
# StatBox 7

## Manuel d'utilisation

### Version 7.5




AgroSolutions
83, avenue de la Grande Armées
75782 Paris cedex 16
France

# SOMMAIRE

# INTRODUCTION

## Plus souple, plus simple à utiliser

La sélection des données est souple, elle peut se faire à la fois de manière automatique (le logiciel détermine sur la feuille active les variables disponibles) soit par sélection manuelle (l'utilisateur sélectionne les plages de données à analyser par sélection à la souris).

L'ergonomie des menus et des boîtes de dialogues a été entièrement revue pour être plus facile à utiliser. StatBox s'intègre désormais totalement dans Excel sous la forme d'un menu ou ruban (Excel 2007 ou supérieur) qui s'ajoute à la suite des menus d'Excel. StatBox pilote entièrement Excel comme application hôte. Cette nouvelle architecture rend StatBox plus stable et plus rapide.

Le système de protection a été profondément modifié de sorte qu'il n'est plus nécessaire d'utiliser une clé de protection physique.

## Des fonctionnalités plus nombreuses

Voici quelques-unes des nouvelles fonctions disponibles :

- calcul vectoriel,
- choix de l'orientation des tests pour la plupart des tests statistiques.

En agriculture :

- duplication de dispositif,
- outils de contrôle de la qualité de plan,
- résultats supplémentaires en analyse de variance (contrôle de la proportionnalité des résidus, résidus par répétition, graphique des moyennes),
- possibilité d'analyser plusieurs variables simultanément en alpha plan.

Enfin de nombreuses options ont été ajoutées pour la personnalisation du logiciel.

# CONFIGURATION MINIMALE REQUISE

- ✓ Microsoft® Windows XP ou supérieur.
- ✓ Microsoft® Excel 2007, Excel 2010 32 bits ou Excel 2013 32 bits, de préférence avec les services pack installés.
- ✓ Un pilote d'imprimante installé.

Un certain nombre de prérequis peuvent être nécessaires à l'installation et au bon fonctionnement de StatBox. Le programme d'installation détecte automatiquement la présence des prérequis sur votre poste et tente de les installer si nécessaire.

Remarque : il peut être nécessaire de disposer d'une connexion active à Internet pour télécharger les prérequis manquants.

## Lancement

Pour lancer StatBox, lancez simplement Excel. StatBox étant un Addin d'Excel, il est chargé automatiquement par Excel lors de son démarrage.

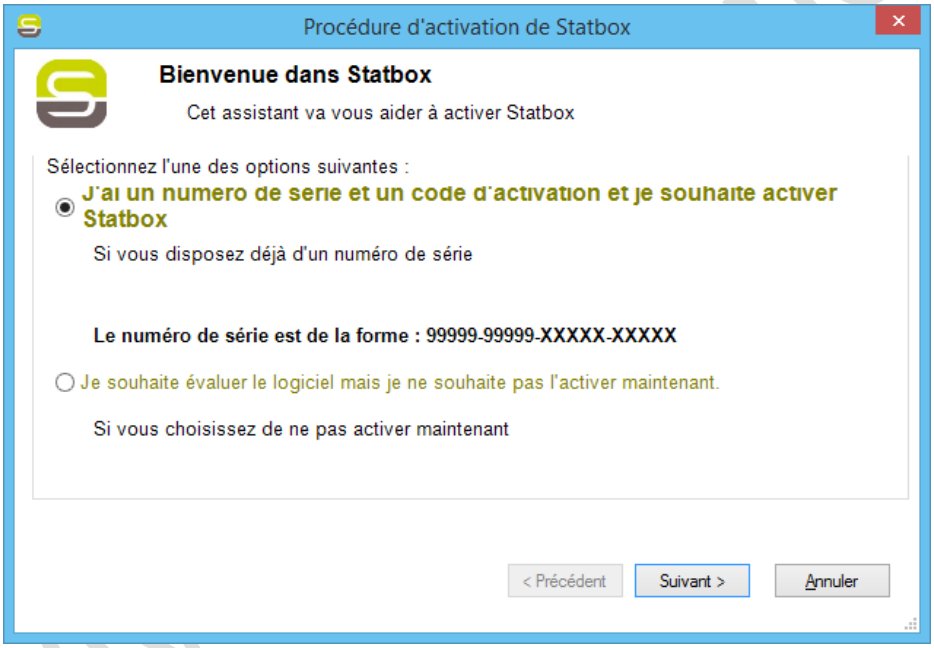
Remarque : il est possible que vous constatiez un léger ralentissement d'Excel au démarrage suite à l'installation de StatBox. Ce temps de chargement supplémentaire est lié à la vérification des règles de sécurité des Addin des applications Microsoft Office, il est incompressible.

## Protection du logiciel

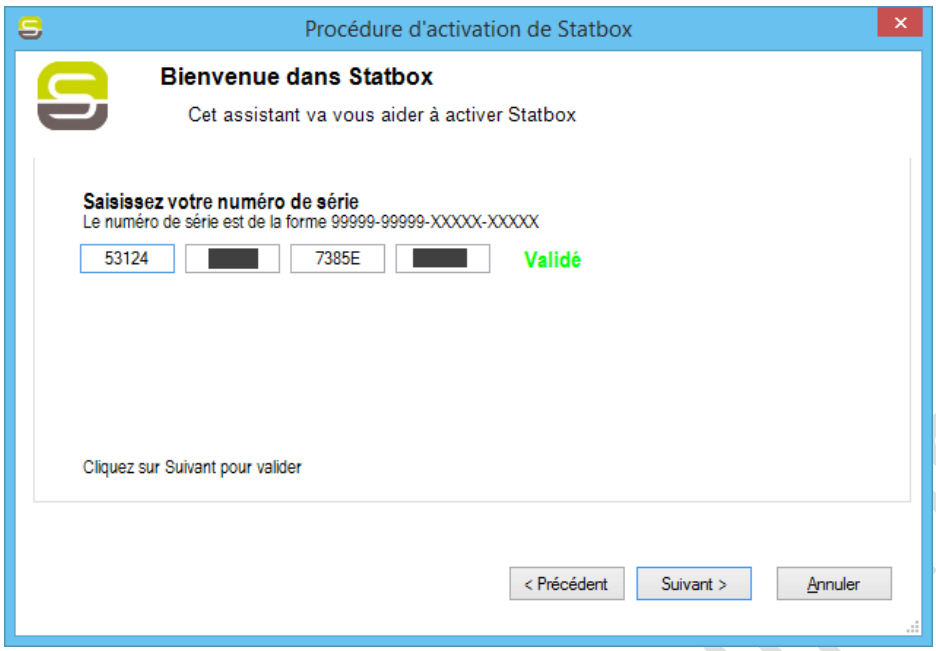
A la première utilisation de l'application, et ce tant que le logiciel ne sera pas activé, le message suivant apparait vous demandant d'activer votre version du logiciel.

Pour activer votre version du logiciel :

- ✓ Cochez l'option « J'ai un numéro de série… » et validez en cliquant sur « Suivant »



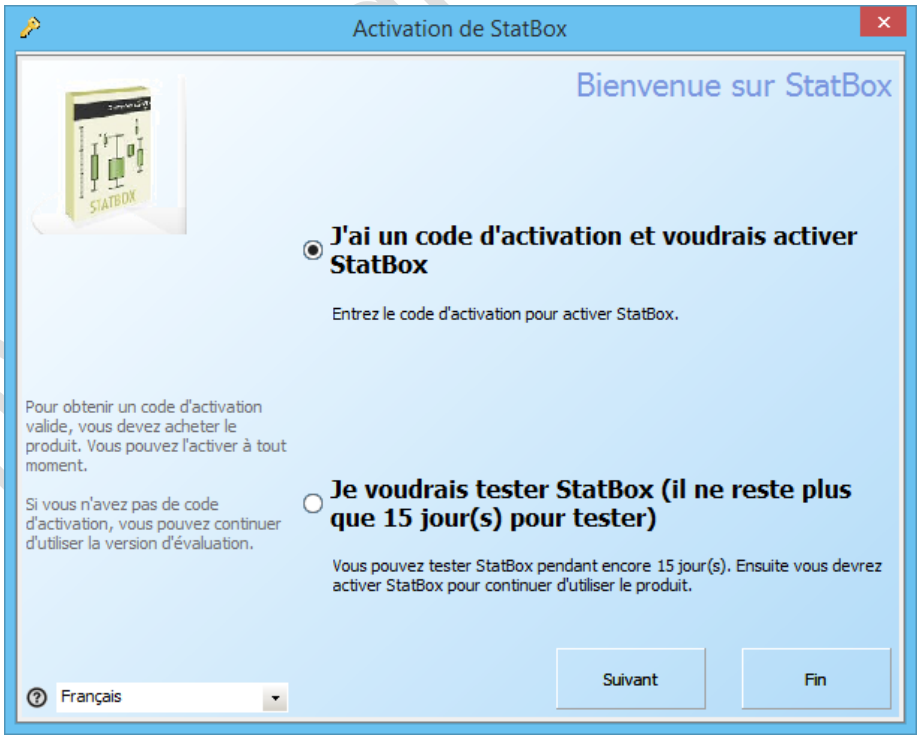

- ✓ Entrez votre numéro de série sous la forme 99999-99999-XXXXX-XXXXX dans la zone de saisie. Un message vert doit apparaitre vous signifiant que le code est valide. Validez en cliquant sur « Suivant ».



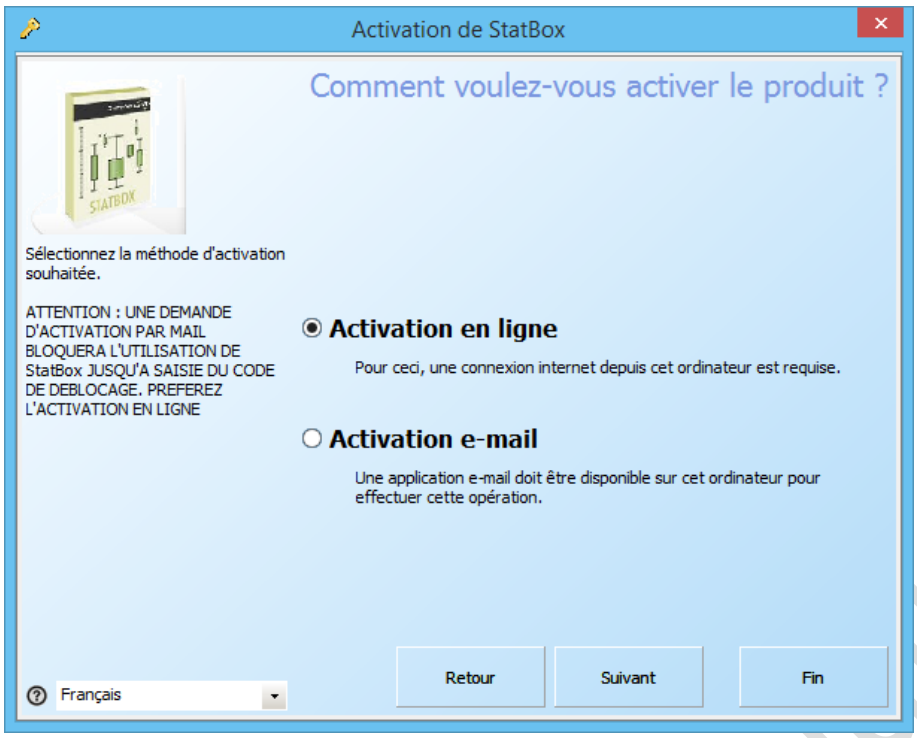

✓ Activez la version du logiciel. Pour cela vous disposez de plusieurs modes d'activation :

- connexion au server Web d'activation (pour cela vous devez disposer d'une connexion active à Internet)
- par email

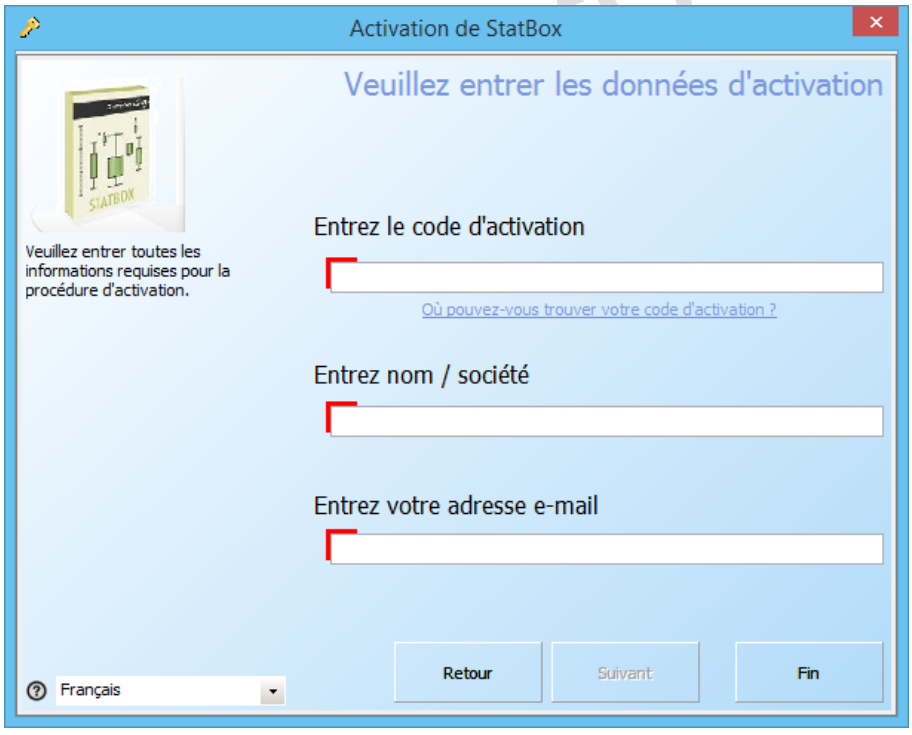
Sur l'écran principal d'activation, sélectionnez l'option « J'ai un code d'activation et je voudrais activer StatBox », puis cliquez sur « Suivant ».



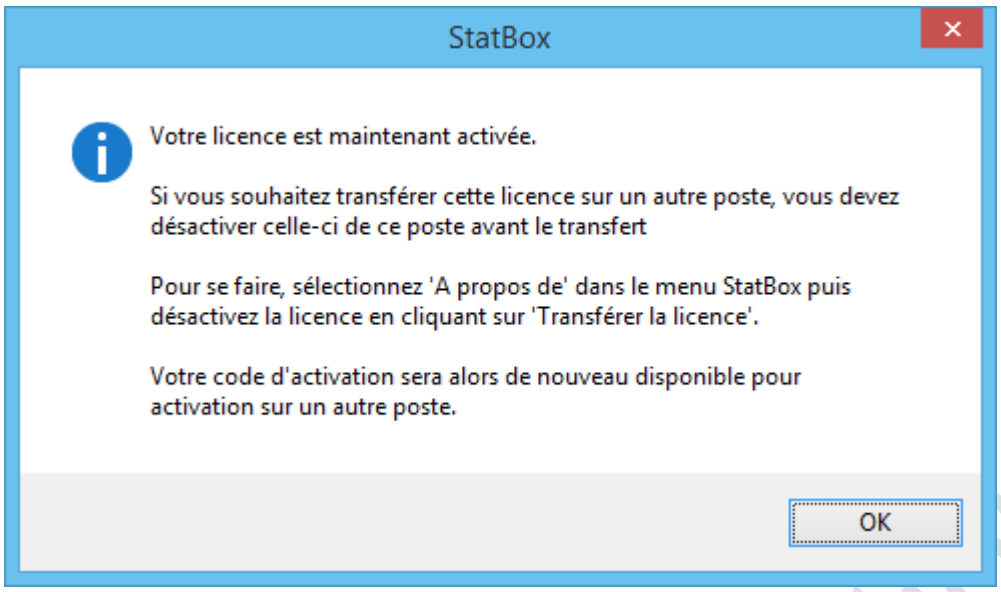

Sélectionnez le mode d'activation souhaité. ATTENTION, il est recommandé d'activer le logiciel en ligne, une demande d'activation par mail impose un délai de quelques jours.




Entrez votre code d'activation sous la forme XXXXX-XXXXX-XXXXX-XXXXX-XXXXX-XXXXX dans la zone prévu à cette effet. Entrez un nom de société et un Email valide. Validez en cliquant sur « Suivant ». Le logiciel tente alors de se connecter au serveur d'activation en vous demandant l'autorisation de se connecter.




Puis vous informe du succès ou non de l'opération en fournissant au besoin un message explicatif.




## Activation et transfert de StatBox sur un autre ordinateur

Le processus d'activation permet de vérifier l'authenticité de votre logiciel. Ce processus permet également de vérifier que StatBox n'a pas été installé sur plus d'ordinateurs que prévu dans votre licence.

## Transfert de votre licence

Le transfert de licence consiste à transférer votre licence d'un ordinateur à un autre.

Avant de transférer votre licence, vous devez au préalable désactiver celle-ci sur l'ordinateur sur laquelle elle est installée.

Le processus de désactivation peut être lancé depuis la fenêtre d'A propos de StatBox (StatBox > A propos)


Un message vous indique que la licence est bien désactivée. Vous pouvez alors lancer l'activation sur un autre ordinateur

Pour tout problème relatif à l'activation du logiciel, n'hésitez pas à contacter notre service assistance.

Si vous souhaitez utiliser le logiciel en mode évaluation, vous disposez d'une période de 15 jours en sélectionnant l'option « Je veux évaluer le logiciel » de l'écran d'accueil. Un message apparait alors vous signifiant le nombre de jour restant de la période d'évaluation. Au terme de cette période d'évaluation, l'utilisation du logiciel sera bloquée dans l'attente d'activation.

## Problème d'imprimante

L'affichage des graphiques sous Excel ne s'effectue pas correctement lorsque Excel ne peut pas imprimer, ce qui est le cas notamment si :

- aucun pilote d'imprimante n'est installé,
- l'imprimante sélectionnée n'est pas connectée ou fonctionne mal.

Essayez d'imprimer depuis Excel afin d'obtenir un diagnostic plus précis.

Installez un pilote d'imprimante sur votre machine en cliquant sur l'icône **Poste de travail** ou en allant dans **Paramètres, Imprimantes**.

## Paramètres régionaux

Deux paramètres régionaux sont essentiels pour StatBox : le séparateur décimal et le séparateur de liste. Pour accéder à ces paramètres :

- Sur Windows XP : Panneau de configuration > Paramètres régionaux > Nombre
- Sur Windows 7 et 8 : Panneau de configuration > Horloge, langue et région > Modifier les formats de date, d'heure ou de nombre

Vous pouvez en outre modifier le séparateur décimal directement sous Excel : allez dans **Outils**, **Options**, **International**, **Gestion des nombres**, décochez l'option « Utilisez les paramètres système », et modifiez le contenu du champ « Séparateur de décimale ».

StatBox fonctionne correctement avec n'importe quel séparateur décimal d'un caractère, y compris lorsque celui-ci est modifié au cours d'une session de travail.

## Données d'exemple

Certains jeux de données proposés dans le fichier data.xls (situé dans le répertoire du dossier d'installation de StatBox) sont issus d'ouvrages cités en références, ce qui permet de :

- comparer les résultats obtenus avec StatBox et les résultats figurant dans les ouvrages cités,
- bénéficier des interprétations développées dans le texte des ouvrages cités.

Toutefois, il est possible de constater de légères différences entre les résultats produits par StatBox et ceux fournis dans les ouvrages dont sont issues les données. Ces écarts proviennent :

- du faible nombre de décimales des valeurs figurant dans les tableaux de données alors que les calculs ont été effectués avec des valeurs comportant davantage de décimales,
- des paramètres d'arrêt des itérations différents dans le cas des méthodes itératives,
- des choix différents dans l'implémentation des algorithmes.

Ces écarts ne sont généralement pas suffisants pour modifier profondément les interprétations des analyses effectuées.

# ORGANISATION DES MENUS

Selon l'édition que vous installez, 1 ou 2 menu / ruban s'ajouteront à la barre de menu Excel : 1 menu commun pour les statistiques standards et 1 menu distinct pour l'édition Agri ou Vegetal.

L'organisation des menus a été repensée dans cette version pour un accès plus intuitif aux différentes fonctions du logiciel. L'apparence des menus est différente selon la version d'Excel que vous utilisez, la version pour Office 2007 et 2010 de StatBox intègre notamment la nouvelle interface par « rubans » de cette version d'Office. Dans cette version, l'incorporation d'icônes sur les menus et le positionnement en premier niveau des fonctions les plus importantes du logiciel facilitent également l'utilisation.

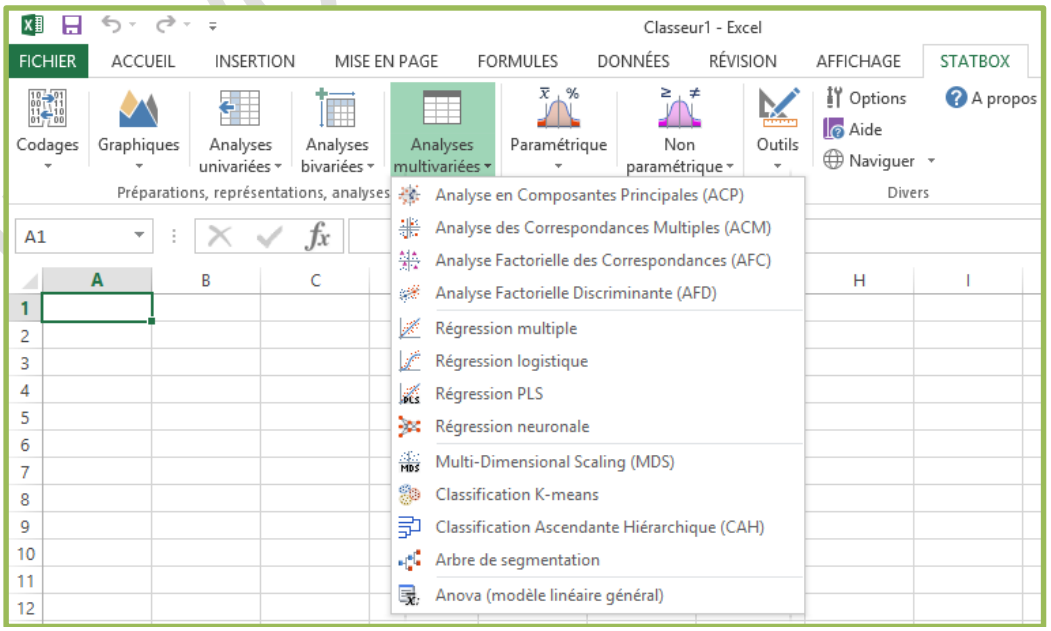
## Partie Standard (menu StatBox)

Les fonctions sont désormais regroupées en 7 grands thèmes représentant le type de rapport qu'il est possible de produire. On retrouve ainsi :

- les codages
- les représentations graphiques
- les analyses univariées
- les analyses bivariées
- les analyses multivariées
- les tests paramétriques
- les tests non paramétriques

Des sous-menus supplémentaires permettent l'accès aux « Outils », « Options » et fonctions de support du logiciel.

**Remarque :** certains rapports sont accessibles à plusieurs endroits des menus car ils s'appliquent à différents thèmes. Par exemple, le type de rapport « Statistiques descriptives » se retrouve à la fois dans le menu « Représentations graphiques » et dans le menu « Analyses univariées ».



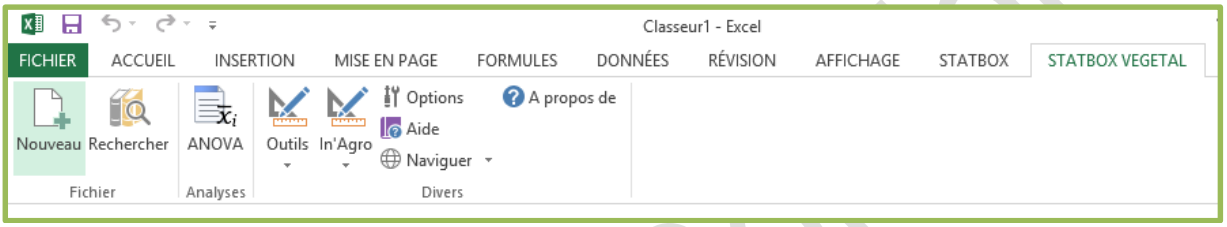
### Edition Agri et Vegetal

L'organisation du menu de l'édition Agri et Vegetal a également été entièrement revue. La création de tous les nouveaux types de plans est désormais accessible à partir du menu « Nouveau ». Les fonctions sont ensuite regroupées selon le type de feuilles auxquelles elles s'appliquent. On retrouve ainsi 4 sous menus principaux :

- Dispositif
- Plan
- Saisie
- Regroupements

Les analyses sont toujours accessibles à partir d'un menu spécifique pour un accès rapide.

Des sous-menus supplémentaires permettent l'accès aux « Options » et fonctions de support du logiciel.




## GESTION DES DONNÉES

### Lecture des données dans la feuille

Avant de choisir dans le menu de StatBox une analyse statistique, assurez-vous que vous êtes positionné sur la feuille contenant les données à analyser.

StatBox propose deux modes de sélection des données : la sélection automatique (le logiciel détermine automatiquement les variables disponibles et leur type) et la sélection manuelle (l'utilisateur sélectionne à la souris les plages contenant les données).

En sélection automatique, plusieurs règles doivent être vérifiées :

- la feuille Excel doit comporter uniquement vos données sous la forme d'un tableau rectangulaire. Rien d'autre ne doit se trouver dans la feuille. N'ajoutez pas en bas du tableau, par exemple, des calculs complémentaires ou des commentaires. StatBox considèrera ces informations comme appartenant au tableau de données.
- StatBox lit les premières lignes pour identifier la nature des données : texte ou numérique. Si vous mélangez le type de données, StatBox ne pourra fonctionner correctement.

Selon les méthodes, tous les types de variables ne seront pas disponibles. En sélection manuelle, il appartiendra à l'utilisateur de contrôler que la sélection de données contient le type de données attendu pour la méthode en cours.

Chaque variable sera identifiée dans les boites de dialogue par son nom précédé d'un code indiquant son type :

- T pour les variables comportant du texte
- N pour les variables numériques
- S pour les variables codées

Remarques :

- Éviter, sur la ligne des libellés, que deux variables aient le même nom. Seule la première sera prise en compte.
- Pour faire une sélection multiple, appuyez sur la touche Ctrl ou la touche Majuscule (Shift).
- Si vous désirez changer de jeu de données, vous pouvez changer de feuille ou effectuer des modifications sur la feuille en cours et recharger les données en cliquant sur « Réinitialiser la boite de dialogue ». Vous perdrez cependant tout le paramétrage effectué.

## Types de données

StatBox contrôle la nature des valeurs des données en fonction de la structure algébrique de la variable attendue :

- quantitative (numérique, continue)
- qualitative

Les variables quantitatives ne peuvent pas comporter de texte. Les variables qualitatives peuvent comporter des valeurs numériques ou du texte, toutes les valeurs étant traitées indifféremment par StatBox sous la forme de chaînes de caractères.

Remarque : lorsque vous sélectionnez une variable nominale comportant des codes numériques, veuillez à ce que le nombre de valeurs différentes soit limité. Exemple : La variable à expliquer dans une analyse factorielle discriminante doit être nominale. Si vous introduisez une variable numérique comme un chiffre d'affaire, le programme ne pourra pas fonctionner correctement. Il s'attend à trouver un nombre limité de valeurs différentes, 2, 3, 4, alors que pour ce chiffre d'affaire, on peut avoir autant de valeurs différentes que d'observations dans le tableau de données.

La valeur d'une cellule d'apparence vide - c'est-à-dire réellement vide ou contenant un ou plusieurs caractères « espace » - ainsi que les valeurs d'erreur retournées par Excel notamment :

- #NOMBRE !
- #DIV/0 !
- #VALEUR !
- #REF !
- #NOM ?

sont interprétées par StatBox comme des valeurs manquantes. Certains traitements de StatBox conduisent éventuellement à des valeurs manquantes, notamment dans le cas d'une transformation effectuée sur des valeurs pour lesquelles la fonction utilisée n'est pas définie (ex. le logarithme d'une valeur négative). La présence de valeurs manquantes n'est généralement pas bloquante pour les modules de StatBox, sauf lorsque le moteur de calcul détecte que la quantité d'information n'est pas suffisante pour effectuer les calculs.

Remarques :

- 0 n'est jamais considéré comme la valeur codant une valeur manquante dans les données. Dans ce cas, faites une recherche/remplacer et substituez le 0 par un vide.
- un poids manquant est assimilé par défaut à un poids nul.

## Libellés des variables codées

### *Principe*

Les variables codées sont des variables nominales ou ordinales représentées par des codes 1, 2, 3,... À chaque modalité 1, 2, 3,… correspond un libellé, ainsi le sexe de la personne est codé 1 pour

« homme » et 2 pour « femme ». Dans la feuille Excel, on trouve les codes 1 ou 2. Bien que composée de chiffres, cette variable n'est pas numérique.

Le libellé des variables peut comporter jusqu'à trois zones :

1) Intitulé ou nom de la variable

Ex :

Q1, Q2, CA1990 situé dans la première ligne de la feuille Excel

Pour toutes les variables : numériques, textes ou codées, il est nécessaire d'avoir un nom.

2) Pour ces différentes variables il est possible d'ajouter en plus, un libellé long

Ex. :

- Chiffre d'affaires de l'année 1990
- Quels produits achetez-vous régulièrement ?
    - Âge de la personne...

3) Pour les variables codées, il est utile d'avoir le libellé des différentes modalités de réponse,

Ex. :

- 1 pour « Homme »
- 2 pour « Femme »

Ou

- 1 pour « Très satisfait »
- 2 pour « Plutôt satisfait »
- 3 pour « Plutôt pas satisfait »
- 4 pour « Pas du tout satisfait »

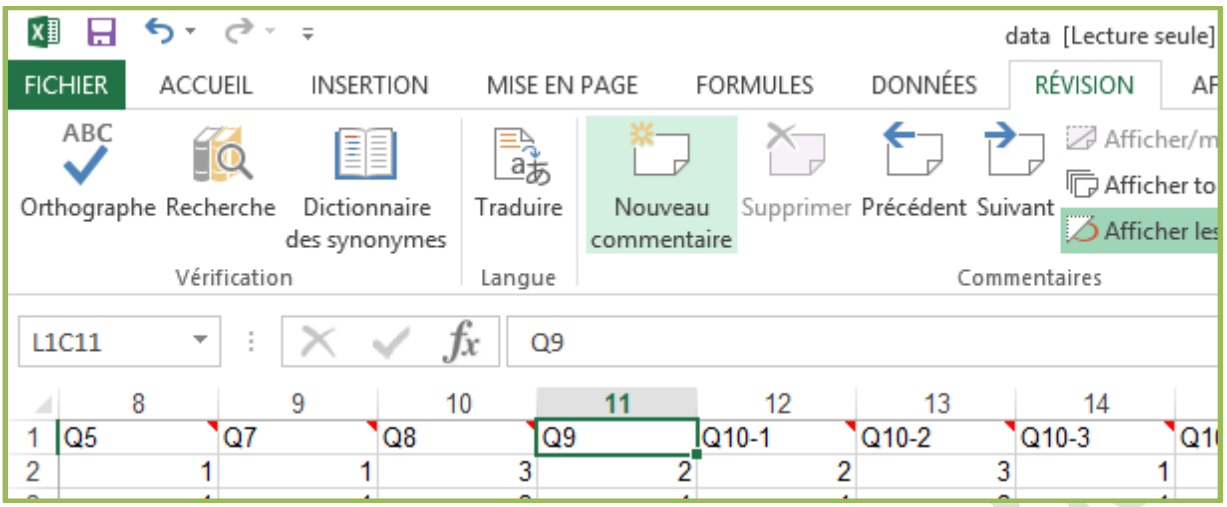
Les libellés longs et les libellés des modalités de réponses sont placés dans la zone commentaire de la cellule comportant le nom de la variable dans la feuille Excel. La zone commentaire est accessible à partir du menu Insertion d'Excel en sélectionnant « Commentaire ».

**Remarque :** Pour les variables numériques, il n'y a pas de libellé de réponse, les valeurs saisies correspondent aux réponses.

### *Saisie des libellés*

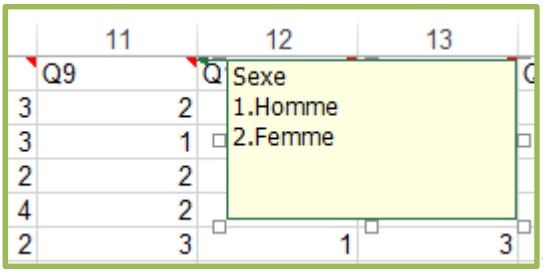
Directement dans la zone Commentaire d'une cellule d'une feuille de données

Placez-vous sur la première ligne de votre feuille de calcul sur le nom de la variable. Sélectionnez l'onglet Révision puis Nouveau commentaire.




Tapez d'abord le libellé long de la variable. Validez ensuite par la touche Entrée.

Sur la 2ème ligne, tapez 1 suivi d'un point ".", puis le libellé de la première modalité et validez. Renouvelez l'opération en incrémentant le numéro de modalité jusqu'à ce que vous ayez saisi tous les libellés. Exemple :




Insérez obligatoirement le numéro de la modalité puis un point devant le libellé.

### *Import d'un fichier texte*

Si vous avez un grand nombre de libellés ou que vous utilisez les mêmes libellés pour plusieurs fichiers de données, nous vous recommandons de les saisir dans un fichier texte. L'objectif est de récupérer des libellés du fichier texte et de les placer automatiquement dans la zone commentaire de la feuille de données.

Dans la feuille de données, vous devez avoir saisi sur la première ligne les noms des variables, identiques à ceux du fichier des libellés. Le fichier de libellés doit respecter la structure suivante :

```
[Q0] où prendriez-vous conseil ?
Auprès de vos relations personnelles
Les salons ou séminaires
La presse
Auprès des entreprises qui en ont eu l'expérience
Auprès des organismes professionnels ou d'un expert comp
Auprès de votre fournisseur habituel
Auprès de spécialistes internes

[Q1] Profession
agriculteur
artisan commerçant
cadre prof. supérieures
Profession intermédiaire
employé
ouvrier
retraité inactif
```

[Q2] âge de l'enquêté

[Q3] Taille de l'agglomération
rurale
2 à 5000
5 à 10000
10 à 20000
20 à 50000
50 à 100 000
100 à 200 000

Plus de 200 000

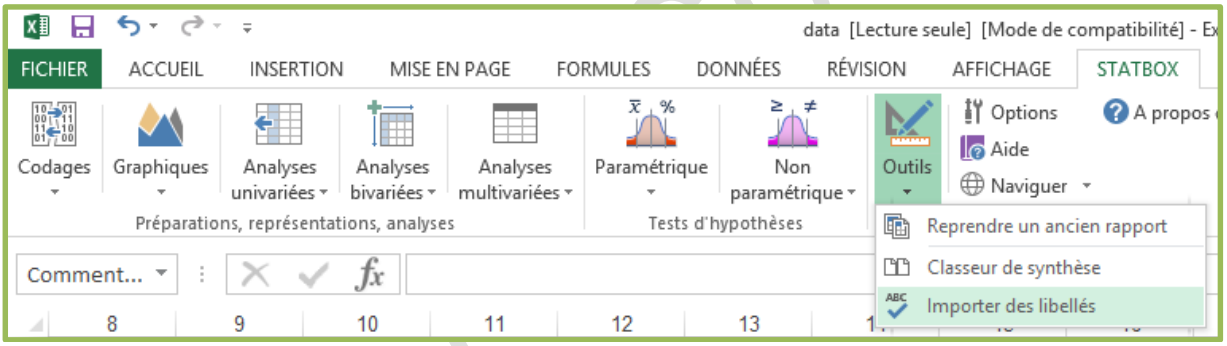
Dans l'exemple précédent, la variable « Age » est numérique, la modalité « Auprès de vos relations personnelles » correspond au code 1, la modalité « Les salons ou séminaires », le code 2, etc.

La structure de saisie doit être la suivante :

- nom de la variable entre crochets suivi d'un espace puis le libellé long de la variable,
- libellés des modalités sur les lignes suivantes,
- une ligne vide sépare les blocs de variables.

L'ordre dans lequel les libellés seront introduits n'a pas d'importance.

Pour importer des libellés, cliquez sur « Importer des libellés » dans le menu Outils,

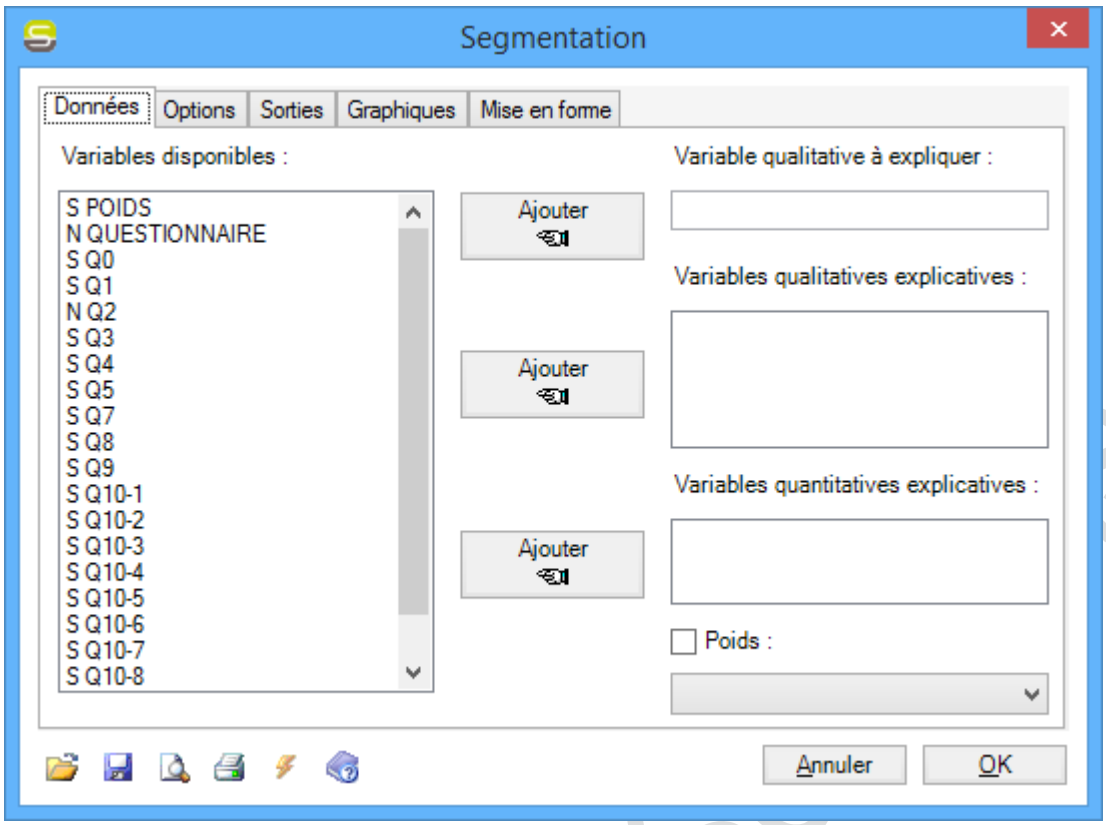



Sélectionnez le fichier à importer, puis validez pour lancer l'importation. Les libellés longs et les modalités de réponse sont alors insérés automatiquement dans les zones commentaires.

## LES BOITES DE DIALOGUE DE RAPPORT

Les boites de dialogues de rapport présentent un mode de fonctionnement et plusieurs fonctions communes.

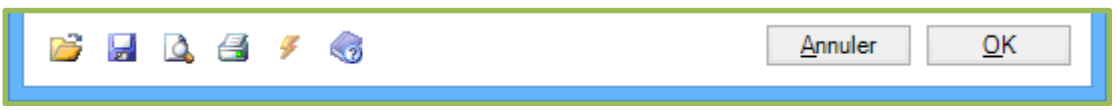
Les paramètres des rapports sont regroupés en 6 thèmes placés dans des onglets spécifiques, par exemple la boite de segmentation propose l'ensemble des onglets disponibles :




- Les « Données » : cet onglet, présent dans toutes les méthodes, regroupe les zones de sélection de données « de base » pour la méthode en cours.
- Les « Variables et observations supplémentaires » : cet onglet, facultatif, présente pour les méthodes où cela est nécessaire les zones de sélection de variable(s) ou d'observations passives. Cet onglet se retrouve uniquement dans les méthodes d'analyses factorielles.
- Les « Options » : cet onglet, présent dans presque toutes les méthodes, propose les options statistiques ou de calculs pour la méthode en cours.
- Les « Sorties » : cet onglet, présent dans toutes les méthodes, présente les différents éléments éditables par la méthode en cours
- Les « Graphiques » : cet onglet, facultatif, présente les options d'affichage et de mise en forme des graphiques pour les méthodes concernées
- La « Mise en forme » : cet onglet, présent dans toutes les méthodes, affiche les options de mise en forme du rapport. Il peut également être le point d'accès aux options générales d'édition des rapports dans StatBox.

L'apparence des boites de dialogue de rapport est modifiée par le choix de l'un ou l'autre des modes de sélection des données. Lorsque vous sélectionnez l'option manuelle, une option « Noms de la variable sur la première ligne » sur l'onglet « Données » vous permet de déterminer si votre sélection de données contient les libellés des variables en première ligne ou bien si la première ligne doit être considérée comme une ligne de données.

Dans la zone située en bas à droite des boites de dialogue de rapport sont proposées 6 fonctionnalités essentielles détaillées ici dans l'ordre d'affichage à l'écran :




*Charger un paramétrage* : cette fonction permet de charger automatiquement dans la boite de dialogue un paramétrage sauvegardé précédemment. Cela est particulièrement utile dans le cas on l'on cherche

à reproduire spécifiquement un paramétrage pour plusieurs analyses sans avoir à tout reparamétrer manuellement. Le fichier contenant le paramétrage est au format *.xml et doit être issu d'une sauvegarde effectué sur la même méthode.

*Sauvegarder le paramétrage* : cette fonction permet de sauvegarder le paramétrage en cours dans la boite de dialogue dans un fichier *.xml.

*Aperçu avant impression* : cette fonction permet de lancer l'édition du rapport et de demander au logiciel d'insérer automatiquement des sauts de page à la fin des sections lorsque la taille d'une page d'impression a été dépassée. Le programme lance ensuite l'aperçu d'Excel afin de vous permettre de juger de la qualité des sauts de page.

*Imprimer le rapport* : cette fonction est presque identique à la précédente, au lieu de l'aperçu avant impression, c'est l'impression elle-même qui est lancée directement. Ce choix est risqué dans le cas de tableaux de grande taille car les sauts de page risquent d'être très espacés. Cette fonction est donc plus adaptée à des rapports de taille limitée (tri à plat, statistiques descriptives,…)

*Réinitialiser la boite de dialogue* : cette fonction permet de réinitialiser l'ensemble du paramétrage en cours dans la boite de dialogue. Les sélections de données sont ainsi vidées et les options statistiques ou de sorties reprennent leurs valeurs par défaut.

*Aide* : cliquez sur ce bouton pour afficher le fichier d'aide principal de l'application

Pour lancer la génération d'un rapport, validez en cliquant sur « OK ».

Pour annuler la génération d'un rapport, cliquez sur « Annuler »

## PERFORMANCES

### Temps de calcul

Les calculs sont généralement assez rapides sauf dans le cas des modules faisant appel à des méthodes itératives d'optimisation (ex. le Multidimensional Scaling) ou à la programmation dynamique (algorithme de Fisher) où les temps de calcul peuvent être élevés, selon le paramétrage utilisé et/ou la taille des jeux de données.

Dans le cas des méthodes itératives, pour vous familiariser avec les temps de réponse de ces méthodes sur votre ordinateur, vous pouvez régler les paramètres contrôlant le nombre de répétitions, le nombre d'itérations maximal et le seuil de convergence avec des valeurs modestes, puis augmenter progressivement le nombre de répétitions, le nombre d'itérations maximal et diminuer le seuil de convergence jusqu'à ce que le temps de calcul vous semble trop élevé.

### Temps d'affichage

L'affichage des tableaux de résultats dans une feuille Excel est assez lent. Aussi, lorsque vous traitez de grands jeux de données, prenez garde aux options qui vous sont proposées dans les boîtes de dialogue au sujet de l'affichage de certains résultats, par exemple :

- la matrice de corrélation dans l'analyse en composantes principales (ACP),
- les matrices d'inertie dans l'analyse factorielle discriminante (AFD),
- la matrice de proximité dans la classification ascendante hiérarchique (CAH).

L'affichage des graphiques est encore plus lent que l'affichage des tableaux de résultats. En particulier, l'affichage des dendrogrammes produits par la CAH peut s'avérer assez long lorsque le nombre d'observations est élevé. De même que pour les tableaux, prenez garde aux options qui vous sont proposées lors de l'affichage. En dehors des problèmes de lisibilité des graphiques, évitez par exemple de représenter 500 observations dans une ACP, car le temps d'affichage sera excessivement long.

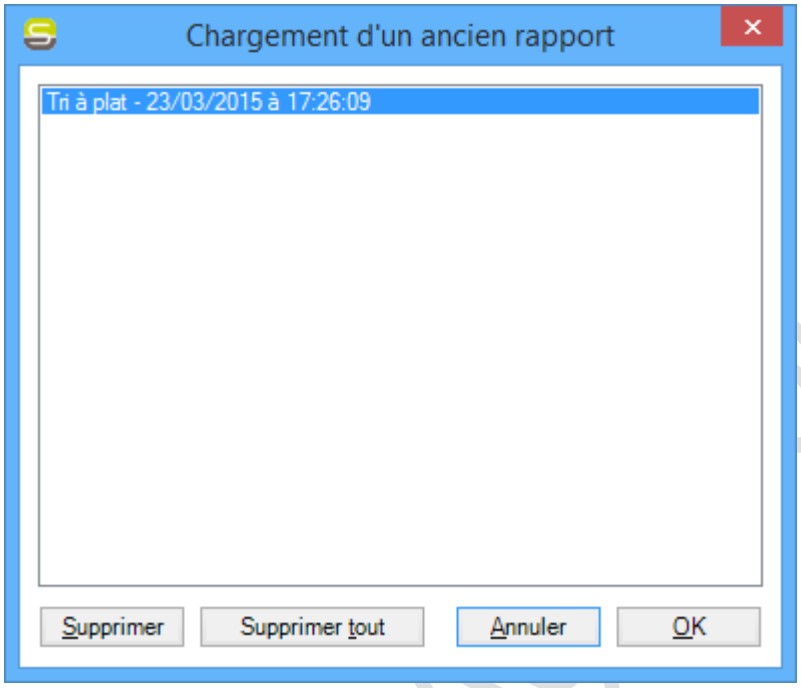
## OUTILS

Plusieurs outils sont proposés afin de faciliter les aspects reporting du logiciel.

### Reprendre un ancien rapport

StatBox garde en mémoire les derniers rapports (données et paramétrage) valides qui ont été édités lors de la session.

Vous pouvez relancer un de ces rapports pour vérifier/modifier par exemple un paramètre statistique ou d'impression. Cliquez sur « Reprendre un ancien rapport », la boite de dialogue suivante apparait :




*Supprimer :* supprime le rapport sélectionné.

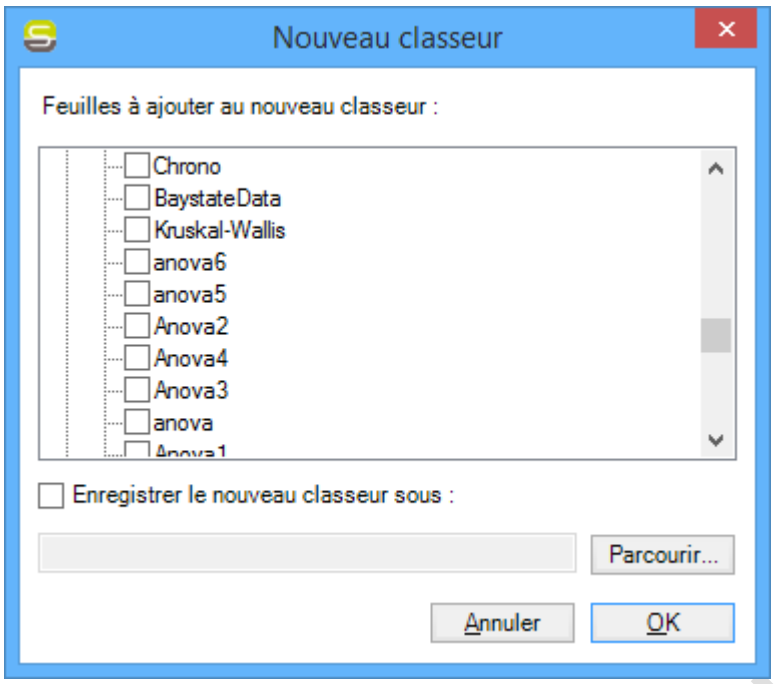
*Supprimer tout :* supprime tous les rapports en mémoire.

Sélectionnez dans la liste des rapports en mémoire le rapport à reprendre puis validez en cliquant sur « OK ». La boite de dialogue correspondant au rapport apparait alors. Le rapport en cours est alors indépendant de la feuille de données en cours dans Excel, ce sont les données chargées précédemment qui seront utilisés (même si le classeur source n'est plus ouvert).

### Classeur de synthèse

Afin de faciliter la constitution de rapports de synthèse un outil « classeur de synthèse » vous est proposé. Plus rapide que la sélection 1 à 1 des feuilles dans Excel il permet de sélectionner parmi toutes les feuilles de tous les classeurs ouverts dans Excel celles que vous souhaitez insérer dans le classeur de synthèse.

Cliquez sur « Classeur de synthèse », la boite de dialogue suivant apparait :




*Feuilles à ajouter au nouveau classeur* : sélectionnez parmi les feuilles disponibles les feuilles qui constitueront la synthèse.

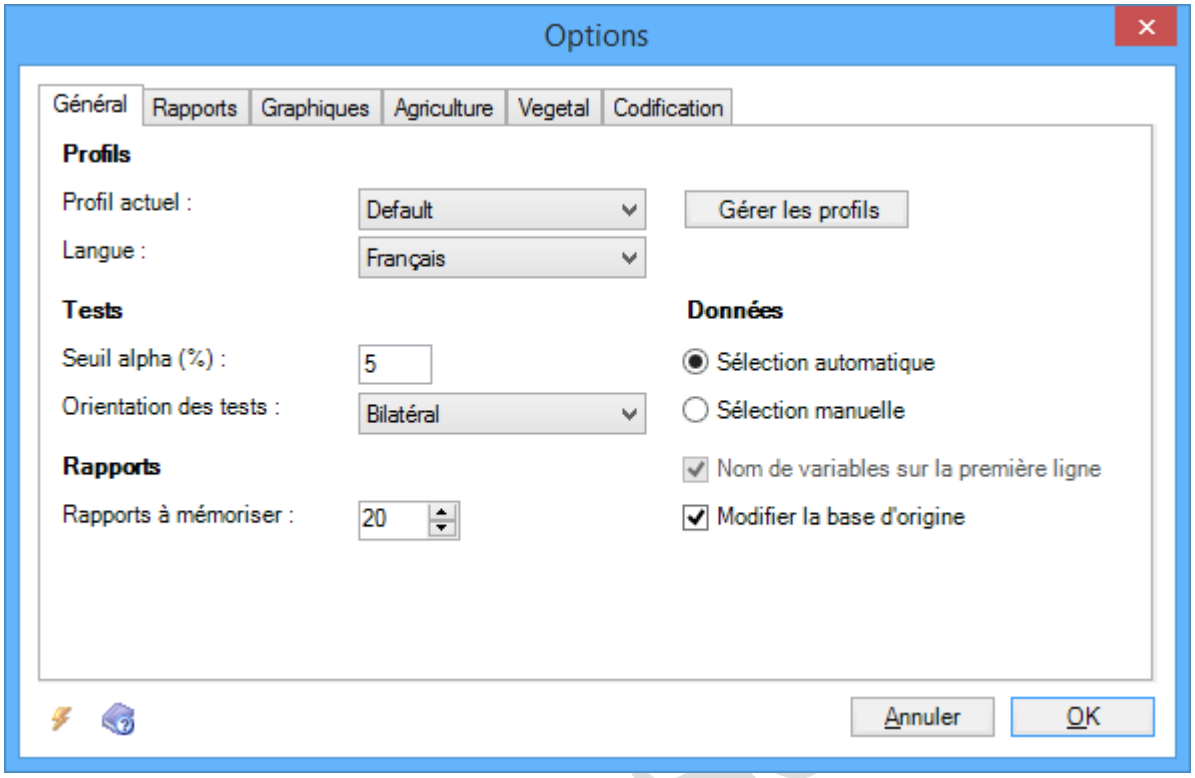
*Enregistrer le nouveau classeur sous :* cochez cette option pour que le classeur généré soit enregistré automatiquement à l'emplacement et avec le nom que spécifierez.

