IBM SPSS Bootstrapping 20

Remarque : Avant d'utiliser ces informations et le produit qu'elles concernent, lisez les informations générales sous Remarques sur p. 43.

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 Bootstrapping fournit les techniques d'analyse supplémentaires décrites dans ce manuel. Le module complémentaire Bootstrapping 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

Annexes

Partie I:
Guide de l'utilisateur

Chapitre
1

Introduction à la méthode des amorces

Lorsque vous collectez des données, vous êtes souvent intéressés à analyser les propriétés de la population parmi laquelle vous avez pris des échantillons. Vous produisez des inférences sur les paramètres de cette population à l'aide d'estimations calculées à partir de l'échantillon. Par exemple, si l'ensemble de données *Employee data.sav* inclus dans le produit est un échantillon aléatoire tiré d'une population d'employés plus large, alors la valeur de la moyenne de l'échantillon du *salaire actuel* de 34 419,57 $ est une estimation du salaire actuel moyen des employés. De plus, cette estimation a une erreur standard de $784 311 pour un échantillon de 474 individus, et un intervalle de confiance de 95% pour le salaire moyen actuel des employés qui est de 32 878,40 $ à 35 960,73 $. Mais à quel point ces estimations sont-elles fiables ? Pour certaines populations « connues » et des paramètres conformes, nous en savons plus sur les propriétés des estimations de l'échantillon et nous pouvons être confiants dans les résultats. La méthode des amorces est destinée à rechercher des informations supplémentaires sur les propriétés des estimateurs pour des populations « inconnues » et des paramètres non conformes.

Figure 1-1
Production d'inférences paramétriques sur la moyenne de la population

			Statistic	Erreur standard
Salaire actuel	Moyenne		$34,419.57	$784.311
	Intervalle de confiance à 95%	Inférieur	$32,878.40	
		Supérieur	$35,960.73	
	Médiane		$28,875.00	

Fonctionnement de l'amorce

Pour un ensemble de données dont la taille est *N*, vous prenez *B* échantillons de « bootstrap » de taille *N* avec remplacement de l'ensemble de données d'origine et calculez l'estimateur de chacun des *B* échantillons de bootstrap. Ces *B* estimations de bootstrap sont un échantillon de taille *B* à partir duquel vous produisez des inférences sur l'estimateur. Par exemple, si vous prenez 1000 échantillons de bootstrap dans l'ensemble de données *Employee data.sav* , vous obtenez une erreur standard de bootstrap estimée de 776,91 $ pour la moyenne de l'échantillon du *salaire actuel*, différente de l'estimation de 784 311 $.

De plus, l'amorce fournit une erreur standard et un intervalle de confiance pour la médiane, pour laquelle les estimations paramétriques ne sont pas disponibles.

Figure 1-2
Production d'inférences par bootstrap sur la moyenne de l'échantillon

			Statistic	Erreur standard	Bootstrap[a]			
					Biais	Erreur standard	Intervalle de confiance à 95%	
							Inférieur	Supérieur
Salaire actuel	Moyenne		$34,419.57	$784.311	$14.66	$776.91	$32,990.38	$36,026.06
	Intervalle de confiance à 95%	Inférieur	$32,878.40					
		Supérieur	$35,960.73					
	Médiane		$28,875.00		$-13.22	$536.63	$27,750.00	$29,850.00

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Prise en charge de l'amorce dans le produit

L'amorce est intégrée en tant que sous-boîte de dialogue dans les procédures la prenant en charge. Reportez-vous à Procédures prenant en charge l'amorce pour obtenir des informations sur les procédures prenant en charge l'amorce.

Lorsque l'amorce est requise dans les boîtes de dialogue, une nouvelle commande distincte `BOOTSTRAP` est collée en plus de la syntaxe usuelle générée par la boîte de dialogue. La commande `BOOTSTRAP` crée les échantillons de bootstrap en fonction de vos spécifications. En interne, le produit traite les échantillons de bootstrap comme des scissions, même si ceux-ci ne sont pas explicitement affichés dans l'éditeur de données. C'est à dire qu'en interne, il existe *B*N* observations, et le compteur d'observations de la barre d'état compte de 1 à *B*N* lors du traitement des données par l'amorce. Le système de gestion des résultats (OMS) est utilisé pour collecter les résultats de l'analyse de chaque « scission de bootstrap ». Ces résultats sont alors regroupés et affichés dans le Viewer, en même temps que le résultat habituel généré par la procédure. Dans certains cas, vous verrez une référence à un "scission 0 de bootstrap", elle correspond à l'ensemble de données d'origine.

Chapitre
2

L'amorce

L'amorce est une méthode consistant à dériver des estimations robustes des erreurs standard et des intervalles de confiance pour des estimations telles que la moyenne, la médiane, le calcul de la proportion, l'odds ratio, le coefficient de corrélation ou de régression. Elle peut aussi être utilisé pour construire des tests d'hypothèse. L'amource est le plus souvent utile comme une alternative aux estimations paramétriques lorsque les hypothèses liées à ces méthodes ne sont pas fiables (comme dans le cas de modèles de regression avec des résidus hétéroscédastiques ajustés à des petits échantillons), ou lorsque l'inférence paramétrique est impossible ou requiert des formules très complexes pour le calcul des erreurs standard (comme dans le cas du calcul d'intervalles de confiance pour la médiane, les quartiles, et autres centiles).

Exemples : Une société en télécommunication perd environ 27% de ses clients chaque mois. Afin de réduire ce taux d'attrition, la direction souhaite savoir si ce taux varie selon les groupes de consommateurs. A l'aide de la méthode de l'amorce, vous pouvez déterminer si un même taux d'attrition décrit de manière appropriée le comportement des quatre types principaux de clients. Pour plus d'informations, reportez-vous à la section Utilisation de l'amorce pour obtenir des intervalles de confiances pour les proportions dans le chapitre 3 dans *IBM SPSS Bootstrapping 20*.

Lors d'une consultation des dossiers des employés, la direction souhaite vérifier leur expérience professionnelle. L'expérience professionnelle est asymétrique, ce qui rend la moyenne moins fiable comme moyen d'estimation de l'expérience antérieure des employés que la médiane. Cependant, les intervalles de confiance paramétriques ne sont pas disponibles pour la médiane dans le produit. Pour plus d'informations, reportez-vous à la section Utilisation de l'amorce pour obtenir des intervalles de confiances pour les médianes dans le chapitre 3 dans *IBM SPSS Bootstrapping 20*.

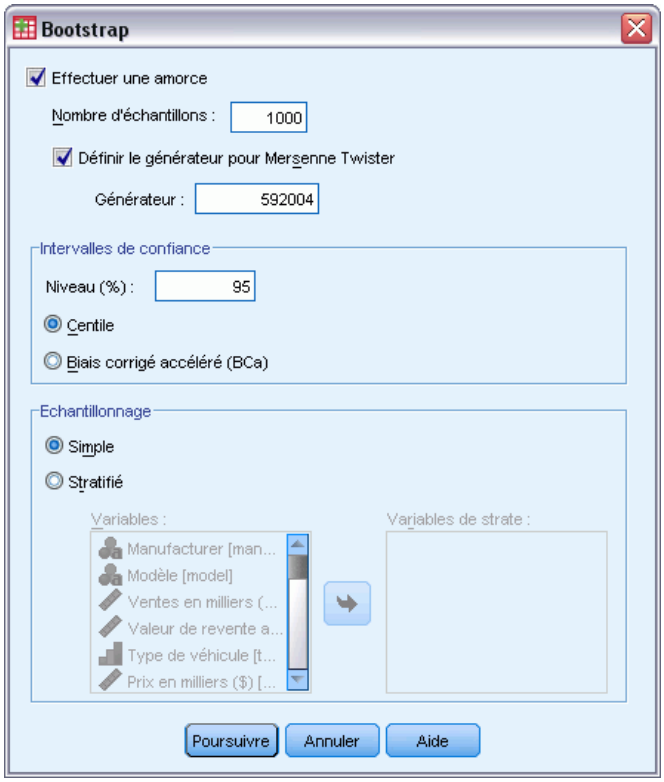
La direction est également intéressée à déterminer les facteurs associés aux augmentations des salaires des employés, en ajustant un modèle linéaire aux différences entre le salaire actuel et le salaire d'embauche. Lorsque la méthode des amorces est appliquée à un modèle linéaire, il est possible d'utiliser des méthodes de ré-échantillonnage (échantillonnage résiduel et wild bootstrap) pour obtenir des résultats plus précis. Pour plus d'informations, reportez-vous à la section Utilisation de l'amorce pour choisir de meilleures valeurs prédites dans le chapitre 3 dans *IBM SPSS Bootstrapping 20*.

