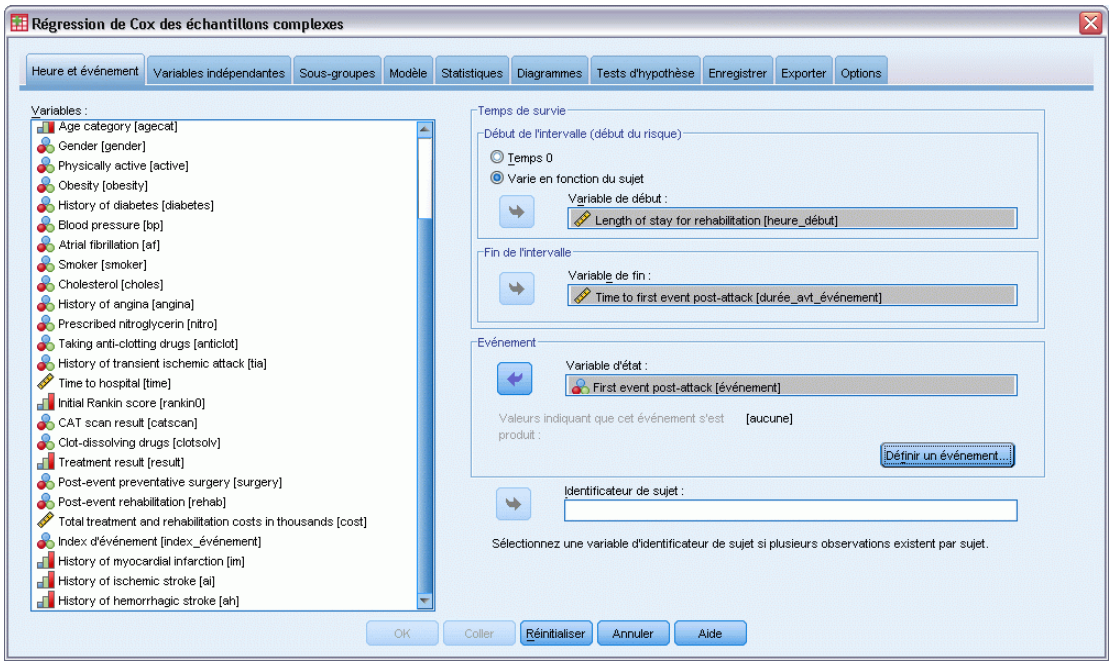
## *Exécution de l'analyse*

- Pour exécuter une analyse de régression de Cox des échantillons complexes, sélectionnez les options suivantes dans le menu :
  Analyse > Echantillonnage > Modèle de Cox

Figure 22-41
*Boîte de dialogue Plan pour régression de Cox*

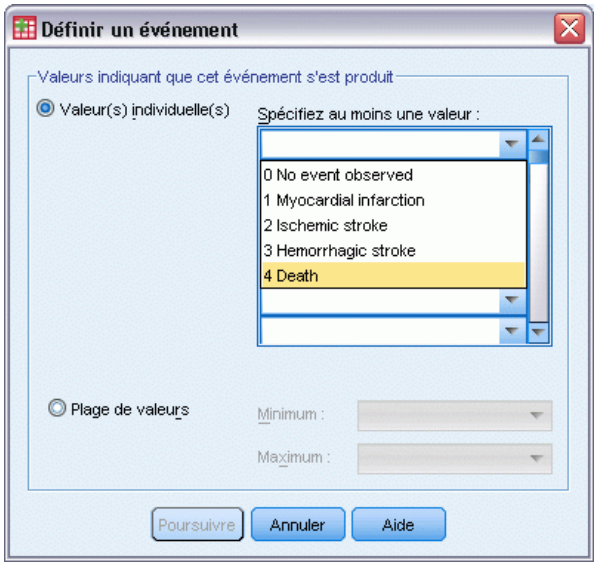

- Accédez à l'emplacement où vous avez enregistré le plan d'analyse d'échantillonnage aléatoire simple, ou au répertoire des fichiers d'exemple, et sélectionnez *srs.csaplan*.
- Cliquez sur Poursuivre.

Figure 22-42
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Heure et événement*



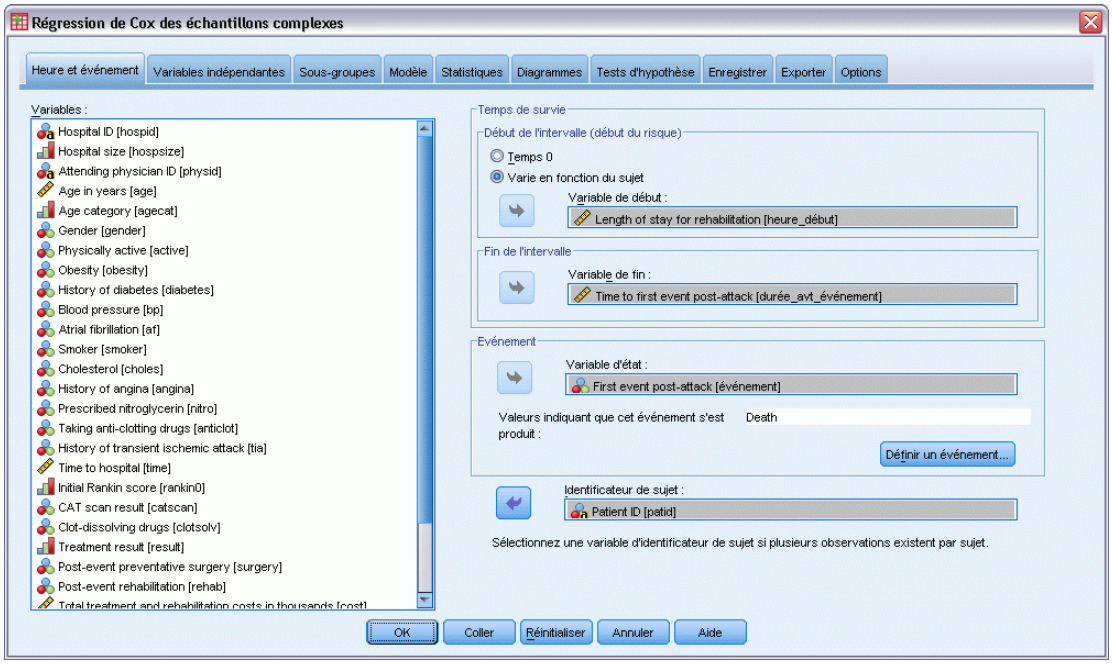

- Sélectionnez Varie en fonction du sujet, puis sélectionnez *Durée du séjour nécessaire à la rééducation [dds_rééduc]* comme variable de début. Notez que la variable restructurée a pris l'étiquette de variable de la première variable utilisée pour la construire, bien que l'étiquette ne soit pas forcément adaptée à la variable conçue.
- Sélectionnez *Durée avant premier événement après attaque [durée_avt_événement]* comme variable de fin.
- Sélectionnez *Premier événement après l'attaque [événement]* comme variable d'état.
- Cliquez sur Définir un événement.