Validez en cliquant sur « OK »

## OPTIONS

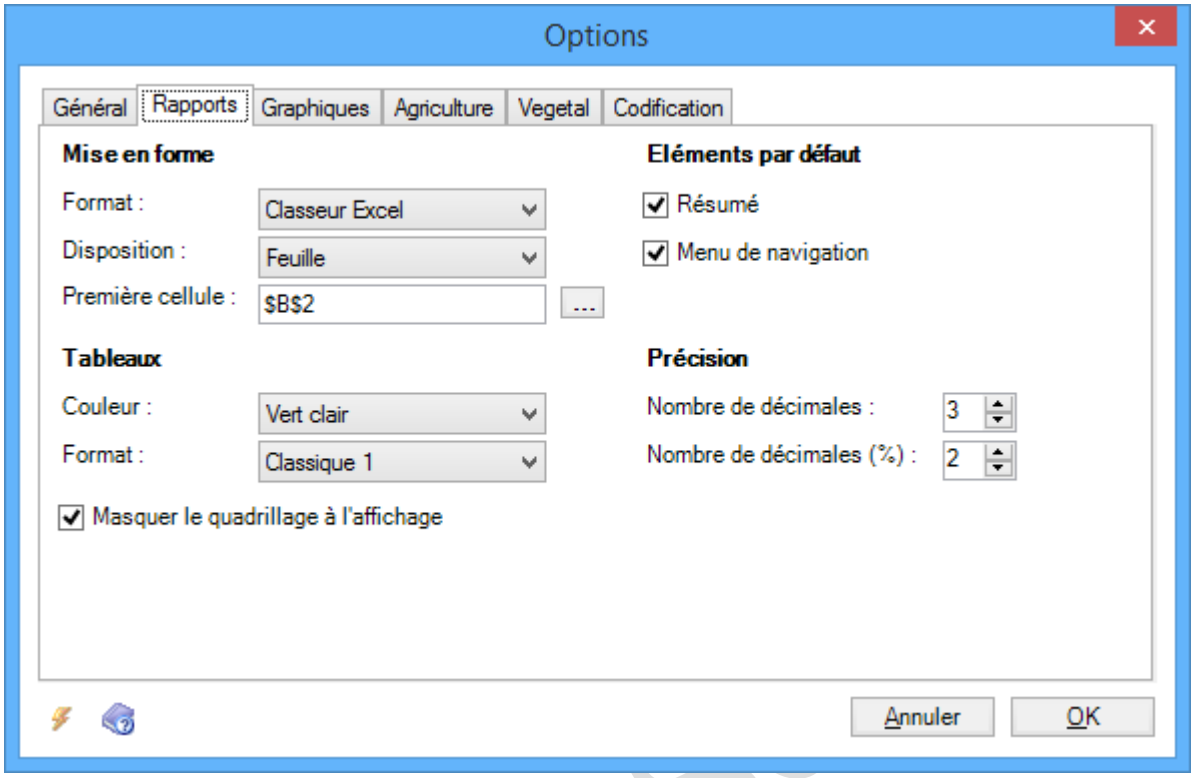
Pour accéder aux options du logiciel, cliquez sur « Options », la boite suivante apparait :

## Onglet « Général »



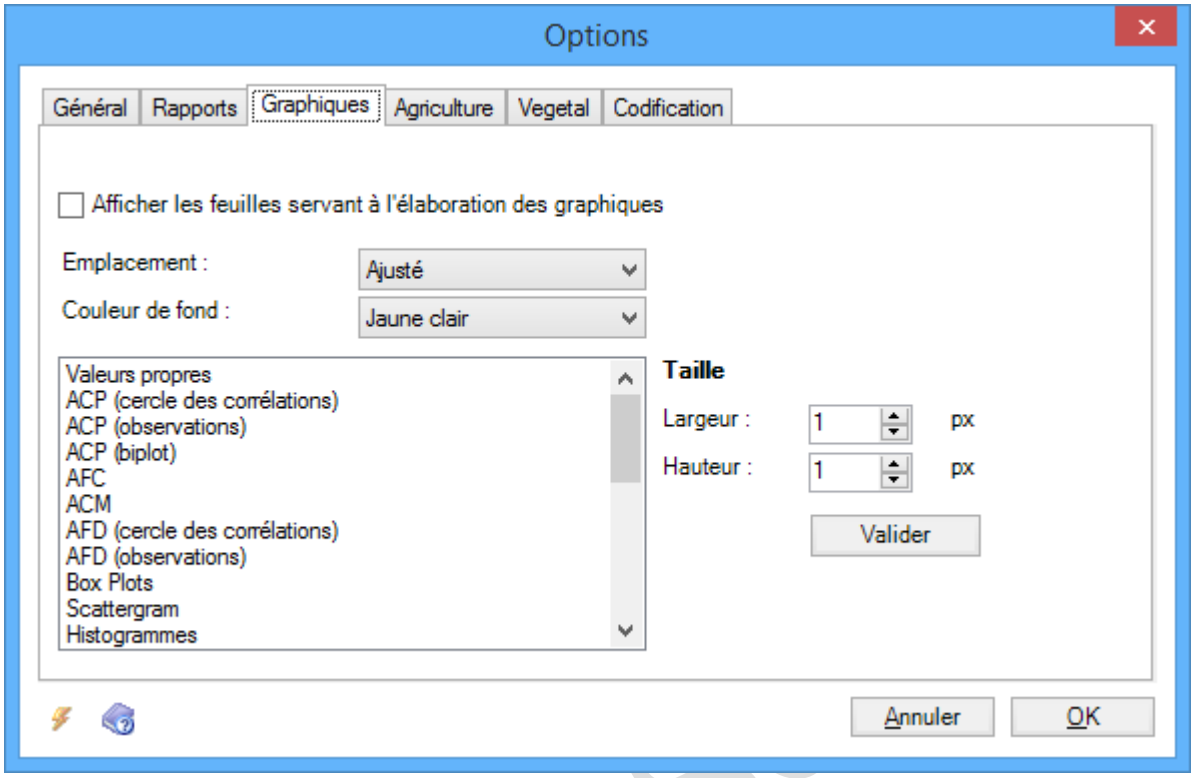

- *Profil actuel :* sélectionnez le nom du profil à charger par défaut.
- *Langue :* sélectionnez la langue du profil en cours.
- *Seuil alpha :* entrez la valeur du risque de première espèce à utiliser par défaut pour les tests.
- *Orientation des tests :* sélectionnez l'orientation par défaut des tests.
- « Sélection automatique » / « Sélection manuelle » : sélectionnez le mode de sélection des données.
- *Nom de variable sur la première ligne :* cochez cette option (en sélection manuelle) afin d'indiquer si par défaut la première ligne de la sélection contient les noms de variable ou bien s'il s'agit d'une ligne contenant des données.
- *Modifier la base d'origine :* cochez cette option pour que les méthodes de codage proposent par défaut d'ajouter les nouvelles colonnes produites à la base d'origine.

## Onglet « Rapports »



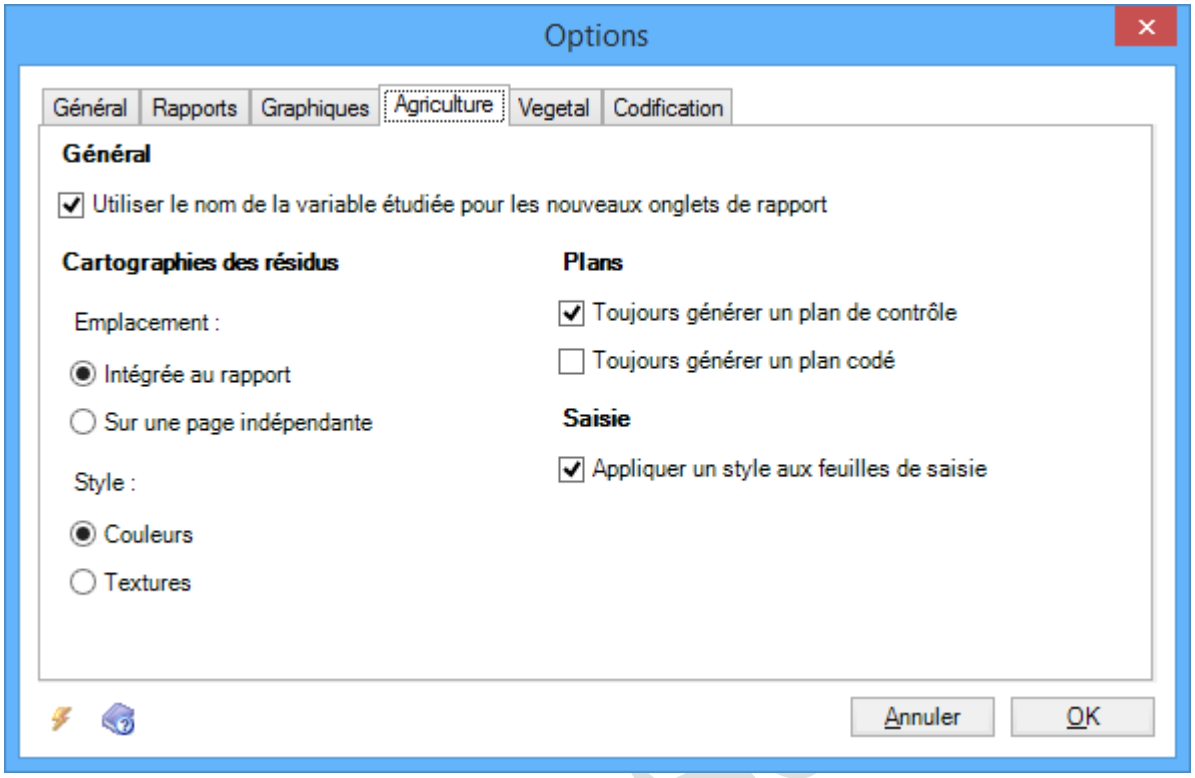

- *Format :* sélectionnez le format de fichier par défaut des nouveaux rapports.
- *Disposition :* sélectionnez la disposition par défaut des nouveaux rapports.
- *Première cellule :* sélectionnez la plage d'origine par défaut des rapports.
- *Couleur :* sélectionnez la couleur d'affichage des tableaux de résultats.
- *Format :* sélectionnez le format d'affichage des tableaux de résultats.
- *Masquer le quadrillage à l'affichage* : masque la grille Excel sur les feuilles de rapports.
- *Résumé :* cochez cette option pour qu'un bref compte rendu sur les variables et les paramètres utilisés dans les rapports soit édité par défaut.
- *Menu de navigation :* cochez cette options pour qu'un menu de navigation rapide soit ajouté au début des rapports pour accéder plus facilement aux différentes sections.
- *Nombre de décimales :* entrez le nombre de décimales par défaut pour les résultats numériques.
- *Nombre de décimales (%) :* entrez le nombre de décimales par défaut pour les résultats en pourcentage.

## Onglets « Graphiques »



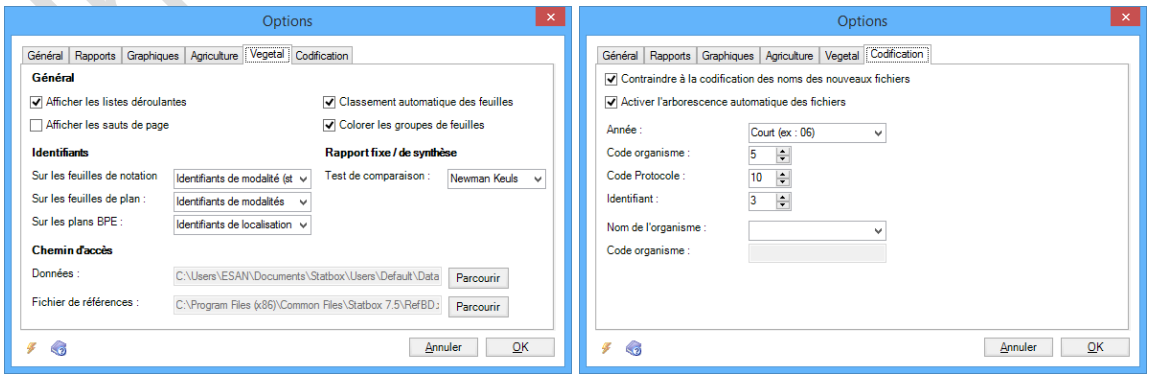

- *Afficher les feuilles servant à l'élaboration des graphiques :* cochez cette option pour que les feuilles contenant les données sources des graphiques soient rendues visibles.
- *Emplacement :* sélectionnez le mode de positionnement des graphiques.
- *Couleur de fond :* sélectionnez la couleur du corps des graphiques.
- *Largeur / Hauteur :* entrez pour le graphique sélectionné dans la liste de gauche les dimensions d'affichage du graphique. Pour que les nouvelles dimensions soit mémorisées, vous devez valider les nouvelles dimensions en cliquant sur « Valider ».

## Onglet « Agriculture »




- *Utiliser le nom de la variable étudiée pour les nouveaux onglets de rapport* : cochez cette option pour que le nom des nouveaux onglets de rapport d'analyse de variance reprenne par défaut le nom des variables étudiées. Si cette option n'est pas cochée un nom incrémentiel est utilisé.
- *« Intégré au rapport » / « sur une page indépendante »* : sélectionnez l'emplacement par défaut des cartographies des résidus. Si vous sélectionnez l'option « page indépendante », une nouvelle feuille sera créée en plus du rapport ou sera placé la cartographique des résidus.
- *« Couleurs » / « Textures »* : sélectionnez le mode d'affichage des cartographies des résidus. Si vous sélectionnez l'option couleur, un gradient de couleur bleu sera utilisé pour identifier les classes des résidus, pour l'option texture, c'est un gradient de texture d'Excel qui sera utilisé.
- *Toujours générer un plan de contrôle :* cochez cette option pour qu'un plan de contrôle soit généré par défaut lors de la génération d'un nouveau plan.
- *Toujours générer un plan codé :* cochez cette option pour qu'un plan codé soit généré par défaut lors de la création d'un nouveau plan.

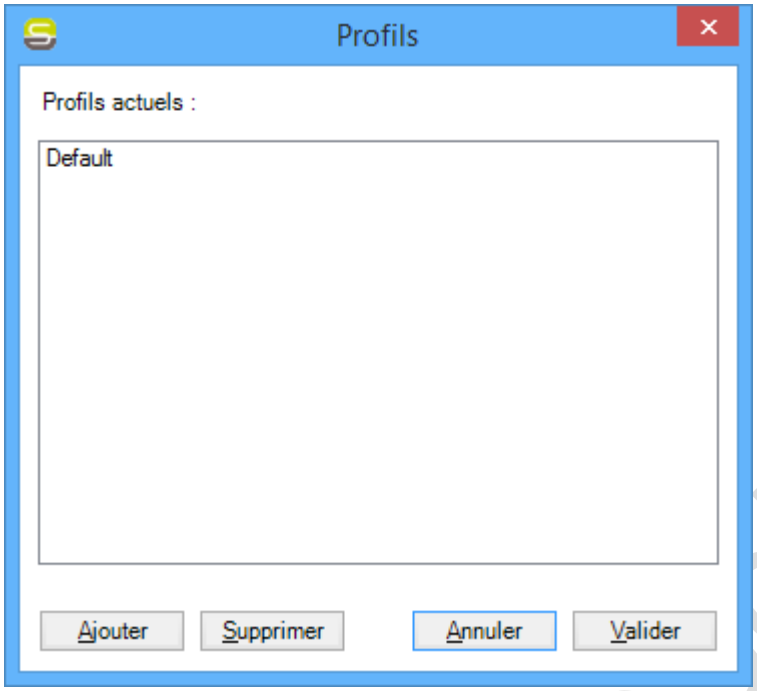
## Onglet « Vegetal » et onglet « Codification » - Edition Vegetal uniquement




- *Se reporter aux sections « **Essais en agriculture** » et « **StatBox Vegetal – Prise en main** »*

## Gestion des profils

Vous avez la possibilité de gérer plusieurs profils d'options pour réaliser par exemple des séries de test à seuils différents ou avec des mises forme différentes. Pour modifier la liste des profils disponibles cliquez sur « Gérer les profils » sur l'onglet « Général ». La boite suivante apparait :




- *Ajouter :* ajoutez un nouveau profil en l'identifiant par un nouveau nom.
- *Supprimer :* supprimer le profil sélectionné.

Validez en cliquant sur « Valider ». Si vous avez supprimé des profils, un message apparaît alors vous demandant si vous désirez également supprimer les répertoires spécifiques aux utilisateurs situées dans le dossier : Mes documents\StatBox\Users.

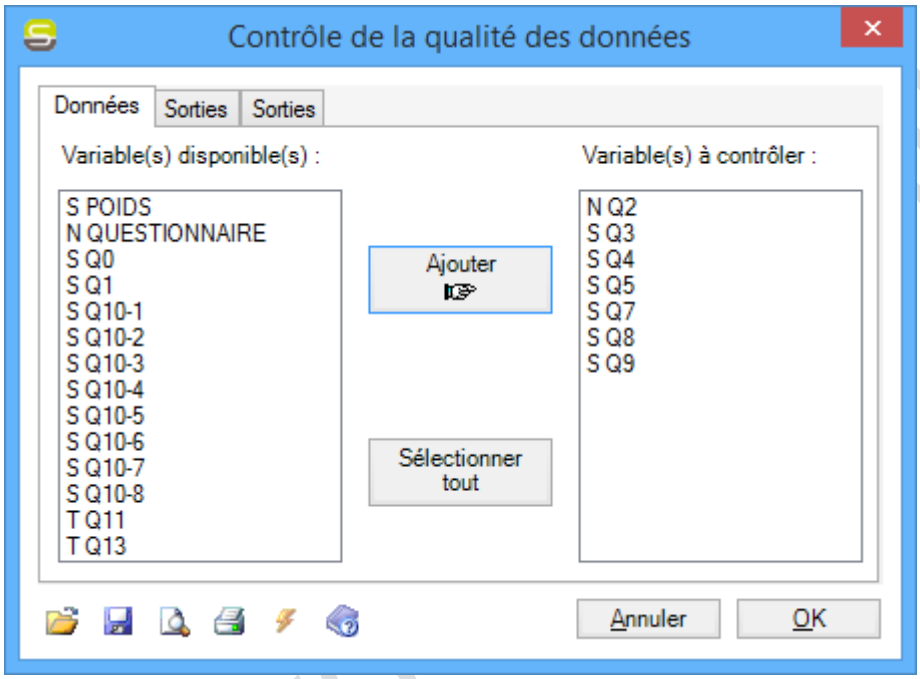
# CODAGE

## CONTRÔLE DE LA QUALITÉ DES DONNÉES

Utilisez ce module pour obtenir rapidement un ensemble d'indicateurs sur le type des données disponibles, le nombre de manquants, la dispersion,…

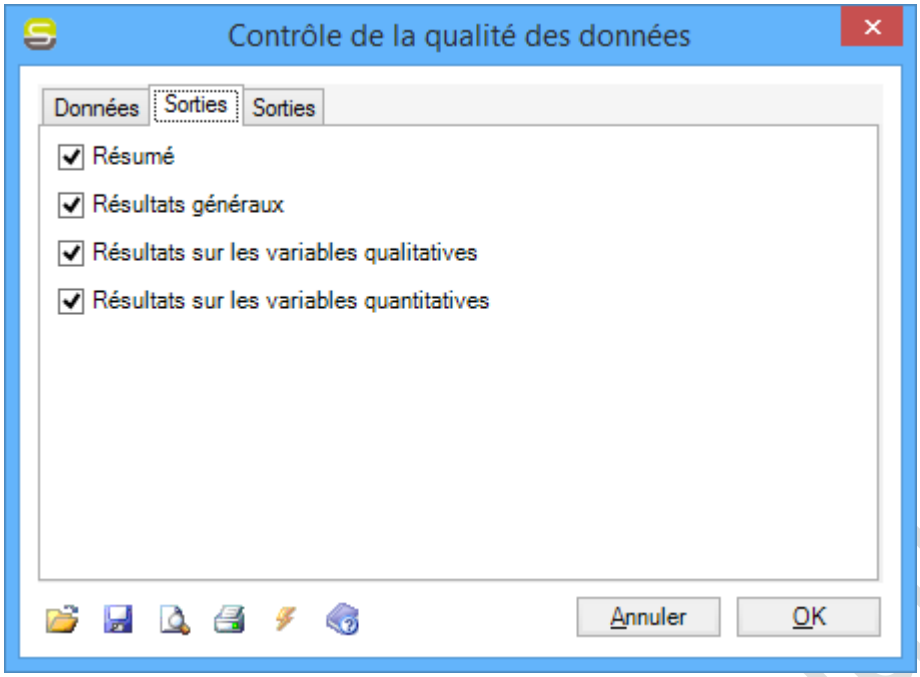
### Mise en œuvre

*Onglet « Données »*




- *Variable(s) à contrôler :* sélectionnez l'ensemble des variables à contrôler en les plaçant dans la liste de droite. Vous pouvez sélectionner automatiquement toutes les variables disponibles en cliquant sur « Sélectionner tout ».

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Résultats généraux :* cochez cette option pour obtenir des résultats généraux sur toutes les variables (nombre de manquants, types,..).
- *Résultats sur les variables qualitatives :* cochez cette option pour obtenir des résultats sur les variables qualitatives (première et dernière modalités).
- *Résultats sur les variables quantitatives :* cochez cette option pour obtenir des résultats sur les variables quantitatives (moyenne, amplitude, écart-type,..).

## CODAGE EN CLASSES

Utilisez ce module pour transformer une variable quantitative en classes de valeurs, c'est-à-dire en une variable ordinale.

### Description

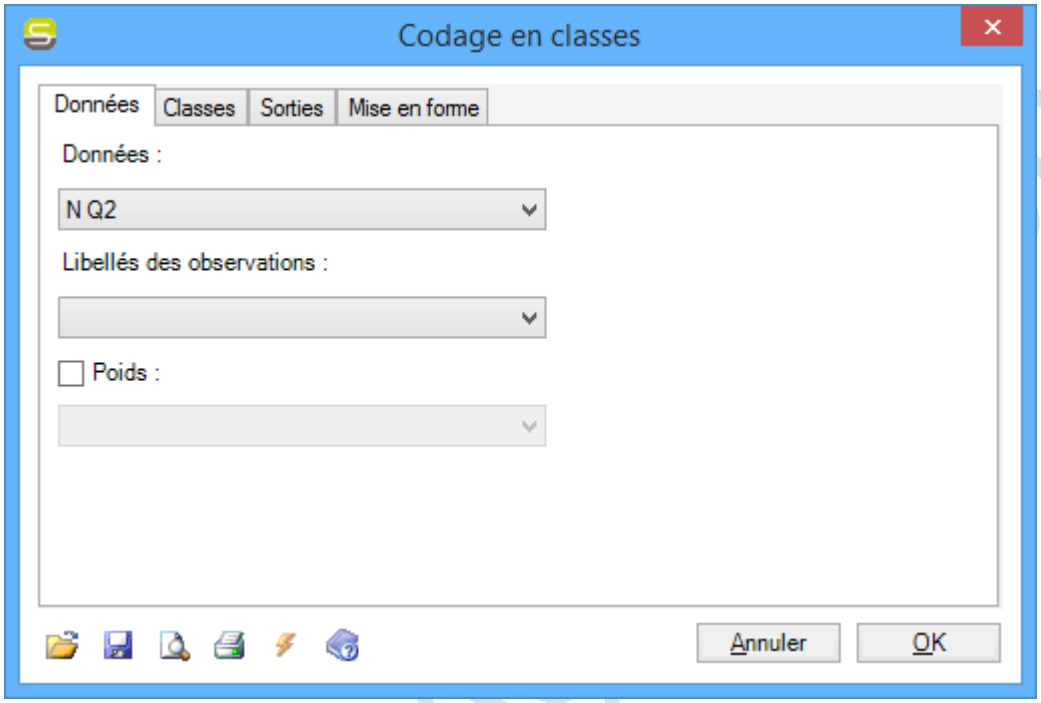
Ce module, très complet, autorise toutes les définitions de classes possibles. Plusieurs modes de discrétisation sont proposés :

- *amplitude constante* : découpage à pas constant entre les valeurs minimale et maximale de la colonne de valeurs sélectionnée,
- *classes optimales* : calcul de classes optimales vis-à-vis de la minimisation de l'inertie intra-classe (les classes sont donc les plus compactes possible). L'algorithme d'Anderberg (algorithme d'amélioration itérative d'une solution initiale) est utilisé.
- *poids égaux* : à effectifs égaux dans le cas de données non pondérées, ou à poids constant, lorsque les données sont pondérées,
- en modifiant manuellement les bornes des classes grâce au module d'édition.
- *importer les bornes* : En important les bornes des classes, exemple pour les classes 18 à moins de 25, 25 à moins de 35,… la liste :

| |
|---|
| 18<br>25<br>35<br>45<br>55<br>99 |

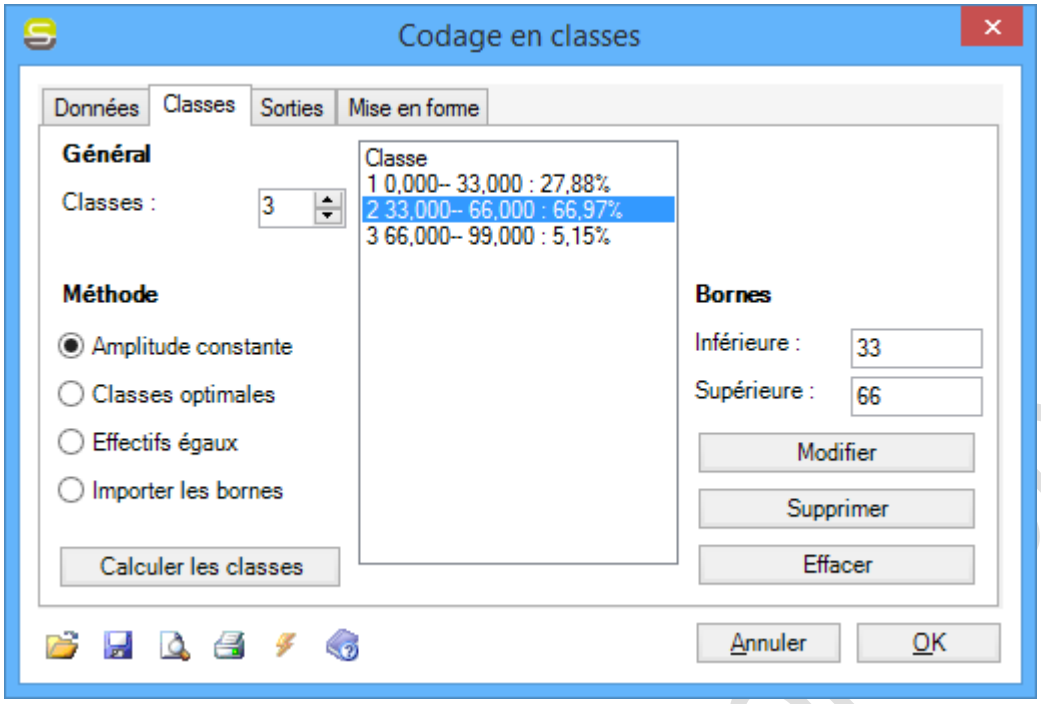
## Mise en œuvre

*Onglet « Données »*




- *Données :* sélectionnez la variable à coder.
- *Observations :* sélectionnez la variable contenant les libellés des observations si vous souhaitez créer un tableau codé avec des libellés particuliers pour les observations. Par défaut, le libellé d'une observation est son numéro de ligne dans le tableau.
- *Poids :* cochez cette case si vous désirez pondérer les données, puis sélectionnez la variable de pondération. Les valeurs manquantes dans les poids sont cumulées avec les valeurs manquantes dans les données.

*Onglet « Classes »*

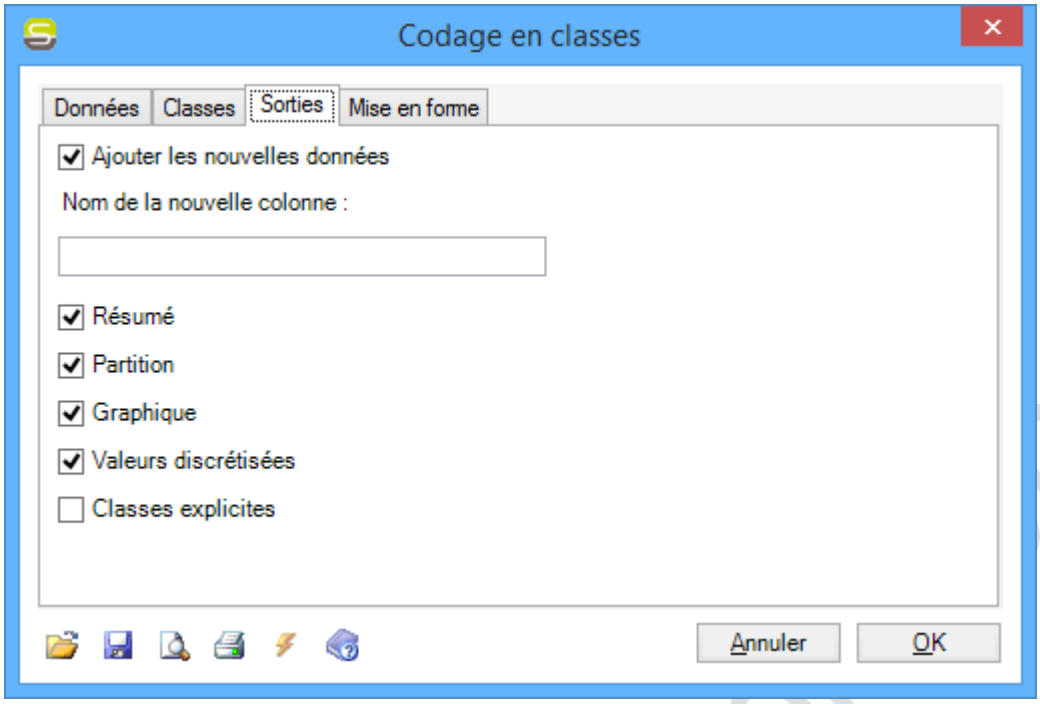



- *Classes :* entrez le nombre d'intervalles à calculer.
- *Amplitude constante* / *Classes optimales* / *Poids égaux* / *Importer les bornes* : choisissez le type de calcul des intervalles.
- Pour effectuer le calcul des classes cliquez sur *« Calculer les classes »*. La liste des bornes des classes calculées s'affiche alors dans la zone centrale.

Vous avez la possibilité de personnaliser la liste des bornes proposée. Pour cela sélectionnez la classe à personnaliser dans la liste centrale, entrez les nouvelles bornes pour cette classe dans les zones *« Inférieure »* et *« Supérieure »* puis cliquez sur *« Valider »*. Les autres bornes sont alors recalculées si nécessaire.

Vous pouvez également supprimer une classe particulière en la sélectionnant dans la liste centrale et en cliquant sur *« Supprimer »*, ou supprimer toutes les classes en cliquant sur *« Effacer »*. La suppression d'un intervalle est en fait une suppression de la borne supérieure, sauf dans le cas du dernier intervalle où il s'agit de la borne inférieure.

*Onglet « Sorties »*




- *Ajouter les nouvelles données* : ajoute la colonne des identifiants de classe à la base d'origine. Vous pouvez donner un nom particulier à la nouvelle colonne ou laisser le logiciel déterminer le nouveau nom automatiquement.
- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Partition :* affiche la table de répartition des observations dans les différentes classes.
- *Graphique :* affiche un histogramme de fréquence des classes.
- *Valeurs discrétisées :* affiche la table d'appartenance des observations aux différentes classes.
- *Classes explicites :* affiche la table d'appartenance des observations, les modalités de la variable ordinale produite correspondent aux bornes des classes et non pas à l'identifiant de la classe.

Remarques :

- Lorsqu'il y a des valeurs manquantes, StatBox propose d'ignorer les lignes concernées. En cas de refus, le traitement est abandonné.
- Si vous ne cliquez pas sur « Calculer les classes », l'affichage du rapport est impossible.

### Références

**Anderberg M.R. (1973)**. Cluster analysis for applications. Academic Press, New York.

**Diday E., J. Lemaire, J. Pouget & F. Testu (1982)**. Eléments d'analyse de données. Dunod, Paris, pp. 32-40, 45-46.

**Fisher W.D. (1958)**. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, **53** : 789-798.

**Frontier** S. (1981). Méthode statistique. Masson, Paris, pp. 42-59.

## REGROUPEMENT DE MODALITÉS

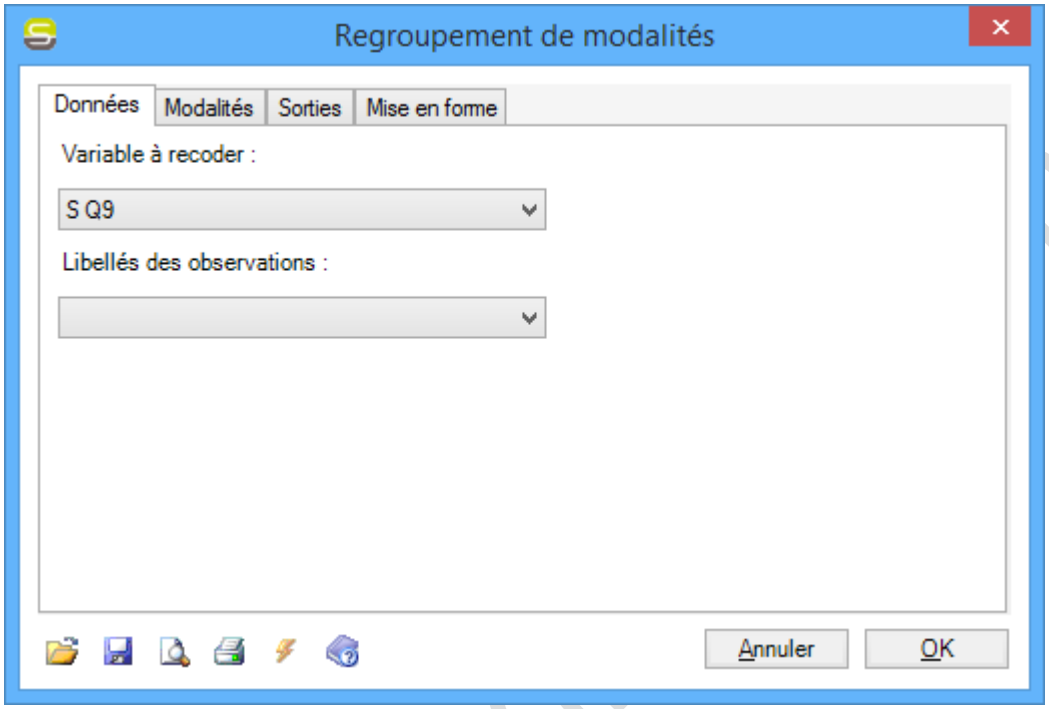
Utilisez ce module pour coder ou recoder les modalités d'une variable qualitative.

## Description

Le regroupement de modalités est une forme de codage particulière dans laquelle un même code est affecté à plusieurs modalités. La procédure de codage produit la variable recodée ainsi que le tableau de correspondance entre les anciens codes et les nouveaux.

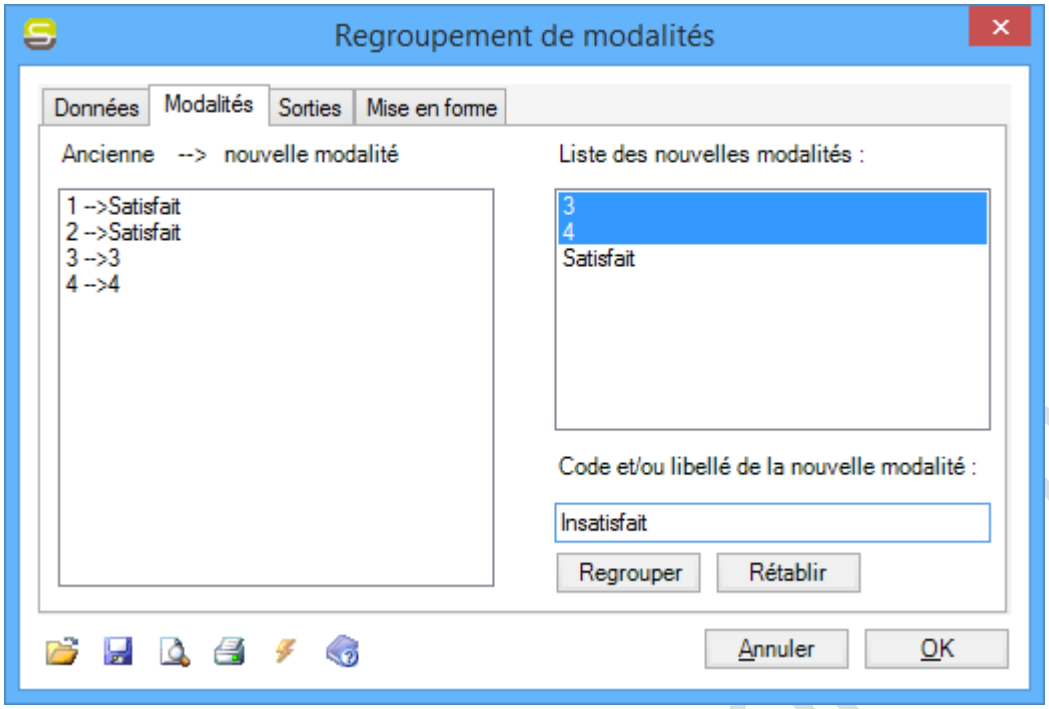
## Mise en œuvre

*Onglet « Données »*




- *Variable à recoder :* sélectionnez une variable qualitative à recoder.
- *Libellés des observations :* sélectionnez la variable contenant les libellés des observations si vous souhaitez créer un tableau de codes avec des libellés particuliers. Par défaut, le libellé d'une observation est son numéro de ligne dans le tableau.

*Onglet « Modalités »*

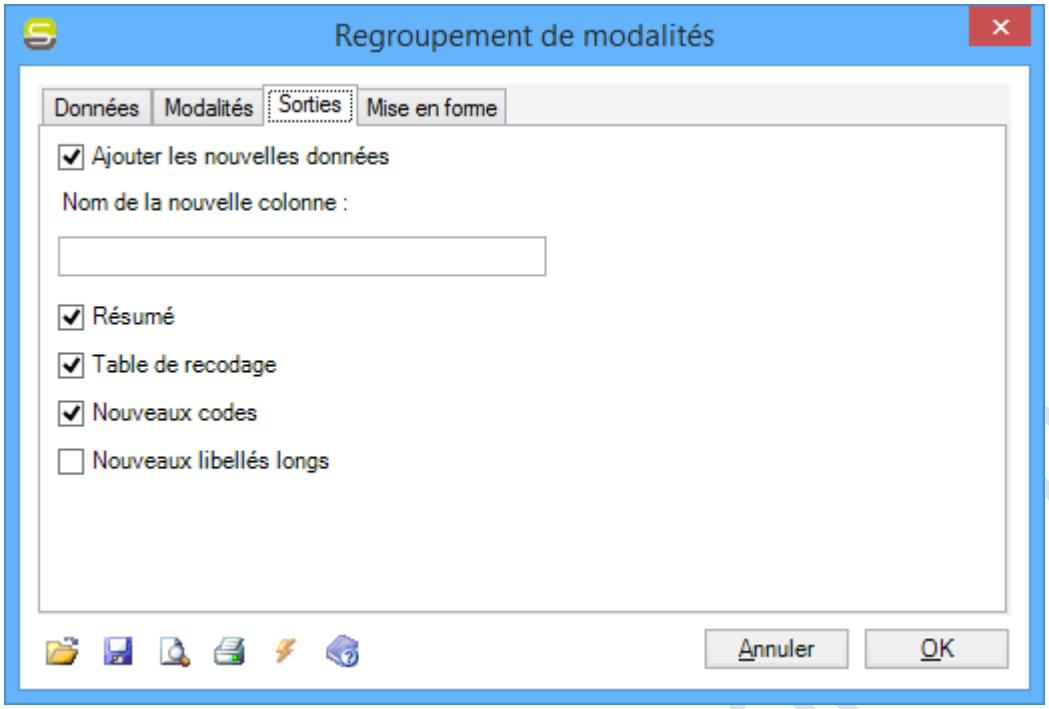



- *« Code et/ou libellé de la nouvelle modalité » :* pour effectuer un recodage, sélectionnez dans la liste de droite les modalités à regrouper. Dans la zone de saisie, entrez le label à affecter à l'ensemble des modalités sélectionnées dans la liste de droite. Cliquez sur le bouton « Regrouper » pour rendre le codage effectif. Les listes de gauche et de droite sont mises à jour et vous pouvez procéder à de nouveaux codages.
- *Rétablir :* vous pouvez revenir en arrière sur les codages effectués. Pour cela, sélectionnez dans la liste de droite la modalité à rétablir puis cliquez sur « Rétablir », un message d'avertissement vous demande alors de valider l'annulation, puis les listes de gauche et de droite sont mises à jour. Le nombre d'étapes de codage et leur annulation n'est pas limité de sorte qu'il est toujours possible de revenir à un état antérieur.

La liste de gauche permet de visualiser la correspondance entre les anciennes modalités et les nouvelles, la liste de droite permet de sélectionner les modalités à recoder.

Les valeurs manquantes sont autorisées et peuvent donc être également recodées. Les valeurs manquantes sont représentées dans la liste des anciennes modalités par un crochet ouvrant suivi d'un crochet fermant : < >.

*Onglet « Sorties »*




- *Ajouter les nouvelles données :* ajoute la colonne des nouvelles modalités à la base d'origine. Vous pouvez donner un nom particulier à la nouvelle colonne ou laisser le logiciel déterminer le nouveau nom automatiquement.
- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Table de recodage :* affiche la table de correspondance entre les anciens et les nouveaux codes.
- *Nouveau codes :* affiche la table des nouveaux codes pour chacune des observations.
- *Nouveau libellés longs :* affiche la table des nouveaux codes pour chacune des observations, les codes sont représentés par le libellé de la nouvelle modalité.

## CODAGE DISJONCTIF (OUI/NON)

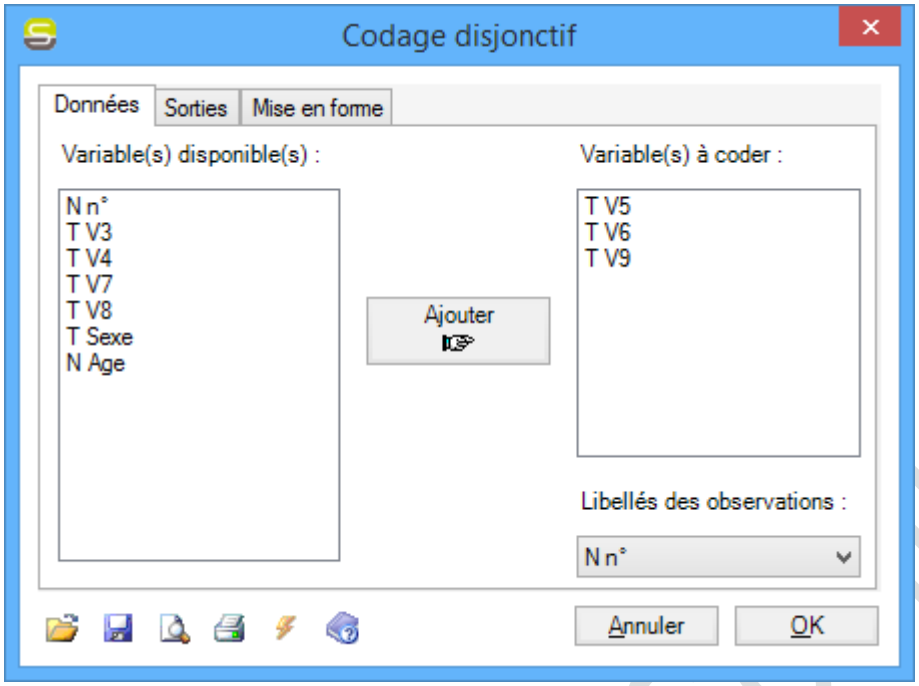
Utilisez ce module pour coder un tableau avec les observations en lignes et des variables qualitatives en colonnes sous la forme d'un tableau binaire (0/1) en utilisant le codage disjonctif complet.

### Description

Le codage disjonctif consiste à affecter la valeur 1 pour la modalité d'une variable qualitative pour l'observation considérée et 0 à toutes les autres modalités de la variable. L'application de ce codage à un ensemble de variables qualitatives consiste à répéter cette procédure pour chaque variable. Le tableau obtenu contient donc autant de colonnes qu'il y a de modalités au total pour l'ensemble des variables qualitatives et autant de 1 pour une observation qu'il y a de variables.

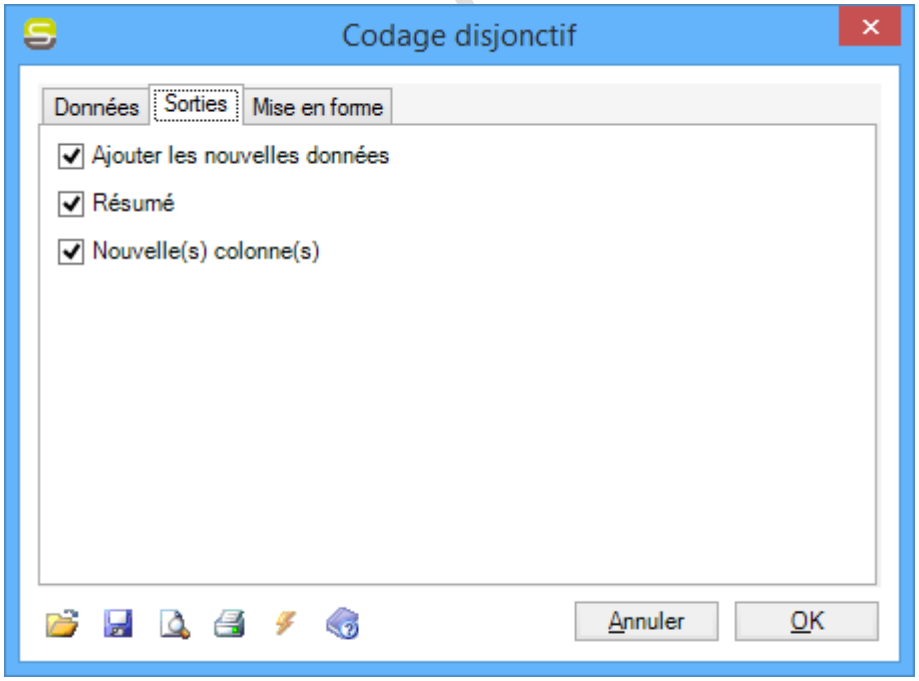
## Mise en œuvre

*Onglet « Données »*




- *Variable(s) à coder :* sélectionnez la/les variables à coder en la/les plaçant dans la liste de droite. En cas de valeur manquante dans une case [*i,j*] (c'est-à-dire pour l'observation en ligne *i* et la variable qualitative en colonne *j*) toutes les modalités de la variable *j* sont mises à 0 pour l'observation *i*.
- *Libellés des observations :* sélectionnez la variable contenant les libellés des observations si vous souhaitez créer un tableau disjonctif avec des libellés particuliers. Par défaut, le libellé d'une observation est son numéro de ligne dans le tableau.

*Onglet « Sorties »*




- *Ajouter les nouvelles données :* ajoute les colonnes disjonctives à la base d'origine.

- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Nouvelle(s) colonne(s) :* affiche la table des données disjonctives

### Références

**Diday E., J. Lemaire, J. Pouget & F. Testu (1982)**. Eléments d'analyse de données. Dunod, Paris, pp. 42-44.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, p. 112.

## CODAGE D'UNE VARIABLE NUMÉRIQUE EN RANGS

Utilisez ce module pour coder en rangs un tableau avec les observations en lignes et les variables en colonnes.

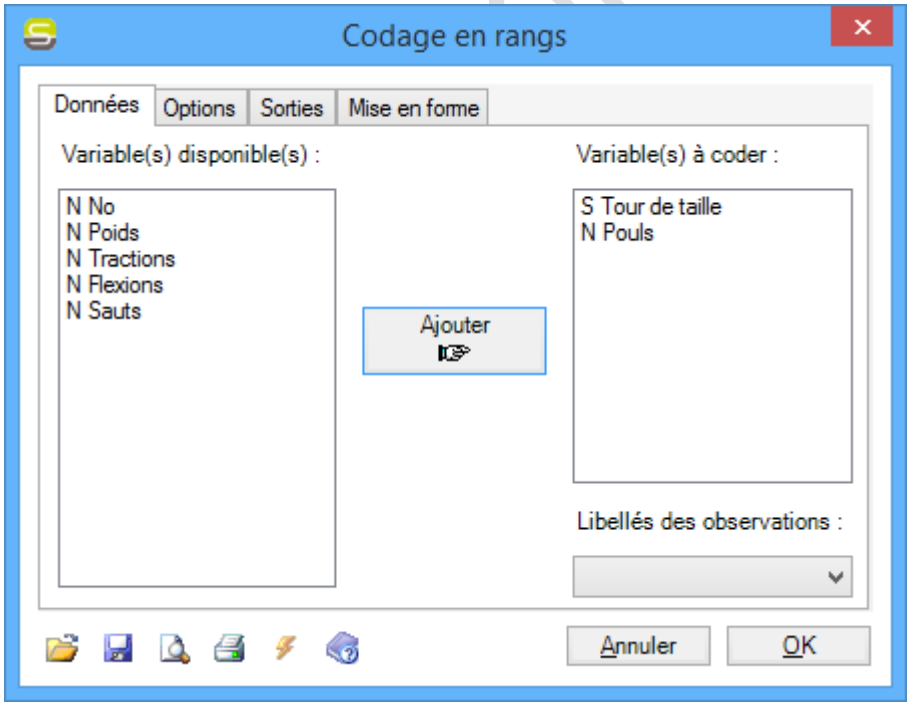
### Description

Pour chaque variable, une observation se voit attribuer le rang de sa valeur par rapport à l'ordre croissant de l'ensemble des valeurs. Le rang des observations ex æquo est calculé comme la moyenne de leurs rangs initiaux ou bien comme le rang de leur valeur commune.

**Remarque :** le premier mode de traitement des observations ex æquo décrit est le seul qui soit valide pour effectuer des tests statistiques (par exemple, tester la corrélation entre deux variables).