De nombreuses procédures prennent en charge l'échantillonnage par bootstrap et le regroupement des résultats d'analyse d'échantillons de bootstrap. Les commandes permettant la spécification des analyses par bootstrap sont intégrées directement comme une sous-boîte de dialogue dans les procédures prenant en charge l'amorce. Les paramètres de la boîte de dialogue du bootstrap sont conservés d'une procédure à l'autre, ainsi si vous exécutez une analyse des effectifs à l'aide de l'amorce dans les boîtes de dialogues, elle sera activée par défaut pour les autres procédure la prenant en charge.

Pour obtenir une analyse par bootstrap

▶ Dans les menus, choisissez une procédure qui prend en charge l'amorce et cliquez sur Bootstrap.

Figure 2-1
Boîte de dialogue Bootstrap



▶ Sélectionnez Effectuer une amorce.

Vous pouvez éventuellement modifier les options suivantes :

Nombre d'échantillons. Pour le centile et les intervalles BCa produits, il est recommandé d'utiliser au moins 1000 échantillons de bootstrap. Spécifiez un nombre entier positif.

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

Intervalles de confiance. Spécifiez un niveau de confiance supérieur à 50 et inférieur à 100. Les intervalles de centile utilisent seulement des valeurs de bootstrap ordonnées correspondant aux centiles d'intervalle de confiance souhaités. Par exemple, un intervalle de confiance de centile de 95 % utilise les 2,5e et 97,5e centiles des valeurs de bootstrap comme bornes inférieure et supérieure de l'intervalle (en interpolant des valeurs de bootstrap si nécessaire). Les intervalles de

biais corrigé et accéléré (BCa) sont des intervalles ajustés plus précis, toutefois ils requièrent plus de temps de calcul.

Echantillonnage. La méthode Simple est le ré-échantillonnage des observations avec remplacement de l'ensemble de données d'origine. La méthode Stratifiée est le ré-échantillonnage des observations avec remplacement de l'ensemble de données d'origine, *au sein* des strates définies par la classification croisée des variables de strate. L'échantillonnage de bootstrap stratifié est utile lorsque les unités au sein des strates sont relativement homogènes, alors qu'elles sont différentes d'une strate à l'autre.

Procédures prenant en charge l'amorce

Les procédures suivants prennent en charge l'amorce.

Remarque :

- L'amorce ne peut pas être utilisée avec des ensembles de données à imputation multiple. Si une variable *Imputation_* est présente dans l'ensemble de données, la boîte de dialogue Bootstrap est désactivée.
- L'amorce utilise l'élimination des observations incomplètes pour déterminer la base de l'observation ; c'est à dire que les observations avec des valeurs manquantes pour n'importe quelle variable de l'analyse sont supprimées de l'analyse. Ainsi lorsque l'amorce est exécutée, l'élimination des observations incomplètes a lieu même si la procédure d'analyse spécifie une autre forme de traitement des valeurs manquantes.

Option Statistiques de base

Effectifs

- Le tableau de statistiques prend en charge les estimations par bootstrap pour la moyenne, l'écart type, la variance, la médiane, l'asymétrie, l'aplatissement et les centiles.
- Le tableau Effectifs prend en charge les estimations par bootstrap pour les pourcentages.

Descriptifs

- Le tableau de statistiques descriptives prend en charge les estimations par bootstrap pour la moyenne, l'écart type, la variance, l'asymétrie et l'aplatissement.

Explorer

- Le tableau Descriptives prend en charge les estimations par bootstrap pour la moyenne, la moyenne tronquée à 5 %, l'écart type, la variance, la médiane, l'asymétrie, l'aplatissement et l'intervalle interquartile.
- Le tableau M-Estimateurs prend en charge les estimations par bootstrap pour le M-Estimateur de Huber, l'estimateur à double pondération de Tukey, le M-estimateur de Hampel, et l'estimateur de Andrew.
- Le tableau Centiles prend en charge les estimations par bootstrap pour les centiles.

Tableaux croisés

- Le tableau Mesures directionnelles prend en charge les estimations par bootstrap pour Lambda, le Tau de Goodman et Kruskal, le coefficient d'incertitude et et le d de Somers.
- Le tableau Mesures symétriques prend en charge les estimations par bootstrap pour Phi, le V de Cramer, le coefficient de contingence, le tau-b de Kendall, le tau-c de Kendall, le Gamma, la corrélation de Spearman, et le R de Pearson.
- Le tableau Estimation du risque prend en charge les estimations par bootstrap pour l'odds ratio.
- Le tableau des odds ratio communs de Mantel-Haenszel prend en charge les estimations par bootstrap et les tests de signification pour In (estimation).

Moyennes

- Le tableau Rapport prend en charge les estimations par bootstrap pour la moyenne, la médiane, la médiane groupée, l'écart type, la variance, l'aplatissement, l'asymétrie, la moyenne harmonique et la moyenne géométrique.

Test T pour échantillon unique

- Le tableau Statistiques prend en charge les estimations par bootstrap pour la moyenne et l'écart type.
- Le tableau Test prend en charge les estimations par bootstrap et les tests de signification pour la différence moyenne.

Test T pour échantillons indépendants

- Le tableau Statistiques de groupe prend en charge les estimations par bootstrap pour la moyenne et l'écart type.
- Le tableau Test prend en charge les estimations par bootstrap et les tests de signification pour la différence moyenne.

Test T pour échantillons appariés

- Le tableau Statistiques prend en charge les estimations par bootstrap pour la moyenne et l'écart type.
- Le tableau Corrélations prend en charge les estimations par bootstrap pour les corrélations.
- Le tableau Test prend en charge les estimations par bootstrap pour la moyenne.

ANOVA à 1 facteur

- Le tableau Descriptive prend en charge les estimations par bootstrap pour la moyenne et l'écart type.
- Le tableau Comparaisons multiples prend en charge les estimations par bootstrap pour la différence moyenne.
- Le tableau Tests de contraste prend en charge les estimations par bootstrap et les tests de signification pour la valeur de contraste.

GLM - Univarié

- Le tableau Descriptive prend en charge les estimations par bootstrap pour la moyenne et l'écart type.

- Le tableau Estimation des paramètres prend en charge les estimations par bootstrap et les tests de signification pour le coefficient, B.
- Le tableau Résultats de contraste prend en charge les estimations par bootstrap et les tests de signification pour la différence.
- Les moyennes marginales estimées : Le tableau Estimations prend en charge les estimations par bootstrap pour la moyenne.
- Les moyennes marginales estimées : Le tableau Comparaisons par paire prend en charge les estimations par bootstrap pour la différence moyenne.
- Les tests post hoc : Le tableau Comparaisons multiples prend en charge les estimations par bootstrap pour la différence moyenne.

Corrélations bivariées

- Le tableau Descriptive prend en charge les estimations par bootstrap pour la moyenne et l'écart type.
- Le tableau Corrélations prend en charge les estimations par bootstrap et les tests de signification pour les corrélations.

Remarques :

Si des corrélations non paramétriques (tau-b de Kendall ou Spearman) sont requises en plus des corrélations de Pearson, la boîte de dialogue colle les commandes `CORRELATIONS` et `NONPAR CORR` avec une commande `BOOTSTRAP` distincte pour chacune d'elles. Les mêmes échantillons de bootstrap seront utilisés pour calculer toutes les corrélations.