Figure 22-43
*Boîte de dialogue Définir un événement*



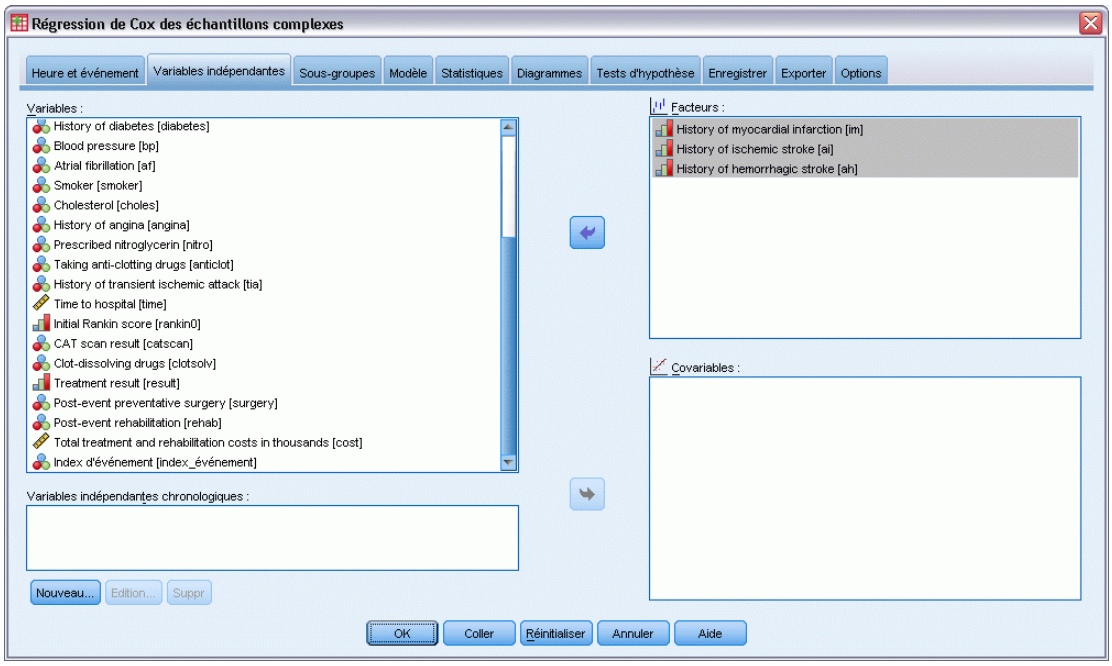

- Sélectionnez 4 décès comme valeur indiquant que l'évènement final s'est produit.
- Cliquez sur Poursuivre.

Figure 22-44
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Heure et événement*



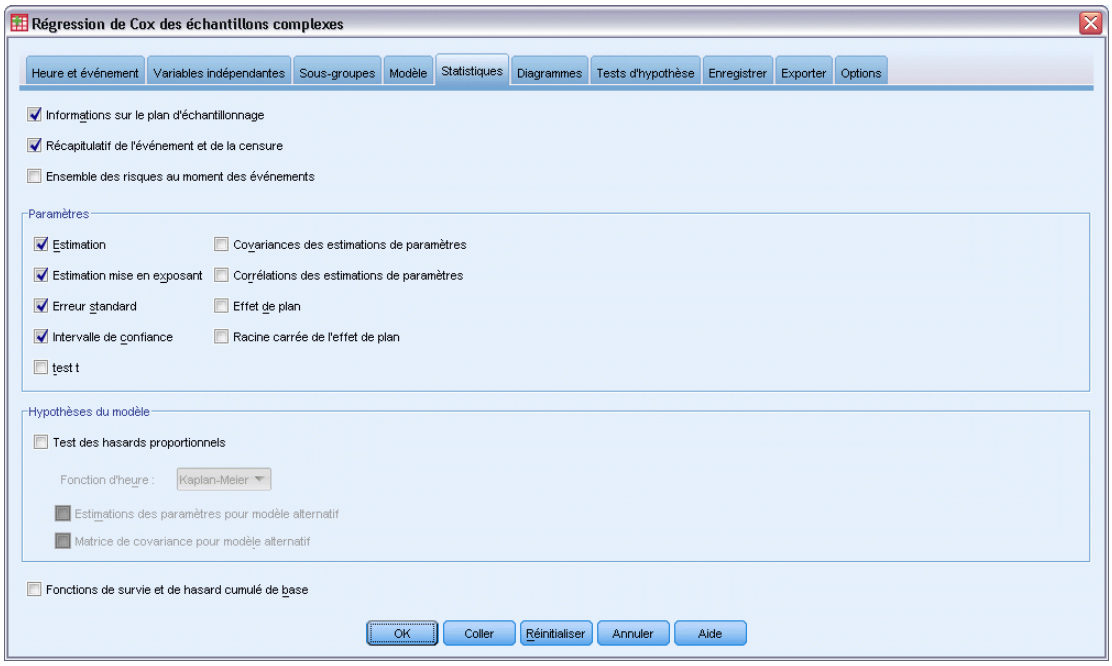

- Sélectionnez *ID patient [idpat]* comme identificateur de sujet.
- Cliquez sur l'onglet Variables indépendantes.

Figure 22-45
*Boîte de dialogue Régression de Cox, onglet Variables indépendantes*



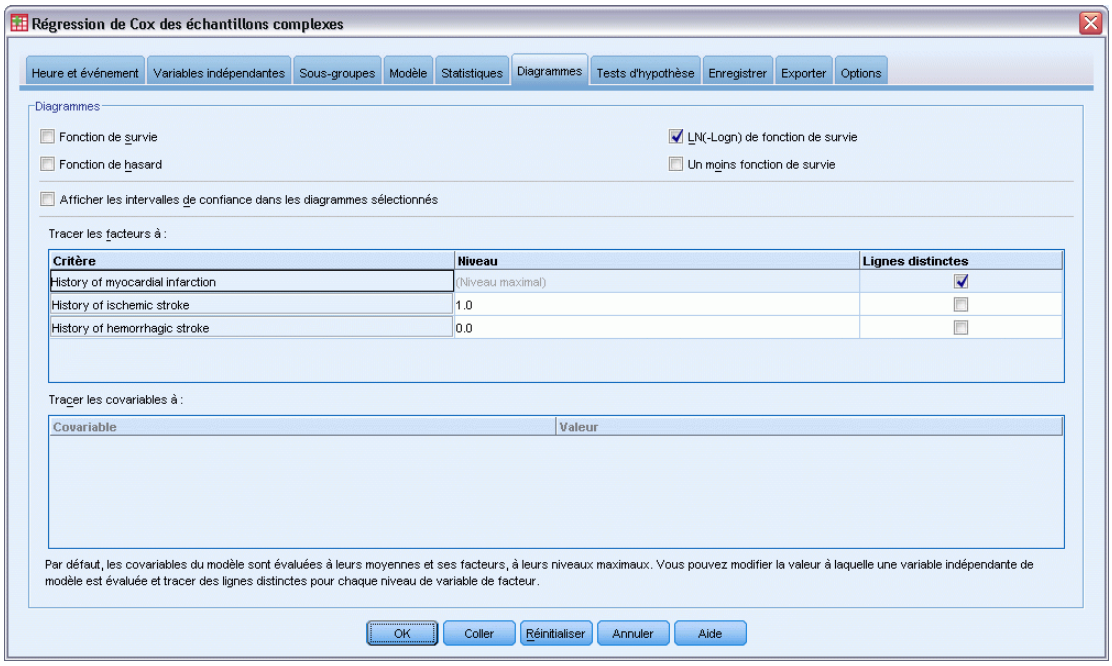

- Sélectionnez les options allant de *Antécédents d'infarctus du myocarde [im]* à *Antécédents d'accidents hémorragiques [ah]* comme facteurs.
- Cliquez sur l'onglet Statistiques.