### Mise en œuvre

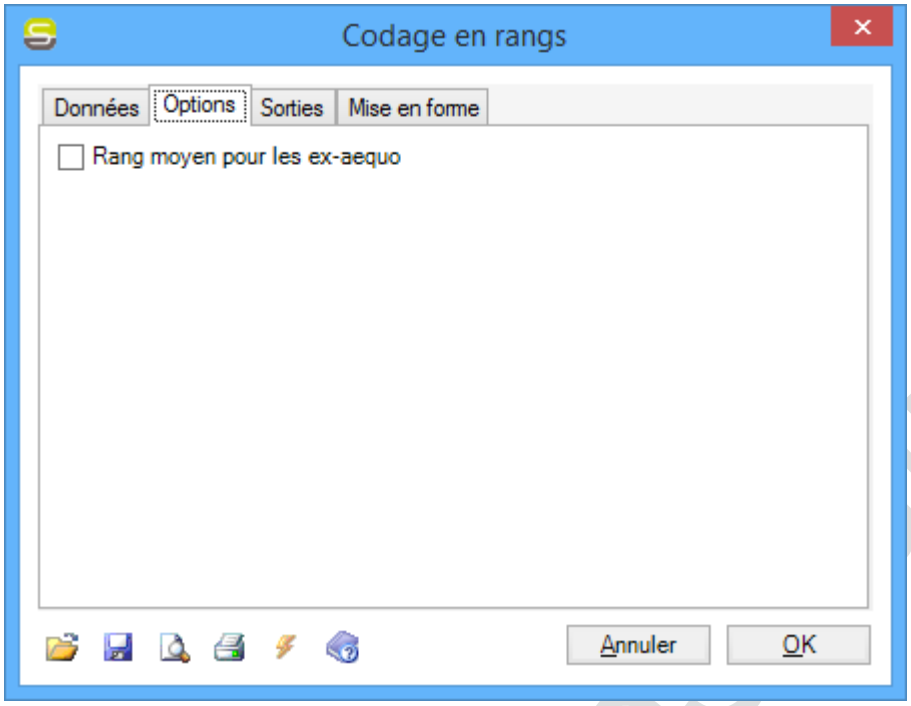
*Onglet « Données »*




- *Variable(s) à coder :* sélectionnez la ou les variables en la/les plaçant dans la liste de droite, le tableau comporte les observations en lignes et les variables quantitatives en colonnes. Les valeurs manquantes sont autorisées et occupent le rang 0.

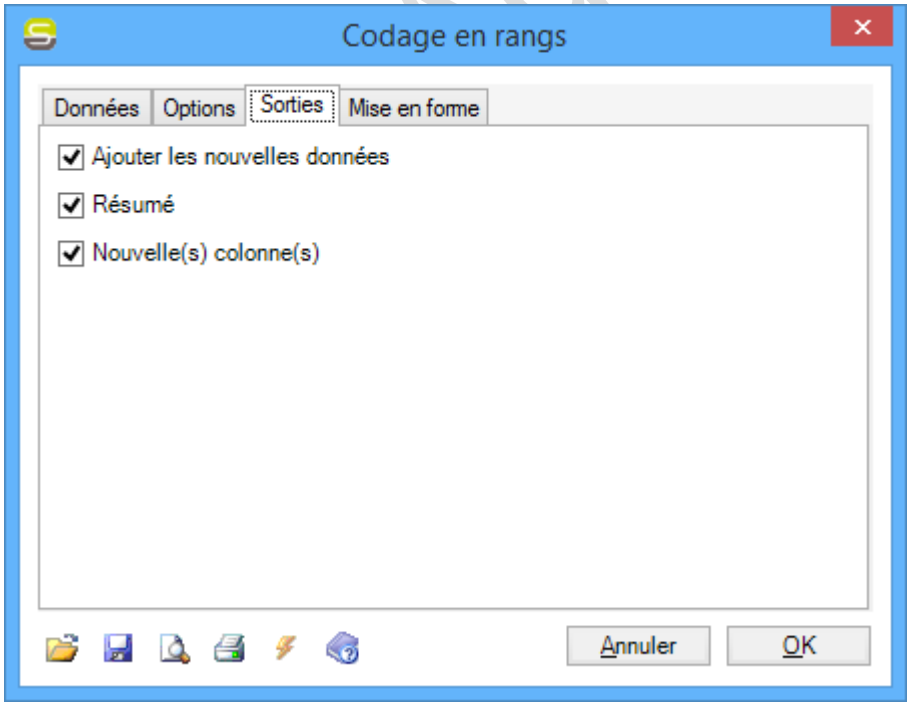
- *Libellés des observations :* sélectionnez la variable contenant les libellés des observations si vous souhaitez créer un tableau de rangs avec des libellés particuliers. Par défaut, le libellé d'une observation est son numéro de ligne dans le tableau.

### *Onglet « Options »*




- *Rangs moyens pour les ex-æquo :* calcule un rang moyen pour les valeurs identiques afin de pouvoir utiliser les rangs pour effectuer des tests statistiques.

### *Onglet « Sorties »*




- *Ajouter les nouvelles données :* ajoute les colonnes de rangs à la base d'origine.
- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Nouvelle(s) colonne(s) :* affiche la table des rangs pour chacune des variables sélectionnées.

## CODAGE D'UNE VARIABLE TEXTE EN CODES

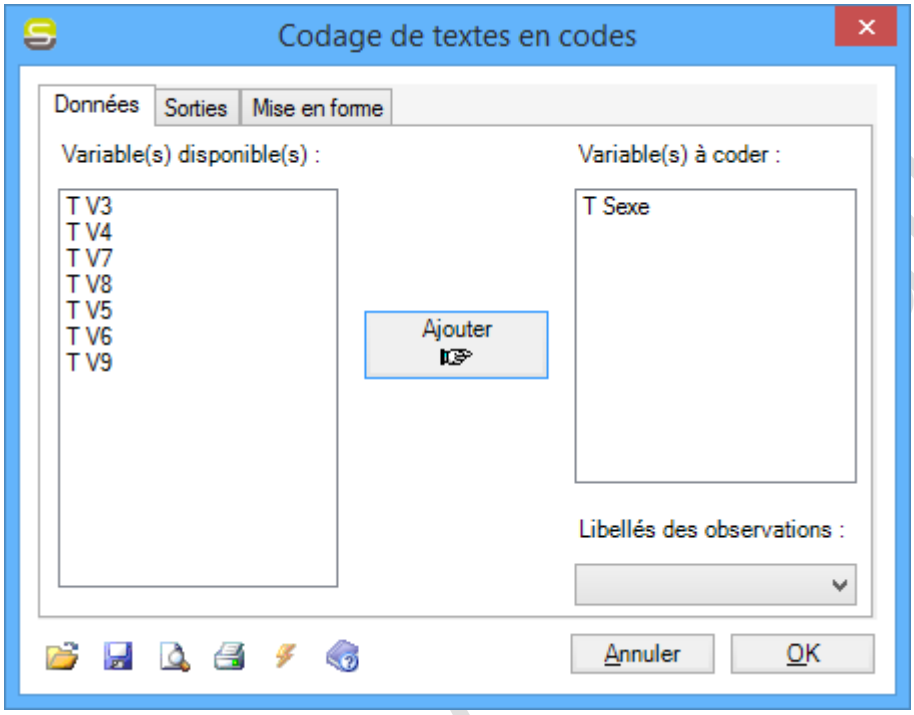
Utilisez ce module pour coder des variables textes en tableau de codes correspondant à l'ordre alphabétique des textes.

### Description

Pour chaque variable, une observation se voit attribué le rang de la chaîne de caractères.

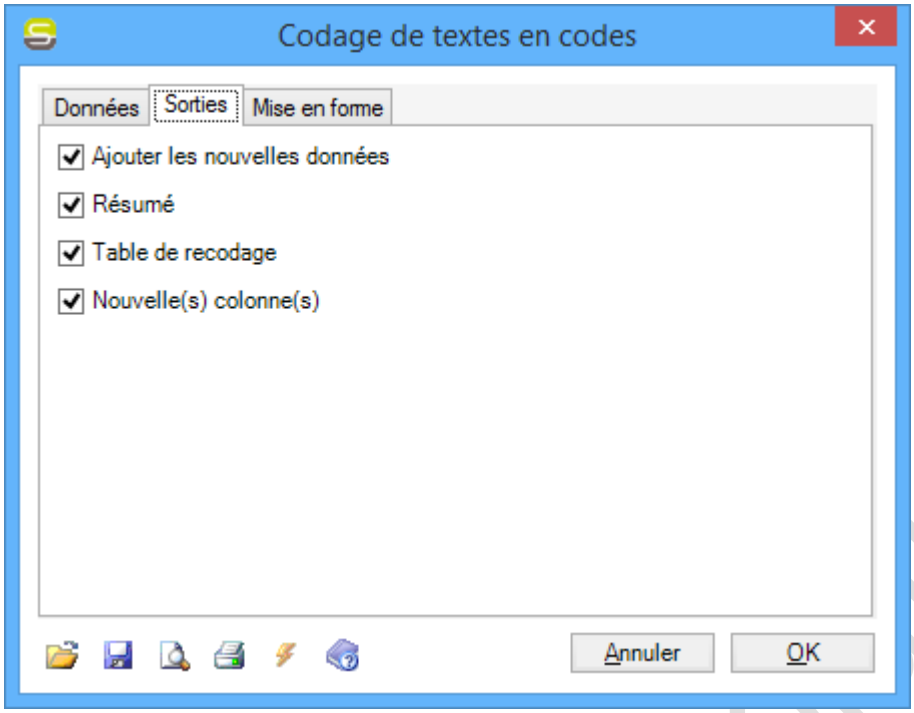
### Mise en œuvre

*Onglet « Données »*




- *Variable(s) à coder :* sélectionnez la/les variable(s) à coder en la/les plaçant dans la liste de droite, le tableau comporte des observations en lignes et les variables qualitatives en colonnes. Les valeurs manquantes sont autorisées et occupent le rang 0.
- *Libellés des observations :* sélectionnez la variable contenant les libellés des observations si vous souhaitez créer un tableau de codes avec des libellés particuliers. Par défaut, le libellé d'une observation est son numéro de ligne dans le tableau.

*Onglet « Sorties »*




- *Ajouter les nouvelles données :* ajoute les colonnes de codes à la base d'origine.
- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Table de recodage :* affiche la table de correspondance entre les textes d'origine et les codes produits. Une table de correspondance est éditée pour chacune des variables recodées.
- *Nouvelle(s) colonne(s) :* affiche la table des codes pour chacune des variables sélectionnées.

## TRANSFORMATION

Utilisez ce module pour transformer une variable quantitative continue au moyen d'une fonction analytique.

### Description

Les transformations disponibles sont :

- *centrer réduire* : les données sont standardisées et ramenées à une variable de moyenne 0 et d'écart-type 1,
- *centrer* : chacune des valeurs est égale à sa valeur dont est soustrait la moyenne de la variable
- *réduire* : chacune des valeurs est divisée par l'écart type de la variable
- *entre 0 et 1* : les valeurs sont transposées entre 0 et 1
- *entre 0 et 100* : les valeurs sont transposées entre 0 et 100
- *log(x)* : logarithmique (base 10), afin de rendre la variance indépendante de la moyenne en cas de proportionnalité entre la variance et la moyenne de la variable initiale, pour les distributions semblables à la distribution lognormale,
- *log(x + 1)* : analogue à la précédente, mais définie pour les données comportant des valeurs nulles,
- *ln(x)* : analogue à log(*x*) mais utilisant le logarithme népérien,

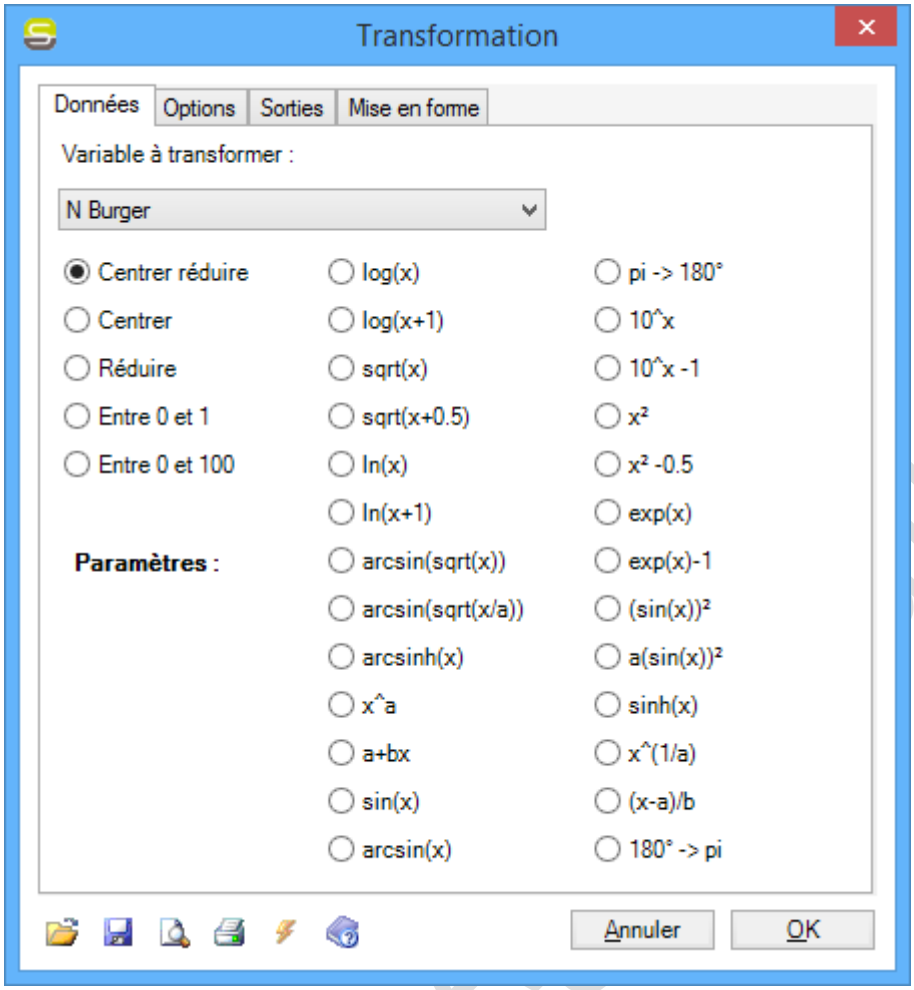
- *ln(x + 1)* : analogue à log(*x* + 1) mais utilisant le logarithme népérien,
- *sqrt(x)* : racine carrée, afin de rendre la variance indépendante de la moyenne en cas de proportionnalité entre la variance et la moyenne de la variable initiale, pour les distributions semblables à la distribution de Poisson,
- *sqrt(x + 0.5)* : analogue à la précédente, mais à préférer dans le cas où les valeurs sont dans l'ensemble relativement faibles,
- *arcsin(sqrt(x))* : angulaire ou arc sinus, concernant les distributions binomiales, et utilisée pour les proportions (valeurs entre 0 et 1), la variable transformée étant alors asymptotiquement normale,
- *arcsin(sqrt(x/a))* : analogue à la précédente, mais pouvant s'appliquer à des pourcentages si *a* = 100 ou directement à des effectifs si *a* est égal à l'effectif total,
- *arcsinh(x)* : arc sinus hyperbolique, concernant les distributions binomiales négatives,
- *x^a* : exponentiation à la puissance *a*,
- *a + bx* : transformation linéaire,
- *π -> 180°* : transformation de radians en degrés.

et les fonctions réciproques, respectivement :

- 10^x
- 10^x - 1
- exp(x)
- exp(x) - 1
- x²
- x² - 0.5
- (sin(x))²
- a(sin(x))²
- sinh(x)
- x^(1/a)
- (x-a)/b
- 180° -> π

## Mise en œuvre

*Onglet « Données »*

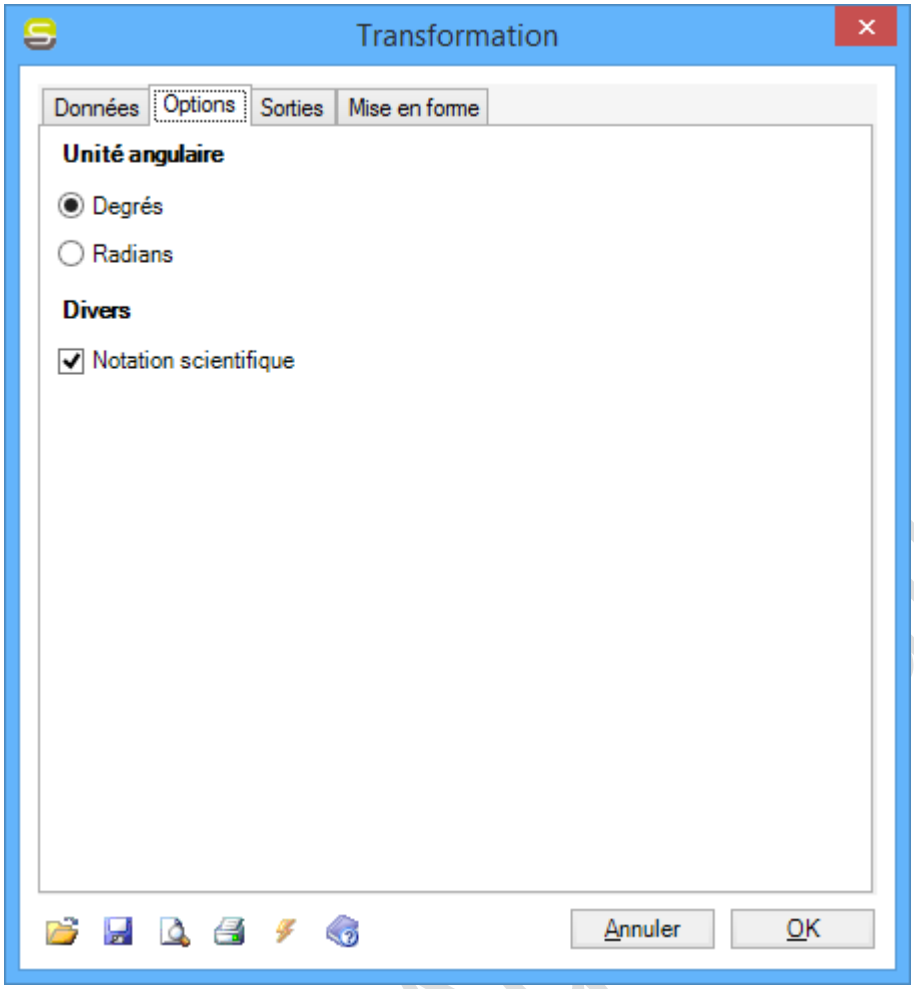



- *Variable à transformer :* sélectionnez la variable quantitative contenant les données source à transformer.

Les valeurs manquantes dans la colonne des données restent manquantes dans la colonne des résultats. Des valeurs manquantes sont également produites lorsque la transformation est impossible (par exemple, le logarithme de valeurs négatives).

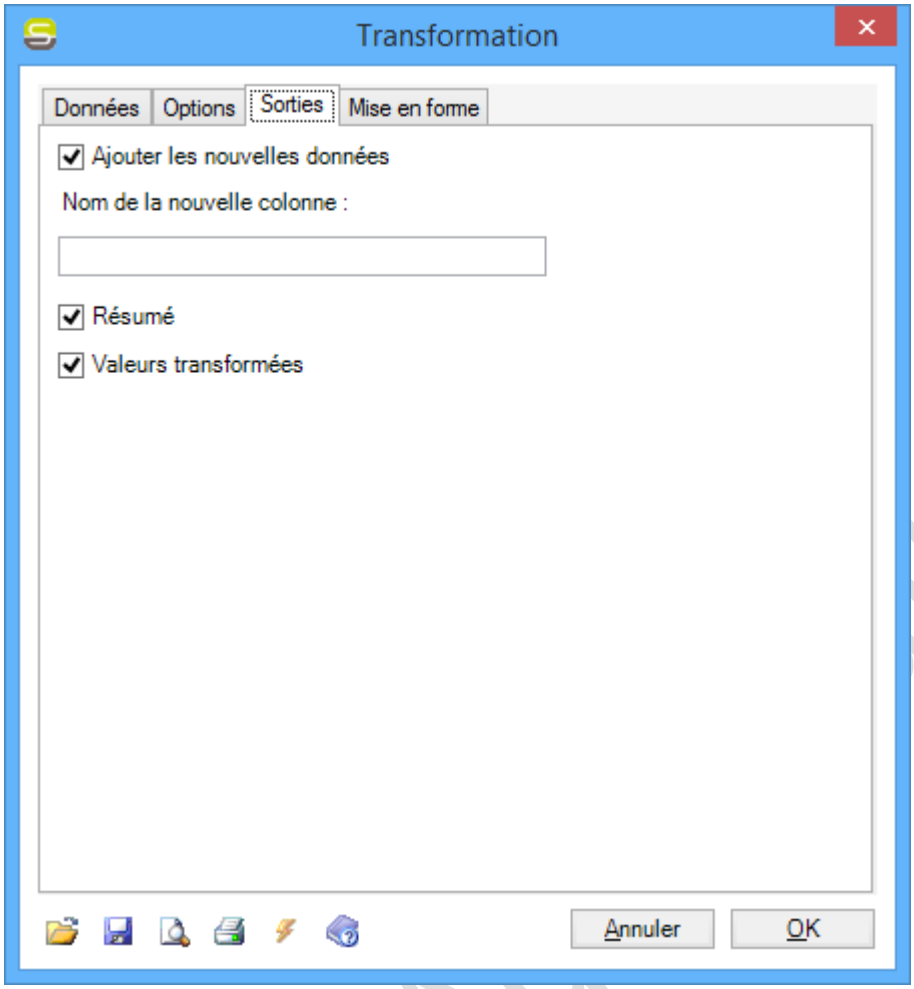
Sélectionnez la fonction à utiliser pour transformer vos données. Lorsque la fonction sélectionnée nécessite un paramètre, un champ de saisie devient visible afin de pouvoir entrer la valeur de ce paramètre.

*Onglet « Options »*




- *Notation scientifique :* cochez cette option si vous désirez que les valeurs trop petites et trop grandes soient affichées en notation scientifique. Une valeur est considérée comme trop petite si la valeur affichée ne comporte aucune décimale différente de 0 et trop grande si la valeur est supérieure à 1E+9.
- *« Degrés » / « Radians » :* sélectionnez « Degrés » si l'argument de sin(*x*) ou le résultat de arcsin(*x*) sont exprimés en degrés et sélectionnez « Radians » si l'argument de sin(*x*) ou le résultat de arcsin(*x*) sont exprimés en radians.

*Onglet « Sorties »*




- *Ajouter les nouvelles données :* ajoute la colonne des données transformées à la base d'origine. Vous pouvez donner un nom particulier à la nouvelle colonne ou laisser le logiciel déterminer le nouveau nom automatiquement.
- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport
- *Valeurs transformées :* affiche la table des valeurs transformées.

### Références

**Dagnelie P. (1986)**. Théorie et méthodes statistiques. Vol. 2. Les Presses Agronomiques de Gembloux, Gembloux, pp. 361-375.

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 409-422.

## CALCUL MATRICIEL

Ce module permet d'effectuer les opérations de base sur des matrices.

### Description

Voici les fonctions disponibles :

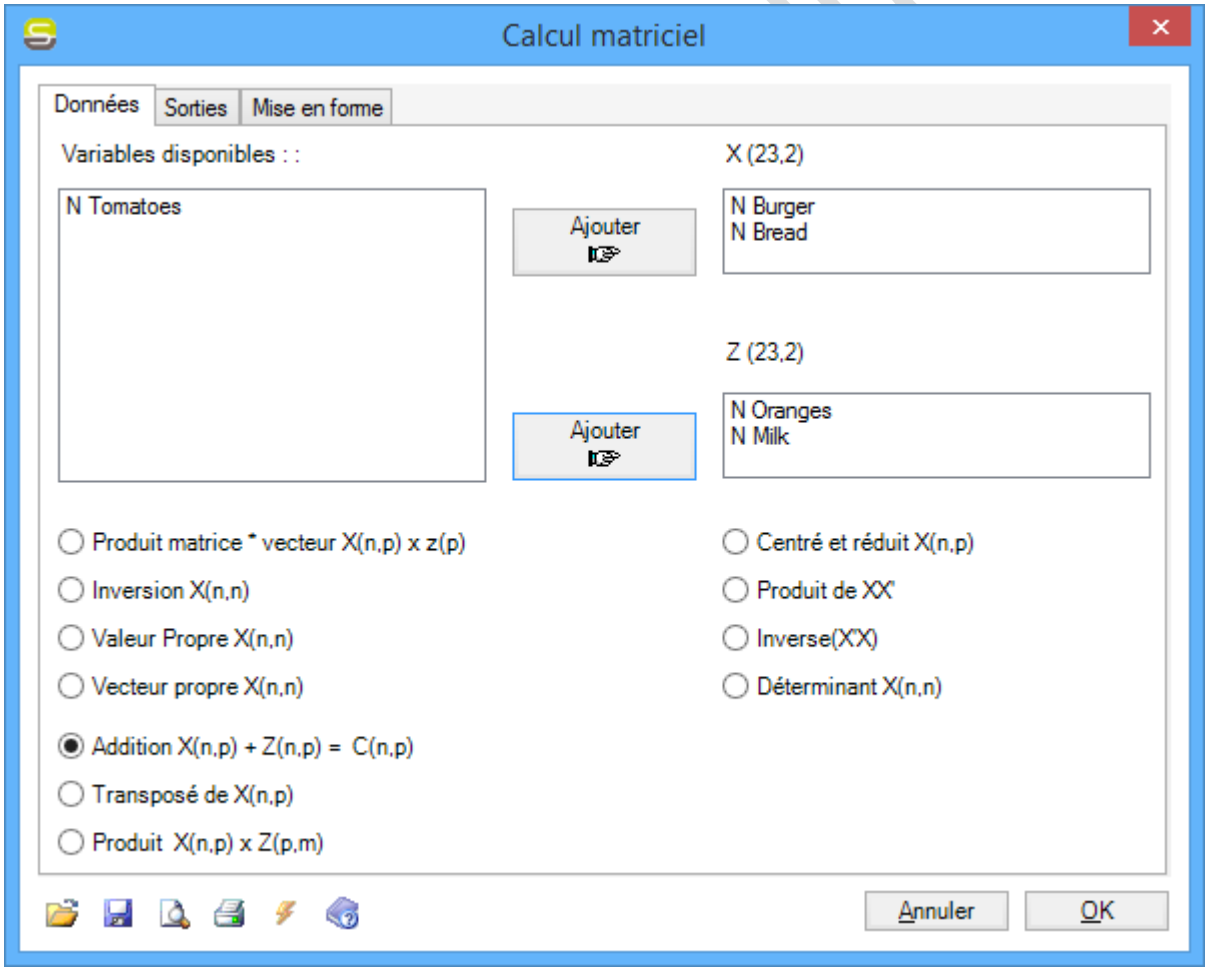
- *Addition de 2 matrices* : X (n, p) + Z (n, p)

- *Transposé* : X (n, p)
- *Produit de 2 matrices* : X (n, p) x Z (p, m)
- Produit d'un vecteur par une matrice : X (n, p) x Z(p)
- Inversion d'une matrice symétrique : X (n, n)
- Valeur propre d'une matrice symétrique : X (n, n)
- Vecteur propre d'une matrice symétrique : X (n, n)
- Matrice Centré et réduite : X (n, p)
- *Produit* de X'X
- Inverse de X'X
- *Déterminant* de X (n, n)

Ces différentes fonctions permettent de procéder au calcul pas à pas de certaines analyses comme l'ACP ou la régression par exemple.

## Mise en œuvre

### *Onglet « Données »*

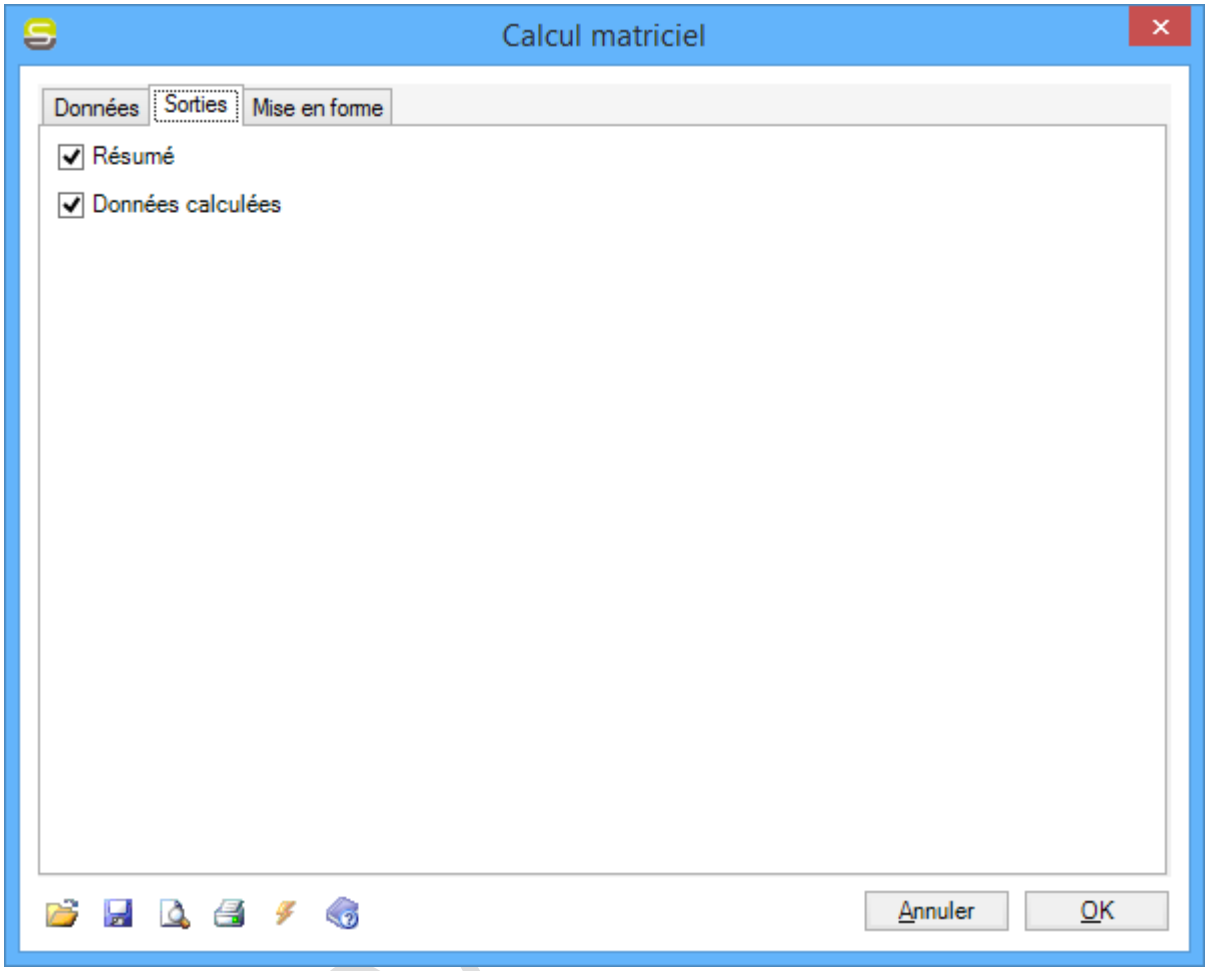



Sélectionnez l'opération à effectuer.

*« X (,) » / « Z (,) » :* sélectionnez les variables à utiliser pour le calcul en les faisant passer dans la/les liste(s) de droite. En fonction de l'opération sélectionnée, la liste des variables de la matrice Z peut être facultative.

Des renseignements sur la taille des matrices d'origine s'affichent en haut des listes qui vous permettent de vérifier les prérequis de taille relative à chacune des opérations (ces prérequis sont indiqués à côté de chacune des opérations)

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Données calculées :* affiche la table des données calculées.

## CALCUL VECTORIEL

Ce module permet d'effectuer les opérations de base sur des vecteurs.

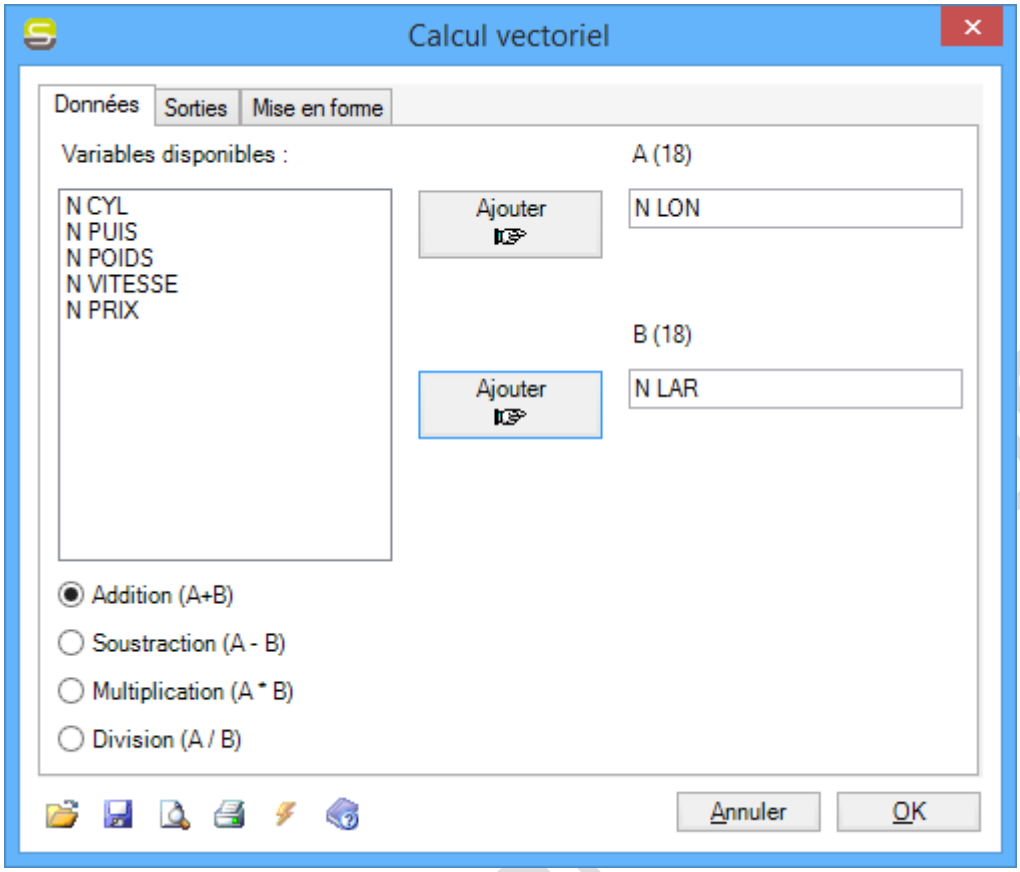
### Description

Voici les fonctions disponibles :

- Addition
- Multiplication
- Soustraction
- Division

### Mise en œuvre

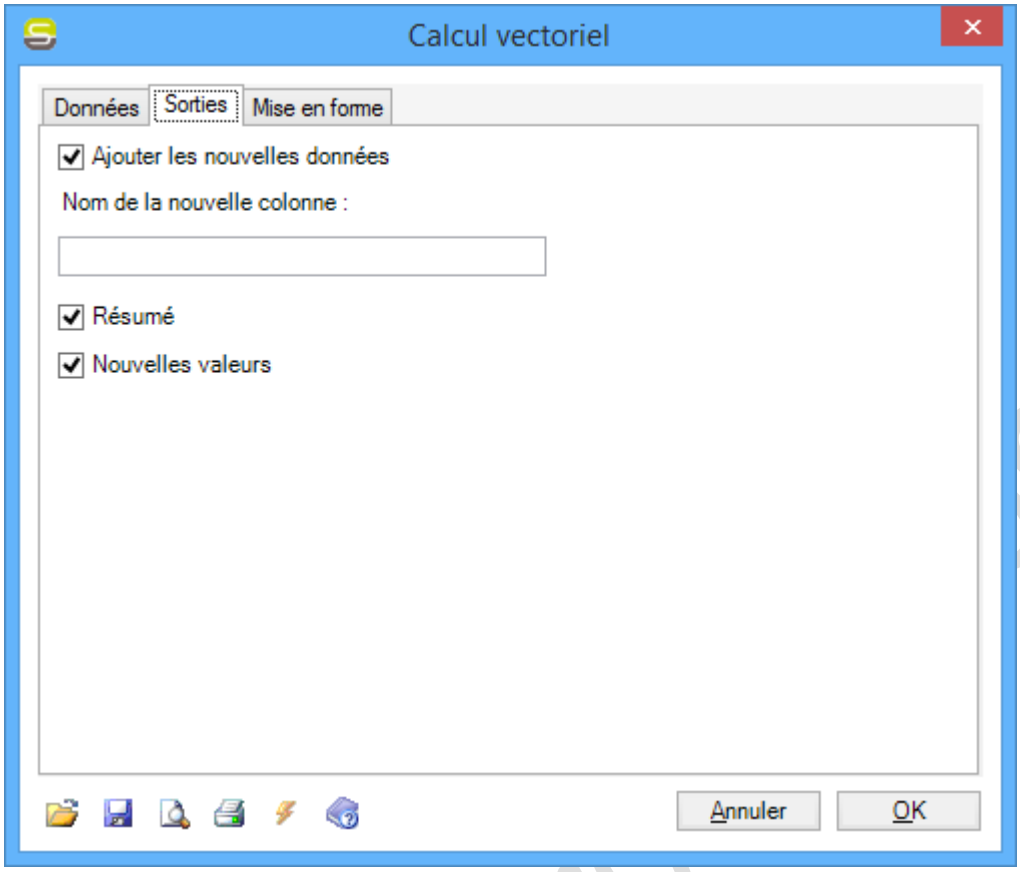
*Onglet « Données »*




Sélectionnez l'opération à effectuer.

Sélectionnez les variables à utiliser pour le calcul en les faisant passer dans les listes de droite.

*Onglet « Sorties »*




- *Ajouter les nouvelles données :* ajoute la colonne des valeurs calculées à la base d'origine. Vous pouvez donner un nom particulier à la nouvelle colonne ou laisser le logiciel déterminer le nouveau nom automatiquement.
- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Nouvelles valeurs :* affiche la table des données calculées.

## ÉCHANTILLONNAGE SIMPLE

### Description

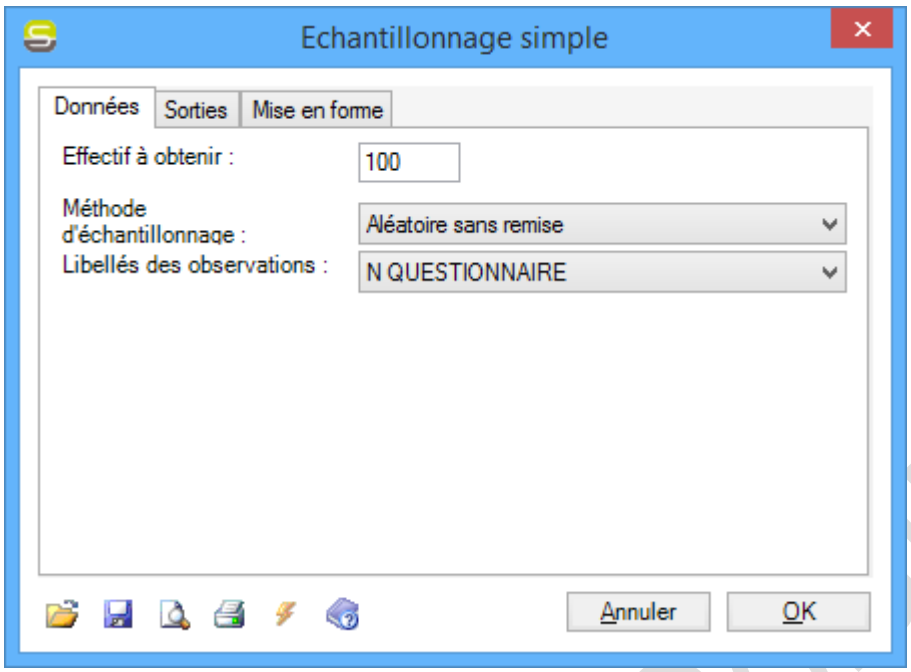
L'échantillonnage simple consiste à extraire un sous-ensemble d'observations du tableau initial par un tirage au hasard.

StatBox va créer une variable comportant les codes 0 et 1 : le code 1 étant celui correspondant à l'échantillon aléatoire et le code 0 pour l'échantillon complémentaire.

L'échantillon complémentaire est utile pour les phases d'apprentissage et de test de la modélisation.

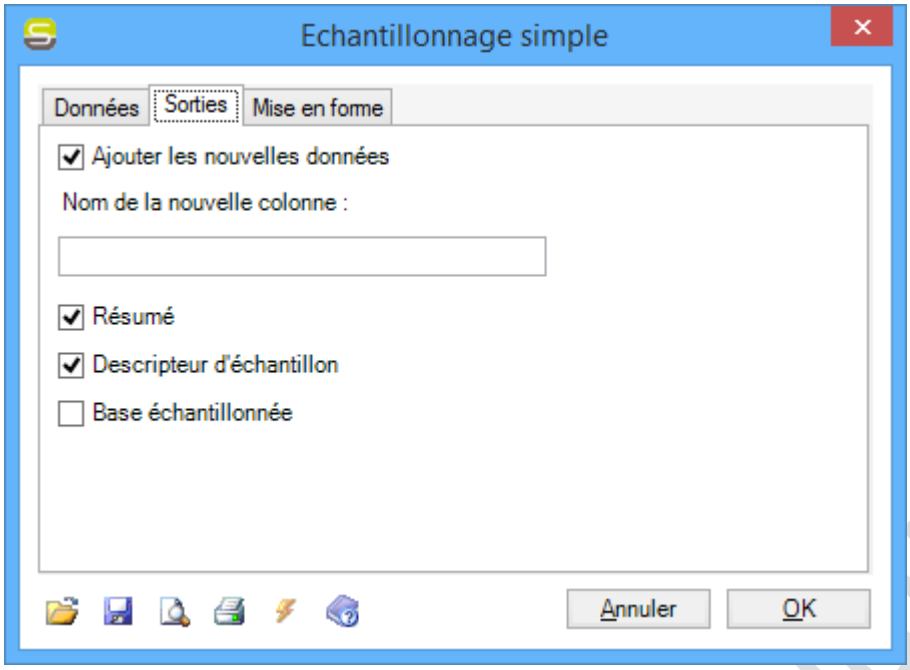
## Mise en œuvre

*Onglet « Données »*




- *Effectif à obtenir :* saisissez l'effectif que vous souhaitez obtenir dans le nouvel échantillon.
- *Méthode d'échantillonnage :* sélectionnez un mode d'échantillonnage parmi ceux proposés :
  - *aléatoire sans remise* : les observations sont sélectionnées au hasard et ne peuvent pas être sélectionnées plus d'une fois
  - *aléatoire avec remise* : les observations sont sélectionnées au hasard et peuvent être sélectionnées plus d'une fois. Les observations sélectionnées plusieurs fois ont alors un code échantillon correspondant au nombre de fois où elles ont été tirées
  - *systématique avec départ aléatoire* : les observations sont sélectionnées de manière consécutive à partir d'une ligne déterminé au hasard
  - *systématique centré* : les observations sont sélectionnées de manière consécutive à partir du centre de la base
  - *des premières valeurs* : les observations sont sélectionnées de manière consécutive à partir du début de la base
  - *des dernières valeurs* : les observations sont sélectionnées de manière consécutive à partir de la fin de la base
  - *aléatoire stratifié à un élément par strate* : la base est découpée en différente strates d'effectifs sensiblement égaux, une observation est alors sélectionnée pour chacune des strates
- *Libellés des observations :* sélectionnez la variable contenant les libellés des observations si vous souhaitez créer un tableau d'échantillonnage avec des libellés. Par défaut, le libellé d'une observation est son numéro de ligne dans le tableau.

*Onglet « Sorties »*




- *Ajouter les nouvelles données :* ajoute la colonne d'échantillonnage à la base d'origine. Vous pouvez donner un nom particulier à la nouvelle colonne ou laisser le logiciel déterminer le nouveau nom automatiquement.
- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Descripteur d'échantillon :* affiche la table des valeurs booléennes indiquant l'appartenance de chacune des observations à l'échantillon demandé.
- *Base échantillonnée :* génère une nouvelle feuille Excel correspondant à la base d'origine où seul figurent les observations retenues pour l'échantillon.

## ÉCHANTILLONNAGE PAR QUOTAS

### Description

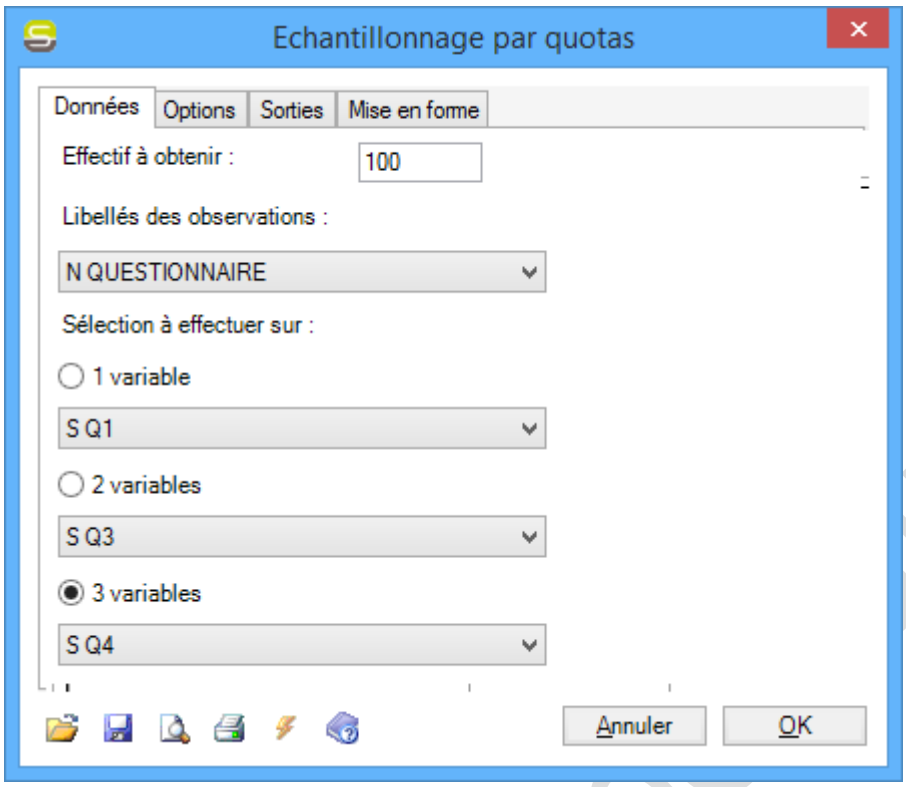
L'échantillonnage par quotas consiste à introduire une contrainte en plus par rapport à l'échantillonnage aléatoire. Il faut respecter une structure particulière sur 1, 2 ou 3 variables. Si on désire avoir un échantillon comportant 50% d'hommes et 50% de femmes, l'extraction devra respecter cette structure.

StatBox va créer une variable comportant les codes 0 et 1 : le code 1 étant celui correspondant à l'échantillon aléatoire et le code 0 pour l'échantillon complémentaire.

L'échantillon complémentaire est utile pour les phases d'apprentissage et de test de la modélisation.

## Mise en œuvre

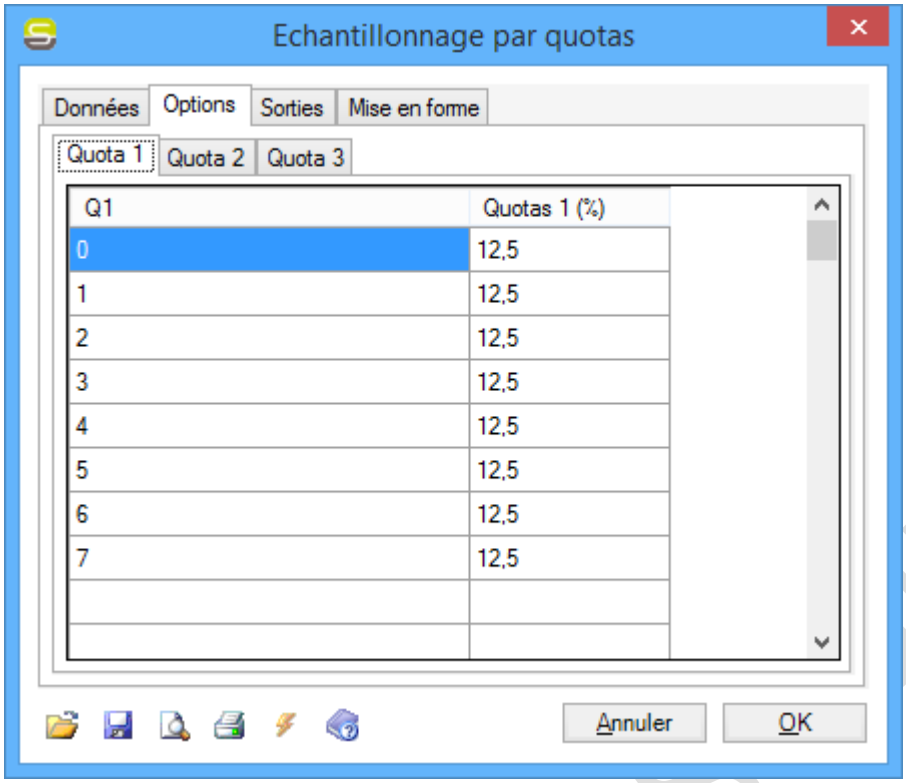
*Onglet « Données »*




- *Effectif à obtenir :* saisissez l'effectif que vous souhaitez obtenir dans le nouvel échantillon et qui respectera les quotas.
- *Libellés des observations :* sélectionnez la variable contenant le libellé des observations si vous souhaitez créer un tableau d'échantillonnage avec des libellés particuliers. Par défaut, le libellé d'une observation est son numéro de ligne dans le tableau.

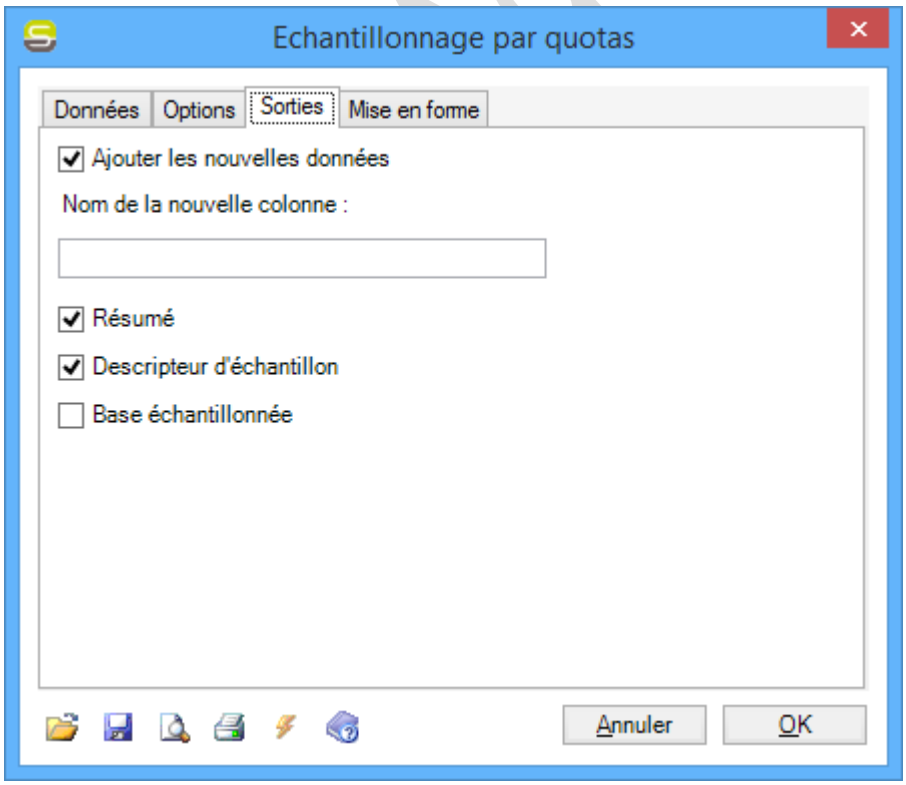
Sélectionnez le nombre de variables servant à l'échantillonnage et sélectionnez le nom de chaque variable.

*Onglet « Options »*




Saisissez pour chacune des variables servant à l'échantillonnage les structures à atteindre pour chacune des modalités en veillant à ce que la somme des quotas pour une variable atteigne 100%. Par exemple 50% d'hommes et 50% de femmes.