Avant le regroupement, la transformation *Z* de Fisher est appliquée aux corrélations. Après le regroupement, la transformation *Z* inverse est appliquée.

Corrélations partielles

- Le tableau Descriptive prend en charge les estimations par bootstrap pour la moyenne et l'écart type.
- Le tableau Corrélations prend en charge les estimations par bootstrap pour les corrélations.

Régression linéaire

- Le tableau Descriptive prend en charge les estimations par bootstrap pour la moyenne et l'écart type.
- Le tableau Corrélations prend en charge les estimations par bootstrap pour les corrélations.
- Le tableau Récapitulatif des modèles prend en charge les estimations par bootstrap pour Durbin-Watson.
- Le tableau Coefficients prend en charge les estimations par bootstrap et les tests de signification pour le coefficient, B.
- Le tableau Coefficients de corrélation prend en charge les estimations par bootstrap pour les corrélations.
- Le tableau Statistiques résiduelles prend en charge les estimations par bootstrap pour la moyenne et l'écart type.

Régression ordinale

- Le tableau Estimation des paramètres prend en charge les estimations par bootstrap et les tests de signification pour le coefficient, B.

Analyse discriminante

- Le tableau Coefficient de fonction de discriminant canonique standardisé prend en charge les estimations par bootstrap des coefficients standardisés.
- Le tableau Coefficient de fonction de discriminant canonique prend en charge les estimations par bootstrap des coefficients non standardisés.
- Le tableau Coefficient de fonction de classification prend en charge les estimations par bootstrap des coefficients.

Option Statistiques avancées

GLM - Multivarié

- Le tableau Estimation des paramètres prend en charge les estimations par bootstrap et les tests de signification pour le coefficient, B.

Modèles mixtes linéaires

- Le tableau Estimations des effets fixes prend en charge les estimations par bootstrap et les tests de signification de l'estimation.
- Le tableau Estimations des paramètres de covariance prend en charge les estimations par bootstrap et les tests de signification de l'estimation.

Modèles linéaires généralisés

- Le tableau Estimation des paramètres prend en charge les estimations par bootstrap et les tests de signification pour le coefficient, B.

Régression de Cox

- Le tableau Variables dans l'équation prend en charge les estimations par bootstrap et les tests de signification pour le coefficient, B.

Option Régression

Régression logistique binaire

- Le tableau Variables dans l'équation prend en charge les estimations par bootstrap et les tests de signification pour le coefficient, B.

Régression logistique multinomiale

- Le tableau Estimation des paramètres prend en charge les estimations par bootstrap et les tests de signification pour le coefficient, B.

Fonctions supplémentaires de la commande BOOTSTRAP

Le langage de syntaxe de commande vous permet aussi de :

- réaliser l'échantillonnage résiduel et par wild bootstrap (sous-commande `SAMPLING`)

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

Partie II: Exemples

Chapitre
3

L'amorce

L'amorce est une méthode consistant à dériver des estimations robustes des erreurs standard et des intervalles de confiance pour des estimations telles que la moyenne, la médiane, le calcul de la proportion, l'odds ratio, le coefficient de corrélation ou de régression. Elle peut aussi être utilisé pour construire des tests d'hypothèse. L'amource est le plus souvent utile comme une alternative aux estimations paramétriques lorsque les hypothèses liées à ces méthodes ne sont pas fiables (comme dans le cas de modèles de regression avec des résidus hétéroscédastiques ajustés à des petits échantillons), ou lorsque l'inférence paramétrique est impossible ou requiert des formules très complexes pour le calcul des erreurs standard (comme dans le cas du calcul d'intervalles de confiance pour la médiane, les quartiles, et autres centiles).

Utilisation de l'amorce pour obtenir des intervalles de confiances pour les proportions

Une société en télécommunication perd environ 27% de ses clients chaque mois. Afin de réduire ce taux d'attrition, la direction souhaite savoir si ce taux varie selon les groupes de consommateurs.

Ces informations sont regroupées dans le fichier *telco.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 32. A l'aide de la méthode de l'amorce, vous pouvez déterminer si un même taux d'attrition décrit de manière appropriée le comportement des quatre types principaux de clients.

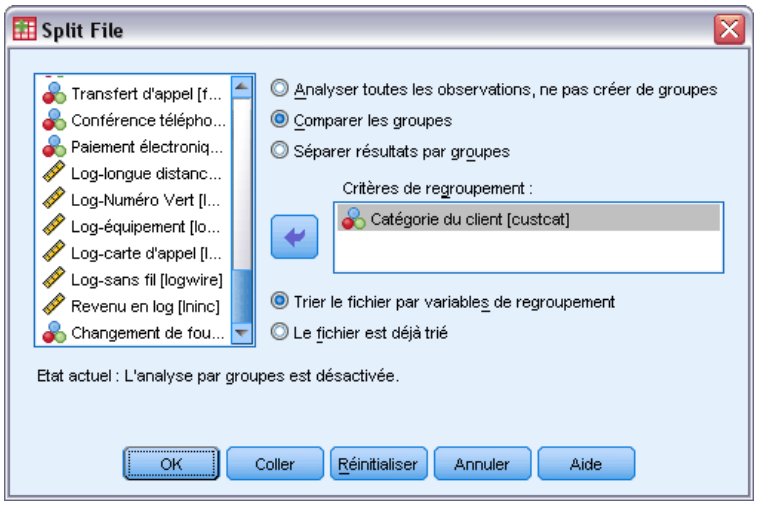
Remarque : Cet exemple utilise la procédure Effectifs et requiert l'option Statistiques de base.

Préparation des données

Vous devez d'abord diviser le fichier en *Catégorie de client*.

- Pour diviser le fichier, dans les menus de l'éditeur de données, choisissez :
  Données > Scinder un fichier

Figure 3-1
Boîte de dialogue Scinder un fichier

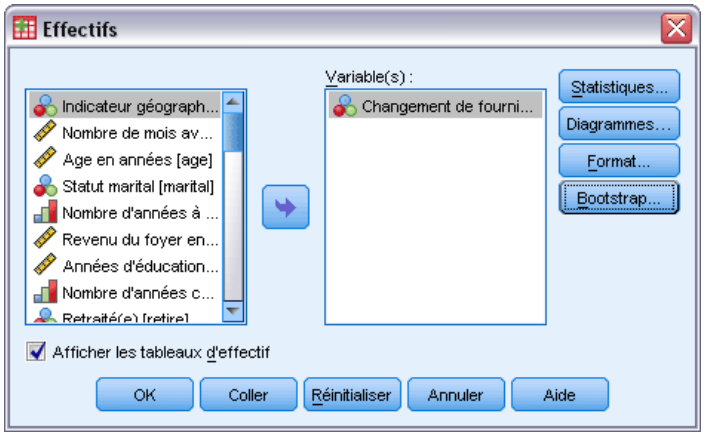


- Sélectionnez Comparer les groupes.
- Sélectionnez *Catégorie de client* comme variable sur laquelle les groupes sont basés.
- Cliquez sur OK.