Figure 22-46
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Statistiques*



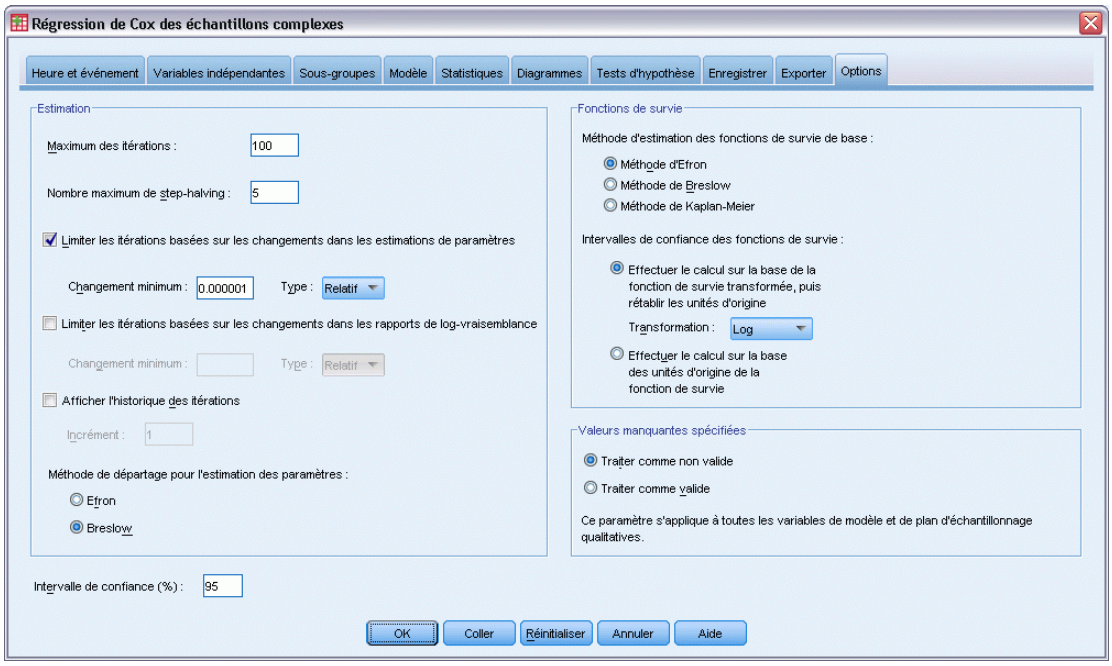

- Sélectionnez Estimation, Estimation mise en exposant, Erreur standard et Intervalle de confiance dans le groupe de paramètres.
- Cliquez sur l'onglet Diagrammes.

Figure 22-47
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Statistiques*




- Sélectionnez LN(-Logn) de fonction de survie.
- Sélectionnez Lignes distinctes pour *Antécédents d'infarctus du myocarde*.
- Sélectionnez 1,0 comme niveau pour *Antécédents d'accidents ischémiques*.
- Sélectionnez 0,0 comme niveau pour *Antécédents d'accidents hémorragiques*.
- Cliquez sur l'onglet Options.

Figure 22-48
*Boîte de dialogue Modèle de Cox///Régression de Cox - Onglet Options*




- Sélectionnez Breslow comme méthode de départage dans le groupe Estimation.
- Cliquez sur OK.

## *Informations sur le plan d'échantillonnage*

Figure 22-49
*Informations sur le plan d'échantillonnage*

| | | N |
|---|---|---|
| Valide | Sujets | 2421 |
| | Observations | 3310 |
| | Observations non valides | 0 |
| | Observations totales | 3310 |
| | Taille de sujet de population | 2421.000 |
| | Strates | 1 |
| | Unités | 2421 |
| | Degrés de liberté de plan d'échantillonnage | 2420 |

Ce tableau contient des informations sur le plan d'échantillonnage se rapportant à l'estimation du modèle.

- Il y a plusieurs observations pour certains sujets, et les 3 310 observations sont toutes utilisées dans l'analyse.
- Le plan possède une strate unique et 2 421 unités (une par sujet). Les degrés de liberté du plan d'échantillonnage sont estimés par 2421−1=2420.

## *Tests des effets de modèle*

Figure 22-50
*Tests des effets de modèle*

| Source | df1 | df2 | Wald F | Sig. |
|---|---|---|---|---|
| mi | 3.000 | 2418.000 | 452.873 | .000 |
| is | 2.000 | 2419.000 | 1064.936 | .000 |
| hs | 2.000 | 2419.000 | 739.197 | .000 |

Variable Durée de survie\ : Length of stay for rehabilitation, Time to first event post-attack
Variable Etat d'événement\ : First event post-attack = 4
Variable ID Sujet\ : Patient ID
Modèle: mi, is, hs

La valeur de signification pour chaque effet est proche de 0, ce qui suggère qu'ils contribuent tous au modèle.

## *Estimations de paramètre*

Figure 22-51
*Estimations des paramètres*

| Paramètre | B | Erreur std. | Intervalle de confiance 95% | | Exp (B) | Intervalle de confiance à 95 % pour Exp(B) | |
|---|---|---|---|---|---|---|---|
| | | | Inférieur | Supérieur | | Inférieur | Supérieur |
| [mi=0] | -6.381 | .283 | -6.935 | -5.827 | .002 | .001 | .003 |
| [mi=1] | -5.589 | .284 | -6.147 | -5.032 | .004 | .002 | .007 |
| [mi=2] | -2.119 | .344 | -2.794 | -1.445 | .120 | .061 | .236 |
| [mi=3] | .000[a] | . | . | . | 1.000 | . | . |
| [is=1] | -6.421 | .202 | -6.817 | -6.024 | .002 | .001 | .002 |
| [is=2] | -2.803 | .222 | -3.239 | -2.366 | .061 | .039 | .094 |
| [is=3] | .000[a] | . | . | . | 1.000 | . | . |
| [hs=0] | -6.148 | .355 | -6.844 | -5.453 | .002 | .001 | .004 |
| [hs=1] | -2.232 | .373 | -2.963 | -1.502 | .107 | .052 | .223 |
| [hs=2] | .000[a] | . | . | . | 1.000 | . | . |

Variable Durée de survie\ : Length of stay for rehabilitation, Time to first event post-attack
Variable Etat d'événement\ : First event post-attack = 4
Variable ID Sujet\ : Patient ID
Modèle: mi, is, hs

a. Paramétré sur zéro car ce paramètre est redondant.

b. Méthode Ex aequo\ : Breslow

La procédure utilise la dernière modalité de chaque facteur comme modalité de référence. L'effet d'autres modalités est fonction de la modalité de référence. Bien que l'estimation soit utile aux tests statistiques, l'estimation exponentielle, ou Exp(B), est plus facile à interpréter en tant que changement prévu du risque relatif à la modalité de référence.

- La valeur de Exp(B) pour *[im=0]* signifie que le risque de décès pour un patient n'ayant aucun antécédent d'infarctus du myocarde est égal à 0,002 fois celui d'un patient ayant subi trois infarctus.

- Les intervalles de confiance pour *[im=1]* et *[im=0]* se chevauchent, ce qui indique que, d'un point de vue statistique, il est impossible de distinguer le risque pour un patient ayant subi un infarctus de celui d'un patient n'ayant subi aucun infarctus.
- Les intervalles de confiance pour *[im=0]* et *[im=1]* et l'intervalle pour *[im=2]* ne se chevauchent pas, et aucun d'entre eux n'inclut 0. Par conséquent, il semble que l'on peut distinguer le risque pour les patients n'ayant subi aucun infarctus ou en ayant subi un seul du risque pour les patients ayant subi deux infarctus. Il est également possible de distinguer le risque pour ces derniers du risque pour les patients ayant subi trois infarctus.

Des relations identiques sont valables pour les niveaux de *ai* et *ah*, où l'augmentation du nombre d'accidents préalables augmente le risque de décès.

## Valeurs des modèles

Figure 22-52
*Valeurs des modèles*

| | | Intervalle de durée de survie | | | | |
|---|---|---|---|---|---|---|
| | | Début | Fin | History of myocardial infarction | History of ischemic stroke | History of hemorrhagic stroke |
| Motif de référence | 1 | .000 | | Three | Three | Two |
| Motif 1.1 | 1 | .000 | | None | One | None |
| Motif 1.2 | 1 | .000 | | One | One | None |
| Motif 1.3 | 1 | .000 | | Two | One | None |
| Motif 1.4 | 1 | .000 | | Three | One | None |