*Onglet « Sorties »*



- *Ajouter les nouvelles données :* ajoute la colonne d'échantillonnage à la base d'origine. Vous pouvez donner un nom particulier à la nouvelle colonne ou laisser le logiciel déterminer le nouveau nom automatiquement.
- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Descripteur d'échantillon :* affiche la table des valeurs booléennes indiquant l'appartenance de chacune des observations à l'échantillon demandé.
- *Base échantillonnée :* génère une nouvelle feuille Excel correspondant à la base d'origine où seul figurent les observations retenues pour l'échantillon.

**Remarque :** il est possible que l'effectif obtenu soit inférieur à celui demandé. Cela veut dire qu'il n'y avait pas suffisamment d'enregistrements répondant aux critères demandés.

## REDRESSEMENT

### Description

Lorsque la structure d'un échantillon ne correspond pas à la structure de la population mère, un redressement consiste à attribuer à chaque observation un poids destiné à contrebalancer l'effet de la sur-représentation ou de la sous-représentation de certains groupes dans l'échantillon.

Supposons qu'un échantillon d'enquête comporte trop d'inactifs. Dans le fichier « redressé », on attribuera aux actifs un poids supérieur à 1 et aux inactifs un poids inférieur à 1.

StatBox permet d'effectuer un redressement d'échantillon sur 1, 2, 3, 4, 5 ou 6 critères. Redresser sur un nombre de critères plus important risquerait de provoquer des distorsions plutôt qu'un redressement (dans la mesure où certaines cases seraient égales à 0 comme par exemple « être retraité » et « avoir moins de 18 ans »).

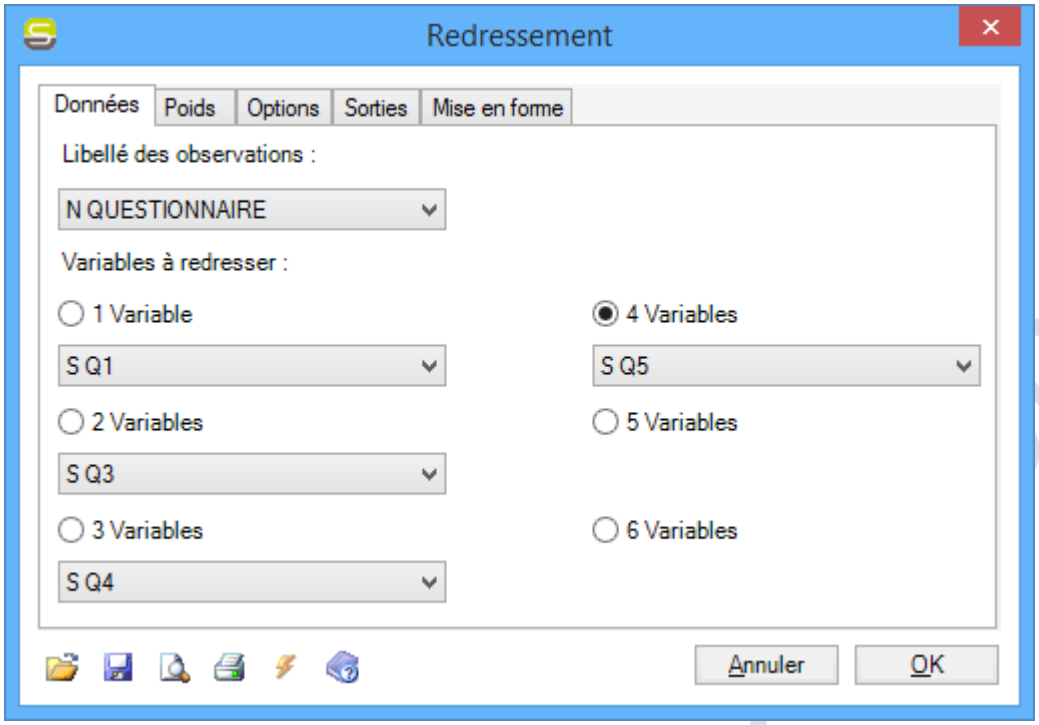
À partir d'une, deux ou trois variables nominales (ou qualitatives), ce module permet de calculer automatiquement le poids de chaque individu ou observation.

Une nouvelle colonne sera créée contenant le poids.

Pour évaluer l'importance du redressement à effectuer, vous pouvez au préalable effectuer un tri à plat des variables utilisées dans le redressement.

## Mise en œuvre

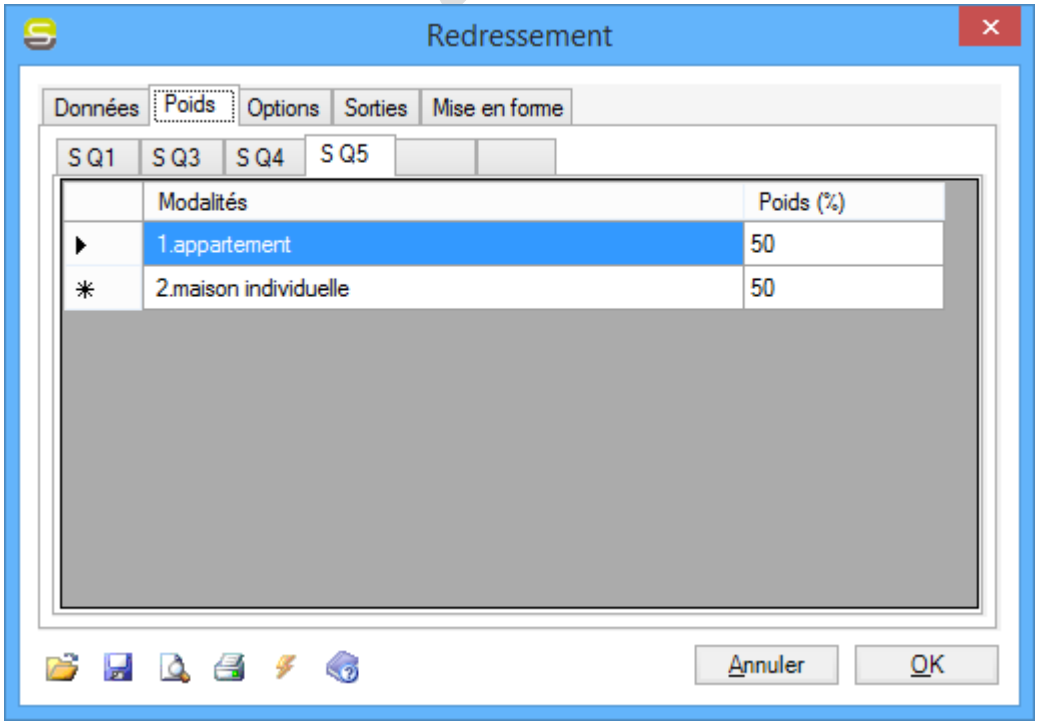
*Onglet « Données »*




- *Libellé des observations :* sélectionnez la variable contenant le libellé des observations si vous souhaitez créer un tableau de poids avec des libellés particuliers pour les observations. Par défaut, le libellé d'une observation est son numéro de ligne dans le tableau.

Cochez le nombre de variable servant au redressement et sélectionnez chacune d'entre elles.

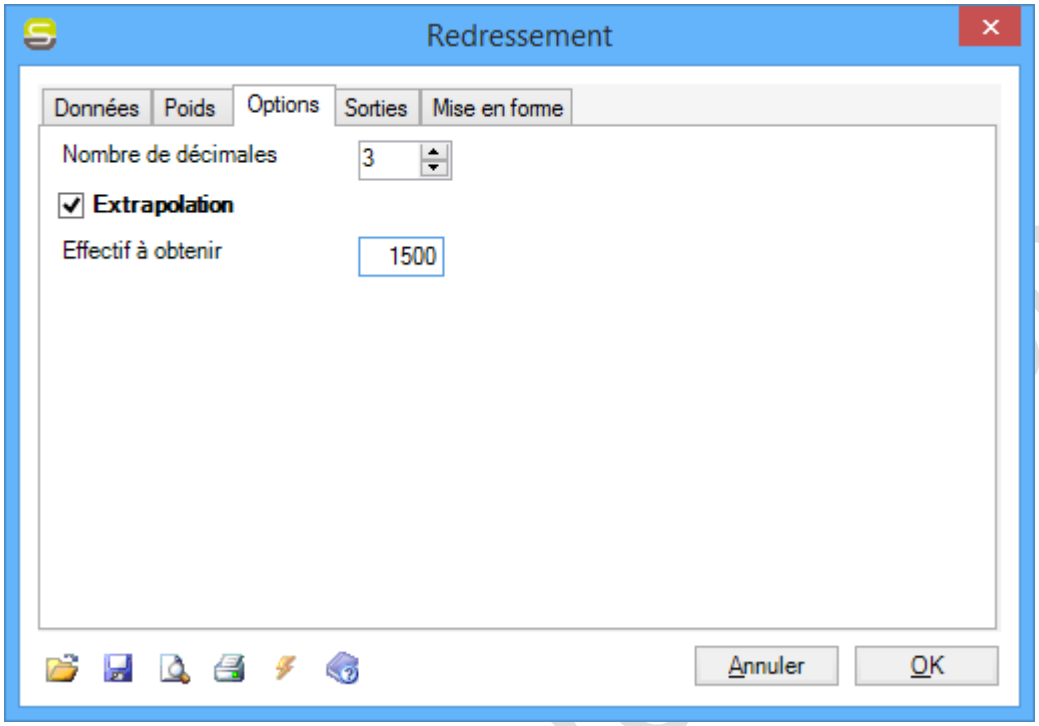
*Onglet « Poids »*



Pour chaque modalité des variables servant au redressement, introduisez les pourcentages théoriques à obtenir. Par exemple : 8% d'agriculteurs, 20% d'ouvriers,…

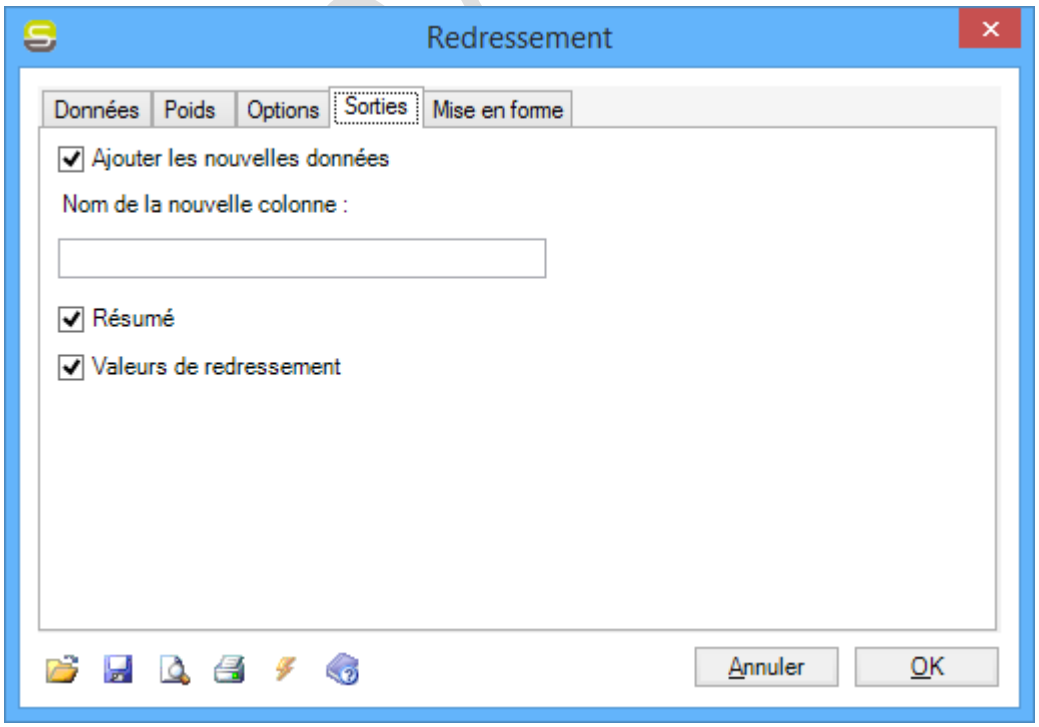
Si la somme de ces pourcentages est supérieure à 100 pour une variable, l'effectif redressé sera supérieur à l'effectif observé, vous pouvez ainsi effectuer des extrapolations.

*Onglet « Options »*




- *Nombre de décimales :* entrez le nombre de décimales de la colonne de poids à éditer.
- *Extrapolation :* cochez cette option pour que les poids édités vérifient les critères demandés pour une population de taille précise. Entrez alors la taille de la population cible.

*Onglet « Sorties »*



- *Ajouter les nouvelles données :* ajoute la colonne des poids à la base d'origine. Vous pouvez donner un nom particulier à la nouvelle colonne ou laisser le logiciel déterminer le nouveau nom automatiquement.
- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Valeurs de redressement :* affiche la table des poids correspondant au redressement demandé.

Remarques :

- Si StatBox fait de nombreuses itérations et qu'il ne trouve pas de solution car l'écart est trop grand, le redressement n'est pas effectué.
- Le nombre de modalités de vos variables ne doit pas être trop grand. Faites, au préalable, un regroupement de modalités.
- N'utilisez pas 2 critères parfaitement dépendants. Par exemple, les départements regroupés de 2 manières différentes. Les critères de redressement doivent en revanche être corrélés avec le phénomène étudié.
- Vous pouvez vérifier le résultat du redressement en effectuant un tri à plat sur les variables ayant servi au redressement et en sélectionnant la colonne générée comme variable de poids.

## CRÉATION D'UNE DISTRIBUTION

Utilisez ce module pour générer des données aléatoires à partir d'une distribution théorique. Vous devez choisir la loi de probabilité et fixer ses paramètres. Vous pouvez également éditer des données vérifiant l'appartenance à une distribution empirique.

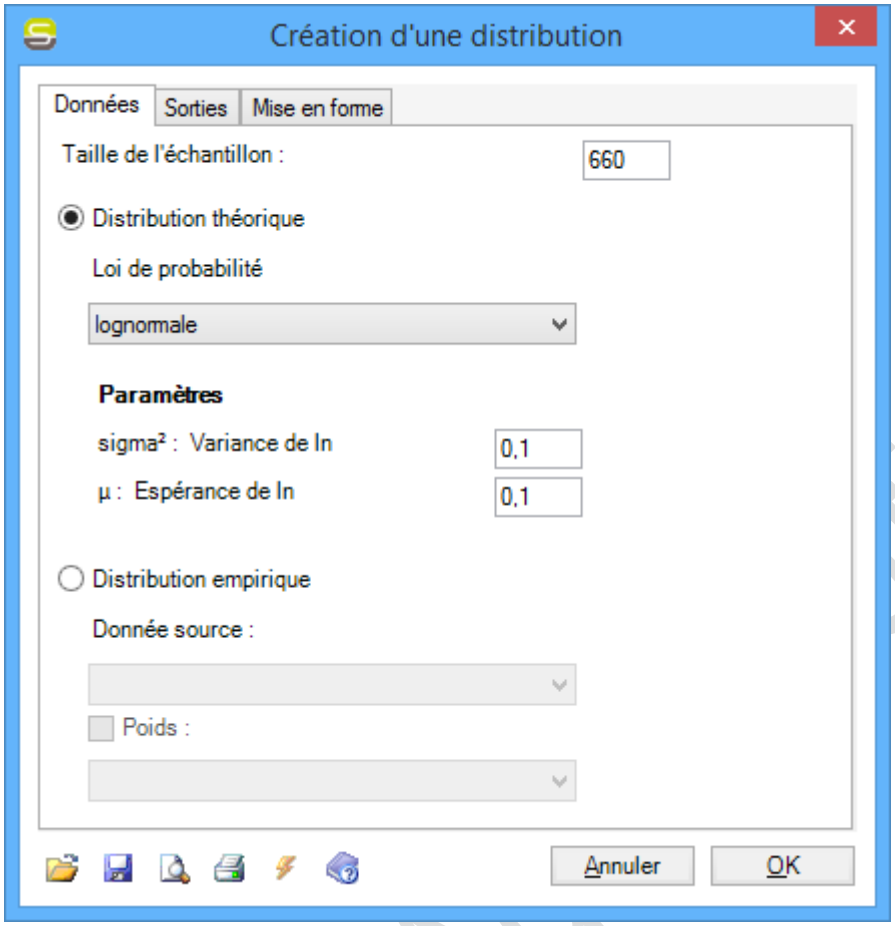
### Description

Plusieurs lois de probabilité sont disponibles :

- uniforme,
- gaussienne standard,
- gaussienne,
- lognormale,
- de Student,
- de Fisher,
- du khi²,
- Bêta,
- exponentielle,
- de Poisson,
- binomiale,
- binomiale négative.

### Mise en œuvre

*Onglet « Données »*




- *Taille de l'échantillon :* entrez le nombre de valeurs à générer
- *« Distribution théorique » / « Distribution empirique » :* cochez si les données à générer doivent vérifier l'appartenance à une distribution théorique ou issue de données que vous fournissez.

Pour une distribution théorique

Loi de probabilité : sélectionnez une loi de distribution et modifiez au besoin les paramètres par défaut :

- uniforme
- a : nombre définissant la borne inférieure de l'intervalle de la loi uniforme
- b : nombre définissant la borne supérieure de l'intervalle de la loi uniforme
- gaussienne standard (ou loi normale centrée et réduite) : loi de Gauss de moyenne nulle et de variance unité
- gaussienne (ou loi normale)
- μ : valeur de l'espérance
- sigma² : valeur de la variance
- log normale (le logarithme de la variable distribuée selon une loi lognormale suit la loi normale de paramètres μ et sigma²)
- μ : valeur de l'espérance de la loi normale selon laquelle est distribué ln(x)
- sigma² : valeur de la variance de la loi normale selon laquelle est distribué ln(x)
- de Student
- ddl : nombre de degrés de liberté de la loi de Student

- de Fisher
- ddl 1 : nombre de degrés de liberté du numérateur du F de Fisher
- ddl 2 : nombre de degrés de liberté du dénominateur du F de Fisher
- du khi²
- ddl : nombre de degrés de liberté de la loi du khi²
- Bêta
- a1 : nombre correspondant au premier paramètre de forme de la loi Bêta
- a2 : nombre correspondant au deuxième paramètre de forme de la loi Bêta
- exponentielle
- Lambda : inverse du temps d'attente moyen entre deux événements d'un phénomène aléatoire pour la loi exponentielle
- de Poisson
- Lambda : valeur moyenne supérieure à 0 pour définir la loi de Poisson
- Binomiale
- n : nombre d'essais définissant la loi binomiale
- p : probabilité de succès définissant la loi binomiale

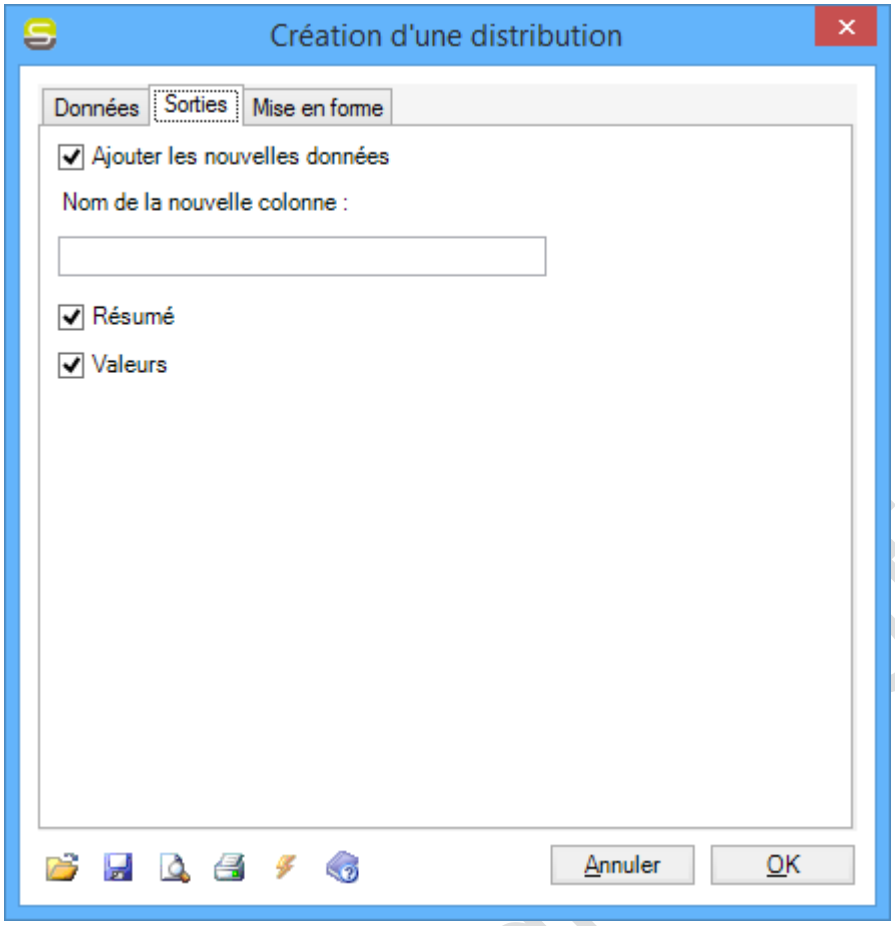
**Remarque :** la loi de Bernoulli est un cas particulier de la loi binomiale pour p = 0,5.

- binomiale négative
- k : nombre de succès définissant la loi binomiale négative
- p : probabilité de succès définissant la loi binomiale négative

Pour une distribution empirique

- *Données sources :* sélectionnez la variable décrivant la distribution à vérifier.
- *Poids :* cochez cette option pour pondérer vos observations et sélectionnez une variable contenant des poids.

*Onglet « Sorties »*




- *Ajouter les nouvelles données :* ajoute la colonne des nouvelles données à la base d'origine. Vous pouvez donner un nom particulier à la nouvelle colonne ou laisser le logiciel déterminer le nouveau nom automatiquement.
- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Valeurs :* affiche la table des données générées.

## Références

**Abramowitz M. & I.A. Stegun (1972)**. Handbook of mathematical functions. Dover Publications, New York, pp. 927-964.

**Aïvazian S., I. Enukov & L. Mechalkine (1986)**. Eléments de modélisation et traitement primaire des données. Mir, Moscou, pp. 126-183.

**Manoukian E.B. (1986)**. Guide de statistique appliquée. Hermann, Paris, pp. 19-68.

**Ripley B.D. (1983)**. Computer generation of random variables: a tutorial. *International Statistical Review*, **51** : 301-319.

**Ripley B.D. (1987)**. Stochastic simulation. John Wiley & Sons, New York.

**Saporta G. (1990)**. Probabilités, analyse des données et statistique. Technip, Paris, pp. 30-56.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 62-65.

# REPRÉSENTATIONS GRAPHIQUES

## STATISTIQUES DESCRIPTIVES

Utilisez ce module pour calculer un ensemble de statistiques descriptives pour une ou plusieurs variables quantitatives, et produire des représentations graphiques ou semi-graphiques utilisées en analyse exploratoire des données.

### Description

Liste des statistiques calculées dans le cas des données quantitatives (les descripteurs qui tiennent compte des poids éventuels sont figurés en gras) :

- **Nombre de valeurs utilisées** : nombre de valeurs effectivement utilisées dans les calculs, c'est-à-dire les valeurs non manquantes et de poids différent de 0,
- **Nombre de valeurs ignorées** : nombre de valeurs ignorées lors des calculs, c'est-à-dire les valeurs manquantes ou de poids nul,
- **Nombre de val**. : min. nombre de valeurs égales à la valeur minimale,
- **% de val. Min.** : pourcentage du nombre de valeurs égales à la valeur minimale,
- **Minimum** : valeur minimale,
- **1er quartile** : valeur en deçà de laquelle se trouvent 25 % des données,
- **Médiane** : valeur en deçà de laquelle se trouvent 50 % des données,
- **3ème quartile** : valeur en deçà de laquelle se trouvent 75 % des données,
- **Maximum** : valeur maximale,
- **Étendue** : différence entre le maximum et le minimum,
- **Somme des poids** : dans le cas de données pondérées, indique la somme des poids des valeurs utilisées dans les calculs,
- **Total** : somme des valeurs, éventuellement pondérée,
- **Moyenne** : somme des valeurs, éventuellement pondérée, divisée par le nombre de valeurs utilisées, ou par la somme des poids si les données sont pondérées,
- **Moyenne géométrique :** moyenne peu influencée par les valeurs élevées. La moyenne géométrique n'est pas définie pour les données contenant des valeurs négatives ou nulles,
- **Moyenne harmonique** : moyenne peu influencée par quelques valeurs beaucoup plus élevées que l'ensemble des autres valeurs, mais sensible aux valeurs beaucoup plus petites. La moyenne harmonique n'est pas définie pour les données contenant des valeurs nulles,
- **Aplatissement (Pearson)** : coefficient caractérisant la forme de pic ou l'aplatissement d'une distribution par rapport à une distribution gaussienne. Pour une distribution gaussienne (loi normale), l'aplatissement vaut 0. Une valeur négative correspond à une distribution plus plate que la loi normale (distribution *platicurtique*) tandis qu'une valeur positive correspond à une distribution plus pointue que la loi normale (distribution *leptocurtique*),
- **Asymétrie (Pearson)** : coefficient caractérisant le degré d'asymétrie d'une distribution par rapport à sa moyenne. Pour une distribution gaussienne (loi normale), l'asymétrie vaut 0. Une valeur négative correspond à la présence d'une queue de distribution vers la gauche tandis qu'une valeur positive correspond à une queue de distribution vers la droite,
- **Aplatissement** : coefficient d'aplatissement tel qu'il est calculé par Excel,

- **Asymétrie** : coefficient d'asymétrie tel qu'il est calculé par Excel,
- **CV (écart-type/moyenne)** : coefficient de variation mesurant la dispersion relative obtenue en divisant l'écart-type par la moyenne. Ce coefficient permet de comparer la dispersion de variables dont les unités sont différentes, ou qui ont des moyennes très différentes,
- **Variance d'échantillon** : variance des données (dans le cas de données non pondérées, le dénominateur est *n*, effectif de l'échantillon),
- **Variance estimée** : estimation de la variance d'une population dont les données constituent un échantillon (estimateur sans biais : dans le cas de données non pondérées, le dénominateur est *n*-1, avec *n* l'effectif de l'échantillon),
- **Écart-type d'échantillon** : racine carrée de la variance des données,
- **Écart-type estimé** : racine carrée de l'estimation de la variance de la population d'origine des données,
- **Écart absolu moyen** : mesure de dispersion indiquant la moyenne des valeurs absolues des écarts de chaque valeur par rapport à la moyenne,
- **Écart-type de la moyenne** : racine carrée du rapport de la variance estimée par le nombre de valeurs utilisées dans les calculs. Cette estimation de la variance de la moyenne n'est valide que si les données constituent un échantillon prélevé au hasard (et sans remise) au sein d'une population infinie (*échantillon aléatoire simple* d'une population infinie).

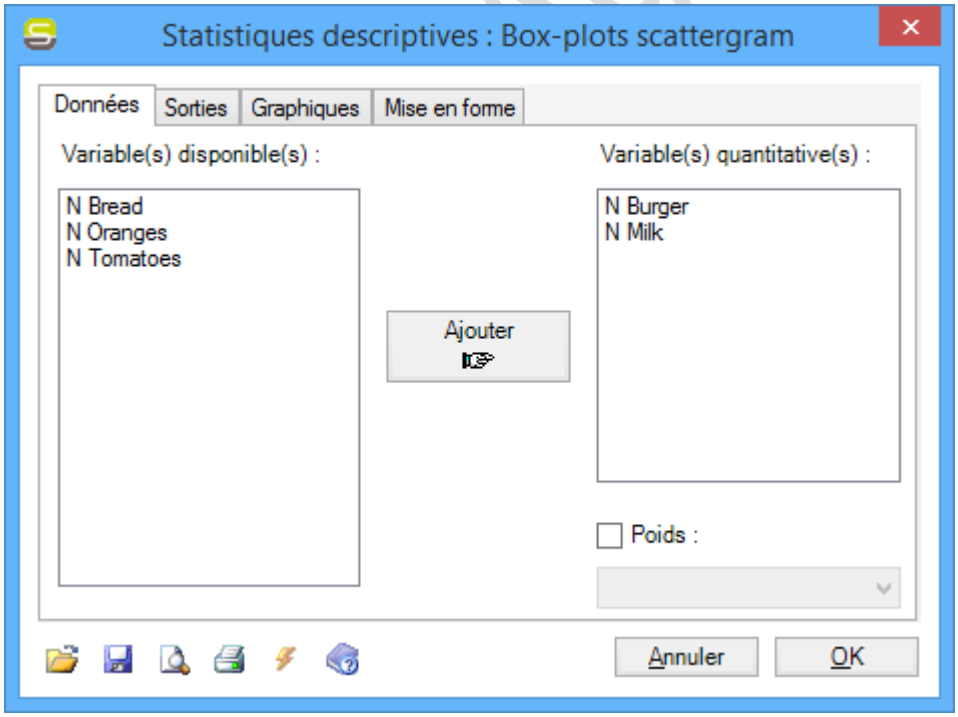
Graphiques produits :

- graphiques « boîte à moustaches » ou box plots,
- nuages de points univariés ou scattergrams,
- diagrammes « tige et feuille » ou stem and leaf plots.

Pour une aide à l'interprétation de ces graphiques, consultez l'annexe *« Graphiques de l'analyse exploratoire ».*

## Mise en œuvre

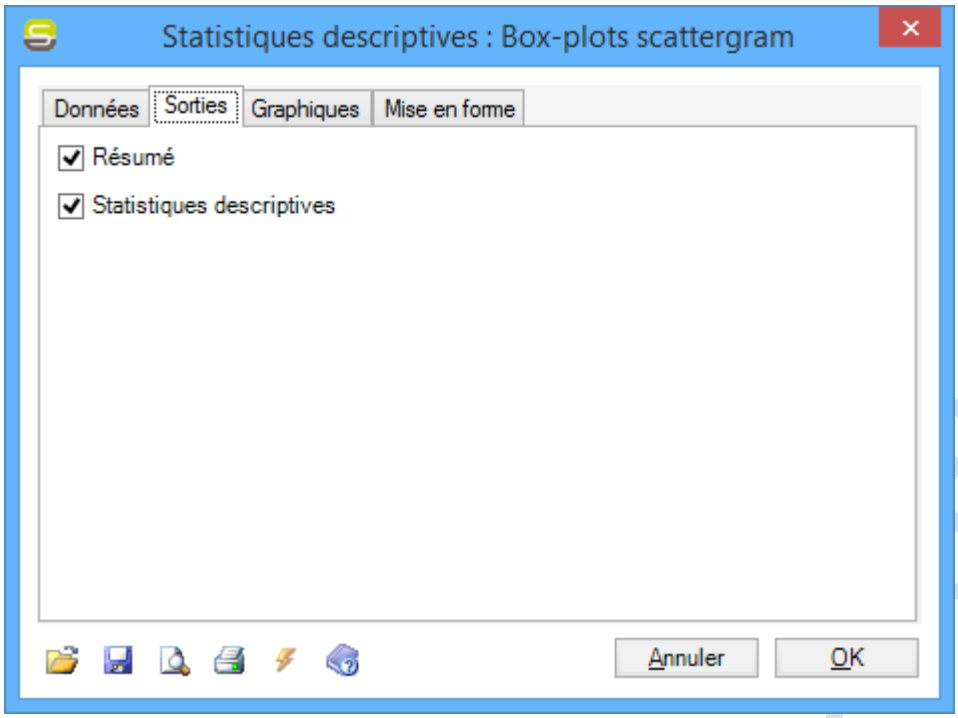
*Onglet « Données »*




➢ *Variable(s) quantitative(s) :* sélectionnez les variables à décrire. Lorsqu'il y a des valeurs manquantes dans une colonne, StatBox propose de les ignorer. En cas de refus, le traitement est abandonné.

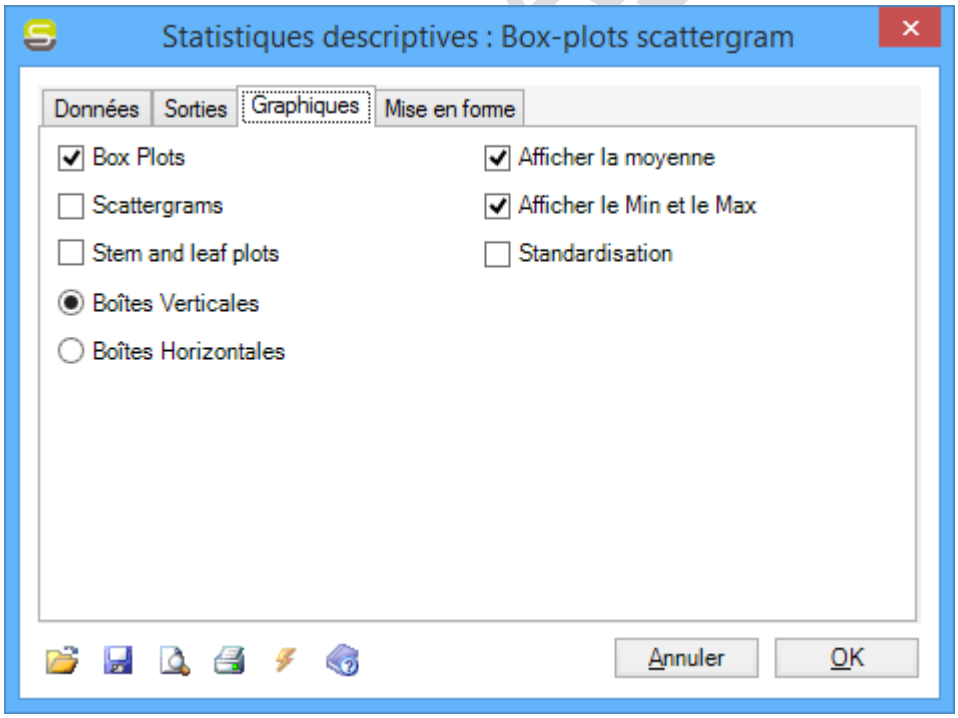
- *Poids :* cochez cette case si vous désirez pondérer les données, puis sélectionnez la variable des poids. Les valeurs manquantes dans les poids sont mises à zéro et conduisent par conséquent à l'inactivation de la ligne correspondante.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Statistiques descriptives :* cochez cette option pour afficher la table des statistiques descriptives pour chacune des variables sélectionnées.

*Onglet « Graphiques »*




- *Box plots :* cochez cette case pour obtenir les graphiques "boîte à moustaches". Ces graphiques ne peuvent pas être affichés s'il y a plus de 16 variables ou plus de 30 000 points.

- *Scattergrams :* cochez cette case pour obtenir les nuages de points univariés. Ces graphiques ne peuvent pas être affichés s'il y a plus de 24 variables ou plus de 30 000 points.
- *Stem and leaf plots :* cochez cette case pour obtenir les diagrammes " tige et feuille ". Ce graphique ne peut pas être produit lorsqu'une variable poids est sélectionné.
- *"Boîtes verticales" / "Boîtes horizontales" :* choisissez l'orientation des box plots et des scattergrams.
- *Afficher la moyenne :* affiche la moyenne sur les box plots et les scattergrams. Cette option n'est pas disponible lorsque l'option « Standardisation » est cochée.
- *Afficher le Min et le Max :* affiche la valeur minimum et la valeur maximum sur les box plots. Cette option n'est pas disponible lorsque l'option « Standardisation » est cochée.
- *Standardisation :* supprime l'effet des différences d'ordre de grandeur entre les variables lors de la production des box plots et des scattergrams, en divisant les valeurs de chaque variable par l'écart-type correspondant.

### Références

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 28-30, 39-60, 151-152.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, p. 115-121.

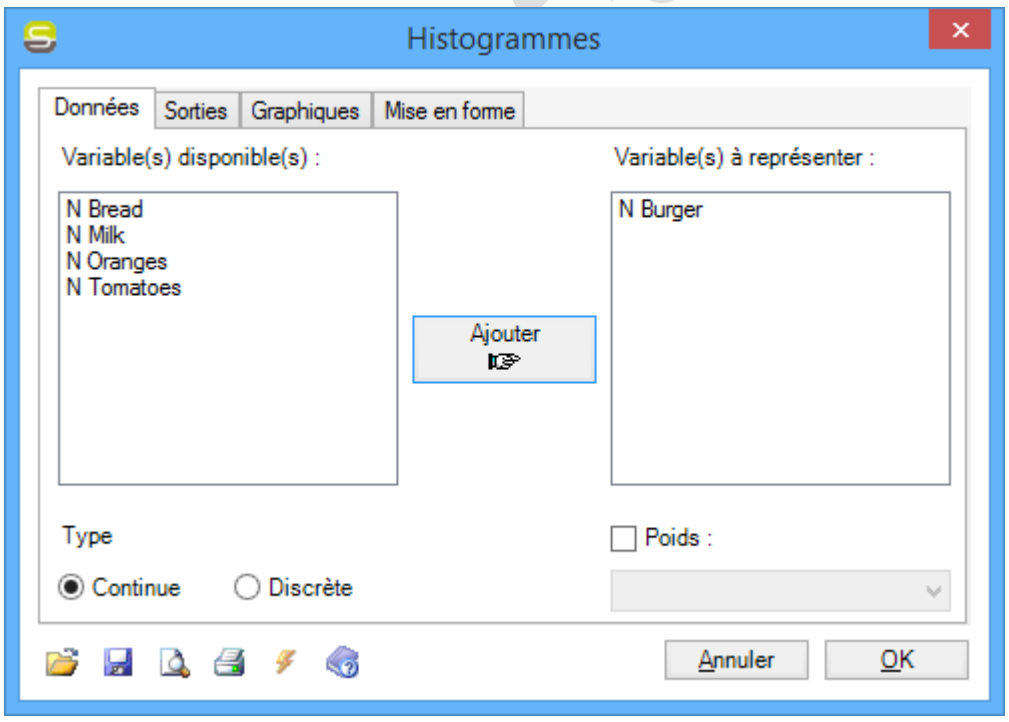
## HISTOGRAMMES

Utilisez ce module pour afficher l'histogramme des fréquences approximant la fonction de densité de probabilité d'une variable quantitative et la distribution des fréquences cumulées approximant sa fonction de répartition.

Le module permet également de produire des histogrammes, en utilisant différentes méthodes, et de modifier les bornes manuellement.

### Mise en œuvre

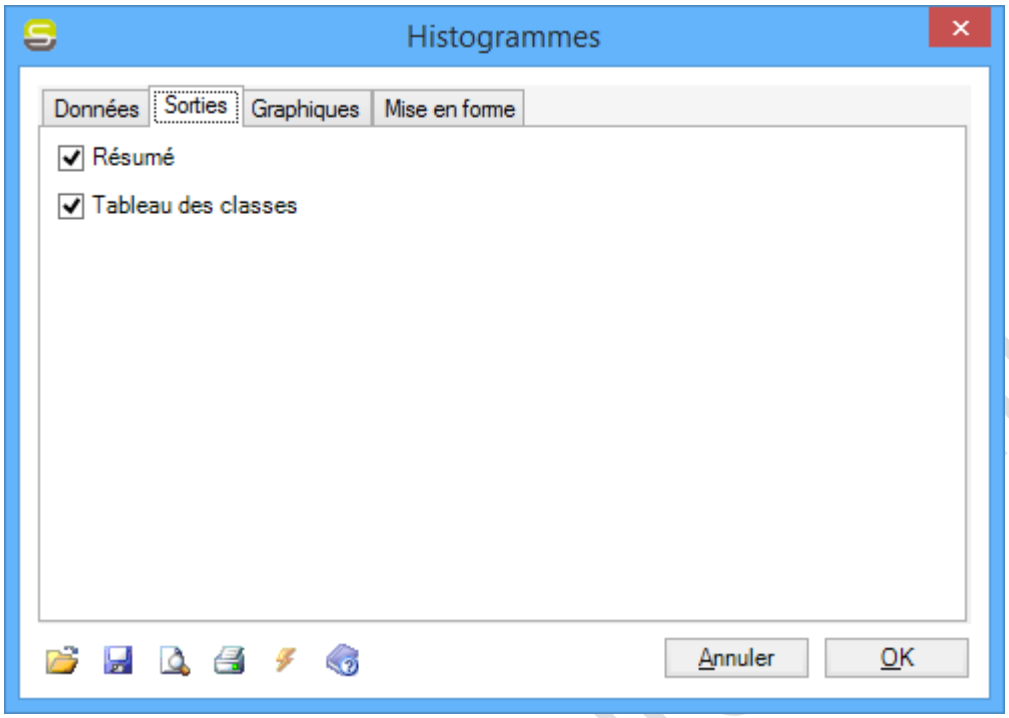
*Onglet « Données »*




- *Type :* sélectionnez si les données à représenter sont des variables continues (numériques) ou discrètes (ordinales).

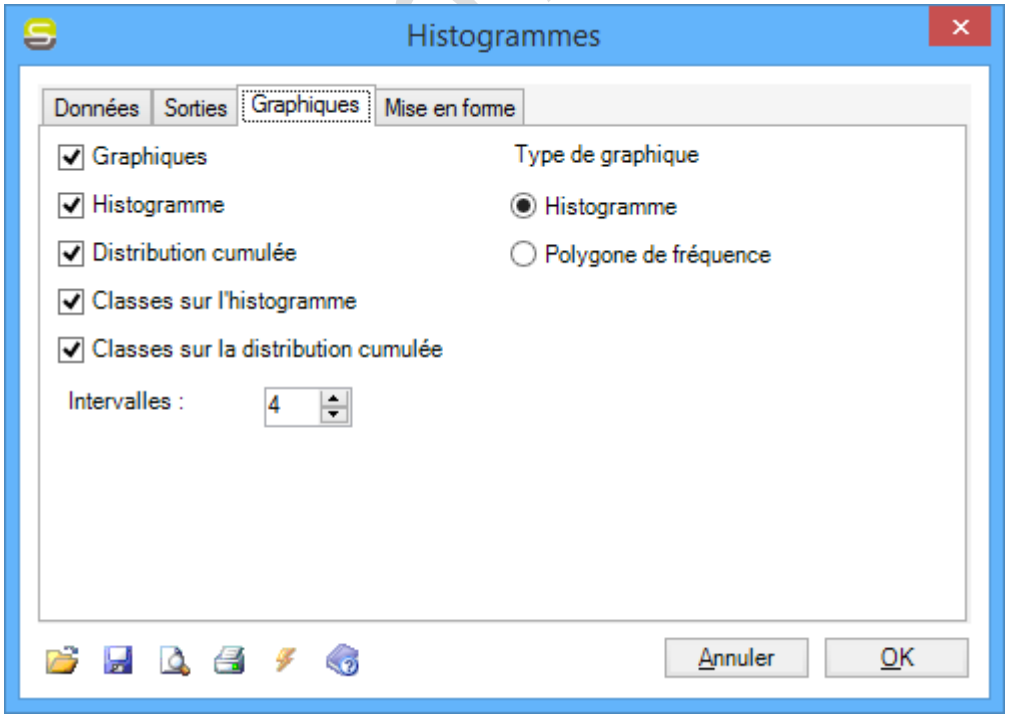
- *Variable(s) à représenter :* sélectionnez la/les variable(s) quantitative(s) à représenter. Lorsqu'il y a des valeurs manquantes, StatBox propose d'ignorer les lignes concernées. En cas de refus, le traitement est abandonné.
- *Poids :* cochez cette case si vous désirez pondérer les données, puis sélectionnez la variable contenant les poids. Les valeurs manquantes dans les poids sont cumulées avec les valeurs manquantes dans les données.

### *Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Tableau des classes :* affiche la table de répartition des observations dans les différentes classes.

### *Onglet « Graphiques »*




- *Graphiques :* cochez cette option pour afficher les graphiques.

- *Histogramme :* cochez cette option pour construire l'histogramme des fréquences par classe.
- *Distribution cumulée :* cochez cette option pour construire la distribution cumulée.
- *Classes sur l'histogramme :* cochez cette option si vous désirez un histogramme avec des barres verticales matérialisant les bornes des intervalles.
- *Classes sur l'histogramme :* cochez cette option si vous désirez une distribution cumulée avec des barres verticales matérialisant les bornes des intervalles.
- *Intervalles :* entrez le nombre d'intervalles d'amplitude constante pour la construction de *l'histogramme.*
- *« Histogramme » / « Polygone de fréquence » :* choisissez le mode de représentation graphique. Le tracé décrit les intervalles lorsque « Histogramme » est sélectionné et joint les centres des intervalles lorsque « Polygone de fréquence » est sélectionné.

### Références

**Frontier S. (1981)**. Méthode statistique. Masson, Paris, pp. 42-59.

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 19-32.

## NUAGES DE POINTS

Utilisez ce module pour calculer un ensemble de statistiques descriptives pour une ou plusieurs variables quantitatives et produire des représentations graphiques en analyse exploratoire des données.

### Description

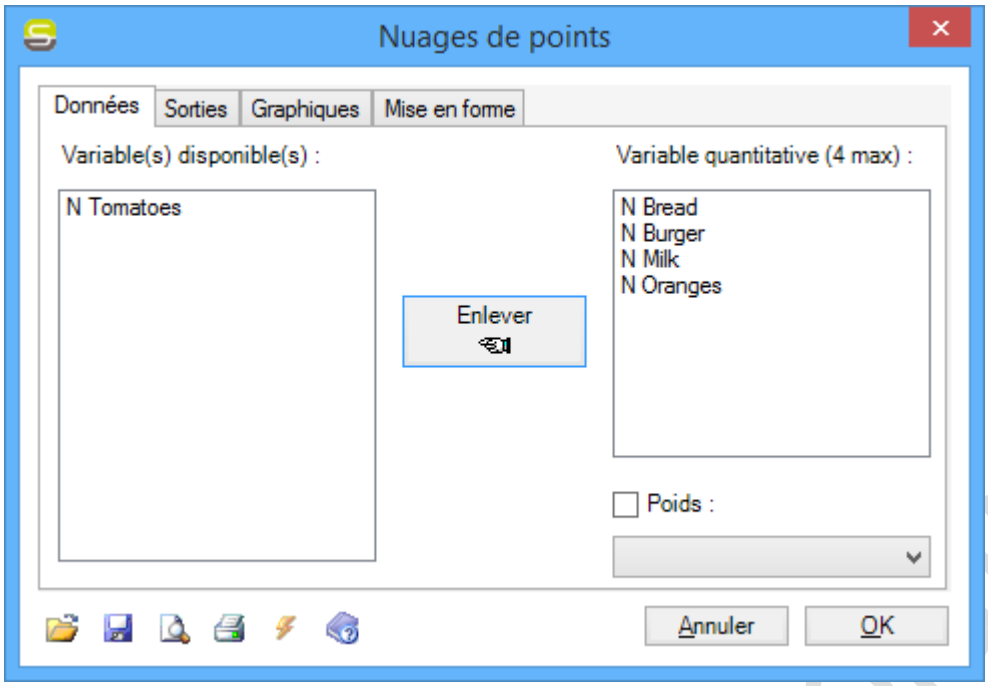
La liste des statistiques calculées est identiques à celle produit dans le cas de la méthode « Statistiques descriptives ».

Graphiques produits :

- collection de nuages de points bivariés XY,
- graphiques « Quantile-Quantile » ou Q-Q plots,
- graphiques « probabilité- probabilité » ou p-p plots,

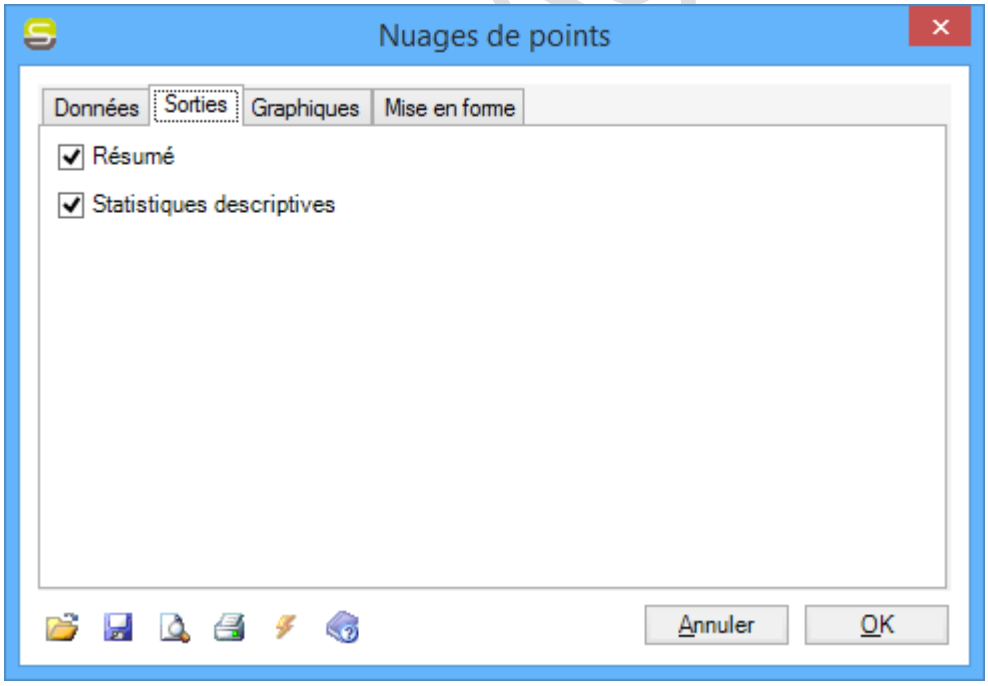
## Mise en œuvre

### *Onglet « Données »*



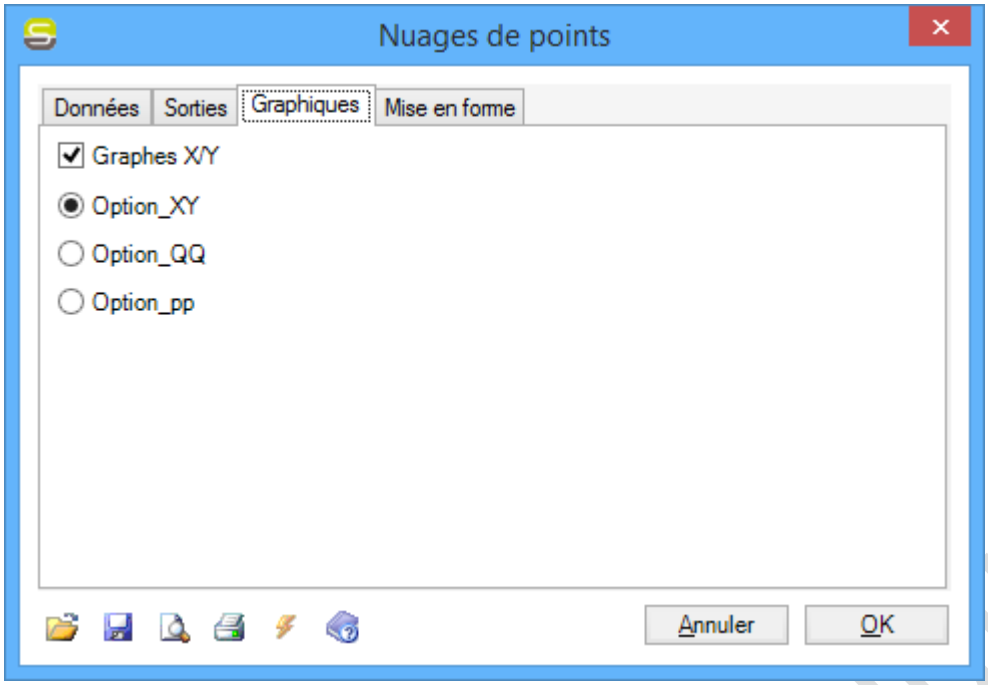

- *Données :* sélectionnez les variables à décrire (2 minimum, 4 maximum). Lorsqu'il y a des valeurs manquantes dans une colonne, StatBox propose de les ignorer. En cas de refus, la boîte de dialogue est fermée et le traitement est abandonné.
- *Poids :* cochez cette case si vous désirez pondérer les données, puis sélectionnez la variable de poids. Les valeurs manquantes dans les poids sont mises à zéro et conduisent par conséquent à l'inactivation de la ligne correspondante.