Exécution de l'analyse

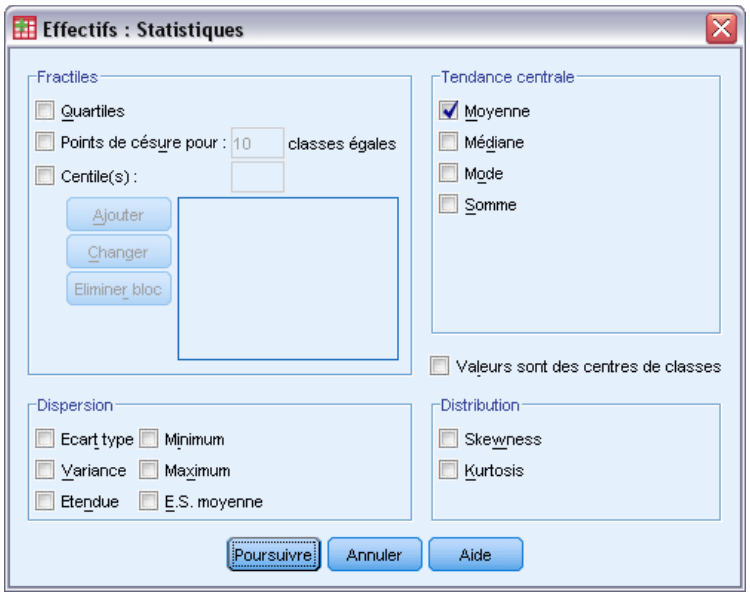
- Pour obtenir des intervalles de confiance bootstrap pour les proportions, choisissez les options suivantes dans les menus :
  Analyse > Statistiques descriptives > Effectifs...

Figure 3-2
Boîte de dialogue Fréquences



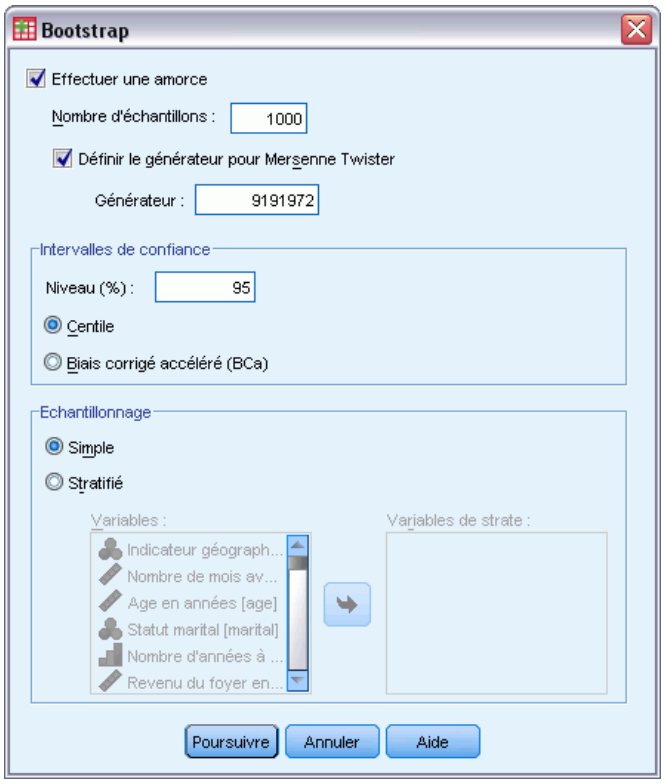
- Sélectionnez *Désabonné au cours du mois dernier [churn]* comme variable dans l'analyse.
- Cliquez sur Statistiques.

Figure 3-3
Boîte de dialogue Statistiques



- Sélectionnez l'option Moyenne dans le groupe Tendance centrale.
- Cliquez sur Poursuivre.
- Cliquez sur Bootstrap dans la boîte de dialogue Fréquences.

Figure 3-4
Boîte de dialogue Bootstrap



- Sélectionnez Effectuer une amorce.
- Afin de reproduire exactement les résultats de cet exemple, sélectionnez Définir le générateur pour Mersenne Twister et saisissez 9191972 comme valeur du générateur.
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Fréquences.

Ces sélections génèrent la syntaxe de commande suivante :

```
SORT CASES  BY custcat.
SPLIT FILE LAYERED BY custcat.
PRESERVE.
SET RNG=MT MTINDEX=9191972.
SHOW RNG.
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES INPUT=churn
  /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
  /MISSING USERMISSING=EXCLUDE.
FREQUENCIES VARIABLES=churn
  /STATISTICS=MEAN
  /ORDER=ANALYSIS.
RESTORE.
```

- Les commandes `SORT CASES` et `SPLIT FILE` divisent le fichier en fonction de la variable *custcat*.

- Les commandes `PRESERVE` et `RESTORE` "mémorisent" l'état actuel du générateur de nombres aléatoires et restaure le système dans cet état, une fois l'amorce terminée.
- La commande `SET` définit le générateur de nombres aléatoires sur le générateur Mersenne Twister et l'index sur 9191972, afin que les résultats de l'amorce puissent être reproduits de manière exacte. La commande `SHOW` affiche l'index dans les résultats pour référence.
- La commande `BOOTSTRAP` requiert 1 000 échantillons de bootstrap pour le ré-échantillonnage simple.
- La variable *churn (désabonnement)* est utilisée pour déterminer la base des observations pour le rééchantillonnage. Les observations contenant des valeurs manquantes sur cette variable sont supprimées de l'analyse.
- La procédure `FREQUENCIES` suivant `BOOTSTRAP` est exécutée sur chacun des échantillons de bootstrap.
- La sous-commande `STATISTICS` produit la moyenne de la variable *churn* des données d'origine. En outre, des statistiques groupées sont produites pour la moyenne et les pourcentages dans le tableau des effectifs.

Spécifications de bootstrap

Figure 3-5
Spécifications de bootstrap

Méthode d'échantillonnage	Simple
Nombre d'échantillons	1000
Niveau d'intervalle de confiance	95.0%
Type d'intervalle de confiance	Centile

Le tableau de spécifications de bootstrap contient les paramètres utilisés lors du rééchantillonnage, et il est une référence utile pour vérifier si l'analyse que vous souhaitiez réaliser a été effectuée.

Statistiques

Figure 3-6
Tableau de statistiques avec intervalle de confiance de bootstrap pour les proportions

Changement de fournisseur internet lors du dernier mois

Catégorie du client			Statistic	Bootstrap[a]			
				Biais	Erreur standard	Intervalle de confiance à 95%	
						Inférieur	Supérieur
Service basic	N	Valide	266	0	0	266	266
		Manquante	0	0	0	0	0
	Moyenne		.31	.00	.03	.26	.37
Service électronique	N	Valide	217	0	0	217	217
		Manquante	0	0	0	0	0
	Moyenne		.27	.00	.03	.21	.34
Service plus	N	Valide	281	0	0	281	281
		Manquante	0	0	0	0	0
	Moyenne		.16	.00	.02	.12	.20
Service Total	N	Valide	236	0	0	236	236
		Manquante	0	0	0	0	0
	Moyenne		.37	.00	.03	.31	.44

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Le tableau des statistiques montre, pour chaque niveau de *Catégorie de client*, la valeur moyenne de la variable *Désabonné au cours du mois dernier*. La variable *Désabonné au cours du mois dernier* ne peut prendre que les valeurs 0 et 1, la valeur 1 signifiant que le client s'est désabonné et la moyenne étant égale à la proportion de désabonnements. La colonne Statistique affiche les valeurs généralement produites par la procédure Effectifs à l'aide de l'ensemble de données d'origine. Les colonnes Bootstrap sont produites par des algorithmes d'amorce.

- Le Biais est la différence entre la valeur moyenne des échantillons de bootstrap et la valeur de la colonne Statistique. Dans ce cas, la valeur moyenne de *Désabonné au cours du mois dernier* est calculée pour les 1000 échantillons de bootstrap, et la moyenne de ces moyennes est alors calculée.
- Ecart- standard représente l'erreur standard de la valeur moyenne de la variable *Désabonné au cours du mois dernier* sur les 1000 échantillons de bootstrap.
- La limite inférieure de l'intervalle de confiance de bootstrap à 95% est une interpolation des 25e et 26e valeurs moyennes de la variable *Désabonné au cours du mois dernier*, si les 1000 échantillons sont classés dans l'ordre croissant. La limite supérieure est une interpolation des 975e et 976e valeurs moyennes.

Les résultats du tableau suggèrent que le taux d'attrition est différent selon les types de clients. En particulier, l'intervalle de confiance des clients *Service Plus* n'en recouvre aucun autre, ce qui suggère que ces clients sont, en moyenne, moins susceptibles de partir.