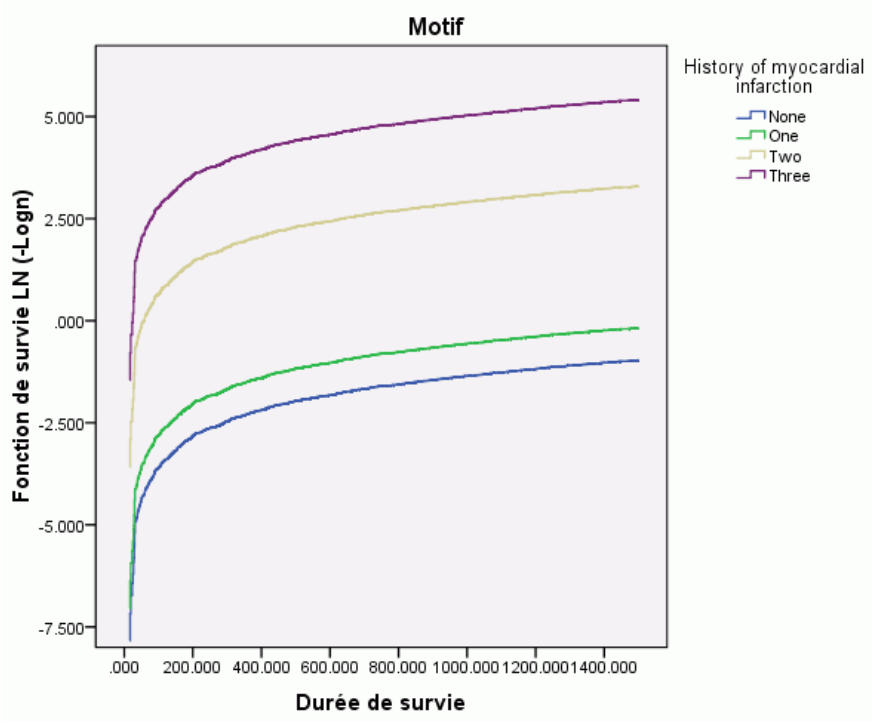
Le prédicteur non spécifié se voit assigner la valeur de ce prédicteur au motif de référence.
Chaque intervalle de durée de survie est défini comme Début < Durée de survie <= Fin.
Modèle\ : mi, is, hs.

Le tableau des valeurs des modèles répertorie les valeurs définissant chaque modèle de variable indépendante. En plus des variables indépendantes du modèle, les heures de début et de fin pour l'intervalle de survie sont affichées. Pour les analyses exécutées à partir des boîtes de dialogue, les heures de début et de fin seront toujours égales à 0 et respectivement non bornées. Vous pouvez indiquer les chemins des variables indépendantes de constante « par morceau » par le biais de la syntaxe.

- Le modèle de référence est défini à la modalité de référence pour chaque facteur et la valeur moyenne de chaque covariable (il n'existe pas de covariable dans ce modèle). Pour cet ensemble de données, la combinaison de facteurs affichés pour le modèle de référence ne peut pas se produire. Nous allons donc ignorer le diagramme LN (-Logn) pour le modèle de référence.
- La seule différence des modèles allant de 1,1 à 1,4 réside dans la valeur de la variable *Antécédents d'infarctus du myocarde*. Un modèle distinct (et une ligne distincte dans le diagramme demandé) est créé pour chaque valeur de la variable *Antécédents d'infarctus du myocarde* tandis que les autres variables restent constantes.

## *Diagramme LN (-Logn).*

Figure 22-53
*Diagramme LN (-Logn)*




Ce diagramme affiche le LN (-Logn) de fonction de survie, ln(−ln(suvival)), par rapport à la durée de survie. Ce diagramme spécifique affiche une courbe distincte pour chaque modalité de *Antécédents d'infarctus du myocarde*, avec *Antécédents d'accidents ischémiques* définie sur *Un* et *Antécédents d'accidents hémorragiques* définie sur *Aucun*, et offre une visualisation utile de l'effet de *Antécédents d'infarctus du myocarde* sur la fonction de survie. Comme le montre le tableau des estimations des paramètres, il semble possible de distinguer la survie des patients ayant subi un infarctus du myocarde ou n'en ayant subi aucun de la survie des patients ayant subi deux infarctus. Il est également possible de distinguer la survie de ces derniers de la survie des patients ayant subi trois infarctus.

## *Récapitulatif*

Vous avez ajusté un modèle de régression de Cox pour la survie suite à une attaque. Ce modèle estime les effets des antécédents variables des patients suite à une attaque. Il ne s'agit que d'un début, dans la mesure où les chercheurs souhaiteraient certainement inclure d'autres variables indépendantes éventuelles dans le modèle. Par ailleurs, dans une analyse ultérieure de cet ensemble de données, vous pouvez envisager des changements plus significatifs de la structure du modèle. Par exemple, le modèle actuel suppose que l'effet d'un événement modifiant les antécédents des patients peut être quantifié par un multiplicateur au hasard de base. Il serait plus raisonnable de

supposer que la forme du hasard de base est modifiée par l'occurrence d'un événement non mortel. Pour ce faire, vous pouvez stratifier l'analyse en fonction de la variable *Index d'événement*.

Annexe

# A

# Fichiers d'exemple

Les fichiers d'exemple installés avec le produit figurent dans le sous-répertoire *Echantillons* du répertoire d'installation. Il existe un dossier distinct au sein du sous-répertoire Echantillons pour chacune des langues suivantes : Anglais, Français, Allemand, Italien, Japonais, Coréen, Polonais, Russe, Chinois simplifié, Espangol et Chinois traditionnel.

Seuls quelques fichiers d'exemples sont disponibles dans toutes les langues. Si un fichier d'exemple n'est pas disponible dans une langue, le dossier de langue contient la version anglaise du fichier d'exemple.

### *Descriptions*

Voici de brèves descriptions des fichiers d'exemple utilisés dans divers exemples à travers la documentation.

- **accidents.sav.**Ce fichier de données d'hypothèse concerne une société d'assurance qui étudie les facteurs de risque liés à l'âge et au sexe dans les accidents de la route survenant dans une région donnée. Chaque observation correspond à une classification croisée de la catégorie d'âge et du sexe.
- **adl.sav.** Ce fichier de données d'hypothèse concerne les mesures entreprises pour identifier les avantages d'un type de thérapie proposé aux patients qui ont subi une attaque cardiaque. Les médecins ont assigné de manière aléatoire les patients du sexe féminin ayant subi une attaque cardiaque à un groupe parmi deux groupes possibles. Le premier groupe a fait l'objet de la thérapie standard tandis que le second a bénéficié en plus d'une thérapie émotionnelle. Trois mois après les traitements, les capacités de chaque patient à effectuer les tâches ordinaires de la vie quotidienne ont été notées en tant que variables ordinales.
- **advert.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un détaillant pour examiner la relation existant entre l'argent dépensé dans la publicité et les ventes résultantes. Pour ce faire, il collecte les chiffres des ventes passées et les coûts associés à la publicité.
- **aflatoxin.sav.** Ce fichier de données d'hypothèse concerne le test de l'aflatoxine dans des récoltes de maïs. La concentration de ce poison varie largement d'une récolte à l'autre et au sein de chaque récolte. Un processeur de grain a reçu 16 échantillons issus de 8 récoltes de maïs et a mesuré les niveaux d'alfatoxine en parties par milliard (PPB).
- **anorectic.sav.** En cherchant à développer une symptomatologie standardisée du comportement anorexique/boulimique, des chercheurs(Van der Ham, Meulman, Van Strien, et Van Engeland, 1997) ont examiné 55 adolescents souffrant de troubles alimentaires. Chaque patient a été observé quatre fois sur une période de quatre années, soit un total de 220 observations. A chaque observation, les patients ont été notés pour chacun des 16 symptômes. En raison de l'absence de scores de symptôme pour le patient 71/visite 2, le patient 76/visite 2 et le patient 47/visite 3, le nombre d'observations valides est de 217.