### *Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport
- *Statistiques descriptives :* affiche la table des statistiques descriptives pour chacune des variables sélectionnées.

*Onglet « Graphiques »*




- *Graphes X/Y :* affiche la collection de nuages bivariés obtenus en croisant deux à deux toutes les variables quantitatives sélectionnées.
- *« Option_XY » / « Option_QQ » / « Option_pp »:* choisissez entre l'affichage de la collection de nuages bivariés, y compris ceux croisant chaque variable avec elle-même, et l'affichage de la collection de nuages bivariés et des *Q-Q plots* ou des *p-p plots* pour toutes les variables. Ces graphiques ne peuvent pas être affichés s'il y a plus de 30 000 points.

### Références

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 28-30, 39-60, 151-152.

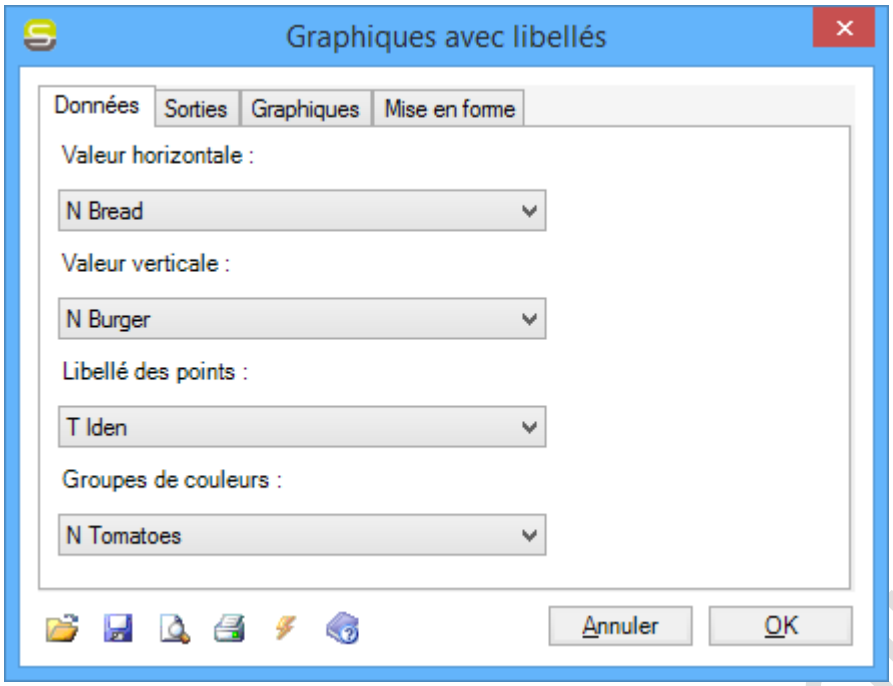
**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, p. 115-121.

## GRAPHIQUE AVEC LIBELLÉS

Utilisez ce module pour représenter simultanément 2 variables quantitatives sous la forme d'un nuage de points bivarié et une variable qualitative ou de « groupe » permettant de colorer chacune des observations selon leur appartenance à tel ou tel groupe. Les observations sont identifiées sur le graphique par leur libellé.

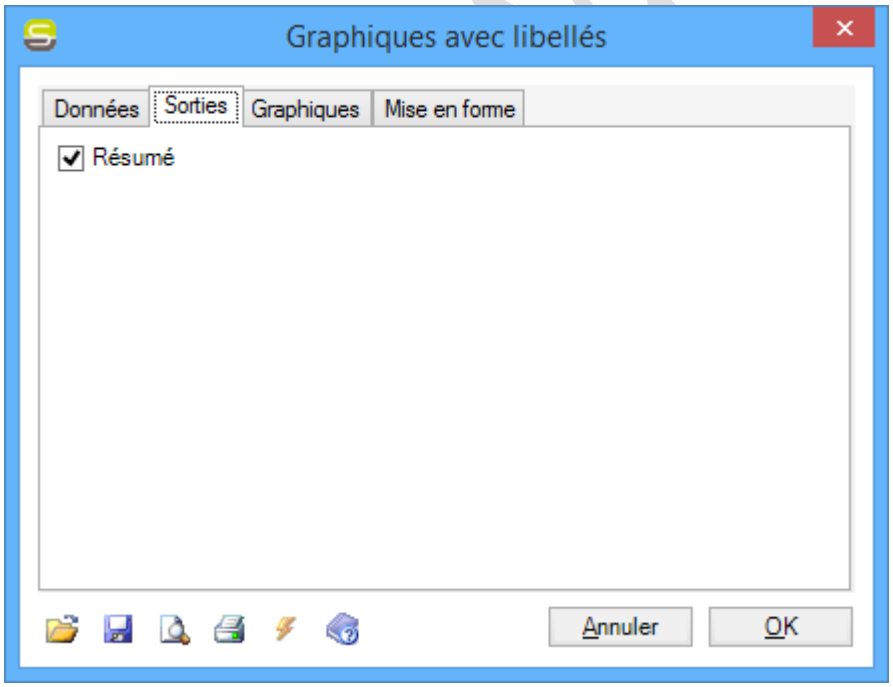
## Mise en œuvre

### *Onglet « Données »*



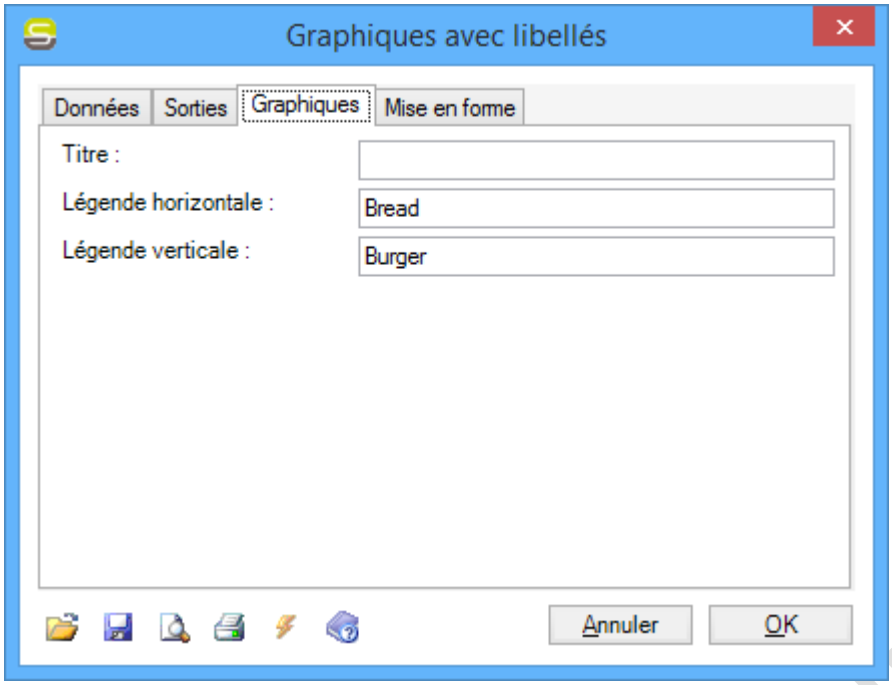

- *Valeur horizontale :* sélectionnez la variable numérique à représenter en abscisse.
- *Valeur vertical :* sélectionnez la variable numérique à représenter en ordonnée.
- *Libellé des points :* sélectionnez la variable contenant les libellés des observations.
- *Groupes de couleurs :* sélectionnez la variable qualitative contenant le descripteur de groupe. Les observations sont colorées en fonction de leur appartenance à telle ou telle modalité de cette variable.

### *Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.

*Onglet « Graphiques »*




- *Titre :* entrez un titre pour le graphique (facultatif).
- *Légende horizontale :* entrez une légende pour l'axe des abscisses (facultatif).
- *Légende verticale :* entrez une légende pour l'axe des ordonnées (facultatif) .

# ANALYSE SUR UNE VARIABLE

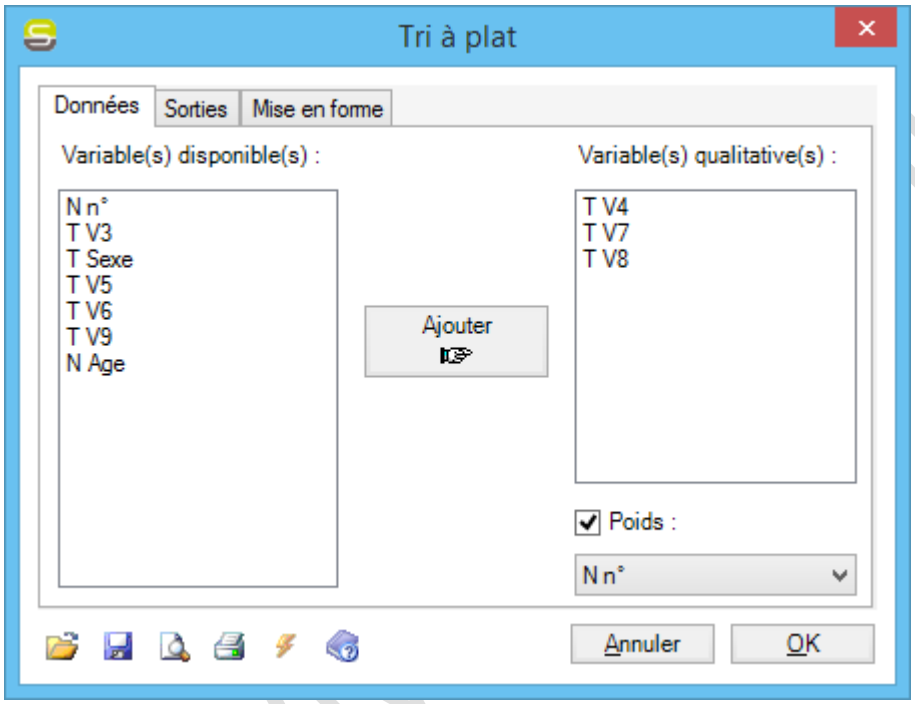
## TRI À PLAT

### Description

Ce module permet de faire un comptage des modalités d'une variable qualitative. Les effectifs et les pourcentages apparaissent dans un tableau de résultats. Des histogrammes et des graphiques en secteurs peuvent être ajoutés aux résultats. En cas de pondération, on nommera poids les fréquences pondérées.

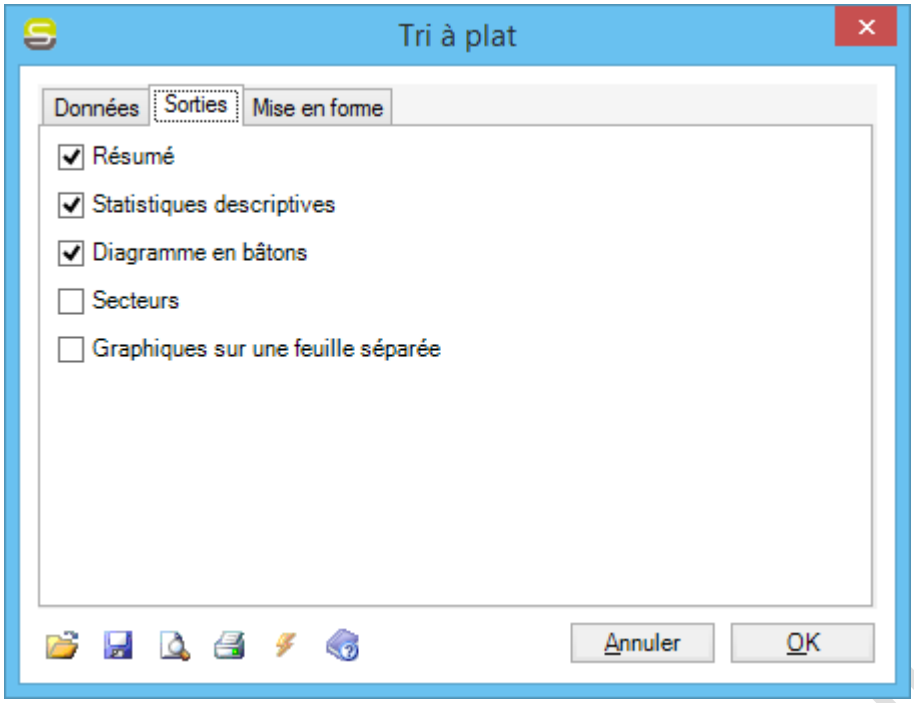
### Mise en œuvre

*Onglet « Données »*




- *Variable(s) qualitative(s) :* sélectionnez les variables à décrire. Lorsqu'il y a des valeurs manquantes dans une colonne, StatBox propose de les ignorer. En cas de refus le traitement est abandonné.
- *Poids :* cochez cette case si vous désirez pondérer les données, puis sélectionnez la variable de poids. Les valeurs manquantes dans les poids sont mises à zéro et conduisent par conséquent à l'inactivation de la ligne correspondante.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Statistiques descriptives :* affiche les tables de dénombrement et de fréquence des différentes modalités de chaque variable.
- *Diagramme en bâtons :* affiche un histogramme de répartition des différentes modalités de chaque variable.
- *Secteurs :* affiche un diagramme en secteurs de répartition des différentes modalités de chaque variable.
- *Graphiques sur une feuille séparée :* affiche tous les graphiques sur une feuille indépendante.

### Références

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 28-30, 39-60, 151-152.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, p. 115-121.

## STATISTIQUES DESCRIPTIVES

Consultez le paragraphe « Statistiques descriptives » de la section « Représentations graphiques ».

## HISTOGRAMMES

Consultez le paragraphe « Histogrammes » de la section « Représentations graphiques ».

## PRÉVISION À COURT TERME

### Principes

La première étape consiste à isoler des chiffres bruts la composante de tendance, de la manière la plus pure possible.

Pour cela un premier filtrage par lissage exponentiel permet de diminuer la composante aléatoire, il est suivi d'un deuxième filtrage par moyenne mobile qui élimine les variations saisonnières.

La deuxième étape est celle de la modélisation de la tendance et de sa prévision.

La méthode de modélisation repose sur l'ajustement de la tendance par son approximation par une série de polynômes orthogonaux.

La troisième étape est celle de la traduction des prévisions de tendance en prévisions brutes.

Pour cela les filtres de la première étape sont appliqués « à l'envers », ils travaillent alors comme des amplificateurs.

Situons d'abord ce que nous entendons par le terme « prévision ».

La prévision au sens statistique est « ce qui doit arriver si tout reste égal par ailleurs »

La prévision part de l'idée que dans le passé il y a des comportements, des lois, qui - si leurs conditions d'application restent conservées - déterminent le futur. « L'histoire, c'est ce qui empêche l'avenir d'être n'importe quoi » disait André Gide.

Termes connexes : extrapolation, prolongement, perspective. La notion de prévision ne doit pas être confondue avec celle d'OBJECTIF !

Objectif : « ce que l'on voudrait voir arriver ». C'est une vision volontariste du futur. On se fixe un futur et on regarde comment y arriver. À ce titre le raisonnement par objectif fait tout pour que la prévision soit fausse !

Techniquement, le terme de « prévision » englobe un ensemble de méthodes très diverses dont le point commun est de chercher à diminuer l'incertitude entraînée par la non-connaissance du futur.

On peut les distinguer en fonction de critères tels que :

- l'horizon : court, moyen ou long terme,
- la finesse : macro-économique ou micro-économique,
- l'approche : reposant sur le jugement humain ou sur la formalisation.

Et aussi :

- la quantité d'informations disponibles,
- la précision souhaitée pour la prévision,
- la part d'aléas dans le phénomène à prévoir.

### Pour prévoir il faut « modéliser »

Prenons deux exemples :

Un chef de produit vous dit que les ventes de son produit seront à la hausse l'année prochaine. Il peut faire cette prévision parce qu'il a vu les ventes augmenter durant les deux dernières années, et pense que, quelles que soient les raisons qui les faisaient croître dans le passé, celles-ci continueront à agir dans le futur.

Un autre chef de produit peut penser que les ventes augmenteront l'année prochaine parce qu'elles sont en relation avec un ensemble de variables économiques à travers des relations complexes. Par exemple, le chef de produit imagine que les ventes sont liées d'une certaine façon au marché, à l'effort publicitaire et au prix de vente, si bien qu'à partir d'hypothèses très probables concernant l'évolution future de ces variables, il est amené à envisager comme vraisemblable une hausse.

Dans les deux cas la prévision est basée sur l'intuition, bien que les modalités de raisonnement diffèrent dans les deux cas cités plus haut. Mais dans chacun de ces raisonnements il y a un certain raisonnement logique implicite. Aucune équation n'a été écrite. Néanmoins, le chef de produit a établi une sorte de modèle implicite :

S'il a établi sa prévision optimiste à partir des taux de croissance du passé, il a bâti un modèle d'extrapolation de série chronologique.

S'il fonde sa prévision sur une connaissance des relations économiques, il a construit, implicitement, un modèle économétrique.

Même inconsciemment, le prévisionniste intuitif construit implicitement des modèles. Une question se pose alors : pourquoi ne pas les construire explicitement, les estimer et les tester ?

Plusieurs raisons incitent à cette démarche de modélisation.

Tout d'abord cela force l'observation à établir clairement et à estimer les inter-relations sous-jacentes. Ensuite, la confiance aveugle dans l'intuition peut amener à l'ignorance de liaisons importantes ou à leur mauvaise utilisation.

De plus, des relations marginales mais néanmoins explicatives, qui ne sont qu'un élément d'un modèle global, doivent être testées et validées afin de les mettre à leur véritable place, ce qui n'est pas fait dans la prévision intuitive.

Enfin, il est nécessaire de fournir en même temps que la prévision une mesure de la confiance que l'utilisateur peut avoir en celle-ci, c'est à dire la précision que l'on peut en attendre. Là encore, l'utilisation de méthodes purement intuitives exclut toute mesure quantitative de la fiabilité d'une prévision.

## Les méthodes de prévision à court terme par extrapolation

Les conditions de mise en œuvre de ces méthodes sont :

- le court terme (jusqu'à un an maximum),
- une quantité d'informations disponibles d'au moins une période et demi à deux périodes,
- La précision souhaitée pour la prévision ne devant pas être inférieure à 1 ou 2%, une part d'aléas dans le phénomène non prédominante.

Ces méthodes par extrapolation, consistent à dégager dans la série elle-même un certain nombre de composantes que l'on peut prolonger dans le futur (en faisant l'hypothèse que leur comportement passé se poursuivra jusqu'à un certain horizon).

| | Logique de l'approche | Avantages | Inconvénients |
|---|---|---|---|
| Méthodes par décomposition<br>(Holt, Winters et Holt) | Basée sur l'analyse des composantes de tendance et de saisonnalité | | Longueur de l'historique<br>Stabilité des lois d'évolution sur plusieurs périodes |
| Méthode de Box et Jenkins | Basée sur l'analyse des aléas et leur auto-corrélation<br>Nécessite plus de 50 observations | | Complexe à mettre en œuvre |
| La méthode par équivalence | Basée sur l'analyse de la tendance | Pas de choix de tendance<br>Pas de choix des coefficients saisonniers<br>Historiques courts<br>S'adapte aux ruptures de tendance | La qualité de la prévision repose sur la seule qualité de la détermination de la tendance |

Un modèle efficace consiste à poser qu'une évolution est le fruit de trois composantes d'importance très variable selon les cas :

- la tendance (l'axe profond de l'évolution, sa ligne directrice)
- les saisonnalités (des variations que l'on retrouve à intervalle de temps constant,
- les aléas (des variations non expliquées par les deux premières composantes)

### *La méthode par équivalence*

La méthode consiste à déterminer la tendance de l'évolution qui en est la partie la plus stable, puis à modéliser et prolonger celle-ci, et enfin à transformer cette prévision de tendance en une prévision en valeur brute (c'est à dire en réinjectant, en particulier, les variations saisonnières)

Partant de la chronique brute, on filtre pour commencer les variations les plus instables : les aléas.

Le moyen utilisé est le lissage exponentiel.

Dans un second temps il s'agit de désaisonnaliser le résultat du filtrage précédent.

Le traitement appliqué est celui d'une moyenne mobile équi-pondérée de longueur égale à la période du phénomène.

La chronique obtenue sera la tendance constatée.

Celle-ci va être modélisée pour permettre son extrapolation. Si les méthodes traditionnelles de régression peuvent être utilisée pour ajuster droites, paraboles, exponentielles,… La démarche retenue ici essaie de pallier aux inconvénients de la régression, à savoir : choix de la taille de l'historique sur lequel l'ajustement se fera. Les tendances constatées aujourd'hui sont rarement homogènes sur de longues périodes.

Le principe utilisé est celui des polynômes orthogonaux : toute fonction être approximée par une série de polynômes, mais cette décomposition peut se faire en particulier sur une base intéressante : des polynômes qui sont dit orthogonaux.

Cette technique, issue de l'analyse numérique, appliquée à notre problème, va assurer un ajustement permanent de la fonction modélisant la tendance : plus de type de fonction à choisir, plus d'historique à sélectionner !

Chaque valeur de tendance modélisée est une combinaison linéaire des trois valeurs de tendance constatées précédentes, les coefficients de la fonction linéaire intégrant, eux, l'ensemble de l'historique.

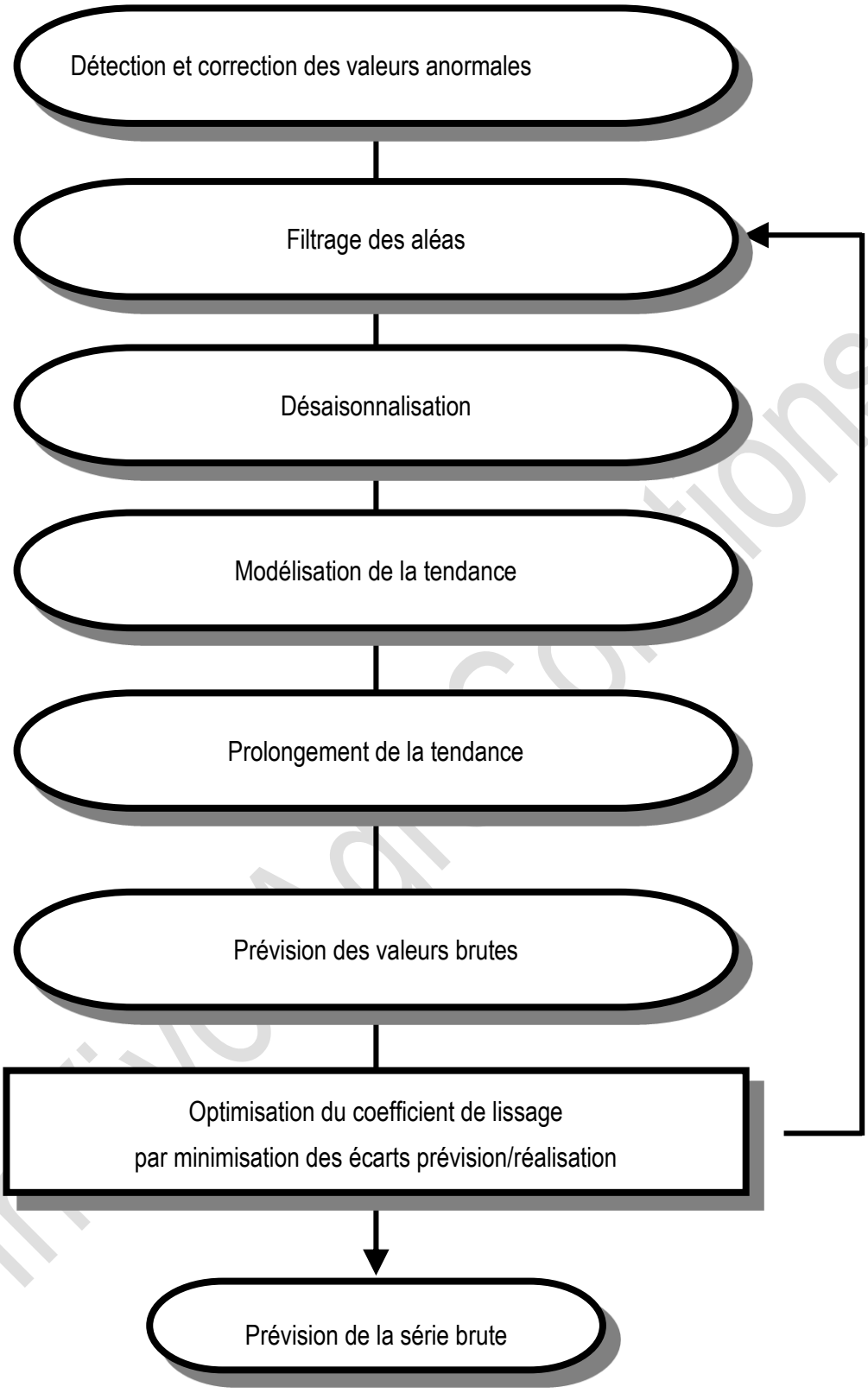
Pour re-saisonnaliser, la recherche de coefficients saisonniers est abandonnée. En effet elle nécessite des historiques longs : plusieurs périodes, dans le cas des chroniques d'entreprise, cela signifie plusieurs années car leur périodicité est souvent annuelle.

Le retour aux données brutes s'effectuera en inversant les processus de moyenne mobile et de lissage. On parlera ainsi de méthode par équivalence car tout au long de l'historique, passé et prévu, les trois niveaux – brut, lissage, moyenne mobile – sont équivalents (on passe de l'un à l'autre dans les deux sens).

Le double avantage est de ne pas avoir à choisir de modèle pour les coefficients saisonniers (additifs, multiplicatifs, mixtes,…) et de pouvoir prévoir à partir d'historiques courts.

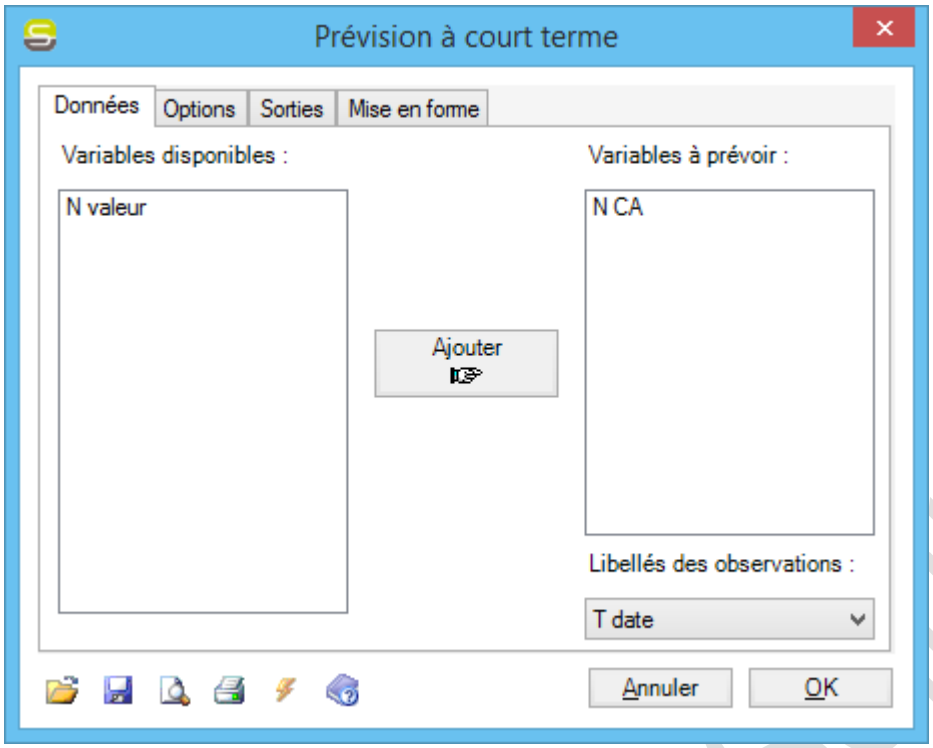
Une optimisation du coefficient de lissage est faite à cette étape, elle permet d'améliorer la qualité de la modélisation de la tendance sur laquelle repose toute la prévision.

SCHEMA DE LA METHODE PAR EQUIVALENCE



## Mise en œuvre

*Onglet « Données »*



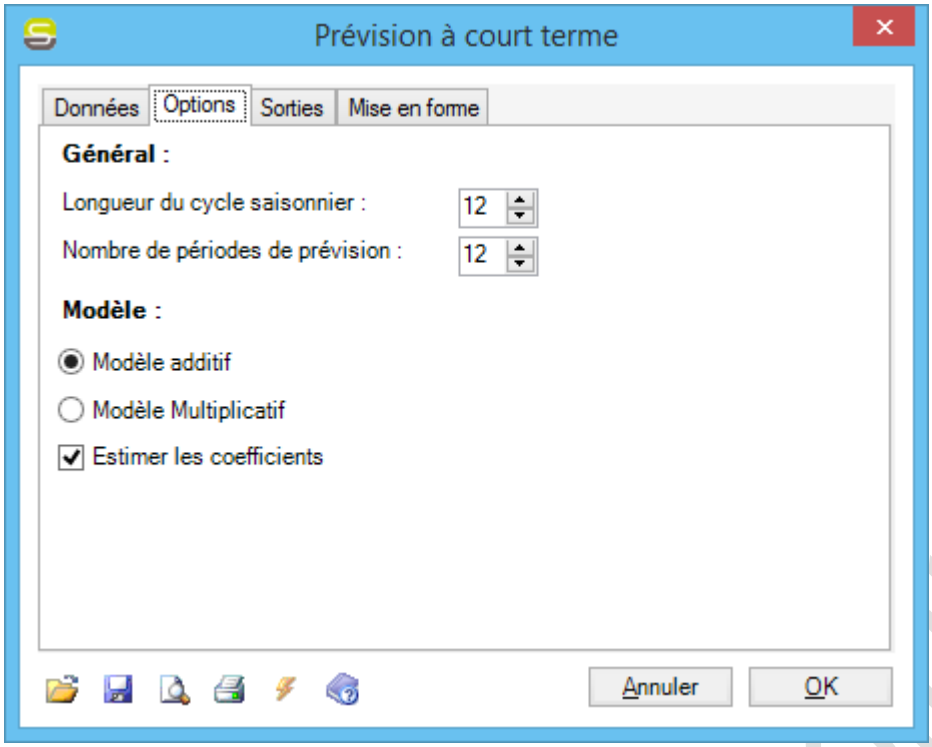

- *Variable à prévoir :* sélectionnez la variable représentant la série chronologique. Si vous sélectionnez plusieurs variables, le logiciel suppose que la variable en colonne représente les différentes années par exemple et qu'en lignes se trouvent les périodes : les 12 mois, les 4 trimestres ou les 52 semaines.
- *Libellés des observations :* sélectionnez la variable contenant le descripteur de période (année, mois,…).

StatBox permet d'analyser des chroniques longues ou courtes. Dans ce cas le nombre de périodes minimum est égal à la longueur de la période+4.

Le logiciel permet également l'analyse de tendance linéaire, parabolique, avec de forts aléas ou avec des ruptures de tendance.

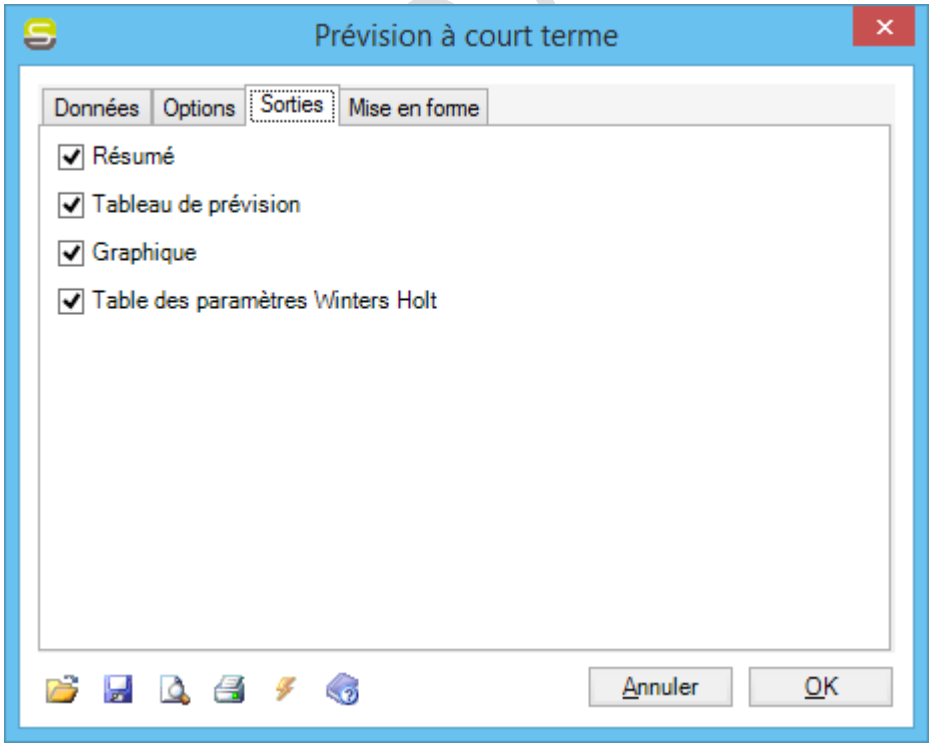
Si vous avez une valeur manquante, faites une analyse sur le sous-ensemble des données antérieur à cette valeur, pour en faire l'estimation. Cela suppose qu'elle ne se situe pas au tout début de la série. Dans ce cas, faites une moyenne des 2 valeurs adjacentes.

*Onglet « Options »*




- *Longueur du cycle saisonnier :* saisissez le nombre de période d'un cycle (par exemple : 12 si vous avez des données mensuelles).
- *Nombre de période de prévision :* indiquez le nombre de période que vous désirez estimer.
- *« Modèle additif » / « Modèle multiplicatif » :* sélectionnez le type de modèle que vous souhaitez utiliser pour l'algorithme de Winters Holt.
- *Estimer les coefficients :* cochez cette option si vous souhaitez que le logiciel estime lui-même les paramètres de l'algorithme de Winters Holt. Si vous souhaitez entrer des valeurs particulières, décochez cette option et entrez une valeur pour chaque paramètre.

*Onglet « Sorties »*



- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Tableau de prévision :* affiche un tableau de synthèse : les données observées (Valeurs brutes), la tendance liée à ses données brutes (Tendance constatée), les prévisions à partir de la fin de la série. Les prévisions incluent également les cinq dernières périodes connues. La comparaison entre les 5 données observées et les 5 valeurs estimées nous donnent un pourcentage d'erreurs. Le tableau présente également en dernière colonne, la tendance prévue.
- *Graphique :* affiche une courbe de tendance associée à la prévision.
- *Table des paramètres de Winters Holt :* affiche la table de synthèse des paramètres utilisés pour l'algorithme de Winters Holt dans le cas où vous avez laissé le logiciel estimer ces paramètres.

### Références

**Bakhvalov N. (1976).** Méthodes numériques, par Analyse, Algèbre, équations différentielles. Ed. Mir, Moscou.

**Bass J. (1964)**. Cours de mathématiques, Tomes 1 et 2. Ed. Masson, Paris.

**Léon Louis (1983)**. TRAITEMENT D'ALGORITHMES PAR ORDINATEUR, Tome 2. ENSTA - Ecole Nationale Supérieures de Techniques avancées, Cepadues-Ed, Toulouse.

**Encyclopaedia Universalis (1997)**. Dictionnaire des mathématiques - Algèbre, Analyse, Géométrie. Ed. Albin Michel, Paris.

## AJUSTEMENT D'UNE LOI DE PROBABILITÉ

Utilisez ce module pour ajuster une loi de probabilité à vos données quantitatives, continues ou discrètes, et vérifier la qualité de l'ajustement effectué.

### Description

L'ajustement d'une loi de probabilité à une distribution de valeurs correspond à la recherche du meilleur modèle paramétrique parmi ceux proposés par StatBox. L'ajustement consiste donc à choisir une loi de probabilité et les valeurs des paramètres de cette loi, de sorte que l'écart entre les valeurs des données et les valeurs du modèle soit le plus faible possible.

Plusieurs lois de probabilité sont disponibles : uniforme, gaussienne, lognormale, de Student, de Fisher, du khi², Bêta, exponentielle, de Poisson, binomiale, binomiale négative. StatBox offre la possibilité de saisir directement les valeurs des paramètres de la loi de probabilité choisie, ou de les estimer automatiquement.

Afin de juger la qualité de l'ajustement, StatBox affiche les valeurs de l'espérance, de la variance, des coefficients d'asymétrie et d'aplatissement, estimées d'après les données, et les valeurs calculées pour la loi de probabilité sélectionnée, compte tenu de ses paramètres (saisis ou estimés). Un accord entre les deux jeux de valeurs constitue un premier élément d'appréciation de l'accord entre la distribution des valeurs et le modèle ajusté.

Deux tests non paramétriques sont également proposés par StatBox :

- le test de Kolmogorov-Smirnov testant l'égalité entre la distribution cumulée et la fonction de répartition de la loi de probabilité ajustée,
- le test du khi² de conformité entre l'histogramme des valeurs observées et l'histogramme des valeurs théoriques.

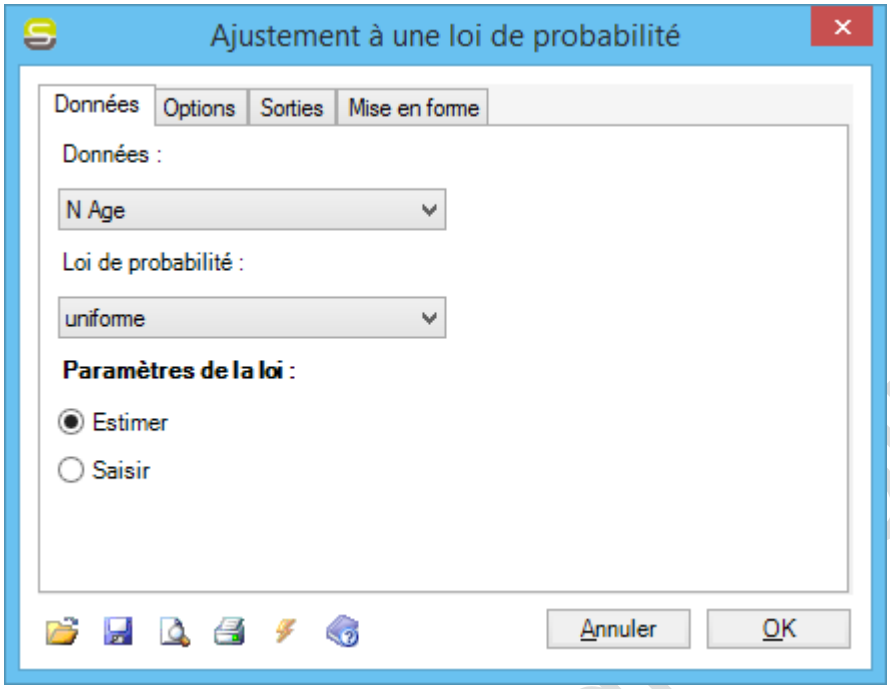
Le test du khi² nécessite de définir l'histogramme :

- en spécifiant le nombre de classes d'amplitude constante,
- en important les bornes des classes,
- en utilisant des bornes discrètes dans le cas d'une loi discrète (loi de Poisson, binomiale et binomiale négative).

Il arrive parfois que le test du khi² conclue à un mauvais ajustement uniquement du fait d'une classe dont la contribution à la valeur du khi² est très élevée. Ceci peut être causé par le découpage en classes de l'histogramme, un autre découpage pouvant changer la conclusion du test. Afin d'apprécier l'impact de la plus forte contribution au khi² dans la conclusion du test, StatBox effectue également le test du khi² sans tenir compte de la plus forte contribution.

## Mise en œuvre

### *Onglet « Données »*




- *Données :* sélectionnez la variable correspondant à la colonne des valeurs à contrôler. Les valeurs manquantes ne sont pas autorisées.
- *Loi de probabilité :* choisissez la loi de probabilité à ajuster parmi celles de la liste.
- *« Estimer » / « Saisir » :* si vous choisissez de saisir les valeurs des paramètres, des champs spécifiques apparaissent dans la boîte de dialogue, en fonction de la loi de probabilité choisie.
    - uniforme
    - a : nombre définissant la borne inférieure de l'intervalle de la loi uniforme
    - b : nombre définissant la borne supérieure de l'intervalle de la loi uniforme
    - gaussienne standard (ou loi normale centrée et réduite) : loi de Gauss de moyenne nulle et de variance unité
    - gaussienne (ou loi normale)
    - $\mu$ : valeur de l'espérance
    - sigma² : valeur de la variance
    - log normale (le logarithme de la variable distribuée selon une loi lognormale suit la loi normale de paramètres $\mu$ et sigma²)
    - $\mu$ : valeur de l'espérance de la loi normale selon laquelle est distribué ln(x)
    - sigma² : valeur de la variance de la loi normale selon laquelle est distribué ln(x)
    - de Student
    - ddl : nombre de degrés de liberté de la loi de Student
    - de Fisher
    - ddl 1 : nombre de degrés de liberté du numérateur du F de Fisher
    - ddl 2 : nombre de degrés de liberté du dénominateur du F de Fisher

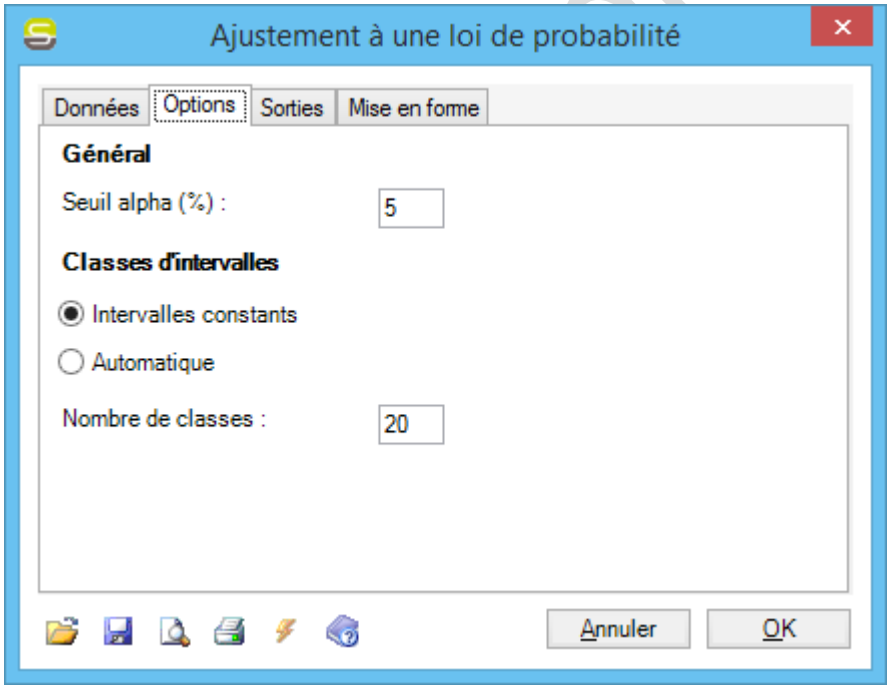
- du khi²
- ddl : nombre de degrés de liberté de la loi du khi²
- Bêta
- a1 : nombre correspondant au premier paramètre de forme de la loi Bêta
- a2 : nombre correspondant au deuxième paramètre de forme de la loi Bêta
- exponentielle
- Lambda : inverse du temps d'attente moyen entre deux événements d'un phénomène aléatoire pour la loi exponentielle
- de Poisson
- Lambda : valeur moyenne supérieure à 0 pour définir la loi de Poisson
- Binomiale
- n : nombre d'essais définissant la loi binomiale
- p : probabilité de succès définissant la loi binomiale

**Remarque :** la loi de Bernoulli est un cas particulier de la loi binomiale pour p = 0,5.

- binomiale négative
- k : nombre de succès définissant la loi binomiale négative
- p : probabilité de succès définissant la loi binomiale négative

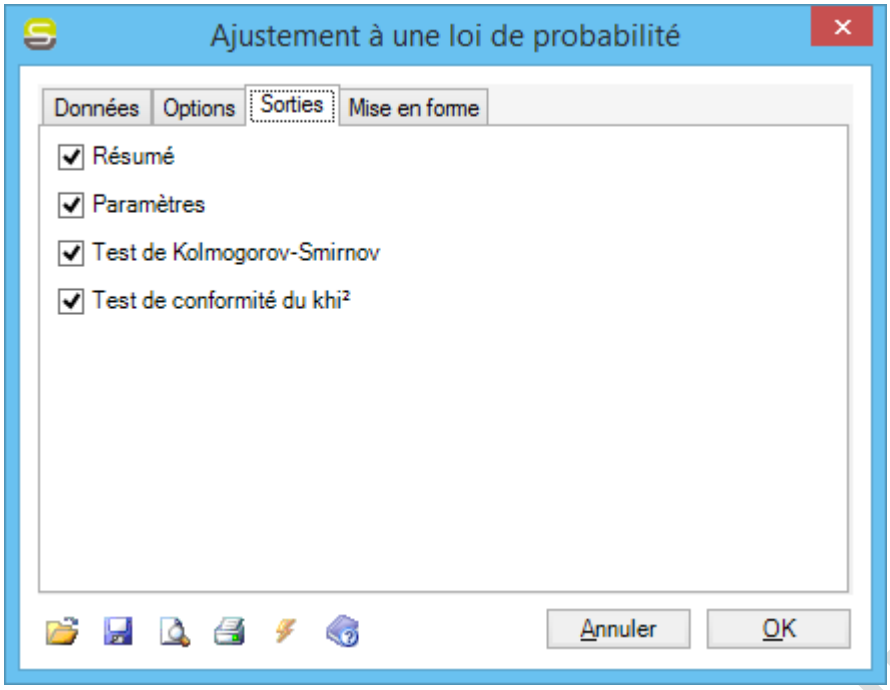
➢ *Méthode itérative :* dans le cas de la loi binomiale, si vous choisissez d'estimer automatiquement les paramètres, cochez cette case pour effectuer une estimation par une méthode itérative (maximum de vraisemblance). Lorsque cette case n'est pas cochée, StatBox demande si vous désirez spécifier la valeur de *n* (nombre d'essais) : si cette valeur est connue, vous obtiendrez alors une meilleure estimation de *p* (probabilité de succès).

*Onglet « Options »*




➢ *Seuil alpha (%) :* entrez la valeur du risque de première espèce des tests.
➢ *« Intervalles constants » / « Automatique » :* sélectionnez le mode de découpage des données en classe pour le test des effectifs.
➢ *« Nombre de classes » / « Nombre maximal de classes » :* entrez le nombre de classes d'amplitude constante de l'histogramme. Dans le cas de l'utilisation de bornes discrètes, StatBox regroupe les bornes au mieux en fonction du nombre maximal de classes.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Paramètres :* affiche les valeurs des paramètres de position (moyenne), de dispersion (variance) et de forme (asymétrie et aplatissement) estimées à partir des données, et les valeurs théoriques calculées pour la loi de probabilité ajustée.
- *Test de Kolmogorov-Smirnov :* effectue le test d'égalité des distributions cumulées empirique et théorique.
- *Test de conformité du khi² :* effectue le test d'égalité des histogrammes des effectifs observés et théoriques.

## Références

**Abramowitz M. & I.A. Stegun (1972)**. Handbook of mathematical functions. Dover Publications, New York, pp. 927-964.

**Aïvazian S., I. Enukov & L. Mechalkine (1986)**. Éléments de modélisation et traitement primaire des données. Mir, Moscou, pp. 126-183.

**Dagnelie P. (1986)**. Théorie et méthodes statistiques. Vol. 2. Les Presses Agronomiques de Gembloux, Gembloux, pp. 61-72.

**Manoukian E.B. (1986)**. Guide de statistique appliquée. Hermann, Paris, pp. 19-68.

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 686-724.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 90-97.

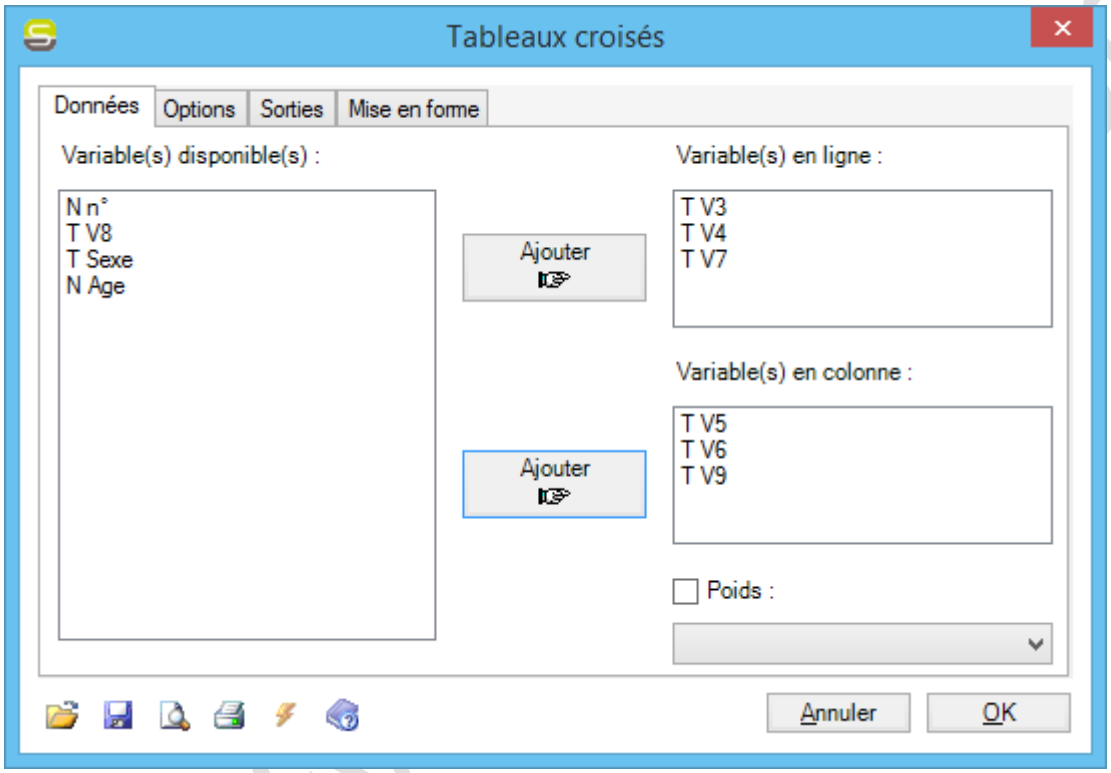
# ANALYSE À DEUX VARIABLES

## DEUX VARIABLES QUALITATIVES : TRIS CROISÉS

Utilisez ce module pour calculer le tableau de contingence (ou tableau croisé) pour deux ensembles de variables qualitatives, ainsi que des tableaux dérivés, et tester l'association entre les lignes et les colonnes.