Lorsque vous utilisez des variables qualitatives à deux valeurs uniquement, les intervalles de confiance sont différents de ceux produits par la procédure Tests non paramétriques à un échantillon ou Test T pour échantillon unique.

Tableau des effectifs :

Figure 3-7
Tableau des effectifs avec intervalle de confiance de bootstrap pour les proportions

Catégorie du client			Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé	Bootstrap pour Pourcentage[a]			
							Biais	Erreur standard	Intervalle de confiance à 95%	
									Inférieur	Supérieur
Service basic	Valide	Non	183	68.8	68.8	68.8	.0	2.8	63.2	74.4
		Oui	83	31.2	31.2	100.0	.0	2.8	25.6	36.8
		Total	266	100.0	100.0		.0	.0	100.0	100.0
Service électronique	Valide	Non	158	72.8	72.8	72.8	.1	3.1	66.4	78.8
		Oui	59	27.2	27.2	100.0	-.1	3.1	21.2	33.6
		Total	217	100.0	100.0		.0	.0	100.0	100.0
Service plus	Valide	Non	237	84.3	84.3	84.3	.0	2.1	80.1	88.3
		Oui	44	15.7	15.7	100.0	.0	2.1	11.7	19.9
		Total	281	100.0	100.0		.0	.0	100.0	100.0
Service Total	Valide	Non	148	62.7	62.7	62.7	.0	3.2	56.4	69.1
		Oui	88	37.3	37.3	100.0	.0	3.2	30.9	43.6
		Total	236	100.0	100.0		.0	.0	100.0	100.0

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Le tableau Effectifs affiche les intervalles de confiance pour les pourcentages (proportion × 100%) pour chaque catégorie, ils sont donc disponibles pour toutes les variables qualitatives. Des intervalles de confiance comparables ne sont pas disponibles ailleurs dans le produit.

Utilisation de l'amorce pour obtenir des intervalles de confiances pour les médianes

Lors d'une consultation des dossiers des employés, la direction souhaite vérifier leur expérience professionnelle. L'expérience professionnelle est asymétrique, ce qui rend la moyenne moins fiable comme moyen d'estimation de l'expérience antérieure des employés que la médiane. Toutefois, sans l'amorce, les intervalles de confiance pour la médiane ne sont généralement pas disponibles dans les procédures statistiques du produit.

Ces informations sont regroupées dans le fichier *Employee data.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 32. Utilisation de l'amorce pour obtenir des intervalles de confiances pour la médiane.

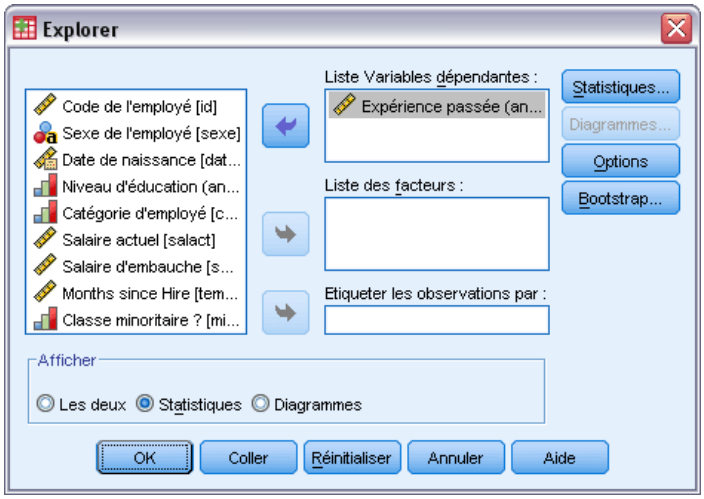
Remarque : cet exemple utilise la procédure Explorer et requiert l'option Statistiques de base.

Exécution de l'analyse

- Pour obtenir des intervalles de confiance de bootstrap pour la médiane, choisissez les options suivantes dans les menus :

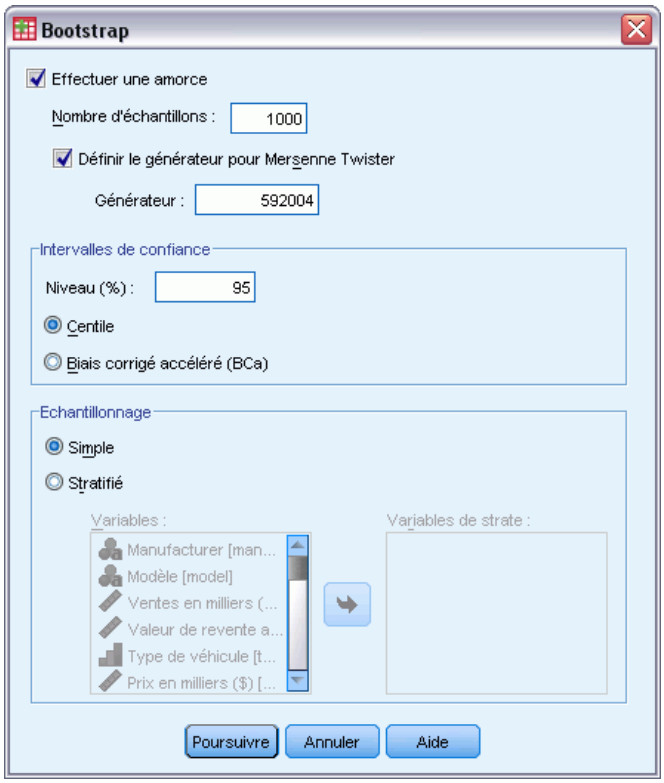
  Analyse > Statistiques descriptives > Explorer

Figure 3-8
Boîte de dialogue principale Explorer



- Sélectionnez *Expérience préalable (mois) [prevexp]* comme variable dépendante.
- Sélectionnez l'option Statistiques dans le groupe Afficher.
- Cliquez sur Bootstrap.

Figure 3-9
Boîte de dialogue Bootstrap



- Sélectionnez Effectuer une amorce.
- Afin de reproduire exactement les résultats de cet exemple, sélectionnez Définir le générateur pour Mersenne Twister et saisissez 592004 comme valeur du générateur.
- Pour obtenir des intervalles plus précis (au prix d'un temps de traitement plus important), sélectionnez Biais corrigé accéléré (BCa).
- Cliquez sur Poursuivre.
- Cliquez sur OK dans la boîte de dialogue Explorer.

Ces sélections génèrent la syntaxe de commande suivante :

```
PRESERVE.
SET RNG=MT MTINDEX=592004.
SHOW RNG.
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES TARGET=prevexp
  /CRITERIA CILEVEL=95 CITYPE=BCA  NSAMPLES=1000
  /MISSING USERMISSING=EXCLUDE.
EXAMINE VARIABLES=prevexp
  /PLOT NONE
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