- **bankloan.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une banque pour réduire le taux de défaut de paiement. Il contient des informations financières et démographiques sur 850 clients existants et éventuels. Les premières 700 observations concernent des clients auxquels des prêts ont été octroyés. Les 150 dernières observations correspondant aux clients éventuels que la banque doit classer comme bons ou mauvais risques de crédit.
- **bankloan_binning.sav.** Ce fichier de données d'hypothèse concerne des informations financières et démographiques sur 5 000 clients existants.
- **behavior.sav.** Dans un exemple classique (Price et Bouffard, 1974), on a demandé à 52 étudiants de noter les combinaisons établies à partir de 15 situations et de 15 comportements sur une échelle de 0 à 9, où 0 = « extrêmement approprié » et 9 = « extrêmement inapproprié ». En effectuant la moyenne des résultats de l'ensemble des individus, on constate une certaine différence entre les valeurs.
- **behavior_ini.sav.** Ce fichier de données contient la configuration initiale d'une solution bidimensionnelle pour *behavior.sav*.
- **brakes.sav.** Ce fichier de données d'hypothèse concerne le contrôle qualité effectué dans une usine qui fabrique des freins à disque pour des voitures haut de gamme. Le fichier de données contient les mesures de diamètre de 16 disques de 8 machines de production. Le diamètre cible des freins est de 322 millimètres.
- **breakfast.sav.** Au cours d'une étude classique (Green et Rao, 1972), on a demandé à 21 étudiants en MBA (Master of Business Administration) de l'école de Wharton et à leurs conjoints de classer 15 aliments du petit-déjeuner selon leurs préférences, de 1= « aliment préféré » à 15= « aliment le moins apprécié ». Leurs préférences ont été enregistrées dans six scénarios différents, allant de « Préférence générale » à « En-cas avec boisson uniquement ».
- **breakfast-overall.sav.** Ce fichier de données contient les préférences de petit-déjeuner du premier scénario uniquement, « Préférence générale ».
- **broadband_1.sav.** Ce fichier de données d'hypothèse concerne le nombre d'abonnés, par région, à un service haut débit. Le fichier de données contient le nombre d'abonnés mensuels de 85 régions sur une période de quatre ans.
- **broadband_2.sav.** Ce fichier de données est identique au fichier *broadband_1.sav* mais contient les données relatives à trois mois supplémentaires.
- **car_insurance_claims.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs (McCullagh et Nelder, 1989) qui concerne des actions en indemnisation pour des voitures. Le montant d'action en indemnisation moyen peut être modelé comme présentant une distribution gamma, à l'aide d'une fonction de lien inverse pour associer la moyenne de la variable dépendante à une combinaison linéaire de l'âge de l'assuré, du type de véhicule et de l'âge du véhicule. Le nombre d'actions entreprises peut être utilisé comme pondération de positionnement.
- **car_sales.sav.** Ce fichier de données contient des estimations de ventes hypothétiques, des barèmes de prix et des spécifications physiques concernant divers modèles et marques de véhicule. Les barèmes de prix et les spécifications physiques proviennent tour à tour de *edmunds.com* et des sites des constructeurs.
- **car_sales_uprepared.sav.** Il s'agit d'une version modifiée de *car_sales.sav* qui n'inclut aucune version transformée des champs.

- **carpet.sav.** Dans un exemple courant (Green et Wind, 1973), une société intéressée par la commercialisation d'un nouveau nettoyeur de tapis souhaite examiner l'influence de cinq critères sur la préférence du consommateur : la conception du conditionnement, la marque, le prix, une étiquette *Economique* et une garantie satisfait ou remboursé. Il existe trois niveaux de critère pour la conception du conditionnement, suivant l'emplacement de l'applicateur, trois marques (*K2R*, *Glory* et *Bissell*), trois niveaux de prix et deux niveaux (non ou oui) pour chacun des deux derniers critères. Dix consommateurs classent 22 profils définis par ces critères. La variable *Préférence* indique le classement des rangs moyens de chaque profil. Un rang faible correspond à une préférence élevée. Cette variable reflète une mesure globale de préférence pour chaque profil.
- **carpet_prefs.sav.** Ce fichier de données repose sur le même exemple que celui décrit pour *carpet.sav*, mais contient les classements réels issus de chacun des 10 clients. On a demandé aux consommateurs de classer les 22 profils de produits, du préféré au moins intéressant. Les variables *PREF1* à *PREF22* contiennent les identificateurs des profils associés, tels qu'ils sont définis dans *carpet_plan.sav*.
- **catalog.sav.** Ce fichier de données contient des chiffres de ventes mensuelles hypothétiques relatifs à trois produits vendus par une entreprise de vente par correspondance. Les données relatives à cinq variables explicatives possibles sont également incluses.
- **catalog_seasfac.sav.** Ce fichier de données est identique à *catalog.sav* mais contient en plus un ensemble de facteurs saisonniers calculés à partir de la procédure de désaisonnalisation, ainsi que les variables de date correspondantes.
- **cellular.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un opérateur téléphonique pour réduire les taux de désabonnement. Des scores de propension au désabonnement sont attribués aux comptes, de 0 à 100. Les comptes ayant une note égale ou supérieure à 50 sont susceptibles de changer de fournisseur.
- **ceramics.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fabricant pour déterminer si un nouvel alliage haute qualité résiste mieux à la chaleur qu'un alliage standard. Chaque observation représente un test séparé de l'un des deux alliages ; le degré de chaleur auquel l'alliage ne résiste pas est enregistré.
- **cereal.sav.** Ce fichier de données d'hypothèse concerne un sondage de 880 personnes interrogées sur leurs préférences de petit-déjeuner et sur leur âge, leur sexe, leur situation familiale et leur mode de vie (actif ou non actif, selon qu'elles pratiquent une activité physique au moins deux fois par semaine). Chaque observation correspond à un répondant distinct.
- **clothing_defects.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de textile. Dans chaque lot produit à l'usine, les inspecteurs prélèvent un échantillon de vêtements et comptent le nombre de vêtements qui ne sont pas acceptables.
- **coffee.sav.** Ce fichier de données concerne l'image perçue de six marques de café frappé (Kennedy, Riquier, et Sharp, 1996). Pour chacun des 23 attributs d'image de café frappé, les personnes sollicitées ont sélectionné toutes les marques décrites par l'attribut. Les six marques sont appelées AA, BB, CC, DD, EE et FF à des fins de confidentialité.
- **contacts.sav.** Ce fichier de données d'hypothèse concerne les listes de contacts d'un groupe de représentants en informatique d'entreprise. Chaque contact est classé selon le service de l'entreprise où il travaille et le classement de son entreprise. Sont également enregistrés le

montant de la dernière vente effectuée, le temps passé depuis la dernière vente et la taille de l'entreprise du contact.

- **creditpromo.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un grand magasin pour évaluer l'efficacité d'une promotion récente de carte de crédit. A cette fin, 500 détenteurs de carte ont été sélectionnés au hasard. La moitié a reçu une publicité faisant la promotion d'un taux d'intérêt réduit sur les achats effectués dans les trois mois à venir. L'autre moitié a reçu une publicité saisonnière standard.
- **customer_dbase.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour utiliser les informations figurant dans sa banque de données et proposer des offres spéciales aux clients susceptibles d'être intéressés. Un sous-groupe de la base de clients a été sélectionné au hasard et a reçu des offres spéciales. Les réponses des clients ont été enregistrées.
- **customer_information.sav.** Un fichier de données d'hypothèse qui contient les informations postales du client, telles que le nom et l'adresse.
- **customer_subset.sav.** Un sous-ensemble de 80 observations de *customer_dbase.sav*.
- **debate.sav.** Ce fichier de données d'hypothèse concerne des réponses appariées à une enquête donnée aux participants à un débat politique avant et après le débat. Chaque observation représente un répondant distinct.
- **debate_aggregate.sav.** Il s'agit d'un fichier de données d'hypothèse qui rassemble les réponses dans le fichier *debate.sav*. Chaque observation correspond à une classification croisée de préférence avant et après le débat.
- **demo.sav.** Ce fichier de données d'hypothèse concerne une base de données clients achetée en vue de diffuser des offres mensuelles. Les données indiquent si le client a répondu ou non à l'offre et contiennent diverses informations démographiques.
- **demo_cs_1.sav.** Ce fichier de données d'hypothèse concerne la première mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à une ville différente. La région, la province, le quartier et la ville sont enregistrés.
- **demo_cs_2.sav.** Ce fichier de données d'hypothèse concerne la seconde mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à un ménage différent issu des villes sélectionnées à la première étape. La région, la province, le quartier, la ville, la sous-division et l'identification sont enregistrés. Les informations d'échantillonnage des deux premières étapes de la conception sont également incluses.
- **demo_cs.sav.** Ce fichier de données d'hypothèse concerne des informations d'enquête collectées via une méthode complexe d'échantillonnage. Chaque observation correspond à un ménage différent et diverses informations géographiques et d'échantillonnage sont enregistrées.
- **dmdata.sav.** Ceci est un fichier de données d'hypothèse qui contient des informations démographiques et des informations concernant les achats pour une entreprise de marketing direct. *dmdata2.sav* contient les informations pour un sous-ensemble de contacts qui ont reçu un envoi d'essai, et *dmdata3.sav* contient des informations sur les contacts restants qui n'ont pas reçu l'envoi d'essai.