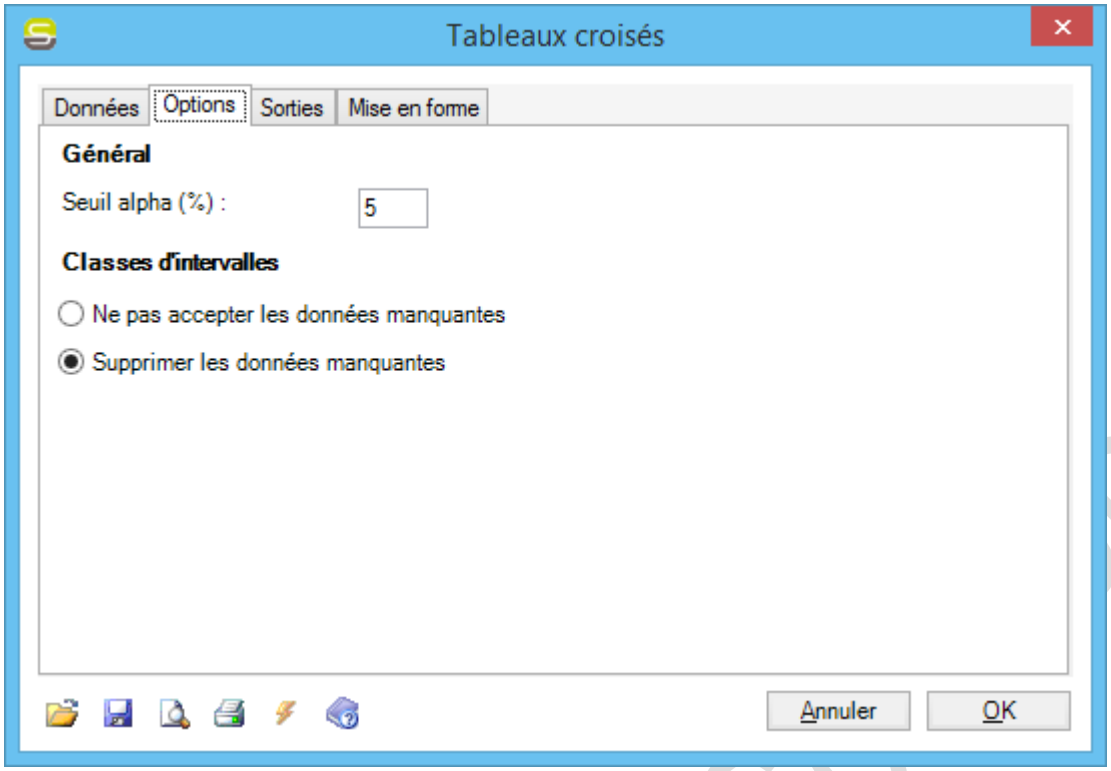
### Mise en œuvre

*Onglet « Données »*



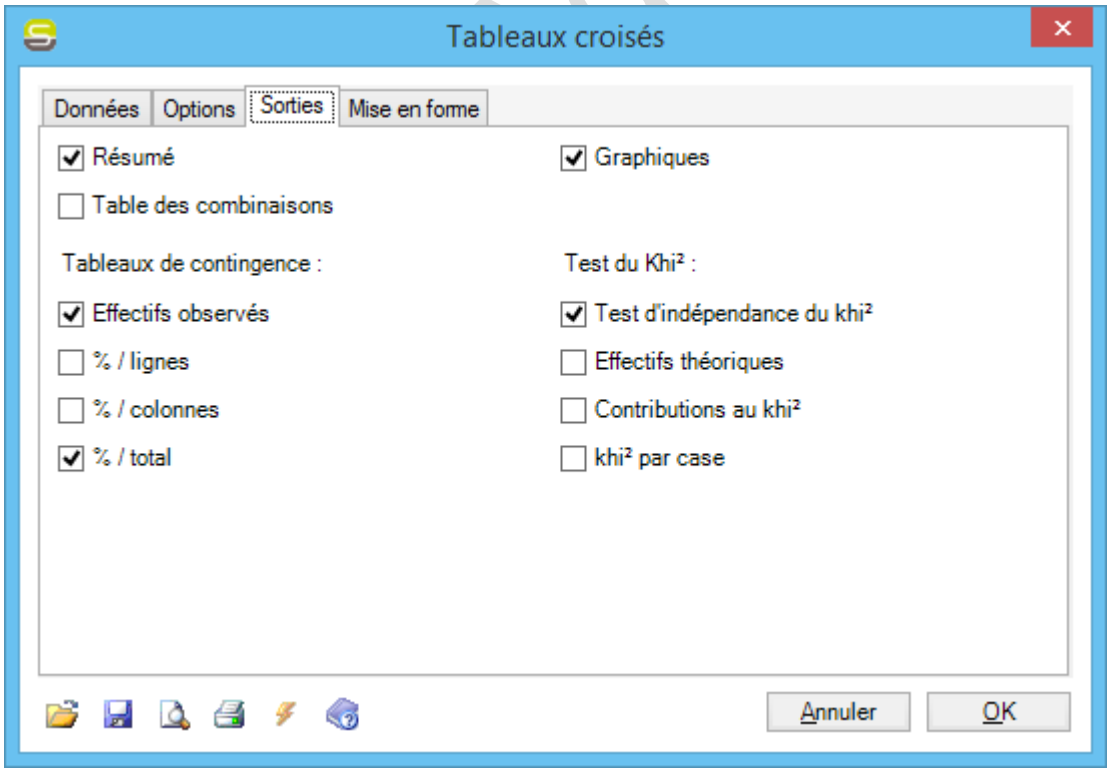

- *Variable(s) en lignes :* sélectionnez les variables qualitatives dont les modalités vont constituer les lignes du tableau de contingence.
- *Variable(s) en colonnes :* sélectionnez les variables qualitatives dont les modalités vont constituer les colonnes du tableau de contingence.
- Lorsqu'il y a des valeurs manquantes, StatBox propose de les ignorer lors de la construction du tableau de contingence. En cas de refus, le traitement est abandonné.
- *Poids :* sélectionnez la variable des poids des observations. Lorsqu'il y a des valeurs manquantes dans les poids, StatBox propose d'ignorer les observations concernées. En cas de refus, le traitement est abandonné.

*Onglet « Options »*




- *Seuil alpha (%) :* entrez la valeur du risque de première espèce des tests.
- *« Ne pas accepter les données manquantes » / « Supprimer les données manquantes »* : choisissez si les données manquantes doivent être supprimées ou si la méthode doit être arrêtée dans le cas de présence de données manquantes.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport
- *Graphiques :* affiche un histogramme de fréquence des croisements de modalités.

- *Table des combinaisons :* affiche un tableau des combinaisons de modalités en lignes et en colonnes. Ce tableau contient la même information que le tableau de contingence et le tableau des pourcentages par rapport à l'effectif total, mais cette information est présentée sous une forme différente.
- *Effectifs observés :* affiche le tableau de contingence auquel ont été ajoutées les sommes marginales ainsi que l'effectif total.
- *% / lignes :* affiche le tableau des pourcentages calculés par rapport aux sommes des lignes.
- *% / colonnes :* affiche le tableau des pourcentages calculés par rapport aux sommes des colonnes.
- *% / total :* affiche le tableau des pourcentages calculés par rapport à l'effectif total.
- *Test d'indépendance du khi² :* testez l'indépendance entre les lignes et les colonnes du tableau de contingence à l'aide d'un test du khi².
- *Effectifs théoriques :* affiche le tableau des effectifs théoriques, calculés sous l'hypothèse d'indépendance des lignes et des colonnes du tableau de contingence.
- *Contributions au khi² :* affiche le tableau des contributions élémentaires de chaque case du tableau de contingence à la valeur du khi² calculée pour l'ensemble du tableau de contingence.
- *Khi² par case :* affiche un tableau montrant, d'une part si l'effectif observé est supérieur, inférieur, ou égal à l'effectif théorique, et d'autre part, le résultat d'un test de khi² partiel dit « khi² par case ». Le khi² par case est un test du khi² calculé sur un tableau à quatre cases : une case correspondant à une case [*i,j*] du tableau de contingence originel, les autres cases correspondants aux effectifs pour la ligne *i* moins la case [*i,j*], pour la colonne *j* moins la case [*i,j*], et pour le reste du tableau.

### Références

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 724-743.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 92-95.

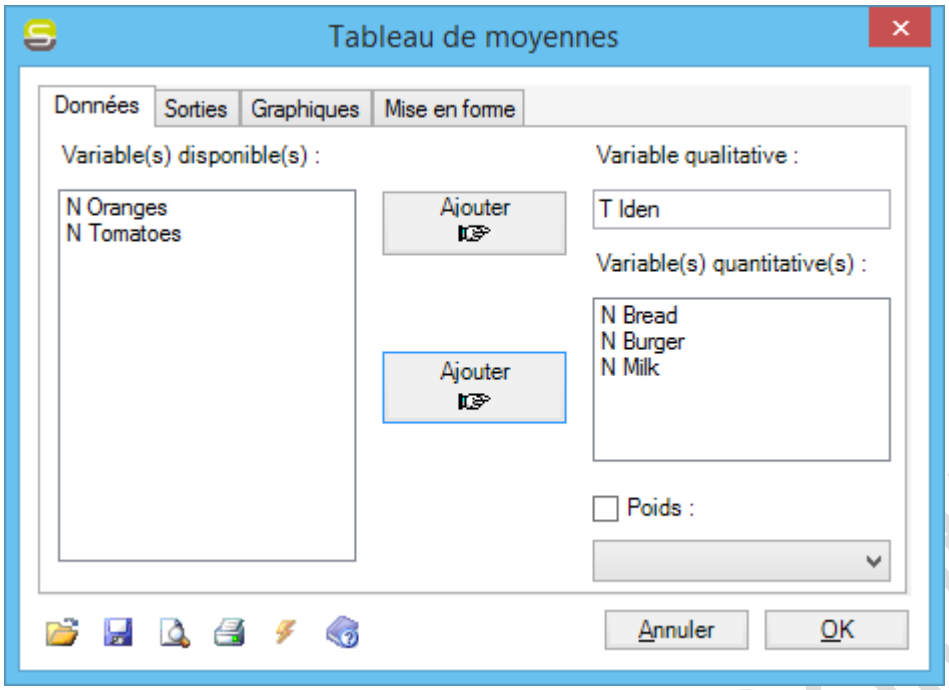
## TABLEAUX DE MOYENNES

### Description

Utilisez ce module pour calculer des statistiques descriptives sur un ensemble de variables quantitatives en les croisant avec les modalités d'une variable qualitative.

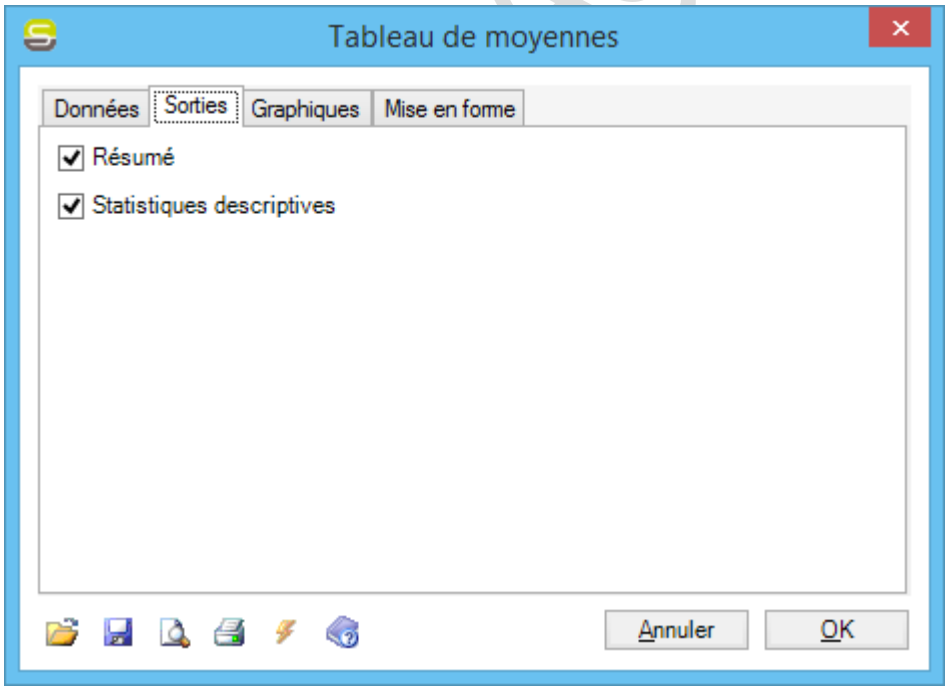
## Mise en œuvre

### *Onglet « Données »*



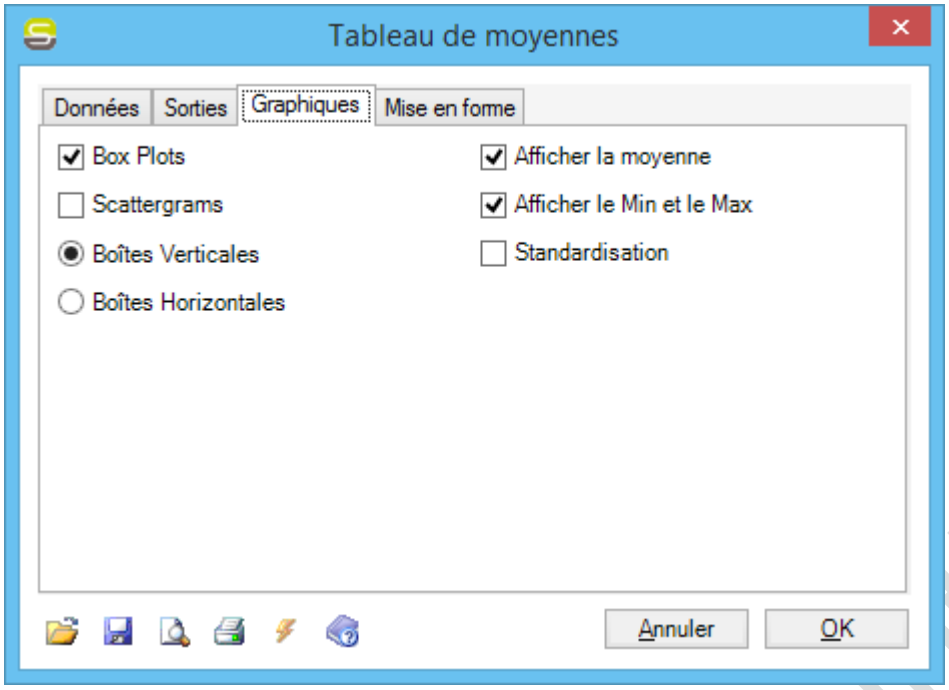

- *Variable qualitative :* sélectionnez la variable dont les modalités permettront de distinguer les « sous-groupes » des variables quantitatives.
- *Variable(s) quantitative(s)* : sélectionnez les variables quantitatives à étudier.
- *Poids :* saisissez la variable des poids des observations. Lorsqu'il y a des valeurs manquantes dans les poids, StatBox propose d'ignorer les observations concernées. En cas de refus, le traitement est abandonné.

### *Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Statistiques descriptives :* cochez cette option pour afficher la table des statistiques descriptives pour chaque croisement entre les variables quantitatives sélectionnées et les modalités de la variable qualitative.

*Onglet « Graphiques »*



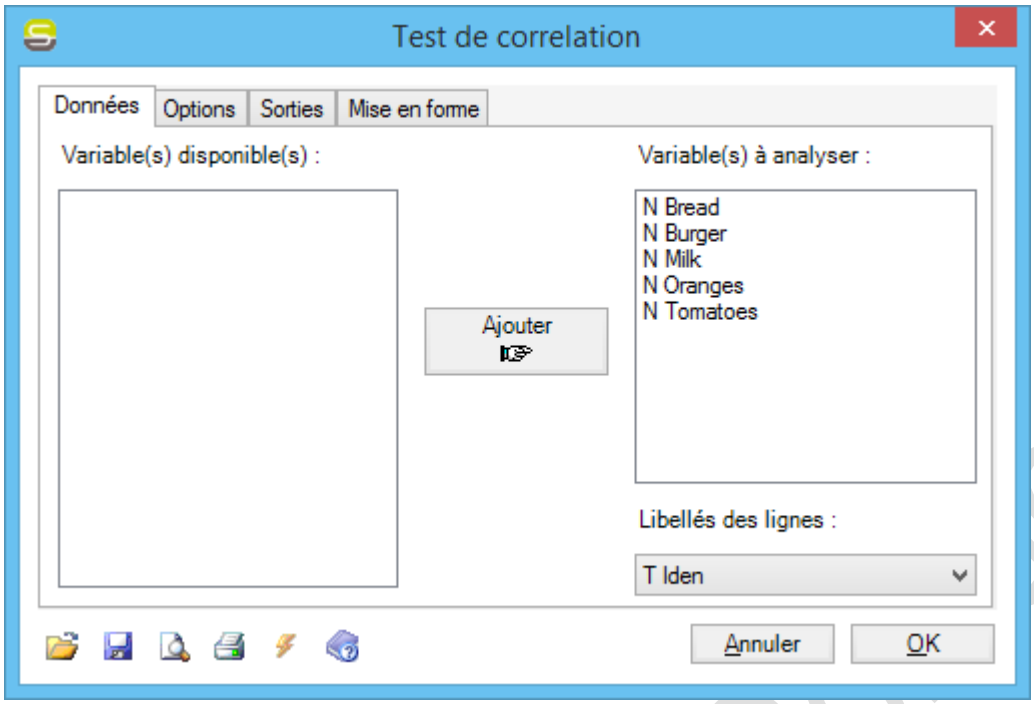

- *Box plots :* cochez cette case pour obtenir les graphiques "boîte à moustaches". Ces graphiques ne peuvent pas être affichés s'il y a plus de 16 variables ou plus de 30 000 points.
- *Scattergrams :* affiche les nuages de points univariés. Ces graphiques ne peuvent pas être affichés s'il y a plus de 24 variables ou plus de 30 000 points.
- *« Boîtes verticales » / « Boîtes horizontales » :* choisissez l'orientation des box plots et des scattergrams.
- *Afficher la moyenne :* affiche la moyenne sur les box plots et les scattergrams. Cette option n'est pas disponible lorsque l'option « Standardisation » est cochée.
- *Afficher le Min et le Max :* affiche la valeur minimum et la valeur maximum sur les box plots. Cette option n'est pas disponible lorsque l'option « Standardisation » est cochée.
- *Standardisation :* cochez cette case afin de supprimer l'effet des différences d'ordre de grandeur entre les variables lors de la production des box plots et des scattergrams, en divisant les valeurs de chaque variable par l'écart-type correspondant.

## MATRICE DE SIMILARITÉ / DISSIMILARITÉ (CORRÉLATIONS)

Utilisez ce module pour calculer une matrice de similarité ou de dissimilarité pour un tableau rectangulaire, en croisant les lignes ou les colonnes, et tester l'hypothèse d'absence de structure de corrélation dans le cas d'une matrice de corrélation paramétrique (corrélation de Pearson) grâce au test de sphéricité de Bartlett.

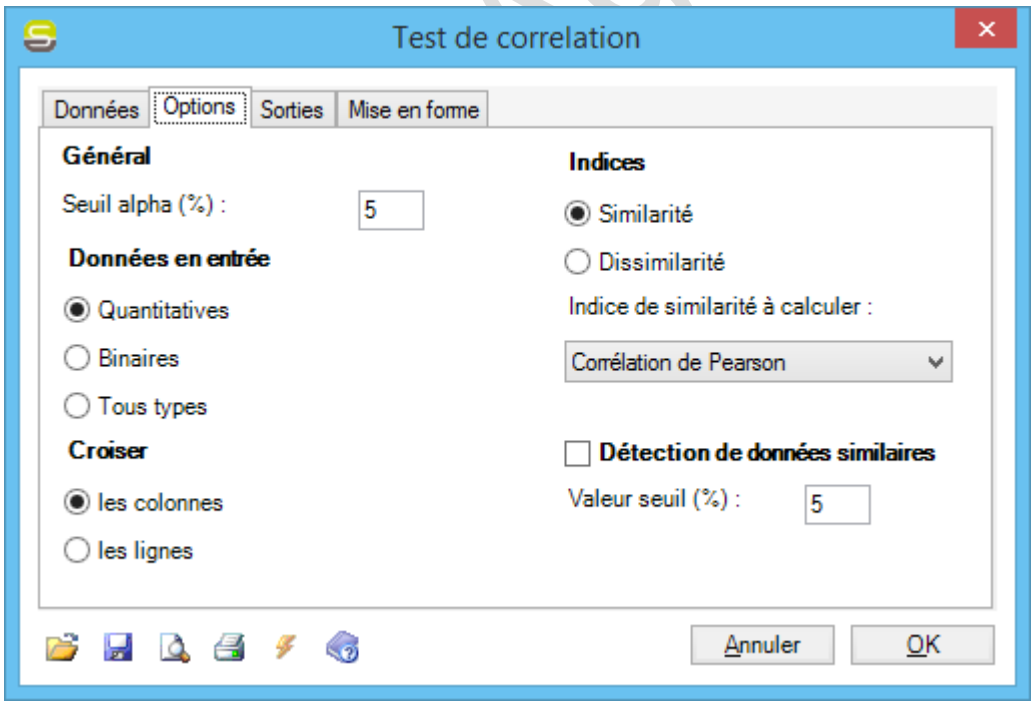
## Mise en œuvre

### *Onglet « Données »*




- *Variables à analyser :* sélectionnez les variables dont la corrélation est à tester. Lorsqu'il y a des valeurs manquantes, StatBox propose tout d'abord d'ignorer les lignes concernées. En cas de refus, StatBox propose alors d'utiliser toute l'information disponible en ignorant simplement les valeurs manquantes (*pairwise deletion*), sinon la boîte de dialogue est fermée et le traitement est abandonné.
- *Libellés des lignes :* sélectionnez la variable contenant les identifiants des observations.

### *Onglet « Options »*




- *Seuil alpha (%) :* entrez la valeur du risque de première espèce pour le test de sphéricité de Bartlett.
- *"Quantitatives" / "Binaires" / "Tous types" :* choisissez le type de données en entrée. Le choix du type de données permet à StatBox d'effectuer des contrôles de validité des données, et d'éviter des erreurs méthodologiques en ce qui concerne le choix d'un indice de similarité/dissimilarité. Dans le cas des variables

quantitatives, seuls les indices définis spécifiquement pour ces types de données sont proposés. Dans le cas de données de tous types (données quantitatives et/ou données qualitatives), un seul indice est proposé, les données étant considérées au niveau le plus bas du point de vue de la structure algébrique, c'est-à-dire au niveau d'une variable nominale : les valeurs ne sont donc plus distinguées entre elles que sur la base de l'égalité/inégalité stricte.

- *« les colonnes » / « les lignes » :* sélectionnez si l'on doit tester la corrélation entre les lignes ou les colonnes du tableau sélectionné.

**Remarque :** Dans le cas d'une variable quantitative, par défaut le calcul d'une similarité s'effectue en croisant les colonnes (par exemple, la corrélation entre les variables d'un tableau observations/variables), et le calcul d'une dissimilarité s'effectue en croisant les lignes (par exemple, la distance euclidienne entre les observations d'un tableau observations/variables). Dans les autres cas, par défaut les calculs s'effectuent en croisant les lignes.

*"Similarité" / "Dissimilarité" :* choisissez si les valeurs calculées doivent être d'autant plus élevées que les données sont ressemblantes (similarité), ou bien d'autant plus faibles que les données sont ressemblantes (dissimilarité). Le choix du type de mesure conditionne la liste des indices proposés.

Pour les données quantitatives :

| Similarité | Dissimilarité |
|---|---|
| Corrélation de Pearson | Distance euclidienne |
| Corrélation de Spearman | Distance du khi² |
| Corrélation de Kendall | Distance de Manhattan |
| Inertie | Dissimilarité de Pearson |
| Covariance (n) | Dissimilarité de Spearman |
| Covariance (n-1) | Dissimilarité de Kendall |

**Remarque :** la " Covariance (n) " et la " Covariance (n-1) " diffèrent uniquement par le dénominateur utilisé, c'est-à-dire soit *n*, soit *n*-1, avec *n* l'effectif (nombre de lignes si vous croisez les colonnes, ou nombre de colonnes si vous croisez les lignes). Dans le second cas, il s'agit de l'estimation sans biais de la matrice de variance-covariance dans le cadre d'un modèle de loi normale multivariée.

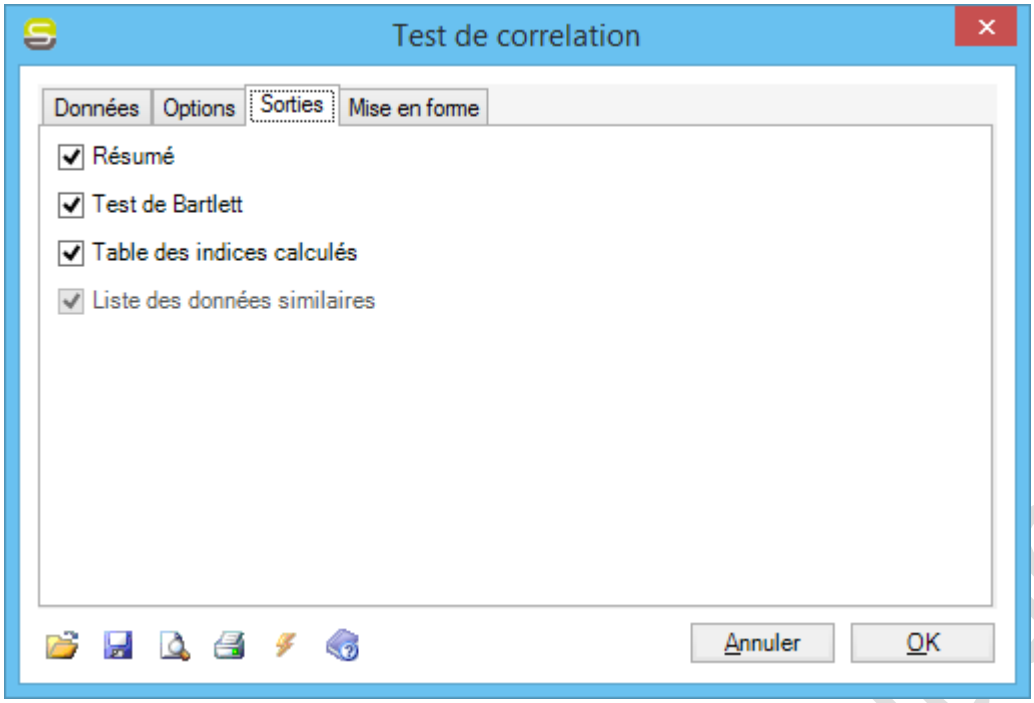
Pour les données de tous types, un seul indice est proposé, qui permet notamment de mettre en évidence des lignes ou des colonnes similaires dans le tableau de données, en fixant un seuil de ressemblance minimale au-delà duquel deux lignes ou deux colonnes sont considérées comme semblables.

*Indice de similarité/dissimilarité à calculer :* sélectionnez parmi les indices proposés l'indice à calculer.

*Détection de données similaires :* lorsque la similarité générale est utilisée (données de tous types), cochez cette case pour mettre en évidence les données similaires (lignes ou colonnes selon l'option choisie précédemment) au seuil spécifié par *" Valeur seuil (%) "*.

*Valeur seuil (%) :* entrez la valeur de la similarité minimale au-delà de laquelle les données sont considérées comme semblables. Les données sont déclarées semblables si la similarité est strictement supérieure à la valeur seuil, ou ce qui revient au même, si la dissimilarité est strictement inférieure à 100 % moins la valeur seuil.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Test de Bartlett :* lorsque la corrélation de Pearson est utilisée (similarité pour données quantitatives), effectue le test de sphéricité de Bartlett testant l'existence d'une structure de corrélation significative au sein de la matrice de corrélation, au seuil de signification spécifié par " Seuil alpha (%) ".
- *Table des indices calculés :* affiche la table des indices de similarité / dissimilarité calculés.
- *Liste des données similaires :* affiche un tableau regroupant les couples de données (lignes ou colonnes) détectées comme étant similaires.

## Références

**Dillon W.R. & M. Goldstein (1984)**. Multivariate analysis. Methods and applications. John Wiley & Sons, New York, pp. 157-167.

**Gower J.C. & P. Legendre (1986)**. Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3** : 5-48.

**Jambu M. (1978)**. Classification automatique pour l'analyse des données. 1 - méthodes et algorithmes. Dunod, Paris, pp. 484-518.

**Jobson J.D. (1992)**. Applied multivariate data analysis. Volume II: categorical and multivariate methods. Springer-Verlag, New York, pp. 345-388.

**Legendre L. & P. Legendre (1984)**. Écologie numérique. Tome 2. La structure des données écologiques. Masson, Paris, pp. 5-50.

**Roux M. (1985)**. Algorithmes de classification. Masson, Paris, pp. 126-134.

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 724-743.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 157-158.

## NUAGES DE POINTS

Consultez le paragraphe « Nuages de points » de la section « Représentations graphiques ».

## GRAPHIQUES AVEC LIBELLÉS

Consultez le paragraphe « Graphique avec libellés » de la section « Représentations graphiques ».

# ANALYSE À N VARIABLES

## ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

Utilisez l'analyse en composantes principales pour résumer la structure de données décrites par plusieurs variables quantitatives, tout en obtenant des facteurs non corrélés entre eux. Ces facteurs peuvent être utilisés comme de nouvelles variables permettant :

- d'éviter la multicolinéarité en régression multiple ou en analyse factorielle discriminante,
- d'effectuer une classification automatique en ne tenant compte que de l'information essentielle, c'est-à-dire en ne conservant que les premiers facteurs.

### Description

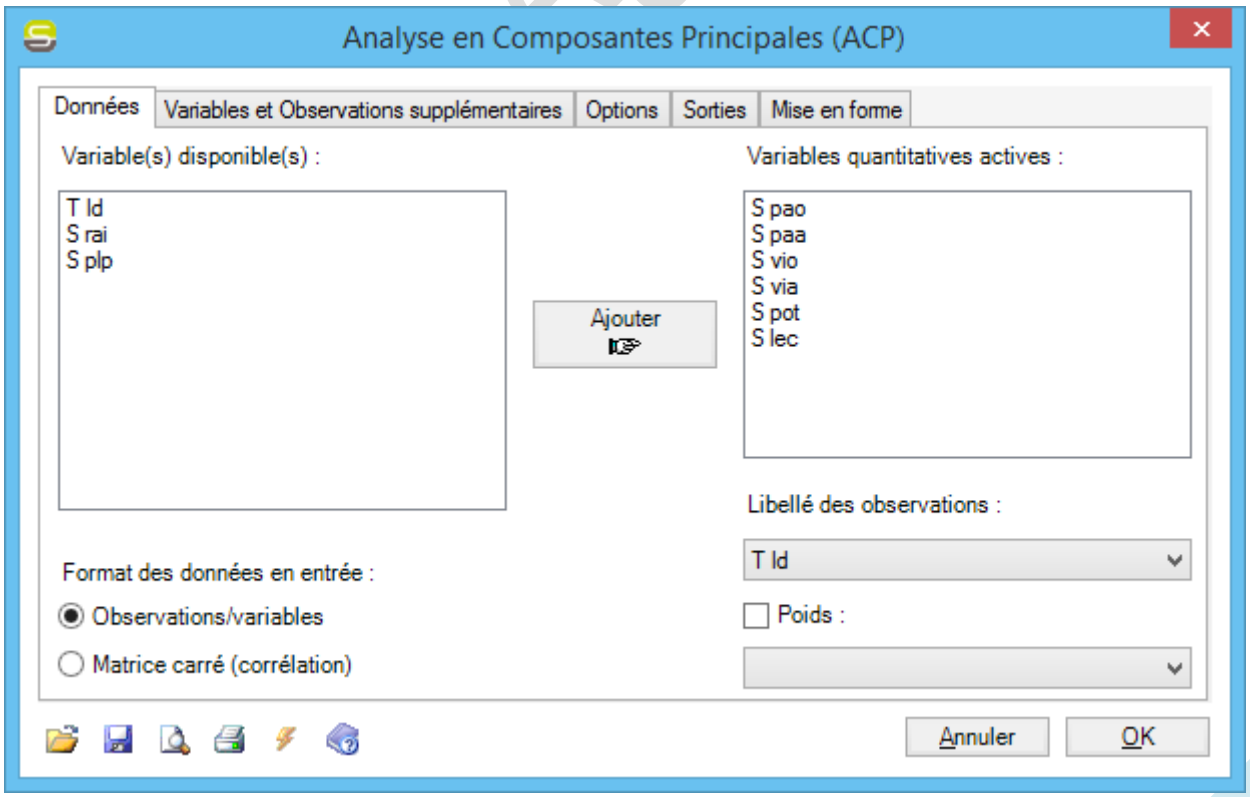
L'analyse en composantes principales (ACP) consiste à exprimer un ensemble de variables en un ensemble de combinaisons linéaires de facteurs non corrélés entre eux, ces facteurs rendant compte d'une fraction de plus en plus faible de la variabilité des données. Cette méthode permet de représenter les données originelles (observations et variables) dans un espace de dimension inférieure à l'espace originel, tout en limitant au maximum la perte d'information. La représentation des données dans des espaces de faible dimension (ici 2 dimensions) en facilite considérablement l'analyse.

L'ACP diffère de l'analyse factorielle en ce qu'elle conduit à un ensemble de facteurs non corrélés entre eux, ce qui correspond au cas particulier des communalités toutes égales à 1 (variances spécifiques nulles).

**Remarque :** ce module accepte jusqu'à 250 variables.

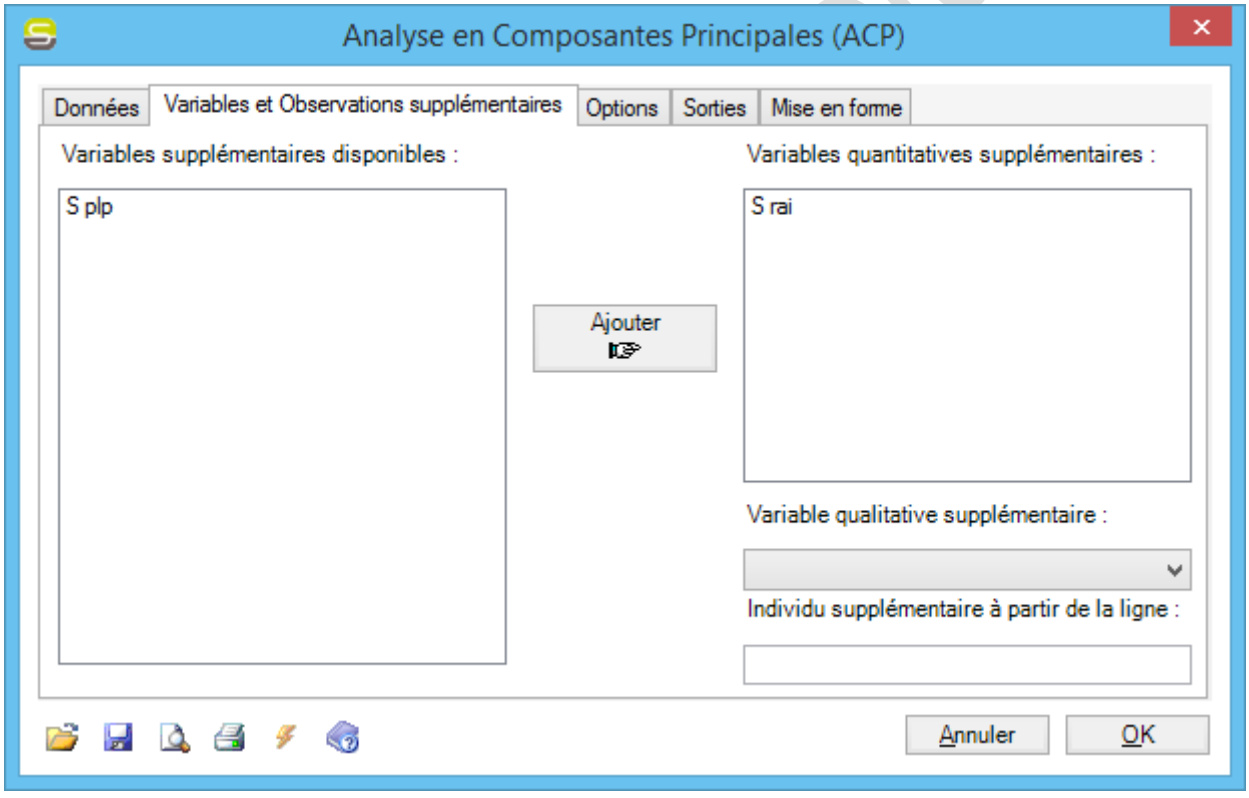
### Mise en œuvre

*Onglet « Données »*



- *« Observations/variables » / « Matrice carrée» :* choisissez la nature des données en entrée, selon qu'il s'agit d'un tableau avec les observations en ligne et les variables en colonnes ou d'une matrice de corrélation.
- *Variables quantitatives actives :* saisissez les variables des données, correspondant à un tableau rectangulaire observations/variables ou à une matrice de corrélation. Dans le cas d'un tableau, lorsqu'il y a des valeurs manquantes StatBox propose tout d'abord d'ignorer les lignes concernées. En cas de refus, StatBox propose alors d'estimer les valeurs manquantes de chaque variable par la moyenne (*cf.* l'option « *Estimation des données manquantes* »), sinon StatBox indique qu'il est possible d'utiliser toute l'information disponible (*pairwise deletion*) grâce au module « Matrice de similarité / dissimilarité », puis la boîte de dialogue est fermée et le traitement est abandonné. Dans le cas d'une matrice de corrélation, les valeurs manquantes ne sont pas autorisées. Cependant, la matrice étant symétrique, il suffit que les données de la sélection permettent de reconstituer correctement la totalité de la matrice.
- *Libellés des observations :* dans le cas d'un tableau observations/variables, saisissez la plage de la colonne de libellés qui correspondent aux lignes du tableau de données.
- *Poids :* dans le cas d'un tableau observations/variables, saisissez la plage de la colonne des poids des observations. Les valeurs manquantes dans les poids sont cumulées avec les valeurs manquantes dans les données : StatBox propose de supprimer les lignes correspondantes ou d'estimer les valeurs manquantes par la moyenne des poids (*cf.* l'option « **Estimation des données manquantes** »), calculée sans tenir compte des éventuels poids nuls.

*Onglet « Variables et observations supplémentaires »*

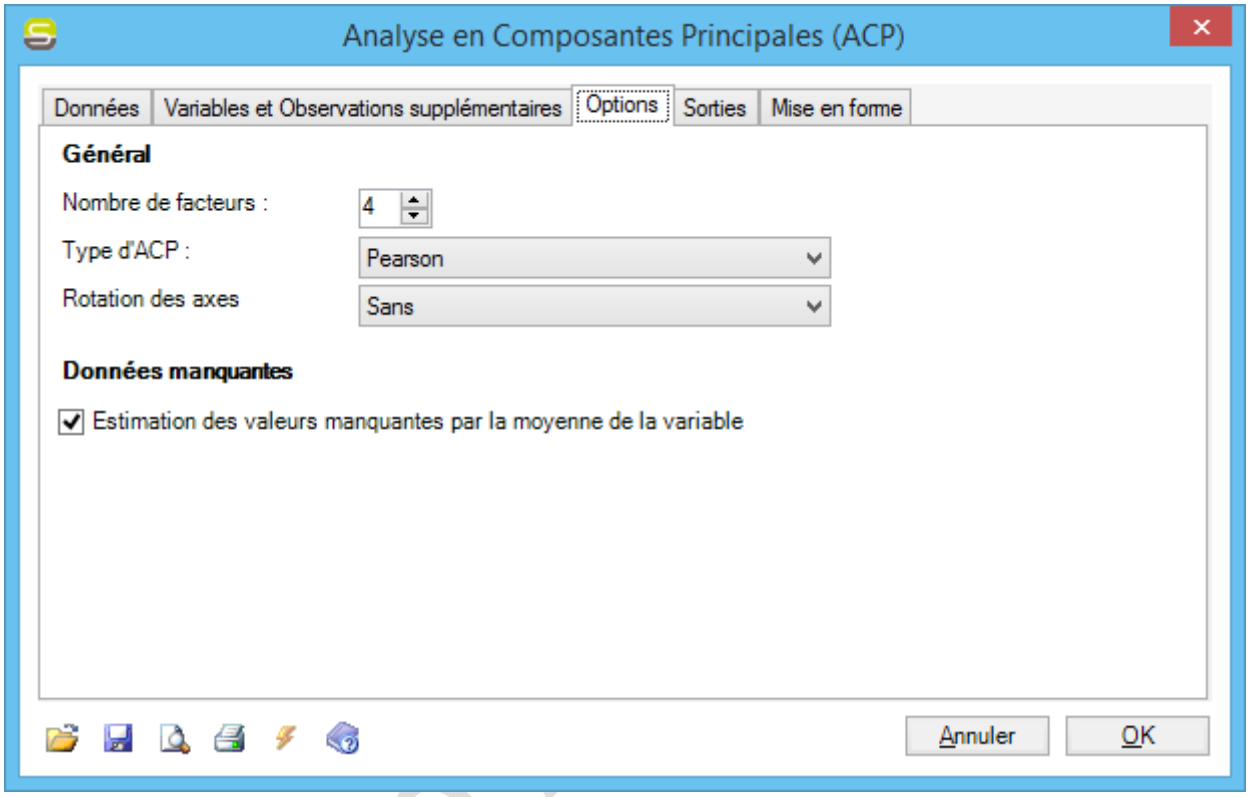



- *Variable(s) quantitative(s) supplémentaire(s) :* dans le cas d'un tableau observations / variables, saisissez les variables supplémentaires ou passives. Les variables passives ne participent pas aux calculs mais sont positionnées sur les plans factoriels avec les variables actives. Les valeurs manquantes sont cumulées avec les valeurs manquantes dans les données actives : StatBox propose d'ignorer les lignes correspondantes ou d'estimer les valeurs manquantes par la moyenne de la variable (*cf.* l'option "*Estimation des données manquantes*").
- *Variable qualitative supplémentaire :* dans le cas d'un tableau observations/variables, saisissez la variable qualitative supplémentaire. Les *m* modalités de cette variable définissent *m* groupes d'observations, chaque groupe étant représenté sur les plans factoriels par son barycentre. Les valeurs manquantes sont cumulées avec les valeurs manquantes dans les données actives. En cas de données manquantes, StatBox propose

de supprimer les lignes correspondantes ou d'estimer les valeurs manquantes par le mode de la variable (*cf.* l'option "*Estimation des données manquantes*").

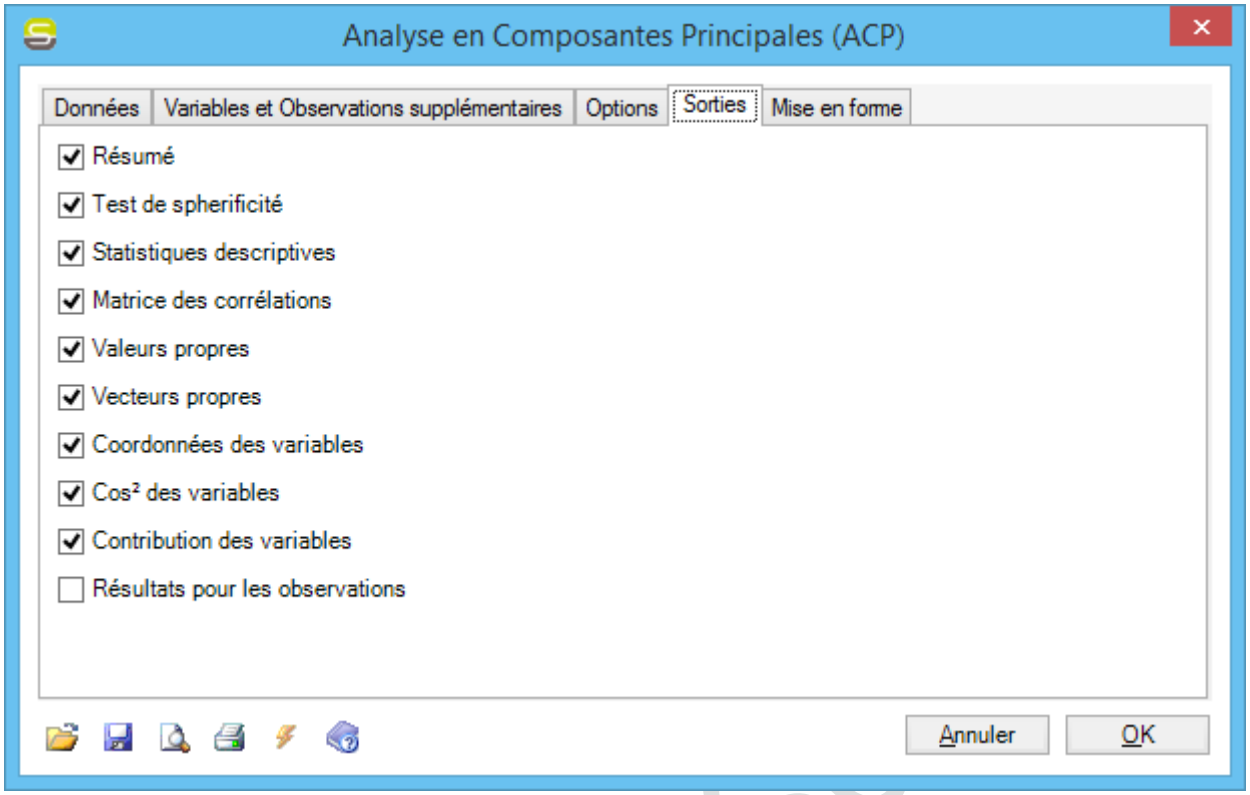
- *Individu supplémentaire à partir de la ligne :* dans le cas d'un tableau observations/variables, saisissez la ligne à partir de laquelle débute la zone des observations supplémentaires ou passives. Les observations passives ne participent pas aux calculs mais sont positionnés sur les plans factoriels avec les observations actives. Les valeurs manquantes sont cumulées avec les valeurs manquantes dans les données actives : StatBox propose d'ignorer les lignes correspondantes ou d'estimer les valeurs manquantes par la moyenne de la variable (*cf.* l'option "***Estimation des données manquantes***"), calculée grâce à la totalité de l'information disponible, c'est-à-dire en tenant compte des observations supplémentaires.

*Onglet « Options »*




- *Nombre de facteurs :* entrez le nombre de facteurs maximal à considérer. Tous calculs faits, StatBox peut éventuellement afficher moins de facteurs que le nombre de facteurs demandé.
- *Type d'ACP :* dans le cas d'un tableau observations/variables, si vous souhaitez effectuer une ACP normée, choisissez le type de corrélation, paramétrique (Pearson), ou non paramétrique (Spearman, Kendall), ou choisissez « Covariance (n) » pour effectuer une ACP non normée.
- *Rotation des axes :* choisissez éventuellement le type de rotation des axes, Varimax ou Quartimax. Pour plus d'information consultez l'annexe consacrée aux rotations des axes.
- *Estimation des données manquantes par la moyenne de la variable :* cochez cette option pour que les données manquantes soient automatiquement estimées par la moyenne des variables concernées.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Test de sphericité :* affiche un test de Bartlett. Ce test permet de vérifier l'hypothèse selon laquelle les variables ne sont pas corrélées.
- *Statistiques descriptives :* affiche pour chaque variable sélectionnée (active ou passive) des statistiques descriptives simples (moyenne et écart type).
- *Matrice des corrélations :* affiche la matrice de corrélation ou de covariance.
- *Valeurs propres :* affiche les valeurs propres, le % de variance expliquée et le graphique correspondant. Le nombre de valeurs propres est égal au nombre de valeurs propres non nulles.
- *Vecteurs propres :* affiche la table des vecteurs propres.
- *Coordonnées des variables :* affiche la table des coordonnées des variables dans le nouvel espace de configuration.
- *Cos² des variables :* affiche la table des cosinus carrés des variables. L'analyse des cosinus carrés permet d'éviter des erreurs d'interprétation dues à des effets de projection.
- *Contribution des variables :* affiche la table des contributions des variables. Les contributions sont une aide à l'interprétation, les variables ayant le plus influencé la construction des axes sont celles dont les contributions sont les plus élevées.
- *Résultats pour les observations :* dans le cas d'un tableau observations/variables, affiche les résultats concernant les observations (coordonnées, cosinus carrés, contributions).

**Remarques :** contrairement aux variables actives, les variables quantitatives supplémentaires ne constituent pas des axes d'origine pour le positionnement des observations, leur représentation sur le graphique observations/variables est donc laissée à l'initiative de l'utilisateur.

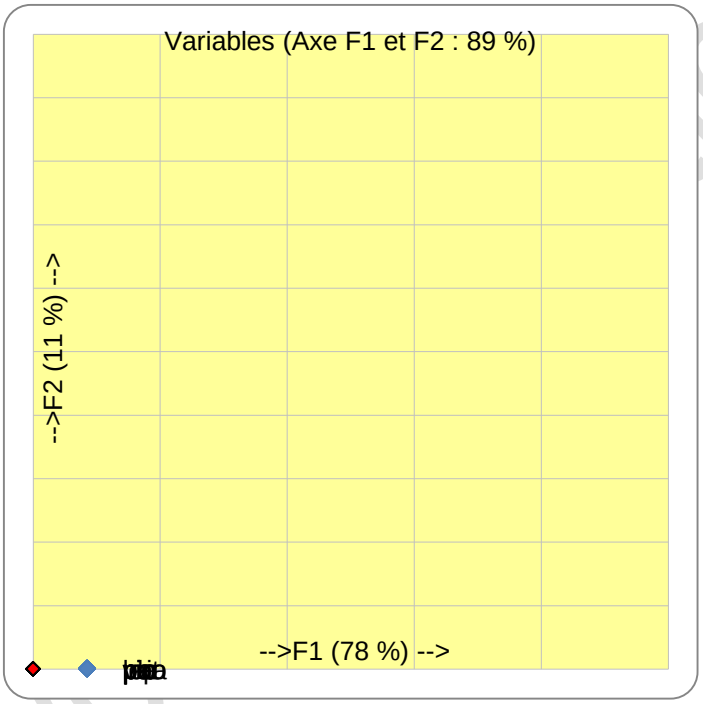
Au lancement de la procédure sélectionnez les options d'affichage des mappings (pour plus d'information consultez l'annexe « Boite d'affichage des graphiques »), et validez.

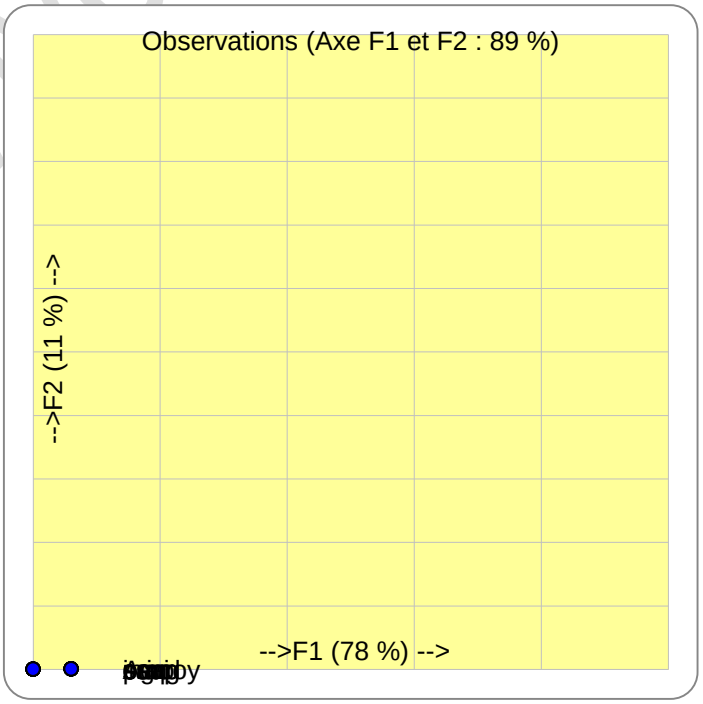
## Exemple

Exemple tiré de l'ouvrage de G. Saporta, Probabilité Analyse des données et statistique, Editions Technip, page 182

| | pao | paa | vio | via | pot | lec | rai | plp |
|---|---|---|---|---|---|---|---|---|
| pao | 1,0 | - 0,774 | 0,926 | - 0,906 | 0,656 | 0,889 | - 0,833 | - 0,856 |
| paa | - 0,774 | 1,0 | - 0,604 | 0,904 | - 0,333 | - 0,673 | 0,959 | 0,771 |
| vio | 0,926 | - 0,604 | 1,0 | - 0,750 | 0,517 | 0,792 | - 0,669 | - 0,828 |
| via | - 0,906 | 0,904 | - 0,750 | 1,0 | - 0,419 | - 0,839 | 0,924 | 0,720 |
| pot | 0,656 | - 0,333 | 0,517 | - 0,419 | 1,0 | 0,603 | - 0,410 | - 0,554 |
| lec | 0,889 | - 0,673 | 0,792 | - 0,839 | 0,603 | 1,0 | - 0,824 | - 0,751 |
| rai | - 0,833 | 0,959 | - 0,669 | 0,924 | - 0,410 | - 0,824 | 1,0 | 0,834 |
| plp | - 0,856 | 0,771 | - 0,828 | 0,720 | - 0,554 | - 0,751 | 0,834 | 1,0 |

En gras valeurs significatives au seuil alpha= 0,05 (test bilatéral)



### Références

**Dillon W.R. & M. Goldstein (1984)**. Multivariate analysis. Methods and applications. John Wiley & Sons, New York, pp. 23-52.