```
RESTORE.
```

- Les commandes PRESERVE et RESTORE "mémorisent" l'état actuel du générateur de nombres aléatoires et restaure le système dans cet état, une fois l'amorce terminée.
- La commande SET définit le générateur de nombres aléatoires sur le générateur Mersenne Twister et l'index sur 592004, afin que les résultats de l'amorce puissent être reproduits de manière exacte. La commande SHOW affiche l'index dans les résultats pour référence.
- La commande BOOTSTRAP requiert 1000 échantillons de bootstrap pour le ré-échantillonnage simple.
- La sous-commande VARIABLES spécifie que la variable *prevexp* est utilisée pour déterminer la base des observations pour le rééchantillonnage. Les observations contenant des valeurs manquantes sur cette variable sont supprimées de l'analyse.
- La sous-commande CRITERIA , en plus de requérir le nombre d'échantillons de bootstrap, requiert des intervalles de confiance de bootstrap de biais corrigé et accéléré à la place des intervalles de centiles par défaut.
- La procédure EXAMINE suivant BOOTSTRAP est exécutée sur chacun des échantillons de bootstrap.
- La sous-commande PLOT désactive les résultats graphiques.
- Toutes les autres options sont définies à leur valeur par défaut.

Descriptives

Figure 3-10
Tableau Descriptives avec intervalles de confiance de bootstrap

					Bootstrap[a]			
							Intervalle de confiance à 95%	
			Statistique	Erreur standard	Biais	Erreur standard	Inférieur	Supérieur
Expérience passée (années)	Moyenne		95.86	4.804	-.01	4.86	86.29	105.41
	Intervalle de confiance à 95% pour la moyenne	Borne inférieure	86.42					
		Borne supérieure	105.30					
	Moyenne tronquée à 5%		84.64		.02	4.94	75.25	94.59
	Médiane		55.00		-.11	3.66	48.00	64.00
	Variance		10938.281		18.783	977.081	8981.729	12908.885
	Ecart-type		104.586		-.015	4.689	94.772	113.617
	Minimum		0					
	Maximum		476					
	Intervalle		476					
	Intervalle interquartile		121		-1	10	101	142
	Asymétrie		1.510	.112	.006	.110	1.290	1.751
	Aplatissement		1.696	.224	.040	.463	.881	2.774

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Le tableau Descriptives contient un nombre de statistiques et des intervalles de confiance de bootstrap pour ces statistiques. L'intervalle de confiance de bootstrap pour la moyenne (86,39 ; 105,20) est similaire à l'intervalle de confiance paramétrique (86,42 ; 105,30) et suggère que l'employé type a environ de 7 à 9 ans d'expérience préalable. Cependant, *Expérience préalable (mois)* possède une distribution asymétrique, ce qui fait de la moyenne un indicateur moins fiable du salaire "type" actuel que la médiane. L'intervalle de confiance de bootstrap pour la médiane (50,00 ; 60,00) est plus restreint et inférieur à l'intervalle de confiance pour la moyenne, et suggère

que l'employé type a environ de 4 à 5 ans d'expérience préalable. L'utilisation de l'amorce a permis d'obtenir une plage de valeurs qui représente mieux l'expérience préalable type.

Utilisation de l'amorce pour choisir de meilleures valeurs prédites

Lors d'une consultation des dossiers des employés, la direction est intéressée à déterminer les facteurs associés aux augmentations des salaires des employés, en ajustant un modèle linéaire aux différences entre le salaire actuel et le salaire d'embauche. Lorsque la méthode des amorces est appliquée à un modèle linéaire, il est possible d'utiliser des méthodes de ré-échantillonnage (échantillonnage résiduel et wild bootstrap) pour obtenir des résultats plus précis.