- **dietstudy.sav.** Ce fichier de données d'hypothèse contient les résultats d'une étude portant sur le régime de Stillman(Rickman, Mitchell, Dingman, et Dalen, 1974). Chaque observation correspond à un sujet distinct et enregistre son poids en livres avant et après le régime, ainsi que ses niveaux de triglycérides en mg/100 ml.
- **dvdplayer.sav.** Ce fichier de données d'hypothèse concerne le développement d'un nouveau lecteur DVD. A l'aide d'un prototype, l'équipe de marketing a collecté des données de groupes spécifiques. Chaque observation correspond à un utilisateur interrogé et enregistre des informations démographiques sur cet utilisateur, ainsi que ses réponses aux questions portant sur le prototype.
- **german_credit.sav.** Ce fichier de données provient de l'ensemble de données « German credit » figurant dans le référentiel Machine Learning Databases (Blake et Merz, 1998) de l'université de Californie, Irvine.
- **grocery_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *grocery_coupons.sav* dans lequel les achats hebdomadaires sont organisés par client distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, le montant dépensé enregistré est à présent la somme des montants dépensés au cours des quatre semaines de l'enquête.
- **grocery_coupons.sav.** Il s'agit d'un fichier de données d'hypothèse qui contient des données d'enquête collectées par une chaîne de magasins d'alimentation qui chercher à déterminer les habitudes de consommation de ses clients. Chaque client est suivi pendant quatre semaines et chaque observation correspond à une semaine distincte. Les informations enregistrées concernent les endroits où le client effectue ses achats, la manière dont il les effectue, ainsi que les sommes dépensées en provisions au cours de cette semaine.
- **guttman.sav.** Bell (Bell, 1961) a présenté un tableau pour illustrer les groupes sociaux possibles. Guttman (Guttman, 1968) a utilisé une partie de ce tableau, dans lequel cinq variables décrivant des éléments tels que l'interaction sociale, le sentiment d'appartenance à un groupe, la proximité physique des membres et la formalité de la relation, ont été croisées avec sept groupes sociaux théoriques, dont les foules (par exemple, le public d'un match de football), l'audience (par exemple, au cinéma ou dans une salle de classe), le public (par exemple, les journaux ou la télévision), les bandes (proche d'une foule, mais qui serait caractérisée par une interaction beaucoup plus intense), les groupes primaires (intimes), les groupes secondaires (volontaires) et la communauté moderne (groupement lâche issu d'une forte proximité physique et d'un besoin de services spécialisés).
- **health_funding.sav.** Ce fichier de données d'hypothèse concerne des données sur le financement des soins de santé (montant par groupe de 100 individus), les taux de maladie (taux par groupe de 10 000 individus) et les visites chez les prestataires de soins de santé (taux par groupe de 10 000 individus). Chaque observation représente une ville différente.
- **hivassay.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un laboratoire pharmaceutique pour développer une analyse rapide de détection d'infection HIV. L'analyse a pour résultat huit nuances de rouge, les nuances les plus marquées indiquant une plus forte probabilité d'infection. Un test en laboratoire a été effectué sur 2 000 échantillons de sang, la moitié de ces échantillons étant infectée par le virus HIV et l'autre moitié étant saine.
- **hourlywagedata.sav.** Ce fichier de données d'hypothèse concerne les salaires horaires d'infirmières occupant des postes administratifs et dans les services de soins, et affichant divers niveaux d'expérience.

- **insurance_claims.sav.** Il s'agit d'un fichier de données hypothétiques qui concerne une compagnie d'assurance souhaitant développer un modèle pour signaler des réclamations suspectes, potentiellement frauduleuses. Chaque observation correspond à une réclamation distincte.
- **insure.sav.** Ce fichier de données d'hypothèse concerne une compagnie d'assurance qui étudie les facteurs de risque indiquant si un client sera amené à déclarer un incident au cours d'un contrat d'assurance vie d'une durée de 10 ans. Chaque observation figurant dans le fichier de données représente deux contrats, l'un ayant enregistré une réclamation et l'autre non, appariés par âge et sexe.
- **judges.sav.** Ce fichier de données d'hypothèse concerne les scores attribués par des juges expérimentés (plus un juge enthousiaste) à 300 performances de gymnastique. Chaque ligne représente une performance distincte ; les juges ont examiné les mêmes performances.
- **kinship_dat.sav.** Rosenberg et Kim (Rosenberg et Kim, 1975) se sont lancés dans l'analyse de 15 termes de parenté (cousin/cousine, fille, fils, frère, grand-mère, grand-père, mère, neveu, nièce, oncle, père, petite-fille, petit-fils, sœur, tante). Ils ont demandé à quatre groupes d'étudiants (deux groupes de femmes et deux groupes d'hommes) de trier ces termes en fonction des similarités. Deux groupes (un groupe de femmes et un groupe d'hommes) ont été invités à effectuer deux tris, en basant le second sur un autre critère que le premier. Ainsi, un total de six "sources" a été obtenu. Chaque source correspond à une matrice de proximité $15 \times 15$, dont le nombre de cellules est égal au nombre de personnes dans une source moins le nombre de fois où les objets ont été partitionnés dans cette source.
- **kinship_ini.sav.** Ce fichier de données contient une configuration initiale d'une solution tridimensionnelle pour *kinship_dat.sav*.
- **kinship_var.sav.** Ce fichier de données contient les variables indépendantes *sexe*, *génér*(ation) et *degré* (de séparation) permettant d'interpréter les dimensions d'une solution pour *kinship_dat.sav*. Elles permettent en particulier de réduire l'espace de la solution à une combinaison linéaire de ces variables.
- **marketvalues.sav.** Ce fichier de données concerne les ventes de maisons dans un nouvel ensemble à Algonquin (Illinois) au cours des années 1999–2000. Ces ventes relèvent des archives publiques.
- **nhis2000_subset.sav.**Le NHIS (National Health Interview Survey) est une enquête de grande envergure concernant la population des Etats-Unis. Des entretiens ont lieu avec un échantillon de ménages représentatifs de la population américaine. Des informations démographiques et des observations sur l'état de santé et le comportement sanitaire sont recueillies auprès des membres de chaque ménage. Ce fichier de données contient un sous-groupe d'informations issues de l'enquête de 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Fichier de données et documentation d'usage public. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accès en 2003.
- **ozone.sav.** Les données incluent 330 observations portant sur six variables météorologiques pour prévoir la concentration d'ozone à partir des variables restantes. Des chercheurs précédents (Breiman et Friedman, 1985), (Hastie et Tibshirani, 1990), ont décelé parmi ces variables des non-linéarités qui pénalisent les approches standard de la régression.

- **pain_medication.sav.** Ce fichier de données d'hypothèse contient les résultats d'un essai clinique d'un remède anti-inflammatoire traitant les douleurs de l'arthrite chronique. On cherche notamment à déterminer le temps nécessaire au médicament pour agir et les résultats qu'il permet d'obtenir par rapport à un médicament existant.
- **patient_los.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux de patients admis à l'hôpital pour suspicion d'infarctus du myocarde suspecté (ou « attaque cardiaque »). Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **patlos_sample.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux d'un échantillon de patients sous traitement thrombolytique après un infarctus du myocarde. Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **poll_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un enquêteur pour déterminer le niveau de soutien du public pour un projet de loi avant législature. Les observations correspondent à des électeurs enregistrés. Chaque observation enregistre le comté, la ville et le quartier où habite l'électeur.
- **poll_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des électeurs répertoriés dans le fichier *poll_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *poll.csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. Toutefois, ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS – Probability-Proportional-to-Size), il existe également un fichier contenant les probabilités de sélection conjointes (*poll_jointprob.sav*). Les variables supplémentaires correspondant à la répartition démographique des électeurs et à leur opinion sur le projet de loi proposé ont été collectées et ajoutées au fichier de données une fois l'échantillon prélevé.
- **property_assess.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur au niveau du comté pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés vendues dans le comté au cours de l'année précédente. Chaque observation du fichier de données enregistre la ville où se trouve la propriété, l'évaluateur ayant visité la propriété pour la dernière fois, le temps écoulé depuis cette évaluation, l'évaluation effectuée à ce moment-là et la valeur de vente de la propriété.
- **property_assess_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur du gouvernement pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés de l'état. Chaque observation du fichier de données enregistre le comté, la ville et le quartier où se trouve la propriété, le temps écoulé depuis la dernière évaluation et l'évaluation alors effectuée.
- **property_assess_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des propriétés répertoriées dans le fichier *property_assess_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *property_assess.csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. La variable supplémentaire *Valeur courante* a été collectée et ajoutée au fichier de données une fois l'échantillon prélevé.