**Escofier B. & J. Pages (1990)**. Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation. Dunod, Paris, pp. 7-24.

**Jobson J.D. (1992)**. Applied multivariate data analysis. Volume II: categorical and multivariate methods. Springer-Verlag, New York, pp. 345-388.

**Johnson R.A. & D.W. Wichern (1992)**. Applied multivariate statistical analysis. Prentice-Hall, Englewood Cliffs, pp. 356-395.

**Lebart L., A. Morineau & M. Piron (1997)**. Statistique exploratoire multidimensionnelle. 2ème édition. Dunod, Paris, pp. 32-66.

**Saporta G. (1990)**. Probabilités, analyse des données et statistique. Technip, Paris, pp. 159-186.

**Sharma S. (1996)**. Applied multivariate techniques. John Wiley & Sons, New York, pp. 58-89.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 134-143.

## ANALYSE FACTORIELLE DES CORRESPONDANCES (AFC)

Utilisez l'analyse factorielle des correspondances afin d'étudier la liaison entre deux ensembles de modalités constituant les lignes et les colonnes d'un tableau de contingence.

### Description

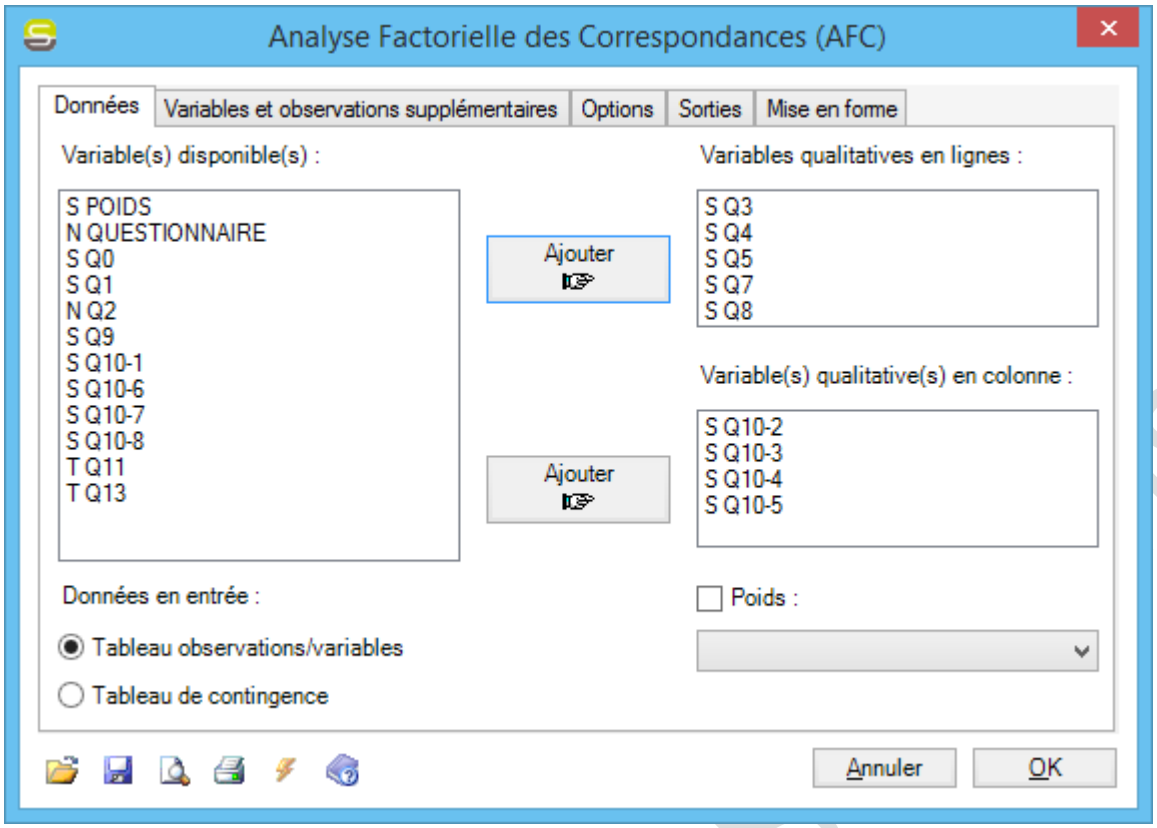
L'analyse factorielle des correspondances (AFC) consiste à rechercher la meilleure représentation simultanée de deux ensembles constituant les lignes et les colonnes d'un tableau de contingence, ces deux ensembles jouant un rôle symétrique. L'AFC peut se ramener à une analyse en composantes principales (ACP) en effectuant les changements de variables appropriés, et constitue également un cas particulier de l'analyse factorielle discriminante (AFD).

### Mise en œuvre

*Onglet « Données »*

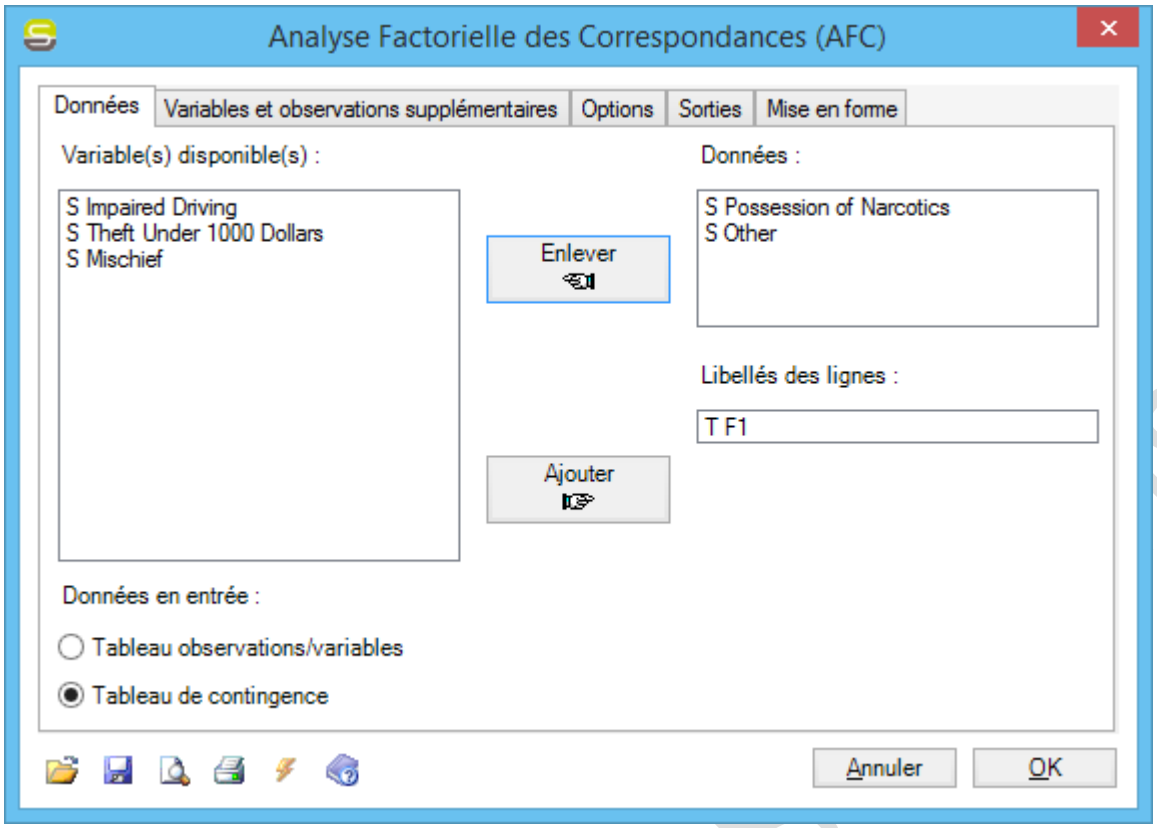
- *« Tableau observations/variables » / « Tableau de contingence »* : choisissez la nature des données, soit sous la forme d'un tableau observations/variables, soit directement sous la forme d'un tableau de contingence.

*Pour un tableau observations / variables* :




- *Variables qualitatives en lignes :* dans le cas d'un tableau observations/variables, saisissez les variables qualitatives dont les modalités constitueront les lignes du tableau de contingence. Lorsqu'il y a des valeurs manquantes, StatBox propose de les ignorer lors de la construction du tableau de contingence. En cas de refus, StatBox propose alors d'estimer les valeurs manquantes par le mode de la variable (*cf.* l'option "*Estimation des données manquantes*"), sinon la boîte de dialogue est fermée et le traitement est abandonné.
- *Variables qualitatives en colonnes* : dans le cas d'un tableau observations/variables, saisissez les variables qualitatives dont les modalités constitueront les colonnes du tableau de contingence. Lorsqu'il y a des valeurs manquantes, StatBox propose de les ignorer lors de la construction du tableau de contingence. En cas de refus, StatBox propose alors d'estimer les valeurs manquantes par le mode de la variable (*cf.* l'option "*Estimation des données manquantes*"), sinon la boîte de dialogue est fermée et le traitement est abandonné.
- *Poids :* dans le cas d'un tableau observations/variables, saisissez la variable poids des observations. Lorsqu'il y a des valeurs manquantes dans les poids, StatBox propose de supprimer les lignes correspondantes ou d'estimer les valeurs manquantes par la moyenne des poids (*cf.* l'option "*Estimation des données manquantes*"), calculée sans tenir compte des éventuels poids nuls.

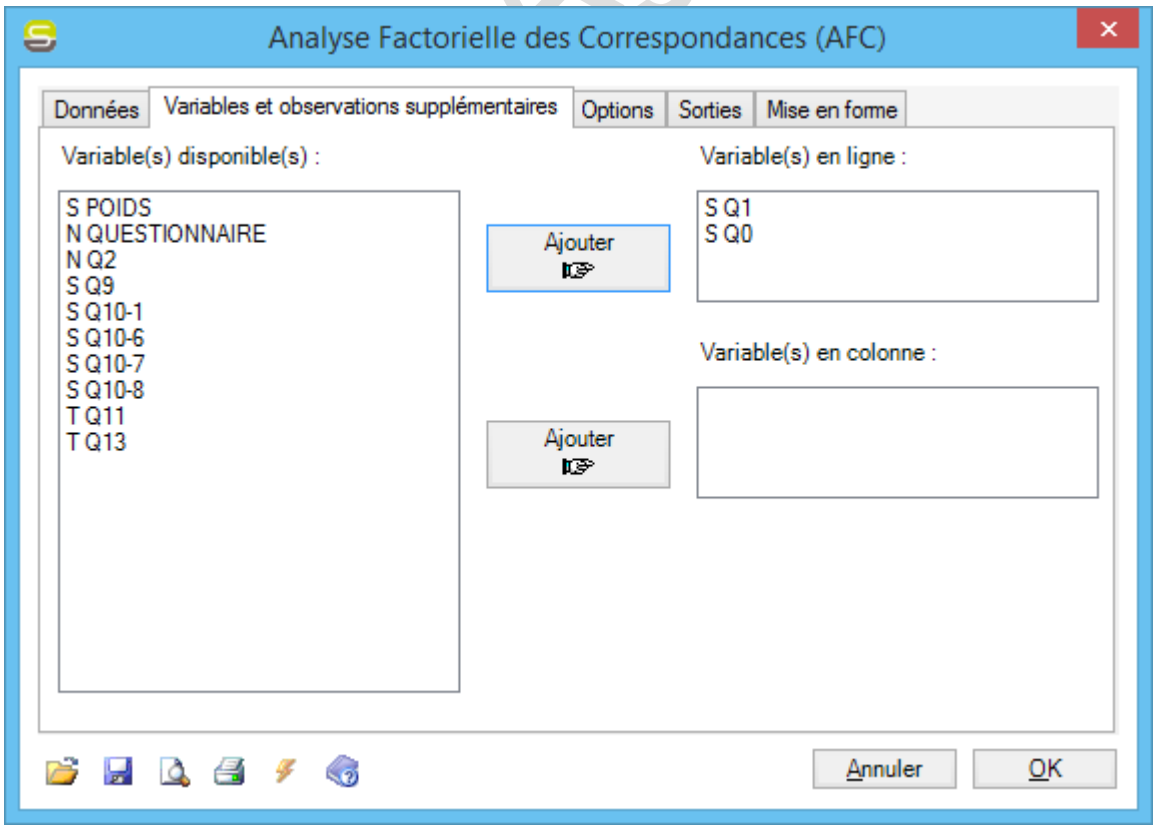
Pour un tableau de contingence :




- *Données :* saisissez les variables colonnes du tableau. Les valeurs manquantes ne sont pas autorisées.
- *Libellés des lignes :* sélectionnez la variable contenant les libellés des lignes du tableau de contingence (facultatif).

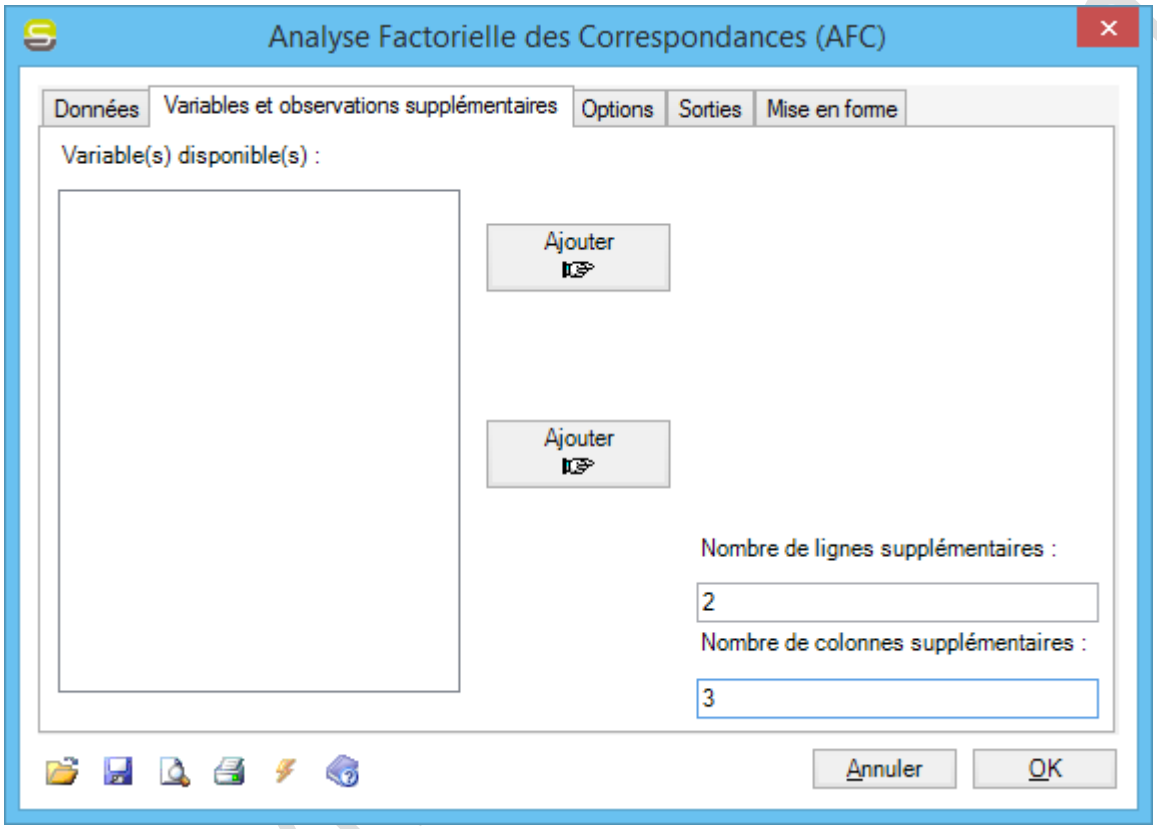
*Onglet « Variables et observations supplémentaires »*

Pour un tableau Observations/ variables :



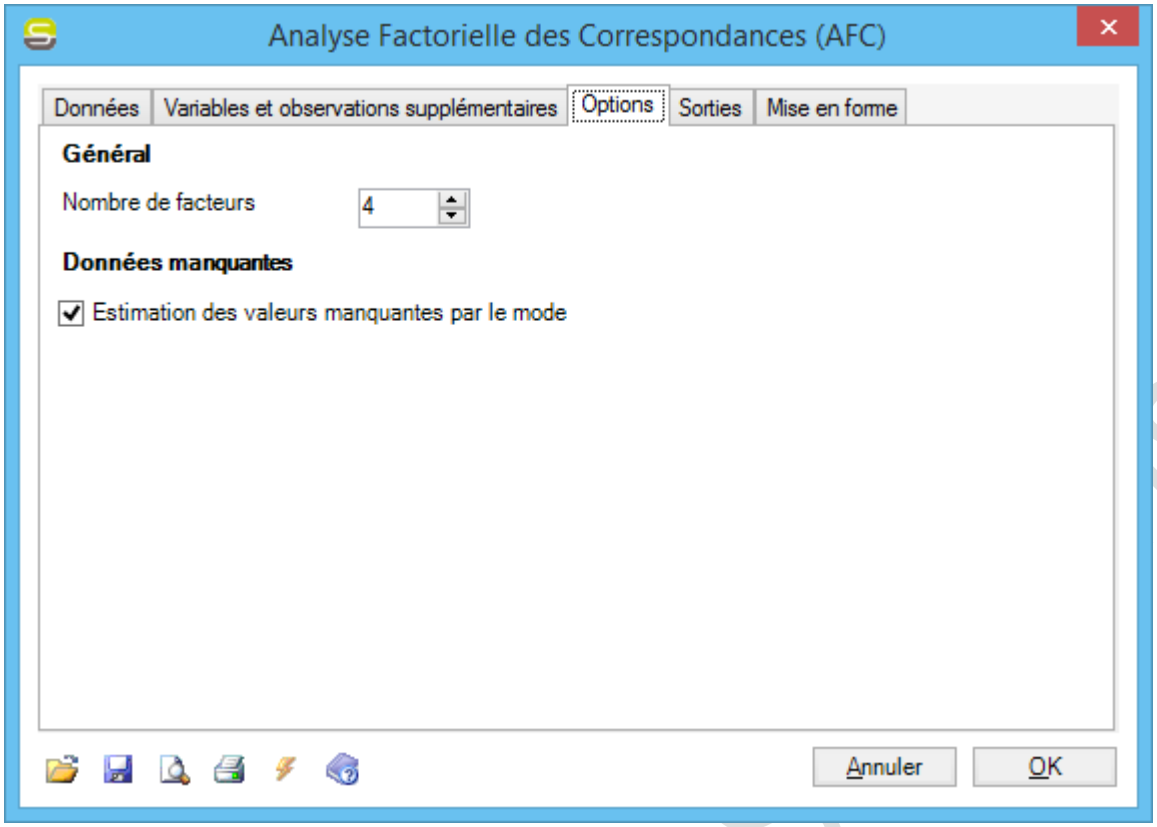
- *Variable(s) en ligne(s) supplémentaire(s)* : dans le cas d'un tableau observations/variables, saisissez la/les variable(s) qualitative(s) supplémentaire(s) dont les modalités constitueront les lignes supplémentaires du tableau de contingence. Les valeurs manquantes sont cumulées avec les valeurs manquantes dans les données actives : StatBox propose de les ignorer lors de la construction du tableau de contingence. En cas de refus, StatBox propose alors d'estimer les valeurs manquantes par le mode de la variable (*cf.* l'option "*Estimation des données manquantes*"), sinon le traitement est abandonné.
- *Variable(s) en colonne(s) supplémentaire(s) :* dans le cas d'un tableau observations/variables, saisissez la/les variable(s) qualitative(s) supplémentaire(s) dont les modalités constitueront les colonnes supplémentaires du tableau de contingence. Les valeurs manquantes sont cumulées avec les valeurs manquantes dans les données actives : StatBox propose de les ignorer lors de la construction du tableau de contingence. En cas de refus, StatBox propose alors d'estimer les valeurs manquantes par le mode de la variable (*cf.* l'option "*Estimation des données manquantes*"), sinon le traitement est abandonné.

Pour un tableau de contingence :



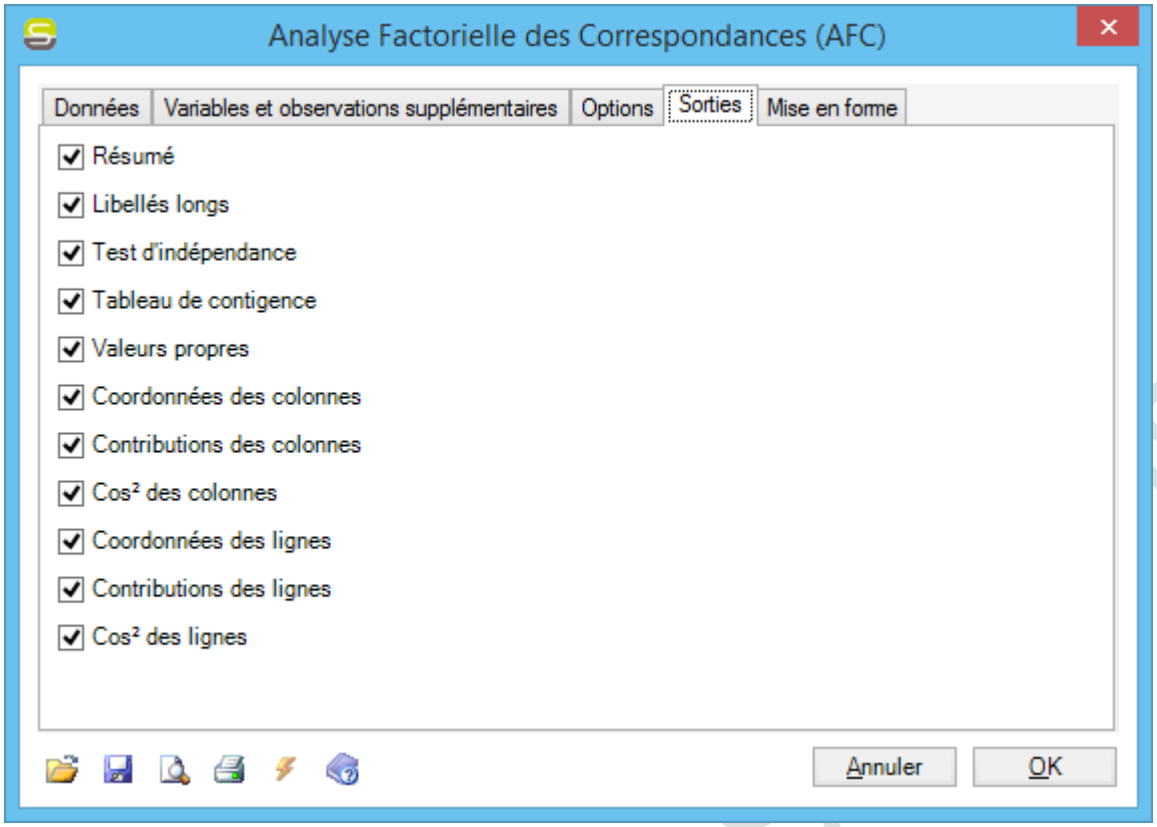

- *Nombre de lignes supplémentaires :* dans le cas d'un tableau de contingence, entrez le nombre de lignes consécutives à la fin du tableau correspondant aux lignes supplémentaires (lignes passives).
- *Nombre de colonnes supplémentaires :* dans le cas d'un tableau de contingence, entrez le nombre de colonnes consécutives à la droite du tableau correspondant aux colonnes supplémentaires (colonnes passives).

*Onglet « Options »*




- *Nombre de facteurs :* entrez le nombre de facteurs maximal à considérer. Tous calculs faits, StatBox peut éventuellement afficher moins de facteurs que le nombre de facteurs demandé.
- *Estimation des valeurs manquantes par le mode :* cochez cette option pour que les données manquantes soient estimées automatiquement par le mode des variables considérées. Si cette option n'est pas cochée et qu'il y a des données manquantes alors le logiciel vous proposera de faire cette estimation au cours de la procédure.

*Onglet « Sorties »*

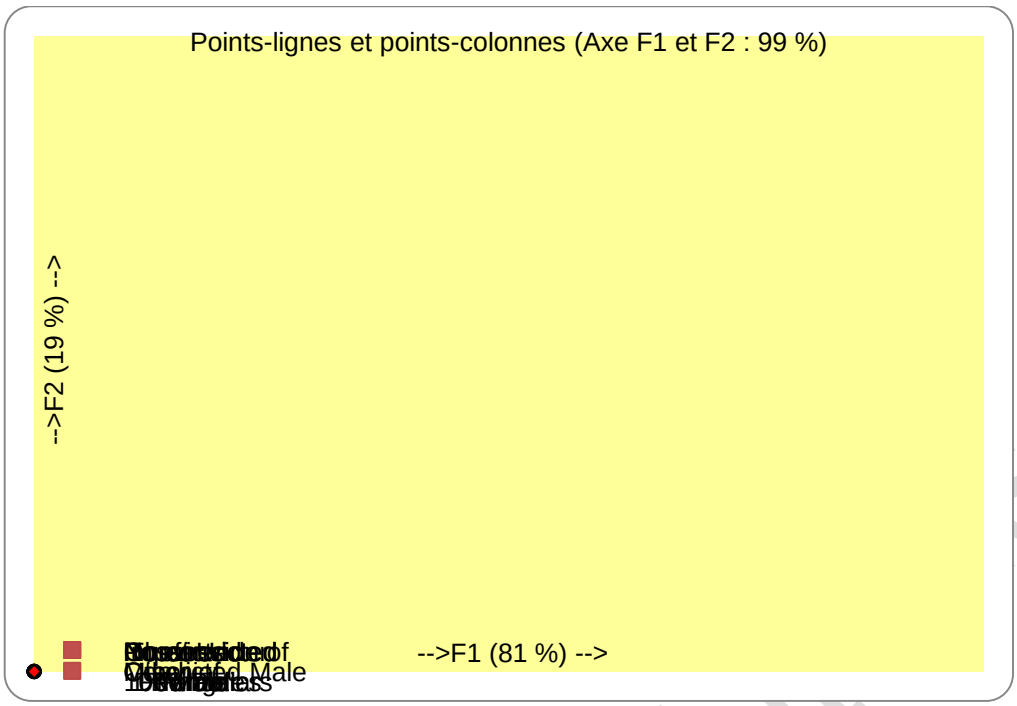



- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Libellés longs :* utilise les libellés longs des variables lorsque ceux-ci sont disponibles.
- *Tests d'indépendance :* affiche un test d'indépendance basé sur la statistique du Khi².
- *Tableau de contingence :* affiche la table de dénombrement des croisements de modalités pour les variables sélectionnées.
- *Valeurs propres :* affiche les valeurs propres, le % de variance expliquée et le graphique correspondant. Le nombre de valeurs propres est égal au nombre de valeurs propres non nulles.
- *Coordonnées des colonnes :* affiche les coordonnées principales des points colonnes dans le plan factoriel.
- *Contributions des colonnes :* affiche les contributions des points colonnes.
- *Cos² des colonnes :* affiche les cosinus carrés des colonnes dans le plan factoriel.
- *Coordonnées des lignes :* affiche les coordonnées principales des points ligne dans le plan factoriel.
- *Contributions des lignes :* affiche les contributions des points lignes.
- *Cos² des lignes :* affiche les cosinus carrés des lignes dans le plan factoriel.

Au lancement de la procédure, sélectionnez les options d'affichage des mappings (pour plus d'information consultez l'annexe « Boite d'affichage des graphiques »), et validez.

## Exemple

Tableau de contingence de la feuille " AFC " du classeur " Data.xls " (Jobson 1992, table 9.39, p. 434)




## Références

**Escofier B. & J. Pages (1990)**. Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation. Dunod, Paris, pp. 25-45.

**Jobson J.D. (1992)**. Applied multivariate data analysis. Volume II: categorical and multivariate methods. Springer-Verlag, New York, pp. 433-462.

**Lebart L., A. Morineau & M. Piron (1997)**. Statistique exploratoire multidimensionnelle. 2ème édition. Dunod, Paris, pp. 67-107.

**Saporta G. (1990)**. Probabilités, analyse des données et statistique. Technip, Paris, pp. 199-216, pp. 199-216.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 143-150.

# ANALYSE DES CORRESPONDANCES MULTIPLES (ACM)

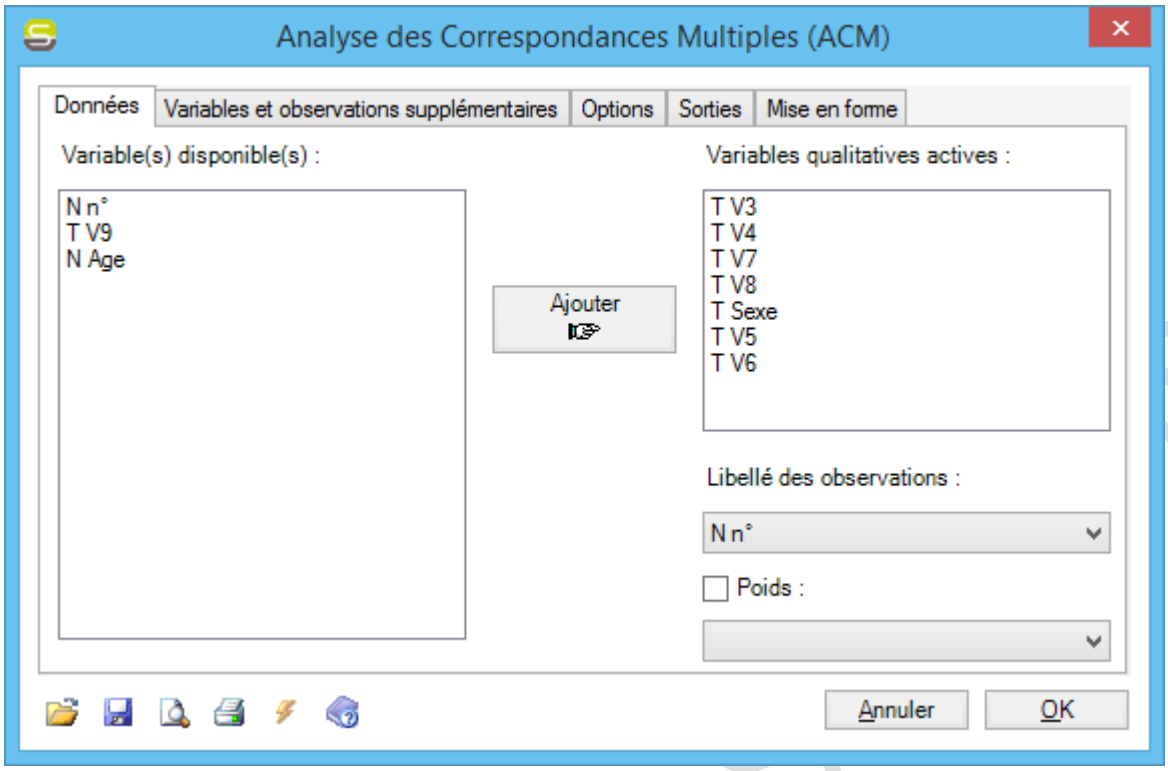
Utilisez l'analyse des correspondances multiples afin d'étudier des données sous la forme d'un tableau d'observations décrits par plusieurs variables qualitatives. Cette méthode est particulièrement adaptée à l'analyse d'enquêtes pour lesquelles les lignes du tableau sont en général des individus (il peut en exister plusieurs milliers) et les colonnes sont des modalités de variables qualitatives, le plus souvent des modalités de réponse à des questions.

## Description

L'analyse des correspondances multiples (ACM) est une extension de l'analyse factorielle des correspondances (AFC) appliquée non plus à un tableau de contingence, mais à un tableau disjonctif complet. Cette méthode peut être vue également comme l'équivalent de l'analyse en composantes principales (ACP) pour des variables qualitatives.

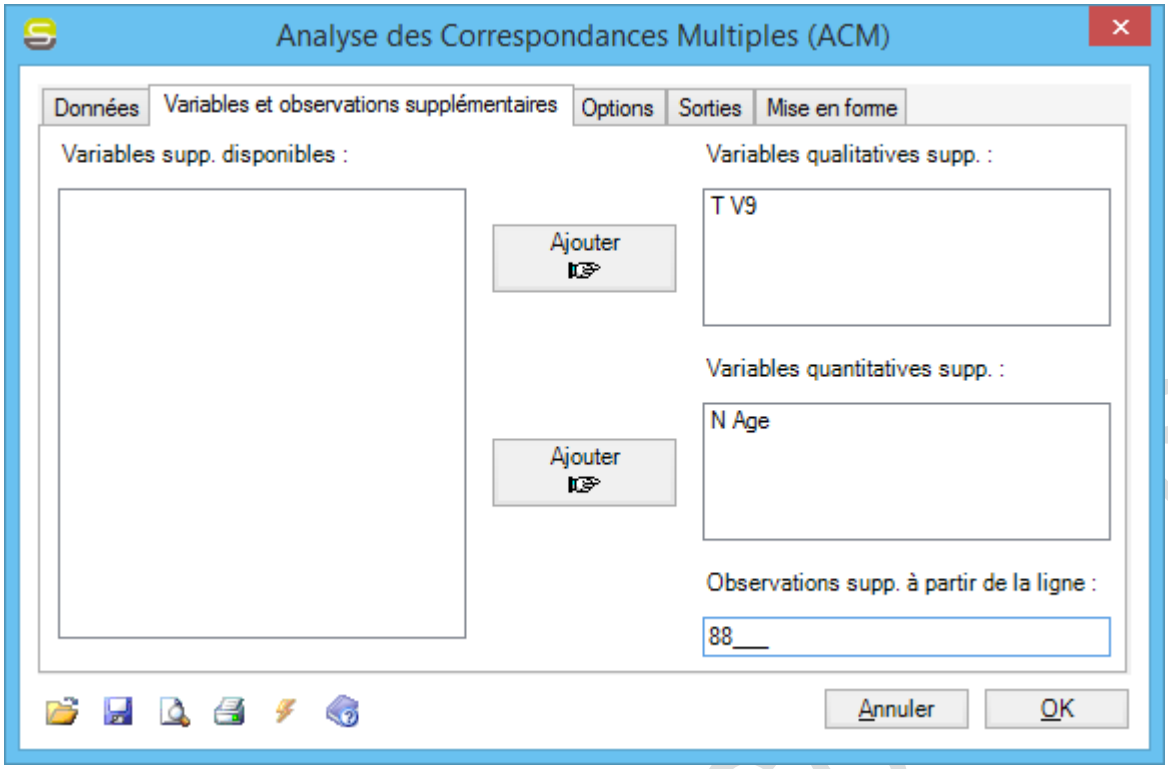
## Mise en œuvre

*Onglet « Données »*



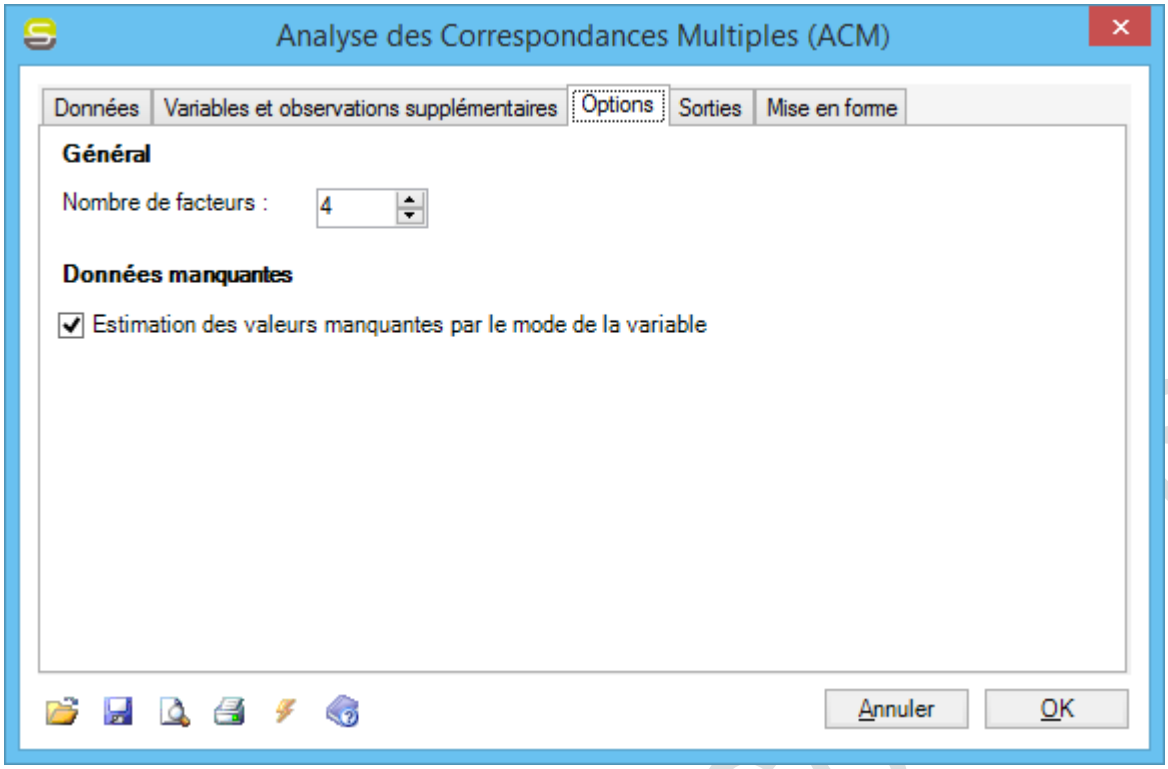

- *Variables qualitatives actives :* saisissez les variables des données, correspondant à un tableau observations/variables. Lorsqu'il y a des valeurs manquantes StatBox propose tout d'abord de les ignorer. En cas de refus, StatBox propose alors d'estimer les valeurs manquantes de chaque variable par le mode (*cf.* l'option "*Estimation des données manquantes*"), sinon la boîte de dialogue est fermée et le traitement est abandonné.
- *Libellés des observations :* sélectionnez la variable contenant les libellés qui correspondent aux lignes du tableau de données.
- *Poids :* cochez cette option pour pondérer vos observations et sélectionnez la colonne des poids des observations. Les poids nuls ne sont pas autorisés. Lorsqu'il y a des valeurs manquantes dans les poids, StatBox propose de les estimer par la moyenne des poids (*cf.* l'option "*Estimation des données manquantes*"), calculée sans tenir compte des éventuels poids nuls. Sinon le traitement est abandonné car des poids manquants sont équivalents à des poids nuls, lesquels sont interdits.

*Onglet « Variables et observations supplémentaires »*



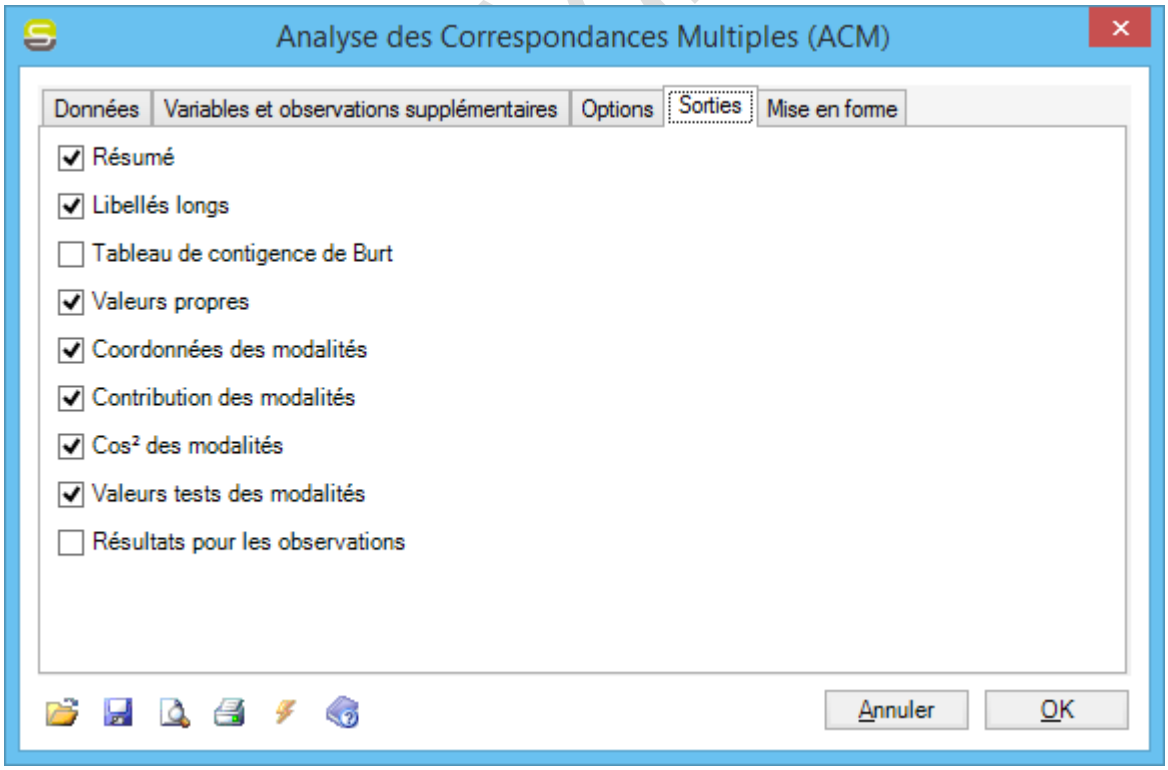

- *Variable(s) qualitative(s) supplémentaire(s)* : saisissez la/les variable(s) supplémentaire(s) ou passive(s). Les variables passives ne participent pas aux calculs mais sont positionnées sur les plans factoriels avec les variables actives. Les valeurs manquantes sont cumulées avec les valeurs manquantes dans les données actives : StatBox propose de les ignorer, et dans le cas d'un tableau observations/variables, de les estimer par le mode de la variable (*cf.* l'option "*Estimation des données manquantes*").
- *Variables quantitatives supplémentaires* : saisissez la/les variable(s). Lorsqu'il y a des valeurs manquantes pour une variable, StatBox propose de les estimer par la moyenne de la variable (*cf.* l'option "*Estimation des données manquantes*"), sinon le traitement est abandonné, parce que les valeurs manquantes pour les variables quantitatives supplémentaires sont interdites.
- *Observations supp. à partir de la ligne* : saisissez la ligne à partir de laquelle débutent les observations supplémentaires ou passives. Les observations passives ne participent pas aux calculs mais sont positionnés sur les plans factoriels avec les observations actives. Les valeurs manquantes sont cumulées avec les valeurs manquantes dans les données actives : StatBox propose de les ignorer, et dans le cas d'un tableau observations/variables, de les estimer par le mode de la variable (*cf.* l'option "*Estimation des données manquantes*"), calculé à partir de la totalité de l'information disponible, c'est-à-dire en tenant compte des observations supplémentaires).

*Onglet « Options »*




- *Nombre de facteurs :* entrez le nombre de facteurs maximal à considérer. Tous calculs faits, StatBox peut éventuellement afficher moins de facteurs que le nombre de facteurs demandé.
- *Estimation des données manquantes par le mode de la variable :* cochez cette option pour que les données manquantes soient estimées automatiquement par le mode des variables concernées.

*Onglet « Sorties »*



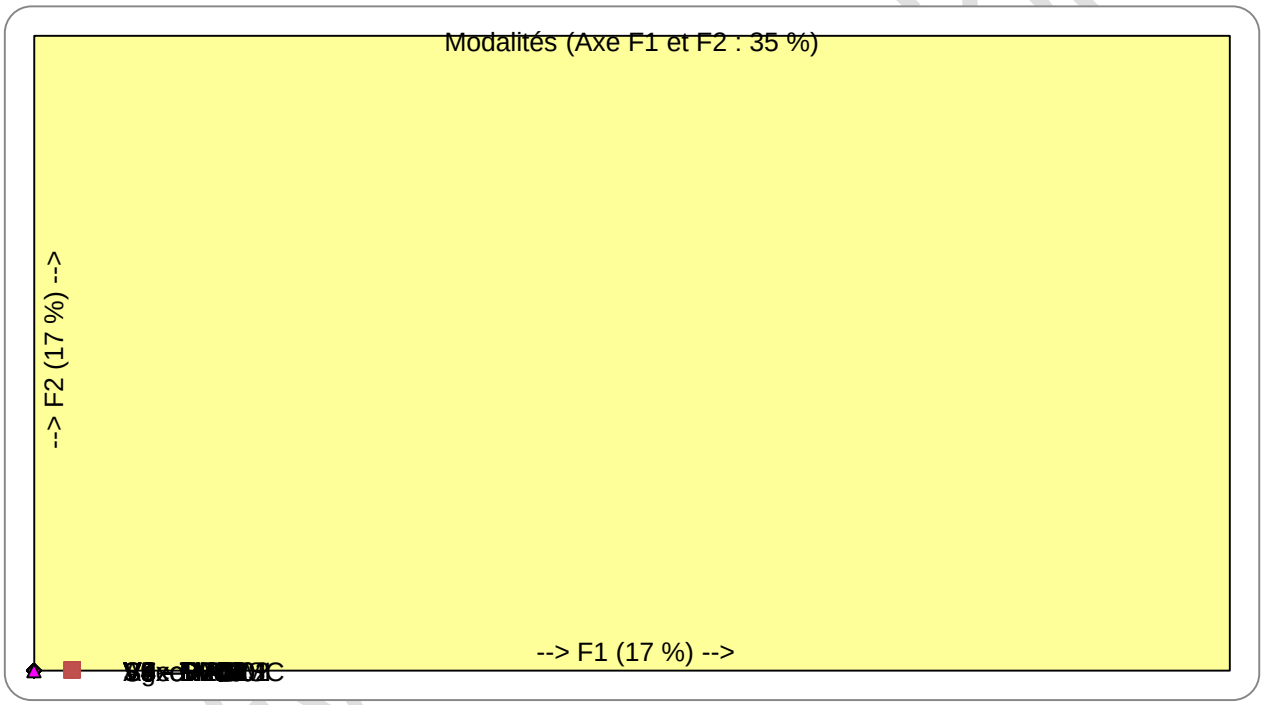

- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Libellés longs :* utilise les libellés longs des variables lorsque ceux-ci sont disponibles.

- *Tableau de contingence de Burt :* affiche la table de contingence de Burt
- *Valeurs propres :* affiche les valeurs propres, le % de variance expliquée et le graphique correspondant. Le nombre de valeurs propres est égal au nombre de valeurs propres non nulles.
- *Coordonnées des variables :* affiche la table des coordonnées des variables dans le nouvel espace de configuration.
- *Cos² des variables :* affiche la table des cosinus carrés des variables. L'analyse des cosinus carrés permet d'éviter des erreurs d'interprétation dues à des effets de projection.
- *Valeurs tests des modalités :* affiche les valeurs test pour les variables.
- *Résultats pour les observations :* dans le cas d'un tableau observations/variables, affiche les résultats concernant les observations (coordonnées, cosinus carrés, contributions).

Au lancement de la procédure, sélectionnez les options d'affichage des mappings (pour plus d'information consultez l'annexe « Boite d'affichage des graphiques »), et validez.

## Exemple

ACM sur le tableau de la feuille « ACM » du classeur « Data.xls » (Lebart et al. 1997, tableau 1.4-2, p. 136).




## Références

**Escofier B. & J. Pages (1990)**. Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation. Dunod, Paris, pp. 47-66.

**Jobson J.D. (1992)**. Applied multivariate data analysis. Volume II: categorical and multivariate methods. Springer-Verlag, New York, pp. 462-465.

**Lebart L., A. Morineau & M. Piron (1997)**. Statistique exploratoire multidimensionnelle. 2ème édition. Dunod, Paris, pp. 108-142.

**Saporta G. (1990)**. Probabilités, analyse des données et statistique. Technip, Paris, pp. 217-239.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 150-155.

# ANALYSE FACTORIELLE DISCRIMINANTE (AFD)

Utilisez l'analyse factorielle discriminante pour classer de nouvelles observations décrites par plusieurs variables quantitatives, connaissant un échantillon d'observations décrits par les mêmes variables, dont les groupes sont connus, et pour analyser la façon dont les variables descriptives contribuent à la constitution des différents groupes.

**Remarque :** l'analyse factorielle discriminante est étroitement liée à l'analyse de variance multivariée (MANOVA).

## Description

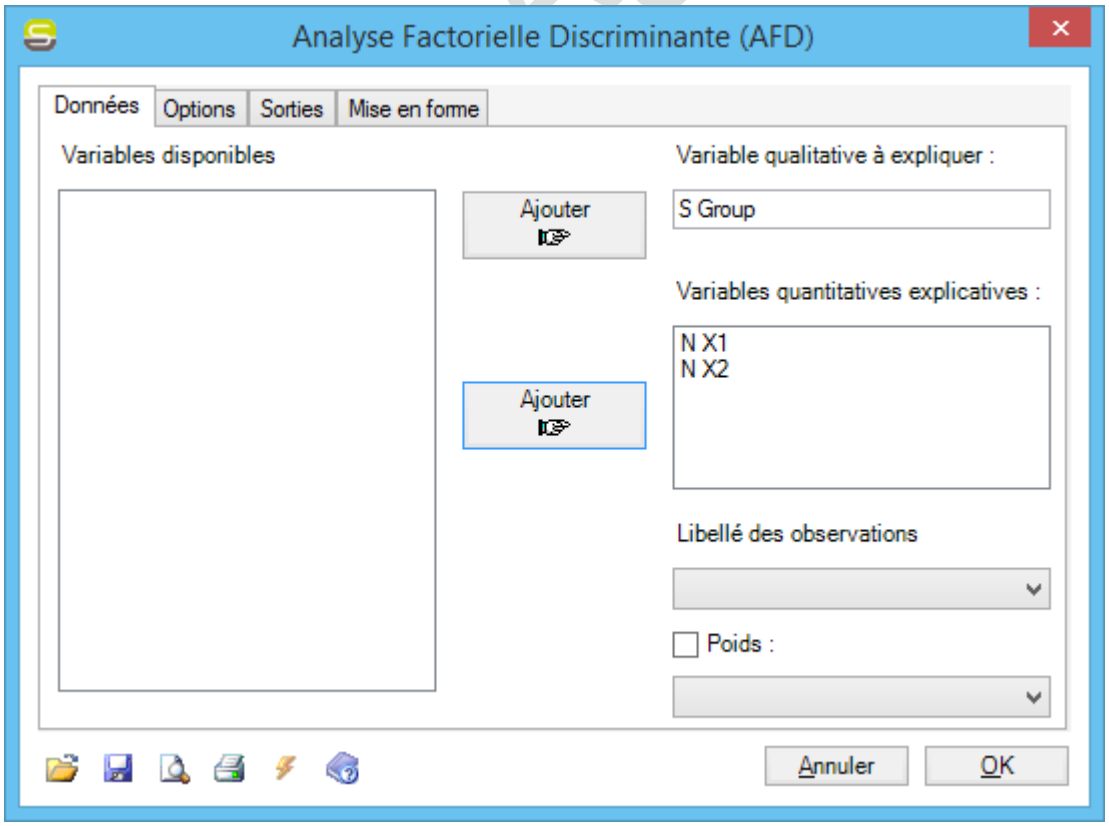
L'analyse factorielle discriminante (AFD) est une méthode permettant de modéliser l'appartenance à un groupe d'observations en fonction des valeurs prises par plusieurs variables, puis de déterminer le groupe le plus probable pour une observation, connaissant uniquement les valeurs des variables qui le caractérisent. Dans StatBox, les variables qui décrivent les observations sont forcément des variables quantitatives, les groupes étant spécifiés par une variable qualitative. L'AFD peut être considérée comme une extension de la régression multiple dans le cas où la variable à expliquer est une variable qualitative décrivant des groupes.