Ces informations sont regroupées dans le fichier *Employee data.sav*. Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A sur p. 32.

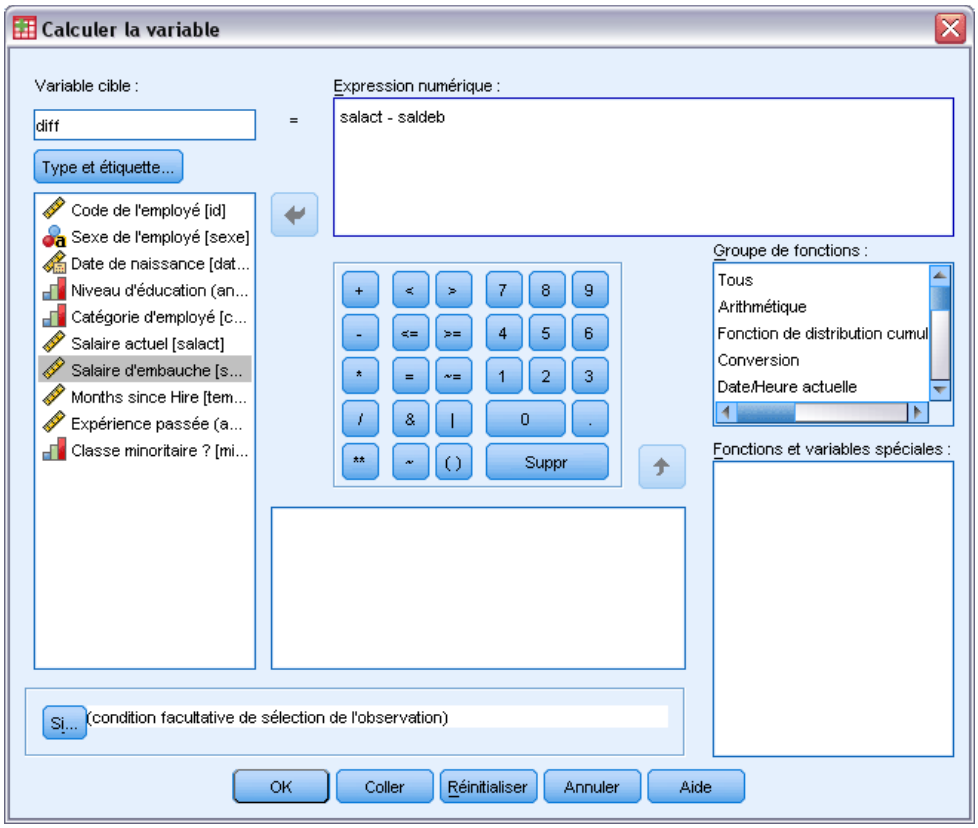
Remarque : cet exemple utilise la procédure GLM Univarié et requiert l'option Statistiques de base.

Préparation des données

Vous devez d'abord calculer la différence entre le salaire actuel et le salaire de départ.

- A partir des menus, sélectionnez :
  Transformer > Calculer la variable...

Figure 3-11
Boîte de dialogue Calculer la variable



- Saisissez diff comme variable cible.
- Saisissez salary-salbegin comme expression numérique.
- Cliquez sur OK.

Exécution de l'analyse

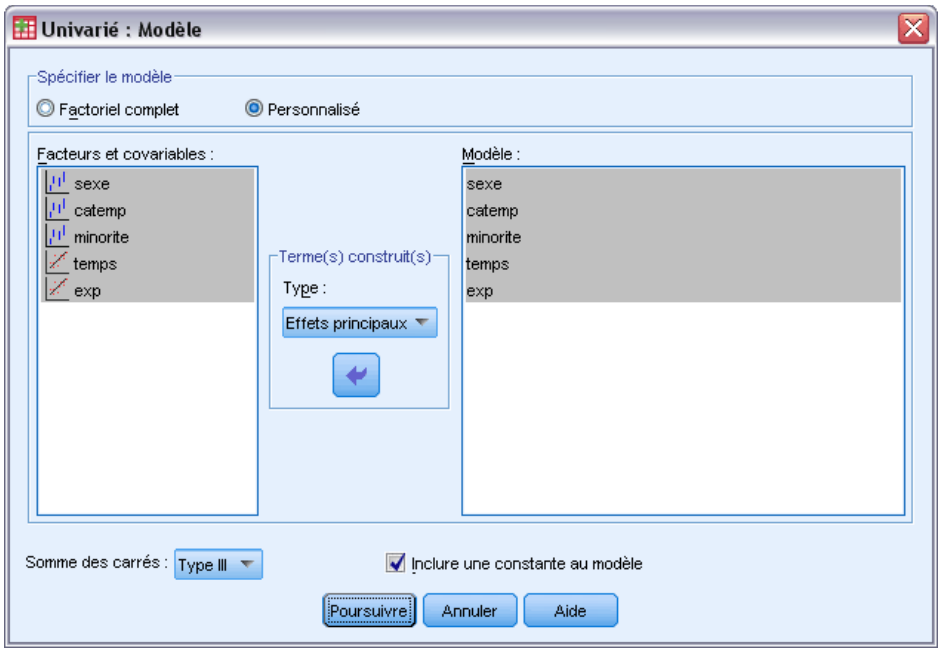
Pour exécuter la procédure GLM Univarié avec une amorce résiduelle sauvage, vous devez d'abord créer des résidus.

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

Figure 3-12
Boîte de dialogue principale GLM Univarié

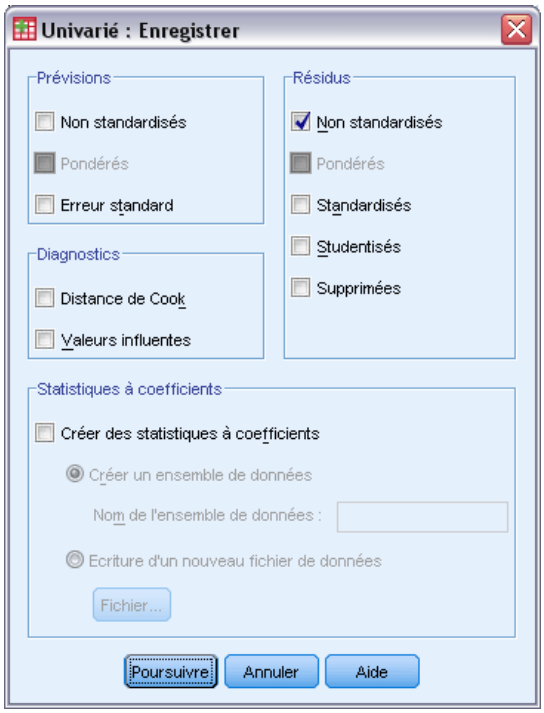
- Sélectionnez *diff* comme variable dépendante.
- Sélectionnez *Sexe [gender]*, *Catégorie d'emploi [jobcat]* et *Classification des minorités [minority]* comme facteurs fixes.
- Sélectionnez *Ancienneté [jobtime]* et *Expérience préalable (mois) [prevexp]* comme covariables.
- Cliquez sur Modèle.

Figure 3-13
Boîte de dialogue Modèle



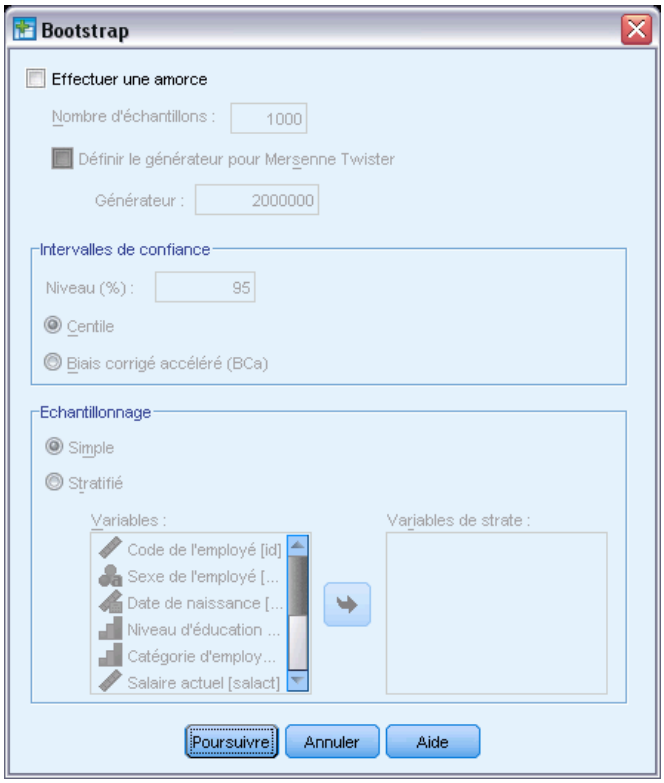
- Sélectionnez Personnalisé puis Effets principaux dans la liste déroulante Termes construits.
- Sélectionnez les variables de *gender* à *prevexp* comme termes de modèle.
- Cliquez sur Poursuivre.
- Cliquez sur le bouton Enregistrer dans la boîte de dialogue GLM Univarié.

Figure 3-14
Boîte de dialogue Enregistrer



- Sélectionnez l'option Non standardisés dans le groupe Résidus.
- Cliquez sur Poursuivre.
- Cliquez sur Bootstrap dans la boîte de dialogue GLM Univarié.

Figure 3-15
Boîte de dialogue Bootstrap

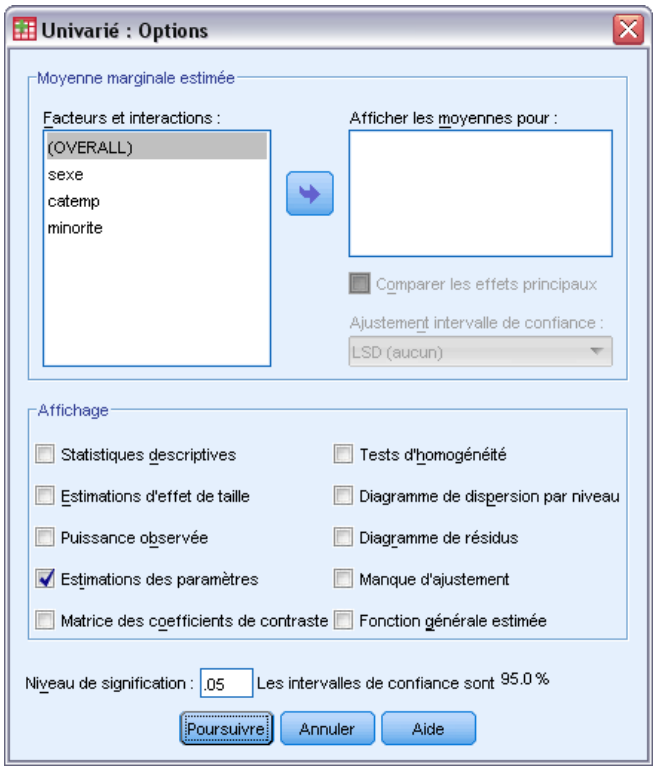


Les paramètres de bootstrap sont conservés dans les boîtes de dialogue qui prennent en charge les amorces. L'enregistrement de nouvelles variables dans l'ensemble de données n'est pas pris en charge lorsque l'amorce est active, de sorte que vous devez vérifier qu'elle est désactivée.

- Si nécessaire, désélectionnez l'option Effectuer une amorce.
- Cliquez sur le bouton OK dans la boîte de dialogue GLM Univarié. L'ensemble de données contient une nouvelle variable *RES_1*, qui comprend les résidus non-standardisés de ce modèle.
- Dans la boîte de dialogue GLM Univarié, cliquez sur Enregistrer.

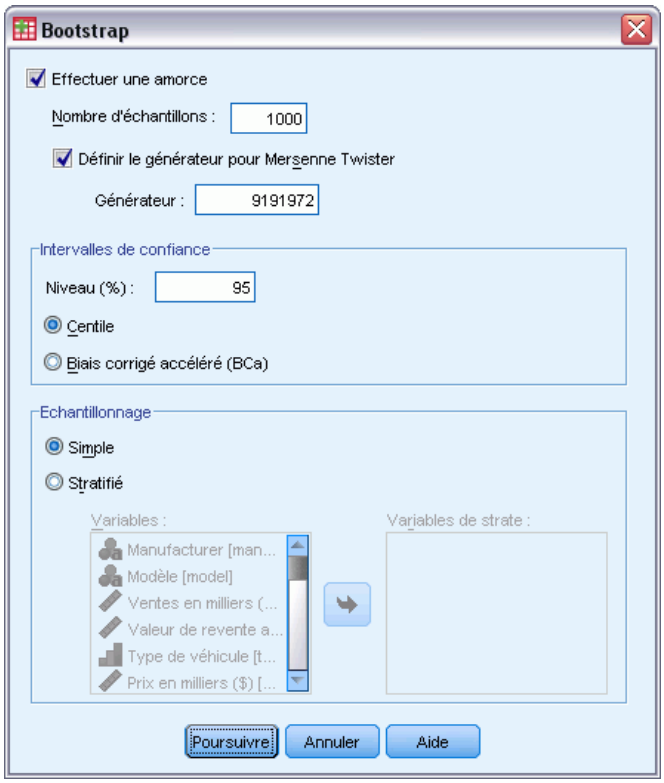
- Désélectionnez l'option Non standardisé, puis cliquez sur Poursuivre et sur Options dans la boîte de dialogue GLM Univarié.

Figure 3-16
Boîte de dialogue Options



- Sélectionnez l'option Estimations des paramètres dans le groupe Afficher.
- Cliquez sur Poursuivre.
- Cliquez sur Bootstrap dans la boîte de dialogue GLM Univarié.

Figure 3-17
Boîte de dialogue Bootstrap



- Sélectionnez Effectuer une amorce.
- Afin de reproduire exactement les résultats de cet exemple, sélectionnez Définir le générateur pour Mersenne Twister et saisissez 9191972 comme valeur du générateur.
- Il n'existe pas d'option pour effectuer une amorce sauvage à partir des boîtes de dialogue, vous devez donc cliquer sur Poursuivre, puis sur Coller dans la boîte de dialogue GLM Univarié.

Ces sélections génèrent la syntaxe de commande suivante :