- **recidivism.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis, ainsi que le temps écoulé jusqu'à la seconde arrestation si elle s'est produite dans les deux années suivant la première.
- **recidivism_cs_sample.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste libéré suite à la première arrestation en juin 2003 et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis et les données relatives à la seconde arrestation, si elle a eu lieu avant fin juin 2006. Les récidivistes ont été choisis dans plusieurs départements échantillonnés conformément au plan d'échantillonnage spécifié dans *recidivism_cs.csplan*. Ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS - Probability proportional to size), il existe également un fichier contenant les probabilités de sélection conjointes (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.**Un fichier de données d'hypothèse qui contient les données de transaction d'achat, y compris la date d'achat, le/les élément(s) acheté(s) et le montant monétaire pour chaque transaction.
- **salesperformance.sav.** Ce fichier de données d'hypothèse concerne l'évaluation de deux nouveaux cours de formation en vente. Soixante employés, divisés en trois groupes, reçoivent chacun une formation standard. En outre, le groupe 2 suit une formation technique et le groupe 3 un didacticiel pratique. A l'issue du cours de formation, chaque employé est testé et sa note enregistrée. Chaque observation du fichier de données représente un stagiaire distinct et enregistre le groupe auquel il a été assigné et la note qu'il a obtenue au test.
- **satisf.sav.** Il s'agit d'un fichier de données d'hypothèse portant sur une enquête de satisfaction effectuée par une société de vente au détail au niveau de quatre magasins. Un total de 582 clients ont été interrogés et chaque observation représente la réponse d'un seul client.
- **screws.sav.**Ce fichier de données contient des informations sur les descriptives des vis, des boulons, des écrous et des clous.(Hartigan, 1975).
- **shampoo_ph.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de produits capillaires. A intervalles réguliers, six lots de sortie distincts sont mesurés et leur pH enregistré. La plage cible est 4,5–5,5.
- **ships.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs (McCullagh et al., 1989) et concernant les dommages causés à des cargos par les vagues. Les effectifs d'incidents peuvent être modélisés comme des incidents se produisant selon un taux de Poisson en fonction du type de navire, de la période de construction et de la période de service. Les mois de service totalisés pour chaque cellule du tableau formé par la classification croisée des facteurs fournissent les valeurs d'exposition au risque.
- **site.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour choisir de nouveaux sites pour le développement de ses activités. L'entreprise a fait appel à deux consultants pour évaluer séparément les sites. Ces consultants, en plus de fournir un rapport approfondi, ont classé chaque site comme constituant une éventualité « bonne », « moyenne » ou « faible ».

- **smokers.sav.** Ce fichier de données est extrait de l'étude National Household Survey of Drug Abuse de 1998 et constitue un échantillon de probabilité des ménages américains. (*http://dx.doi.org/10.3886/ICPSR02934*) Ainsi, la première étape dans l'analyse de ce fichier doit consister à pondérer les données pour refléter les tendances de population.
- **stocks.sav** Ce fichier de données hypothétiques contient le cours et le volume des actions pour un an.
- **stroke_clean.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois celle-ci purgée via des procédures de l'option Validation de données.
- **stroke_invalid.sav.** Ce fichier de données d'hypothèse concerne l'état initial d'une base de données médicales et comporte plusieurs erreurs de saisie de données.
- **stroke_survival.** Ce fichier de données d'hypothèse concerne les temps de survie de patients qui quittent un programme de rééducation à la suite d'un accident ischémique et rencontrent un certain nombre de problèmes. Après l'attaque, l'occurrence d'infarctus du myocarde, d'accidents ischémiques ou hémorragiques est signalée, et le moment de l'événement enregistré. L'échantillon est tronqué à gauche car il n'inclut que les patients ayant survécu durant le programme de rééducation mis en place suite à une attaque.
- **stroke_valid.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois les valeurs vérifiées via la procédure Validation de données. Elle contient encore des observations anormales potentielles.
- **survey_sample.sav.** Ce fichier de données concerne des informations d'enquête dont des données démographiques et des mesures comportementales. Il est basé sur un sous-ensemble de variables de la 1998 NORC General Social Survey, bien que certaines valeurs de données aient été modifiées et que des variables supplémentaires fictives aient été ajoutées à titre de démonstration.
- **telco.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société de télécommunications pour réduire les taux de désabonnement de sa base de clients. Chaque observation correspond à un client distinct et enregistre diverses informations démographiques et d'utilisation de service.
- **telco_extra.sav.** Ce fichier de données est semblable au fichier de données *telco.sav* mais les variables de permanence et de dépenses des consommateurs transformées log ont été supprimées et remplacées par des variables de dépenses des consommateurs transformées log standardisées.
- **telco_missing.sav.** Ce fichier de données est un sous-ensemble du fichier de données *telco.sav* mais certaines des valeurs de données démographiques ont été remplacées par des valeurs manquantes.
- **testmarket.sav.** Ce fichier de données d'hypothèse concerne une chaîne de fast foods et ses plans marketing visant à ajouter un nouveau plat à son menu. Trois campagnes étant possibles pour promouvoir le nouveau produit, le nouveau plat est introduit sur des sites sur plusieurs marchés sélectionnés au hasard. Une promotion différente est effectuée sur chaque site et les ventes hebdomadaires du nouveau plat sont enregistrées pour les quatre premières semaines. Chaque observation correspond à un site-semaine distinct.
- **testmarket_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *testmarket.sav* dans lequel les ventes hebdomadaires sont organisées par site distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, les ventes

enregistrées sont à présent la somme des ventes réalisées au cours des quatre semaines de l'enquête.

- **tree_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_credit.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire.
- **tree_missing_data.sav** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire avec un grand nombre de valeurs manquantes.
- **tree_score_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_textdata.sav.** Ce fichier de données simples ne comporte que deux variables et vise essentiellement à indiquer l'état par défaut des variables avant affectation du niveau de mesure et des étiquettes de valeurs.
- **tv-survey.sav.** Ce fichier de données d'hypothèse concerne une enquête menée par un studio de télévision qui envisage de prolonger la diffusion d'un programme ou de l'arrêter. On a demandé à 906 personnes si elles regarderaient le programme dans diverses situations. Chaque ligne représente un répondant distinct et chaque colonne une situation distincte.
- **ulcer_recurrence.sav.** Ce fichier contient des informations partielles d'une enquête visant à comparer l'efficacité de deux thérapies de prévention de la récurrence des ulcères. Il fournit un bon exemple de données censurées par intervalle et a été présenté et analysé ailleurs (Collett, 2003).
- **ulcer_recurrence_recoded.sav.** Ce fichier réorganise les informations figurant dans le fichier *ulcer_recurrence.sav* pour que vous puissiez modéliser la probabilité d'événement pour chaque intervalle de l'enquête plutôt que la probabilité d'événement de fin d'enquête. Il a été présenté et analysé ailleurs (Collett et al., 2003).
- **verd1985.sav.** Ce fichier de données concerne une enquête (Verdegaal, 1985). Les réponses de 15 sujets à 8 variables ont été enregistrées. Les variables présentant un intérêt sont divisées en trois ensembles. Le groupe 1 comprend l'*âge* et la *situation familiale*, le groupe 2 les *animaux domestiques* et la *presse*, et le groupe 3 la *musique* et l'*habitat*. A la variable *animal domestique* est appliqué un codage nominal multiple et à *âge*, un codage ordinal ; toutes les autres variables ont un codage nominal simple.
- **virus.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fournisseur de services Internet pour déterminer les effets d'un virus sur ses réseaux. Il a suivi le pourcentage (approximatif) de trafic de messages électroniques infectés par un virus sur ses réseaux sur la durée, de la découverte à la circonscription de la menace.
- **wheeze_steubenville.sav.** Il s'agit d'un sous-ensemble d'une enquête longitudinale des effets de la pollution de l'air sur la santé des enfants (Ware, Dockery, Spiro III, Speizer, et Ferris Jr., 1984). Les données contiennent des mesures binaires répétées de l'état asthmatique d'enfants de la ville de Steubenville (Ohio), âgés de 7, 8, 9 et 10 ans, et indiquent si la mère fumait au cours de la première année de l'enquête.
- **workprog.sav.** Ce fichier de données d'hypothèse concerne un programme de l'administration visant à proposer de meilleurs postes aux personnes défavorisées. Un échantillon de participants potentiels au programme a ensuite été prélevé. Certains de ces participants ont

été sélectionnés au hasard pour participer au programme. Chaque observation représente un participant au programme distinct.

- **worldsales.sav** Ce fichier de données hypothétiques contient les revenus des ventes par continent et par produit.

Annexe

# B

# Remarques

Ces informations ont été développées pour les produits et services offerts dans le monde.

Il est possible qu'IBM n'offre pas dans les autres pays les produits, services et fonctionnalités décrits dans ce document. Contactez votre représentant local IBM pour obtenir des informations sur les produits et services actuellement disponibles dans votre région. Toute référence à un produit, programme ou service IBM n'implique pas que les seuls les produits, programmes ou services IBM peuvent être utilisés. Tout produit, programme ou service de fonctionnalité équivalente qui ne viole pas la propriété intellectuelle IBM peut être utilisé à la place. Cependant l'utilisateur doit évaluer et vérifier l'utilisation d'un produit, programme ou service non IBM.

IBM peut posséder des brevets ou des applications de brevet en attente qui couvrent les sujets décrits dans ce document. L'octroi de ce document n'équivaut aucunement à celui d'une licence pour ces brevets. Vous pouvez envoyer par écrit des questions concernant la licence à :

*IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, États-Unis*

Pour obtenir des informations de licence concernant la configuration de caractères codés sur deux octets (DBCS), veuillez contacter dans votre pays le département chargé de la propriété intellectuelle chez IBM ou envoyez vos commentaires par écrit à :

*Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japon.*

**Le paragraphe suivant ne s'applique pas au Royaume-Uni ni à aucun pays dans lequel ces dispositions sont contraires au droit local :** INTERNATIONAL BUSINESS MACHINES FOURNIT CETTE PUBLICATION « EN L'ETAT » SANS GARANTIE D'AUCUNE SORTE, IMPLICITE OU EXPLICITE, Y COMPRIS, MAIS SANS ETRE LIMITE AUX GARANTIES IMPLICITES DE NON VIOLATION, DE QUALITE MARCHANDE OU D'ADAPTATION POUR UN USAGE PARTICULIER. Certains états n'autorisent pas l'exclusion de garanties explicites ou implicites lors de certaines transactions, par conséquent, il est possible que cet énoncé ne vous concerne pas.

Ces informations peuvent contenir des erreurs techniques ou des erreurs typographiques. Ces informations sont modifiées de temps en temps ; ces modifications seront intégrées aux nouvelles versions de la publication. IBM peut apporter des améliorations et/ou modifications des produits et/ou des programmes décrits dans cette publications à tout moment sans avertissement préalable.

Toute référence dans ces informations à des sites Web autres qu'IBM est fournie dans un but pratique uniquement et ne sert en aucun cas de recommandation pour ces sites Web. Le matériel contenu sur ces sites Web ne fait pas partie du matériel de ce produit IBM et l'utilisation de ces sites Web se fait à vos propres risques.

IBM peut utiliser ou distribuer les informations que vous lui fournissez, de la façon dont il le souhaite, sans encourir aucune obligation envers vous.

Les personnes disposant d'une licence pour ce programme et qui souhaitent obtenir des informations sur celui-ci pour activer : (i) l'échange d'informations entre des programmes créés de manière indépendante et d'autres programmes (notamment celui-ci) et (ii) l'utilisation mutuelle des informations qui ont été échangées, doivent contacter :

*IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, États-Unis*.

Ces informations peuvent être disponibles, soumises à des conditions générales, et dans certains cas payantes.

Le programme sous licence décrit dans ce document et toute la documentation sous licence disponible pour ce programme sont fournis par IBM en conformité avec les conditions de l'accord du client IBM, avec l'accord de licence du programme international IBM et avec tout accord équivalent entre nous.

les informations concernant les produits autres qu'IBM ont été obtenues auprès des fabricants de ces produits, leurs annonces publiques ou d'autres sources publiques disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances, leur compatibilité ou toute autre fonctionnalité associée à des produits autres qu'IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits.

Ces informations contiennent des exemples de données et de rapports utilisés au cours d'opérations quotidiennes standard. Pour les illustrer le mieux possible, ces exemples contiennent des noms d'individus, d'entreprises, de marques et de produits. Tous ces noms sont fictifs et toute ressemblance avec des noms et des adresses utilisés par une entreprise réelle ne serait que pure coïncidence.

Si vous consultez la version papier de ces informations, il est possible que certaines photographies et illustrations en couleurs n'apparaissent pas.

### *Marques commerciales*

IBM, le logo IBM, ibm.com et SPSS sont des marques commerciales d'IBM Corporation, déposées dans de nombreuses juridictions du monde entier. Une liste à jour des marques IBM est disponible sur Internet à l'adresse *http://www.ibm.com/legal/copytrade.shtml*.

Adobe, le logo Adobe, PostScript et le logo PostScript sont des marques déposées ou des marques commerciales de Adobe Systems Incorporated aux Etats-Unis et/ou dans d'autres pays.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, et Pentium sont des marques commerciales ou des marques déposées de Intel Corporation ou de ses filiales aux Etats-Unis et dans d'autres pays.

Java et toutes les marques et logos Java sont des marques commerciales de Sun Microsystems, Inc. aux Etats-Unis et/ou dans d'autres pays.

Linux est une marque déposée de Linus Torvalds aux Etats-Unis et/ou dans d'autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques commerciales de Microsoft Corporation aux Etats-Unis et/ou dans d'autres pays.

UNIX est une marque déposée de The Open Group aux Etats-Unis et dans d'autres pays.

Ce produit utilise WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com/*.

Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés.

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

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

# Bibliographie


Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

Blake, C. L., et C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., et J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, .

Cochran, W. G. 1977. *Sampling Techniques*, 3rd éd. New York: John Wiley and Sons.

Collett, D. 2003. *Modelling survival data in medical research*, 2 éd. Boca Raton: Chapman & Hall/CRC.

Cox, D. R., et E. J. Snell. 1989. *The Analysis of Binary Data*, 2nd éd. Londres: Chapman and Hall.

Green, P. E., et V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, P. E., et Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, .

Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

Hastie, T., et R. Tibshirani. 1990. *Generalized additive models*. Londres: Chapman and Hall.

Kennedy, R., C. Riquier, et B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, .

Kish, L. 1965. *Survey Sampling*. New York: John Wiley and Sons.

Kish, L. 1987. *Statistical Design for Research*. New York: John Wiley and Sons.

McCullagh, P., et J. A. Nelder. 1989. *Generalized Linear Models*, 2nd éd. Londres: Chapman & Hall.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. Dans : *Frontiers in Economics,* P. Zarembka, éd. New York: Academic Press.

Murthy, M. N. 1967. *Sampling Theory and Methods*. Calcutta, Inde: Statistical Publishing Society.

Nagelkerke, N. J. D. 1991. A note on the general definition of the coefficient of determination. *Biometrika*, 78:3, .

Price, R. H., et D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

Rickman, R., N. Mitchell, J. Dingman, et J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, .

Rosenberg, S., et M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, .

Särndal, C., B. Swensson, et J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Van der Ham, T., J. J. Meulman, D. C. Van Strien, et H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, .

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (en néerlandais)*. Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, et B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, .

# Index