```
PRESERVE.
SET RNG=MT MTINDEX=9191972.
SHOW RNG.
BOOTSTRAP
 /SAMPLING METHOD=SIMPLE
 /VARIABLES TARGET=diff INPUT=gender jobcat minority jobtime prevexp
 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
 /MISSING USERMISSING=EXCLUDE.
UNIANOVA diff BY gender jobcat minority WITH jobtime prevexp
 /METHOD=SSTYPE(3)
 /INTERCEPT=INCLUDE
 /PRINT=PARAMETER
 /CRITERIA=ALPHA(.05)
 /DESIGN=gender jobcat minority jobtime prevexp.
```

```
RESTORE.
```

Afin de réaliser l'échantillonnage de wild bootstrap, éditez le mot-clé `METHOD` de la sous-commande `SAMPLING` de la façon suivante : `METHOD=WILD(RESIDUALS=RES_1)`.

Le groupe de syntaxe de commande " final " apparaît comme suit :

```
PRESERVE.
SET RNG=MT MTINDEX=9191972.
SHOW RNG.
BOOTSTRAP
 /SAMPLING METHOD=WILD(RESIDUALS=RES_1)
 /VARIABLES TARGET=diff INPUT=gender jobcat minority jobtime prevexp
 /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
 /MISSING USERMISSING=EXCLUDE.
UNIANOVA diff BY gender jobcat minority WITH jobtime prevexp
 /METHOD=SSTYPE(3)
 /INTERCEPT=INCLUDE
 /PRINT=PARAMETER
 /CRITERIA=ALPHA(.05)
 /DESIGN=gender jobcat minority jobtime prevexp.
RESTORE.
```

- Les commandes `PRESERVE` et `RESTORE` "mémorisent" l'état actuel du générateur de nombres aléatoires et restaure le système dans cet état, une fois l'amorce terminée.
- La commande `SET` définit le générateur de nombres aléatoires sur le générateur Mersenne Twister et l'index sur 9191972, afin que les résultats de l'amorce puissent être reproduits de manière exacte. La commande `SHOW` affiche l'index dans les résultats pour référence.
- La commande `BOOTSTRAP` requiert 1000 échantillons de bootstrap pour l'échantillonnage sauvage et *RES_1* comme la variable contenant les résidus.
- La sous-commande `VARIABLES` spécifie que *diff* est la variable cible dans le modèle linéaire. Cette variable cible et les variables *gender*, *jobcat*, *minority*, *jobtime*, et *prevexp* sont utilisées pour déterminer la base des observations pour le rééchantillonnage. Les observations contenant des valeurs manquantes sur ces variables sont supprimées de l'analyse.
- La sous-commande `CRITERIA` , en plus de requérir le nombre d'échantillons de bootstrap, requiert des intervalles de confiance de bootstrap de biais corrigé et accéléré à la place des intervalles de centiles par défaut.
- La procédure `UNIANOVA` suivant `BOOTSTRAP` est exécutée sur chacun des échantillons de bootstrap et produit des estimations de paramètre pour les données d'origines. En outre, des statistiques groupées sont produites pour les coefficients du modèle.

Estimations des paramètres

Figure 3-18
Estimations des paramètres

Variable dépendante:diff

Paramètre	A	Erreur standard	t	Sig.	Intervalle de confiance à 95%	
					Borne inférieure	Limite supérieure
Constante	22789.014	2920.700	7.803	.000	17049.673	28528.355
[sexe=f]	-4085.253	726.416	-5.624	.000	-5512.701	-2657.804
[sexe=m]	0[a]	.	.	.	.	.
[catemp=1]	-17717.706	939.798	-18.853	.000	-19564.463	-15870.949
[catemp=2]	-13101.918	1780.683	-7.358	.000	-16601.061	-9602.776
[catemp=3]	0[a]	.	.	.	.	.
[minorite=0]	1332.363	819.349	1.626	.105	-277.705	2942.431
[minorite=1]	0[a]	.	.	.	.	.
temps	145.539	32.586	4.466	.000	81.505	209.572
exp	-21.423	3.575	-5.993	.000	-28.447	-14.398

a.Ce paramètre est mis à zéro car il est redondant.

Le tableau Estimations des paramètres affiche les estimations des paramètres, habituelles et non-amorcées pour les termes du modèle. La valeur de signification de 0,105 pour *[minority=0]* est supérieure à 0,05, et suggère que la *Classification des minorités* n'a aucun effet sur l'augmentation du salaire.

Figure 3-19
Estimations des paramètres de bootstrap

Variable dépendante:diff

Paramètre	A	Bootstrap[a]				
		Biais	Erreur standard	Sig. (bilatéral)	Intervalle de confiance à 95%	
					Inférieur	Supérieur
Constante	22789.014	-95.084	3280.762	.001	16079.630	28835.063
[sexe=f]	-4085.253	32.480	622.971	.001	-5365.321	-2892.131
[sexe=m]	0	0	0		0	0
[catemp=1]	-17717.706	46.324	1454.230	.001	-20671.451	-14889.507
[catemp=2]	-13101.918	47.958	1753.311	.001	-16658.596	-9671.891
[catemp=3]	0	0	0		0	0
[minorite=0]	1332.363	-10.592	651.144	.012	57.831	2642.534
[minorite=1]	0	0	0		0	0
temps	145.539	.707	35.285	.001	79.081	217.761
exp	-21.423	-.065	2.859	.001	-27.533	-16.055

a.Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Consultez maintenant le bootstrap pour le tableau Estimations des paramètres. Dans la colonne Erreur standard, vous pouvez voir que les erreurs standard paramétriques de certains coefficients, comme la constante, sont trop faibles comparés aux estimations de bootstrap et que les intervalles de confiance sont plus larges. Pour certains coefficients, comme *[minority=0]*, les erreurs standard paramétriques sont trop grandes et la valeur de signification de 0.006 rapportée dans les résultats de bootstrap, inférieure à 0,05, montre que la différence observée dans les augmentations de salaire entre les employés qui sont classés en tant que minorités et ceux qui ne le sont pas n'est

pas due au hasard. La direction est désormais au courant de cette différence et peut pousser son investigation plus loin pour en déterminer les causes.

Lectures recommandées

Reportez-vous aux documents suivants pour plus d'informations sur l'amorce :

Davison, A. C., et D. V. Hinkley. 2006. *Bootstrap Methods and their Application*. : Cambridge University Press.

Shao, J., et D. Tu. 1995. *The Jackknife and Bootstrap*. New York: Springer.

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 15×15, 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, .

Collett, D. 2003. *Modelling survival data in medical research*, 2 éd. Boca Raton: Chapman & Hall/CRC.

Davison, A. C., et D. V. Hinkley. 2006. *Bootstrap Methods and their Application*. : Cambridge University Press.

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, .

McCullagh, P., et J. A. Nelder. 1989. *Generalized Linear Models*, 2nd éd. Londres: Chapman & Hall.

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, .

Shao, J., et D. Tu. 1995. *The Jackknife and Bootstrap*. New York: Springer.

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