**Remarque :** les calculs de l'AFD ne peuvent pas s'exécuter si les variables explicatives sont linéairement dépendantes (*multicolinéarité*). En conséquence, aucune variable ne doit pouvoir être déduite des autres par une relation linéaire. Par exemple, dans un jeu de variables explicatives correspondant aux pourcentages de votes exprimés pour un ensemble de candidats, il convient de ne pas inclure parmi les variables explicatives le pourcentage de votes non exprimés puisque cette variable se déduit linéairement de toutes les autres (100 % moins la somme des pourcentages de votes exprimés). Jusqu'à 50 variables explicatives, StatBox propose de vérifier automatiquement que les variables explicatives sont bien linéairement indépendantes, en calculant la corrélation multiple de chaque variable avec toutes les autres. Vous pouvez également détecter le problème de la multicolinéarité avec le module « Matrice de similarité / dissimilarité », en calculant la matrice de corrélation entre les variables et en vérifiant qu'il n'y a pas de couples de variables fortement corrélées.

**Remarque :** ce module accepte jusqu'à 250 variables explicatives.

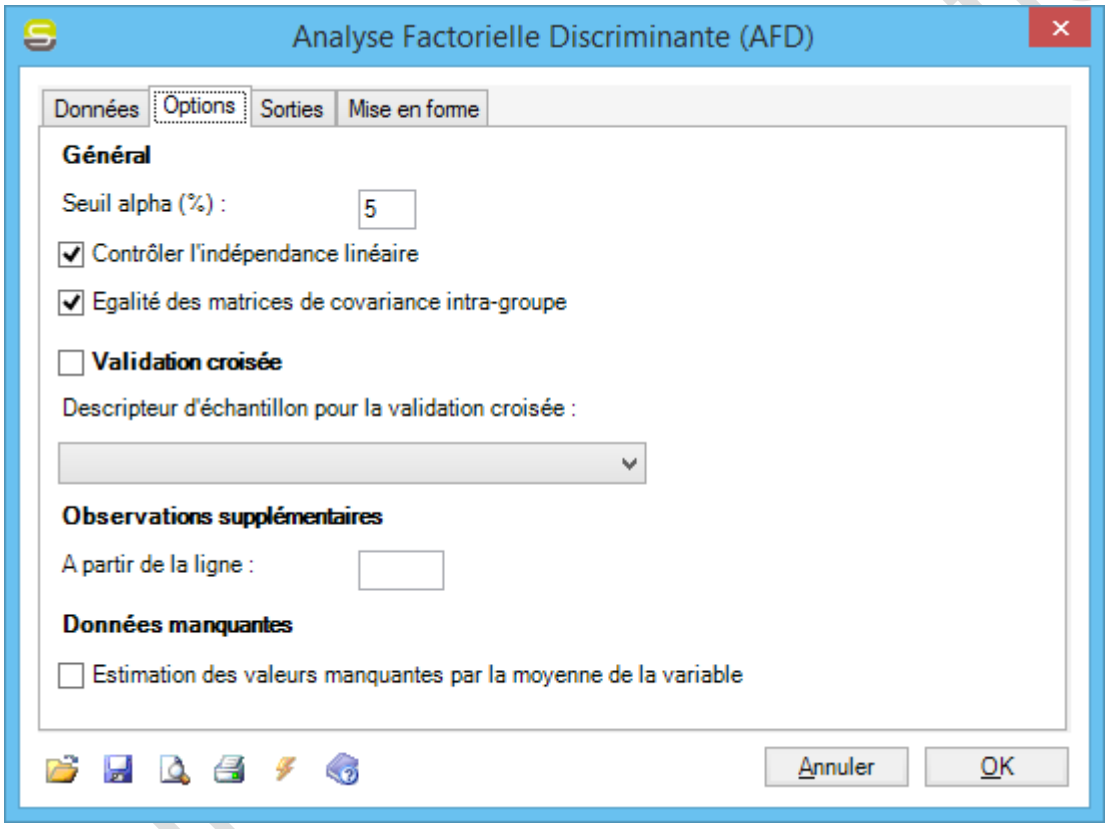
## Mise en œuvre

*Onglet « Données »*



- *Variable qualitative à expliquer :* saisissez la variable qualitative décrivant les groupes des observations. Pas de donnée manquante dans la variable groupe. Lorsqu'il y a des valeurs manquantes StatBox propose tout d'abord d'ignorer les lignes concernées. En cas de refus, StatBox propose alors d'estimer les valeurs manquantes par le mode de la variable (*cf.* l'option "Estimation des données manquantes"), sinon le traitement est abandonné.
- *Variables quantitatives explicatives* : saisissez les variables quantitatives qui doivent expliquer l'appartenance aux groupes. Les valeurs manquantes sont cumulées avec les éventuelles valeurs manquantes de la variable à expliquer. StatBox propose d'ignorer les lignes correspondantes ou d'estimer les valeurs manquantes de chaque variable par la moyenne (*cf.* l'option "Estimation des données manquantes").
- *Libellés des observations :* saisissez la variable de libellés qui correspondent aux lignes du tableau de données.
- *Poids :* saisissez la variable des poids des observations. Les valeurs manquantes dans les poids sont cumulées avec les valeurs manquantes dans les données actives : StatBox propose d'ignorer les lignes correspondantes ou d'estimer les valeurs manquantes par la moyenne des poids (*cf.* l'option "*Estimation des données manquantes*"), calculée sans tenir compte des éventuels poids nuls.

### *Onglet « Options »*



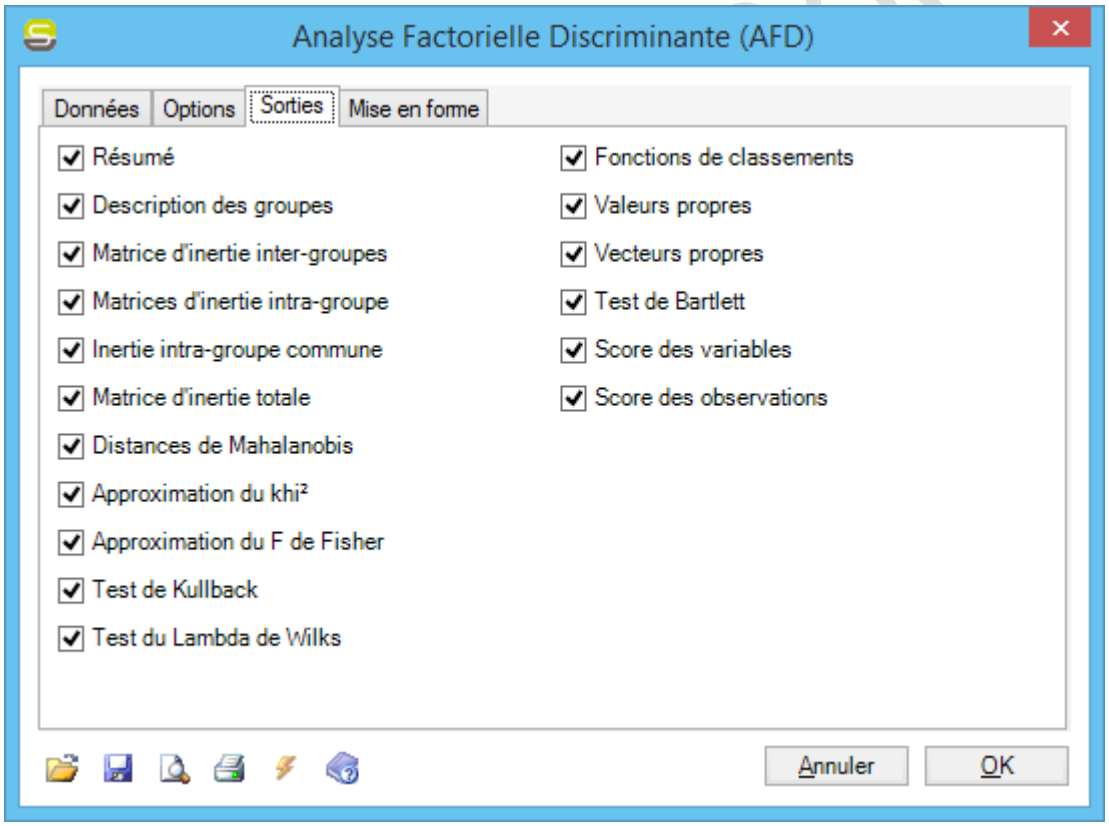

- *Seuil alpha (%) :* entrez la valeur du risque de première espèce des tests.
- *Contrôler l'indépendance linéaire :* cochez cette case afin que StatBox contrôle l'indépendance linéaire entre les variables (jusqu'à 50 variables). Lorsque le contrôle est désactivé et/ou lorsque le nombre de variables explicatives dépasse 50, le problème de la multicolinéarité est détecté lors des calculs de l'AFD eux-mêmes. L'analyse est alors interrompue : le message d'erreur affiché ne spécifie pas l'origine de l'échec de l'AFD mais signale que les calculs ne peuvent pas être effectués avec les données sélectionnées.
- *Egalité des matrices de covariance intra-groupe : cochez* cette case si vous faites l'hypothèse que les matrices de covariance pour les différents groupes ne sont pas significativement différentes. Un test est réalisé par StatBox afin de vous permettre de vérifier que votre hypothèse est raisonnable. Lorsque cette option est décochée, le tableau des carrés des distances de Mahalanobis entre groupe est différent, les *F* de Fisher associés et les *p*-values ne sont pas disponibles, les fonctions de classement sont différentes. Les autres calculs sont néanmoins effectués avec la matrice de covariance intra-groupe commune.

- *Validation croisée :* cochez cette case pour calculer le taux d'erreur de classement sur un échantillon-test, l'AFD étant effectuée sur un échantillon d'apprentissage, et saisissez la plage de la variable binaire indicatrice (1/0) désignant les observations de l'échantillon d'apprentissage (valeur 1) et les observations de l'échantillon-test (valeur 0). Les valeurs manquantes ne sont pas autorisées pour la variable indicatrice.

**Remarque :** le taux d'erreur de classement calculé uniquement sur l'échantillon d'apprentissage (c'est-à-dire sans validation croisée) augmente automatiquement avec le nombre de variables explicatives et peut s'avérer excellent si le nombre de variables est élevé, sans pour autant assurer que le modèle permette de prédire correctement les groupes des observations supplémentaires. Le taux de resubstitution calculé sur les données d'apprentissage ou *taux d'erreur apparent* s'avère donc plutôt optimiste puisqu'il sous-estime systématiquement le taux d'erreur réel. Il est préférable d'utiliser la validation croisée afin d'estimer le taux d'erreur par le taux de resubstitution calculé sur l'échantillon-test, en prenant par exemple 75 % des observations pour l'apprentissage et les 25 % qui restent pour l'estimation du taux d'erreur.

- *À partir de la ligne :* saisissez la ligne à partir de laquelle les observations supplémentaires ou passives commencent. Les observations passives ne participent pas aux calculs mais sont positionnés sur les plans factoriels avec les observations actives, et leur appartenance aux groupes est prédite par le modèle. Les valeurs manquantes sont cumulées avec les valeurs manquantes dans les données actives : StatBox propose d'ignorer les lignes correspondantes ou d'estimer les valeurs manquantes par la moyenne de la variable (*cf.* l'option "*Estimation des données manquantes*"), calculée grâce à la totalité de l'information disponible, c'est-à-dire en tenant compte des observations supplémentaires.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Description des groupes :* affiche des statistiques de base sur les groupes étudiés (fréquence, moyenne et écart-type).
- *Matrice d'inertie inter-groupe :* affiche la matrice d'inertie inter-groupe.
- *Matrice d'inertie intra-groupe :* affiche les matrices d'inertie intra-groupe
- *Inertie intra-groupe commune :* affiche la matrice d'inertie inter-groupe commune
- *Matrices d'inertie totale :* affiche la matrice d'inertie totale.

- *Distance de Mahalanobis :* affiche la table des distances de Mahalanobis qui permet de mesurer la distance entre les classes en tenant compte de la structure de covariance.
- *Approximation du Khi² :* effectue une approximation du Khi²
- *Approximation du F de Fisher :* effectue une approximation du F de Fisher.
- *Test de Kullback :* affiche un test de Kullback, ce test permet de tester l'hypothèse d'égalité des matrices de covariance intra-classe
- *Test du Lambda de Wilks :* affiche un test du Lambda de Wilks qui permet de tester l'hypothèse d'égalité des vecteurs moyens des différentes classes
- *Fonctions de classements* : affecte chaque observation à la classe pour laquelle la fonction de classement est la plus élevée. Les fonctions de classement sont utilisées pour déterminer à quelle classe doit être affectée une observation sur la base des valeurs prises pour les différentes variables explicatives.
- *Valeurs propres :* dans ce tableau sont affichées les valeurs propres associées aux différents facteurs, ainsi que les pourcentages et pourcentages cumulés de discrimination correspondant. En analyse discriminante, le nombre de valeurs propres non nulles est au plus égal à (k-1) où k est le nombre de classes.
- *Vecteurs propres :* affiche la table des vecteurs propres servant aux calculs des corrélations.
- *Test de Bartlett :* affiche un test de Bartlett. Ce test de permet de vérifier l'hypothèse selon laquelle les variables ne sont pas significativement corrélées.
- *Score des variables :* affiche les coordonnées des variables.
- *Score des observations :* affiche les coordonnées des observations.

Au lancement de la procédure, sélectionnez les options d'affichage des mappings (pour plus d'information consultez l'annexe « Boite d'affichage des graphiques »), et validez.

### Références

**Dillon W.R. & M. Goldstein (1984)**. Multivariate analysis. Methods and applications. John Wiley & Sons, New York, pp. 360-429.

**Jobson J.D. (1992)**. Applied multivariate data analysis. Volume II: categorical and multivariate methods. Springer-Verlag, New York, pp. 209-278.

**Johnson R.A. & D.W. Wichern (1992)**. Applied multivariate statistical analysis. Prentice-Hall, Englewood Cliffs, pp. 246-284.

**Lebart L., A. Morineau & M. Piron (1997)**. Statistique exploratoire multidimensionnelle. 2ème édition. Dunod, Paris, pp. 251-277.

**Saporta G. (1990)**. Probabilités, analyse des données et statistique. Technip, Paris, pp. 403-428.

**Sharma S. (1996)**. Applied multivariate techniques. John Wiley & Sons, New York, pp. 287-316.

Tomassone R., M. Danzart, J.J. Daudin & J.P. Masson (1988). Discrimination et classement. Masson, Paris.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 348-352, 358-367.

## RÉGRESSION MULTIPLE

### Description

Le programme de régression multiple permet d'expliquer la variation d'une variable en fonction de plusieurs autres. Les variables à expliquer et les variables explicatives doivent être de nature numérique.

Il est d'autre part, envisageable d'utiliser la transformation binaire disjonctive pour coder une question nominale en plusieurs variables pseudo-numériques.

Une seconde contrainte réside dans l'indépendance des variables explicatives. Souvent, elles sont corrélées entre elles. Si c'est le cas, on effectue d'abord une ACP et on sélectionne les questions qui sont les plus typiques des

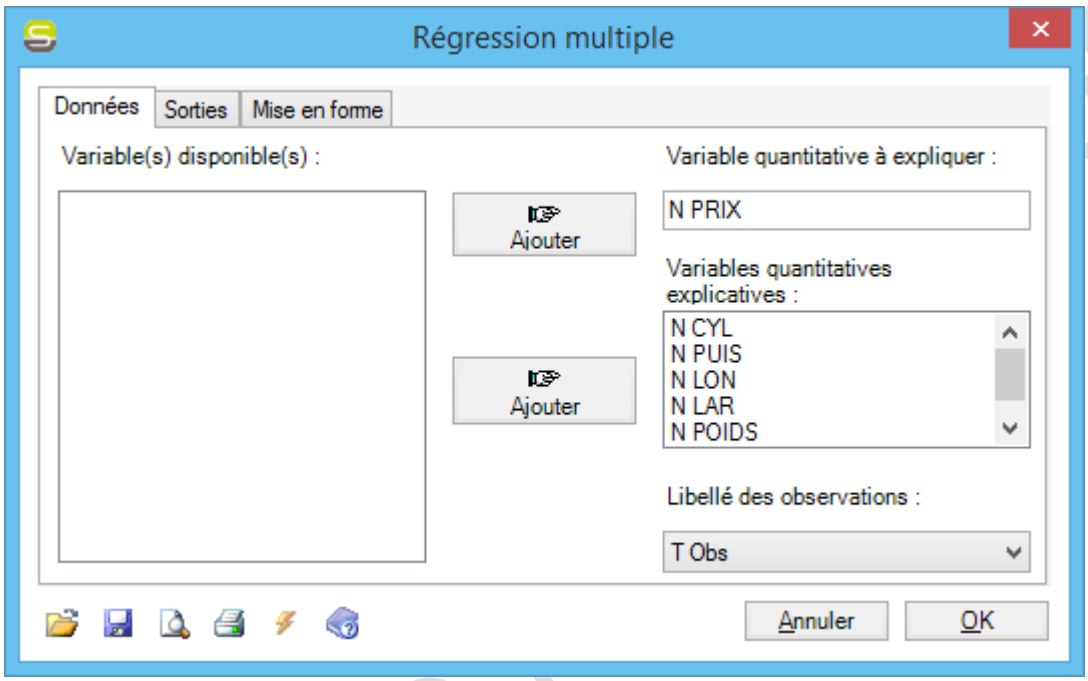
différents axes factoriels. Si les axes factoriels ont une signification claire, on peut les utiliser directement en tant que variables explicatives, ou en tant que variables à expliquer.

Le modèle est le suivant :

- y = a1 x1 + a2 x2 + ... + an xn + C
- Où **y** est la variable à expliquer
- Où **x1, x2, x3, ..., xn** sont les variables explicatives
- Où **a1, a2, a3,...,an** sont les coefficients de régression
- Où **C** est une constante

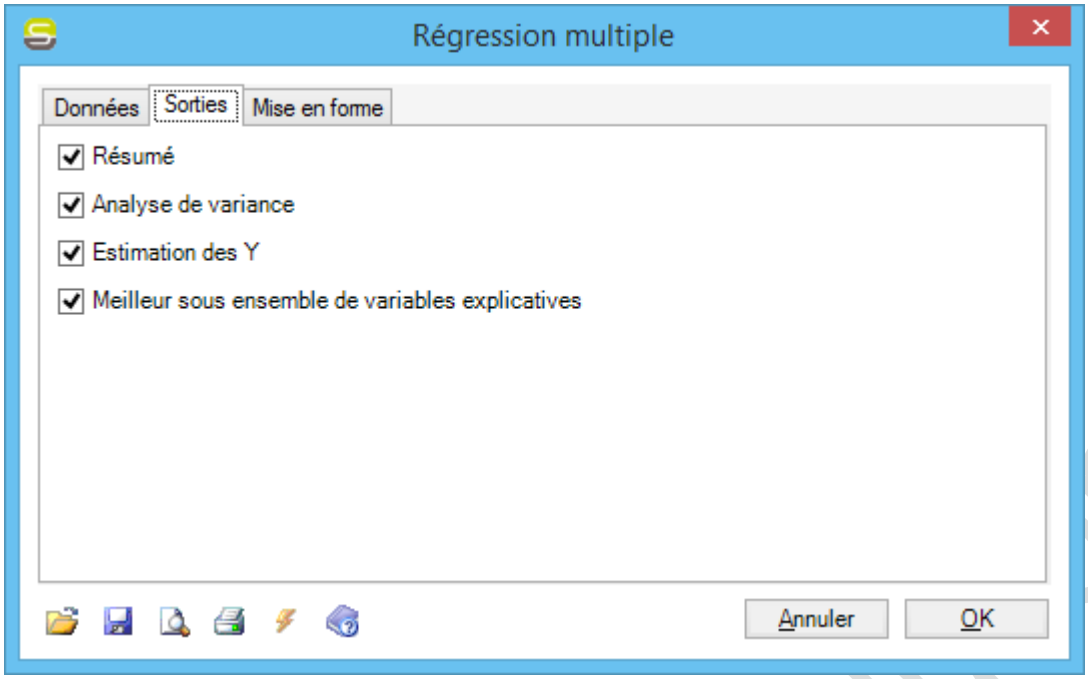
## Mise en œuvre

### *Onglet « Données »*




- *Variable quantitative à expliquer* : sélectionnez la variable quantitative à expliquer.
- *Variable(s) quantitative(s) explicative(s) :* sélectionnez dans la liste, celles que vous désirez intégrer dans le modèle. Elles doivent être toutes de nature numérique.
- *Libellé des observations :* sélectionnez la variable contenant le libellé des observations.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Analyse de variance :* affiche la table de décomposition de la variance expliquée par les variables sélectionnées.
- *Estimation des Y :* affiche pour chaque observation l'estimation de Y, la valeur Y, l'erreur et la distance de Cook. Lorsque cette distance est supérieure à 1, il est probable que le point influence trop les paramètres de la régression. Pour vérifier que l'erreur est bien une variable aléatoire distribuée normalement, vous pouvez effectuer un histogramme ou dans le module « Ajustement à une loi de probabilité » comparer la distribution à une loi normale.
- *Meilleur sous-ensemble de variables explicatives :* L'option « meilleur sous-ensemble » (best subset) permet de trouver le meilleur modèle comportant le moins de variables explicatives. Par exemple, avec au départ 5 variables explicatives, le logiciel évalue toutes les combinaisons de 4 variables parmi 5, de 3 variables parmi 5, etc.

## Exemple

L'exemple suivant est tiré de l'ouvrage de G. Saporta, Probabilités Analyse des données et statistique, Edition Technip, 1990, page 394

| Meilleur sous ensemble de variables explicatives : | | | | | | | |
|---|---|---|---|---|---|---|---|
| | CYL | PUIS | LON | LAR | POIDS | VITESSE | R2 ajusté |
| 1 variable | | X | | | | | 0,615 |
| 2 variables | | X | | | X | | 0,645 |
| 3 variables | X | X | | | X | | 0,634 |
| 4 variables | X | X | | X | X | | 0,610 |
| 5 variables | X | X | | X | X | X | 0,587 |
| 6 variables | X | X | X | X | X | X | 0,550 |

StatBox vous présente les différentes solutions possibles, associées à un R2 ajusté. Il s'agit alors de trouver un compromis entre la simplicité du modèle (c'est à dire le nombre de variables que l'on intègre au modèle) et son pouvoir explicatif (plus le R2 ajusté est élevé, plus le pouvoir explicatif du modèle est fort). En effet, dans le cadre d'une régression multiple, le meilleur modèle est le plus compact. Dans notre exemple les modèles à 2 ou 3 variables sont plus efficaces que les autres modèles. Cette méthode est probablement plus efficace que les méthodes pas à pas.

Une fois que vous avez déterminé le modèle le plus compact, vous pourrez refaire le traitement avec les variables les plus pertinentes.

Le modèle est :

PRIX = -8239,363 -3,505 CYL + 282,169 PUIS -15,038 LON + 208,694 LAR + 12,575 POIDS -111,114 VITESSE

| | Coef | Stdev | Std Coef | t-ratio | P |
|---|---|---|---|---|---|
| Constante | - 8239,363 | 42718,423 | 0,000 | - 0,193 | 0,425 |
| CYL | - 3,505 | 5,551 | - 0,199 | - 0,631 | 0,270 |
| PUIS | 282,169 | 174,883 | 0,875 | 1,613 | 0,067 |
| LON | - 15,038 | 129,747 | - 0,051 | - 0,116 | 0,455 |
| LAR | 208,694 | 412,048 | 0,169 | 0,506 | 0,311 |
| POIDS | 12,575 | 24,622 | 0,262 | 0,511 | 0,310 |
| VITESSE | - 111,114 | 222,257 | - 0,205 | - 0,500 | 0,313 |

R2 = 0,709 R2 ajusté = 0,55

Analyse de variance :

| | DDL | SCE | CM | F | P |
|---|---|---|---|---|---|
| Régression | 6 | 520591932,388 | 86765322,065 | 4,469 | 0,016 |
| Erreur Résiduelle | 11 | 213563857,889 | 19414896,172 | | |
| Total | 17 | 734155790,278 | | | |

Le tableau précédent donne les résultats de la régression multiple :

Le **R2** et le **R2 ajusté** : part de la variance expliquée par le modèle.

**Coef** : Cette colonne vous donne les résultats bruts de la régression multiple. Ce sont ces valeurs qu'il faut prendre en compte si vous voulez estimer la valeur Y d'une nouvelle observation.

**Std Coef** : Cette colonne vous donne les résultats sur des variables centrées et réduites de votre régression multiple (dans ce cas, il n'y a pas de constante).

**t-ratio et P** : Pour chacune des variables explicatives, la valeur du t de Student permet de savoir si elles participent d'une manière significative à l'explication du modèle. Pour des effectifs supérieurs à 60, un t de Student supérieur à 1,96 est significatif à P= 0.05. La colonne P donne la probabilité correspondant à la valeur de t.

**Tableau d'analyse de variance** : Il permet de savoir si, globalement, le modèle est statistiquement significatif.

Si vous avez coché dans la fenêtre de paramétrage estimation de Y, vous obtiendrez les résultats suivants :

Estimation :

| | PRIX | PRIX estimé | Résidu | Cook Dist. |
|---|---|---|---|---|
| Alphasud | 30570,000 | 29616,109 | 953,891 | 0,009 |
| audi | 39990,000 | 36259,655 | 3730,345 | 0,573 |
| simca | 29600,000 | 31411,149 | - 1811,149 | 0,017 |
| citroen | 28250,000 | 26445,751 | 1804,249 | 0,012 |
| fiat | 34900,000 | 37042,997 | - 2142,997 | 0,014 |
| lancia | 35480,000 | 34972,834 | 507,166 | 0,002 |
| peugeot | 32300,000 | 33749,145 | - 1449,145 | 0,005 |
| renault16 | 32000,000 | 26579,957 | 5420,043 | 0,230 |
| renault30 | 47700,000 | 44445,577 | 3254,423 | 0,600 |
| toyota | 26540,000 | 24650,241 | 1889,759 | 0,046 |
| alfetta | 42395,000 | 38270,462 | 4124,538 | 0,204 |
| princess | 33990,000 | 34830,418 | - 840,418 | 0,002 |
| datsun | 43980,000 | 44872,423 | - 892,423 | 0,019 |

| | | | | |
|---|---|---|---|---|
| taunus | 35010,000 | 36343,489 | - 1333,489 | 0,007 |
| rancho | 39450,000 | 35638,065 | 3811,935 | 0,070 |
| mazda | 27900,000 | 32233,420 | - 4333,420 | 0,139 |
| opel | 32700,000 | 37103,495 | - 4403,495 | 0,106 |
| lada | 22100,000 | 30389,814 | - 8289,814 | 0,533 |

## RÉGRESSION LOGISTIQUE

### Description

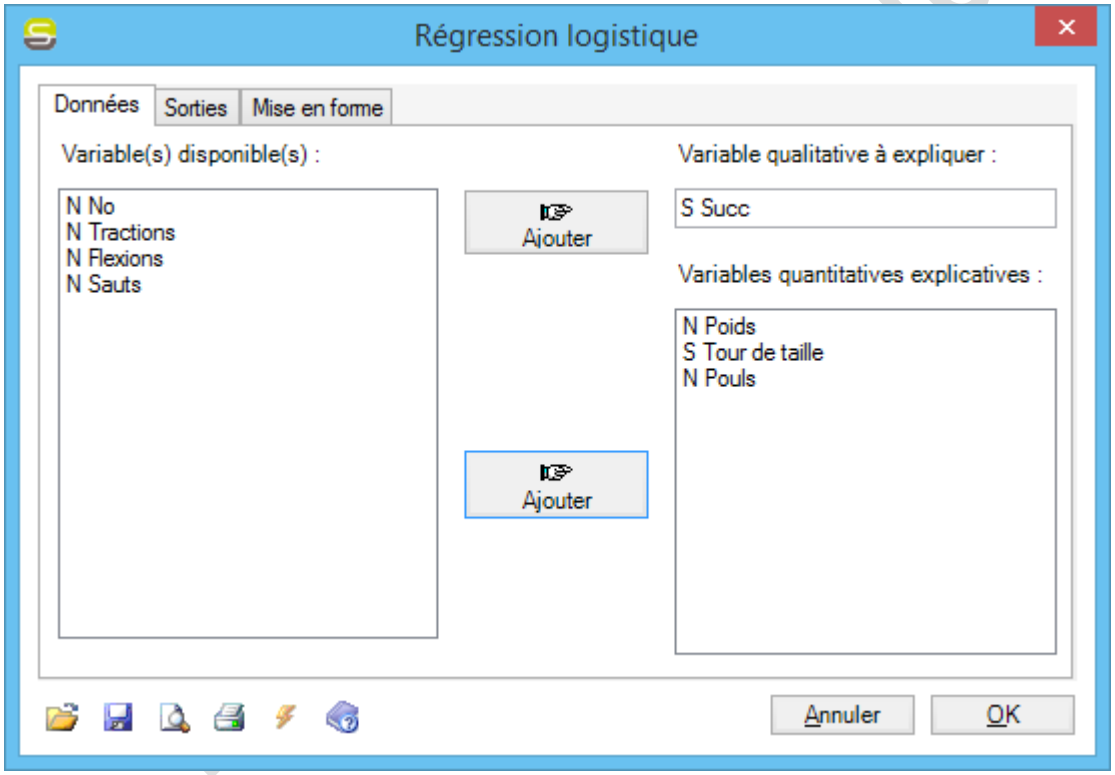
Dans la régression logistique la variable à expliquer prend les valeurs 0 ou 1, absence ou présence, vrai ou faux etc.

Comme pour la régression multiple, les variables explicatives sont numériques.

La méthode de calcul basée sur les moindres carrés n'est plus utilisable. La régression logistique utilise la méthode du maximum de vraisemblance pour estimer les coefficients de régression.

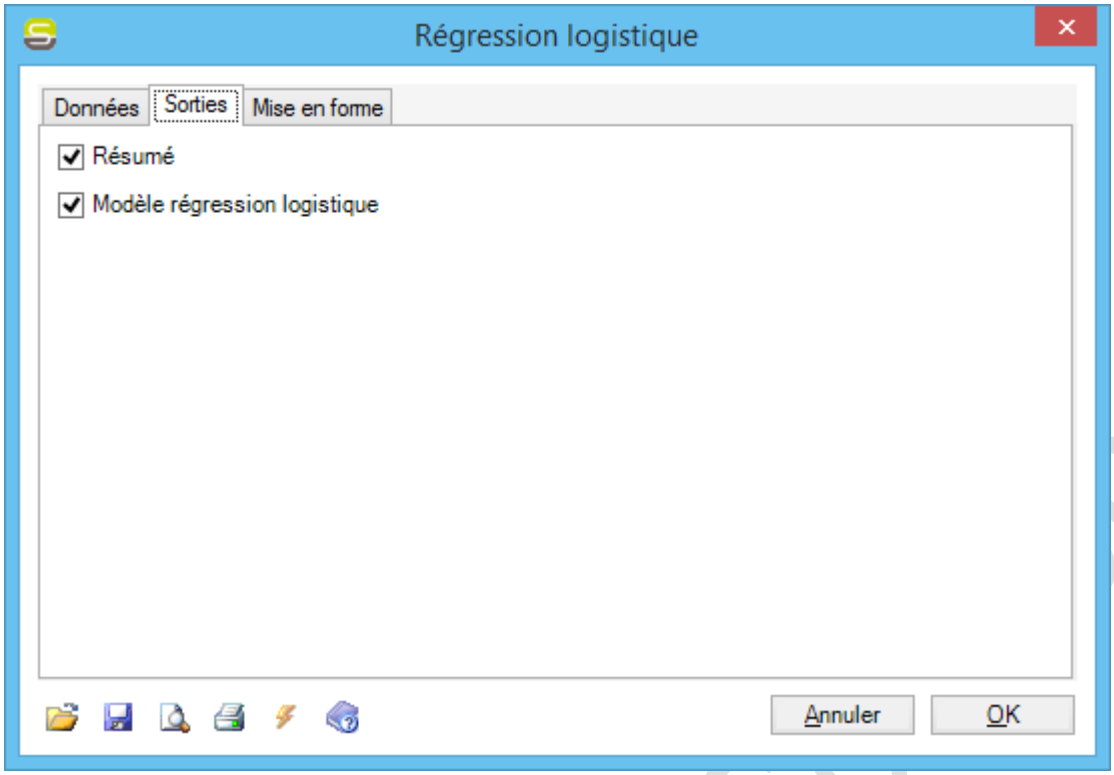
### Mise en œuvre

*Onglet « Données »*




- *Variable qualitative à expliquer :* sélectionnez la variable qualitative à expliquer.
- *Variables quantitatives explicatives :* sélectionnez les variables quantitatives explicatives.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Modèle de régression logistique* : affiche la table des coefficients associés à chaque variable explicative ainsi que les erreurs associées.

## Exemple

Voici un exemple de tableau de résultats sur un jeu de données issus de l'ouvrage de David W.Homer et Stanley Lemeshow, Applied Logistic Regression John Wiley&Sons, page 30.

Régression logistique :

| | Coef | Std.Error | Wald test | Pvalue |
|---|---|---|---|---|
| Constante | 1,295 | 1,071 | 1,209 | 0,228 |
| age | - 0,024 | 0,034 | 0,706 | 0,481 |
| lwt | - 0,014 | 0,007 | 2,178 | 0,031 |
| race1 | 1,004 | 0,498 | 2,016 | 0,045 |
| race2 | 0,433 | 0,362 | 1,196 | 0,233 |
| ftv | - 0,049 | 0,167 | 0,295 | 0,768 |

Log-Likelihood : -111,286
G : 12,099
Pvalue : 0,0335

Nombre d'itérations : 6
Nombre d'observations : 189

On trouve les coefficients de régression, l'écart type, le test de Wald pour évaluer la significativité des variables dans le modèle et la probabilité associée.

# RÉGRESSION PLS

## Description

Il est fréquent d'avoir à explorer rapidement les rapports existant entre deux groupes de variables décrivant les mêmes unités statistiques. On peut imaginer par exemple un ensemble d'observations décrites d'un côté par un certain nombre de caractéristiques socio-économiques et de l'autre par leur emploi du temps (durée dévolue à différentes activités), ou un ensemble de produits alimentaires de même type décrit, d'une part à l'aide de leur composition chimique, et d'autre part à l'aide de notes décernées par une équipe de goûteurs relativement à plusieurs composantes du goût.

Lorsque les variables des deux groupes sont qualitatives, il est tout indiqué de calculer le tableau croisant toutes les variables du groupe 1 avec toutes celles du groupe 2. On procède alors à l'Analyse des Correspondances Simples de ce tableau. Ce dernier est en effet une juxtaposition de tableaux de contingence ordinaires ventilant la même population.

Lorsque les variables des deux groupes sont quantitatives, on cherchera à visualiser rapidement les liaisons linéaires entre variables des deux groupes à l'aide de la régression PLS.

Cette méthode permet de visualiser les liaisons linéaires entre 2 tableaux de variables quantitatives X et Y décrivant les mêmes observations.

La régression PLS cherche à trouver dans X les grands axes qui expliquent le mieux Y.

Prenons l'exemple des résultats des deux tours d'un scrutin dans l'ensemble des régions d'un pays. Au premier tour, l'électeur avait J choix possibles. Au second tout il ne reste que K choix. On cherche à expliquer les résultats du second tour à l'aide de ceux du premier tour, c'est-à-dire capter l'essentiel du mécanisme de report des votes. De manière générale on cherche à expliquer (puis éventuellement à prédire) globalement les variables du groupe Y (groupe à expliquer) à l'aide de celles du groupe X (groupe explicatif).

Le problème qu'on se pose : trouver les facteurs (combinaisons linéaires) des X d'une part et ceux des Y d'autre part, tels que :

- les facteurs des X résument le mieux possible les X (propriété des axes factoriels),
- les facteurs des Y résument le mieux possible les Y (propriété des axes factoriels),
- les facteurs des X soient les meilleures variables explicatives possibles de ceux des Y, sous les contraintes précédentes. En particulier, les facteurs des X seront deux à deux décorrélés, alors que ceux des Y, a priori non.

On voudrait représenter ensuite les variables et les observations sur les paires d'axes correspondant à ces facteurs (on utilisera seulement les facteurs explicatifs, i.e. ceux des X, pour la représentation des variables). L'explication de variables par des facteurs, en termes géométriques, c'est justement la projection de ces variables sur le sous-espace de ces facteurs.

De plus, puisqu'il s'agit d'expliquer, et éventuellement de prédire, on cherche aussi à obtenir des équations de régression des Y sur les facteurs des X (à partir desquelles on peut retrouver, éventuellement, des équations de régression des Y en fonction des X).

Comme en ACP, on peut juger de la corrélation de deux variables selon l'angle que font leurs vecteurs.

La projection des observations est double, i.e. chaque observation est projetée deux sur l'axe : une fois en tant que décrit par le groupe Y, et une fois en tant que décrit par le groupe X. Si les deux projections d'une même observation sont proches l'une de l'autre dans un plan, cette observation concourt aux liaisons entre les deux groupes dépistées par le plan. A contrario, une observation dont les deux projections sont éloignées, voire très opposées est une observation qui va contre la liaison générale entre les 2 ensemble de variables. Dans le cas des votes, il s'agit par exemple d'un département dont le report des votes s'est effectuer différemment.

Chacun des facteurs de X résument les disparités des observations du point de vue des X, ils sont par ailleurs indépendants. Ils captent une part de la variance totale du groupe X. Ces parts s'additionnent. On peut donc juger du nombre de facteurs à conserver.

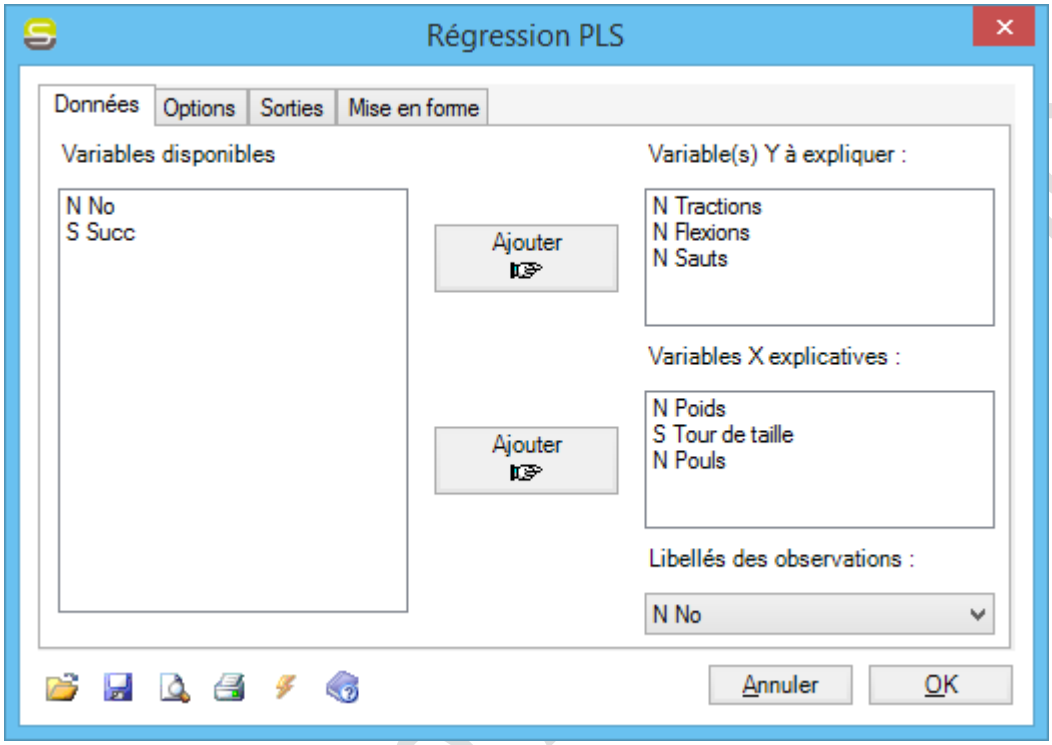
Puisqu'il s'agit d'expliquer, et éventuellement de prédire le groupe Y, on obtient aussi les équations de régression des Y sur les facteurs des X.

La régression PLS permet de s'affranchir des limites de la régression multiple :

- les variables explicatives du groupe X peuvent être très corrélées entre elles,
- le nombre d'observations peut être inférieur au nombre de variables explicatives,
- la régression PLS permet d'isoler le bruit dans le modèle,
- elle accepte plusieurs variables Y à expliquer.

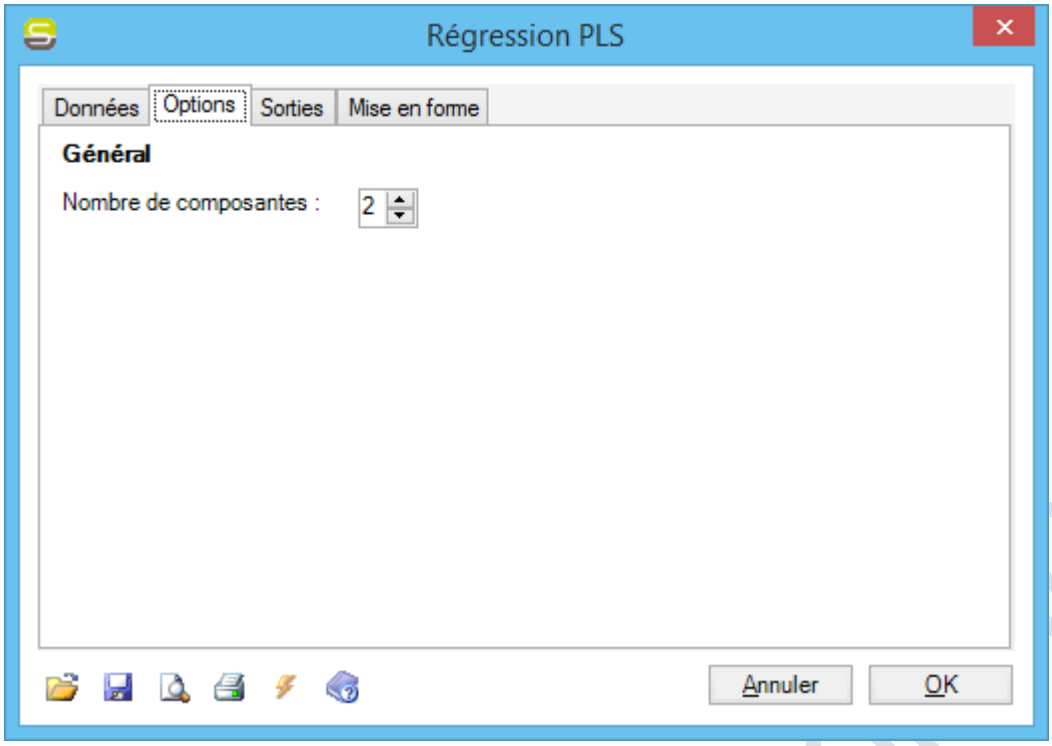
## Mise en œuvre

### *Onglet « Données »*



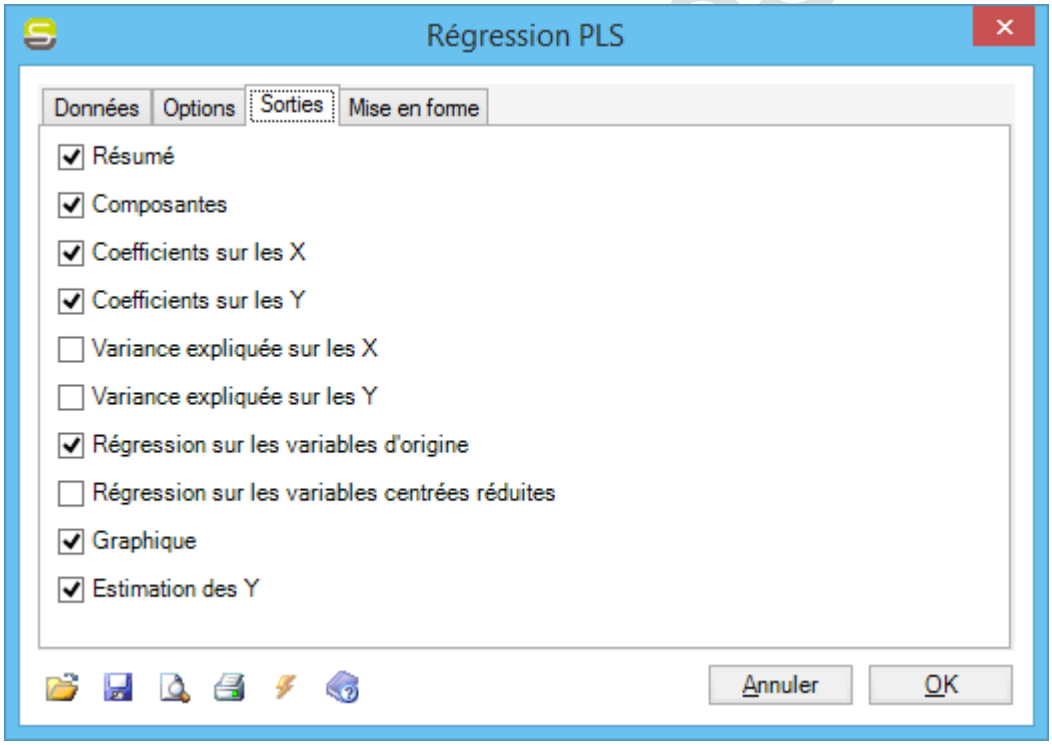

- *Variable(s) Y à expliquer* : sélectionnez les variables quantitatives dépendantes à expliquer.
- *Variables X explicatives* : sélectionnez les variables quantitatives explicatives.
- *Libellés des observations* : sélectionnez la variable contenant les libellés des observations.

*Onglet « Options »*




- *Nombre de composantes :* entrez le nombre maximal de composantes à prendre en compte dans le modèle.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Composantes :* affiche la table des composantes du modèle
- *Coefficients sur les X :* affiche les coefficients des variables X sur les composantes du modèle.
- *Coefficients sur les Y :* affiche les coefficients des variables Y sur les composantes du modèle.
- *Variance expliquée sur les X :* affiche la table de la variance expliquée sur les X.
- *Variance expliquée sur les Y :* affiche la table de la variance expliquée sur les Y.
- *Régression sur les variables d'origine :* affiche le modèle de régression sur les variables d'origine.

- *Régression sur les variables centrées réduites :* affiche le modèle de régression sur les variables centrées réduites.
- *Graphique :* affiche les cartes des erreurs d'estimation par le modèle pour les variables et les observations.
- *Estimation des Y :* affiche pour chaque observation les valeurs prédites par le modèle des variables à expliquer.

### Exemple

Les résultats suivants ont été obtenus avec les données Linerud. Le nombre de facteur est égal à 2. Ils ont été également traités dans l'ouvrage de M.Tenenhaus, La régression PLS Théorie et pratique. Éditions Technip, 1998.

Coefficients des variables explicatives X sur les composantes t de l'ensemble X :

| | w*1 | w*2 |
|---|---|---|
| Tractions | - 0,827 | - 0,074 |
| Flexions | - 0,449 | - 0,595 |
| Sauts | - 0,337 | 0,803 |

Coefficients des variables à expliquer Y sur les composantes t de l'ensemble X :

| | c1 | c2 |
|---|---|---|
| Poids | 0,311 | 0,383 |
| Tour de taille | 0,406 | 0,740 |
| Pouls | - 0,119 | - 0,319 |

Équation de régression sur variables initiales :
Poids = 205,448 - 1,334 Tractions - 0,145 Flexions + 0,098 Sauts
Tour de taille = 40,273 - 0,237 Tractions - 0,032 Flexions + 0,029 Sauts
Pouls = 52,581 + 0,167 Tractions + 0,028 Flexions - 0,03 Sauts

| | Poids | Tour de taille | Pouls |
|---|---|---|---|
| Const. | 205,448 | 40,273 | 52,581 |
| Tractions | - 1,334 | - 0,237 | 0,167 |
| Flexions | - 0,145 | - 0,032 | 0,028 |
| Sauts | 0,098 | 0,029 | - 0,030 |

Vous trouverez également le mapping des variables et des observations, les composantes et les estimations.

## RÉGRESSION NEURONALE

### Les réseaux de neurones

Les réseaux de neurones permettent d'effectuer des analyses multivariées et de compléter un certain nombre de méthodes statistiques classiques comme :

- l'Analyse en Composantes Principales,
- la Régression Multiple,
- l'Analyse Factorielle Discriminante
- la Classification.

Les réseaux de neurones de StatBox ont été adaptés pour être utilisés de la même manière que les méthodes statistiques classiques.

Avec les réseaux de neurones, l'ajustement étant non linéaire, la prédiction sera souvent meilleure que les techniques classiques.

La régression neuronale va vous permettre de prédire la valeur d'une variable numérique en fonction de plusieurs autres.

Avec StatBox vous pouvez utiliser les méthodes neuronales et comparer les résultats obtenus avec les méthodes statistiques d'analyse des données. L'intérêt des réseaux de neurones est d'aller plus loin que les méthodes classiques. En particulier grâce à leur algorithme de traitement non-linéaire. En revanche les réseaux de neurones ne fournissent pas les résultats habituels (coefficients de régression, test de significativité, etc.)

Si le processus de convergence des réseaux de neurones est complexe à suivre parce qu'il s'agit d'un algorithme itératif mettant en jeu de nombreux neurones, les principes de base sont d'une grande simplicité.

C'est en effectuant des traitements que l'utilisateur va comprendre le fonctionnement des réseaux de neurones et en particulier le processus de convergence. La pratique est ici indispensable pour la maîtrise de ces nouvelles techniques.

StatBox comporte les réseaux dont l'apprentissage est supervisé du type rétropropagation (backpropagation).

On distingue deux étapes : la phase d'apprentissage pendant laquelle les poids sont calculés pour que le modèle s'ajuste au mieux aux données et une phase de test pendant laquelle on peut évaluer le modèle sur de nouveaux jeux de données.

StatBox affiche la courbe d'apprentissage et les valeurs estimées. L'utilisateur peut modifier le taux d'apprentissage, le nombre de neurones dans la couche cachée et le nombre d'itérations maximum.

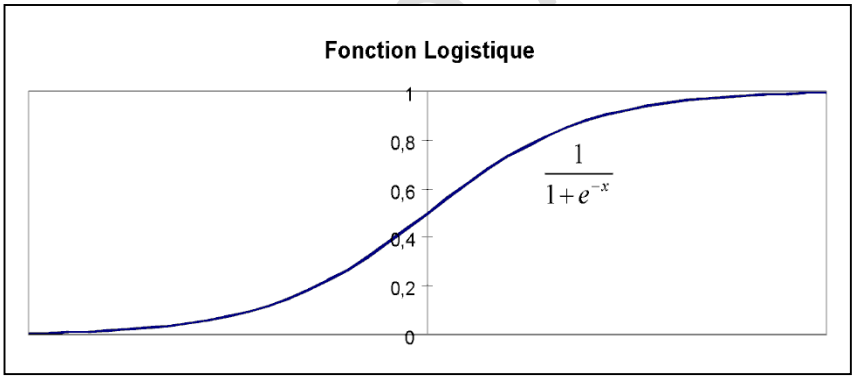
Les réseaux de neurones complètent les analyses statistiques des données présentes dans StatBox. L'utilisateur pourra ainsi obtenir ces résultats avec les deux méthodes : statistique et neuronale. C'est à partir de cette comparaison que l'on évalue l'apport des algorithmes non linéaires des réseaux de neurones.

Les réseaux de neurones de StatBox intéresseront ceux qui pratiquent déjà l'analyse des données. Ils intéresseront également ceux dont les exigences ne sont pas satisfaites avec les méthodes statistiques classiques. Enfin StatBox constitue un outil d'une grande simplicité pour la formation aux méthodes d'analyses neuronales et à l'analyse des données.

## Les principes de base

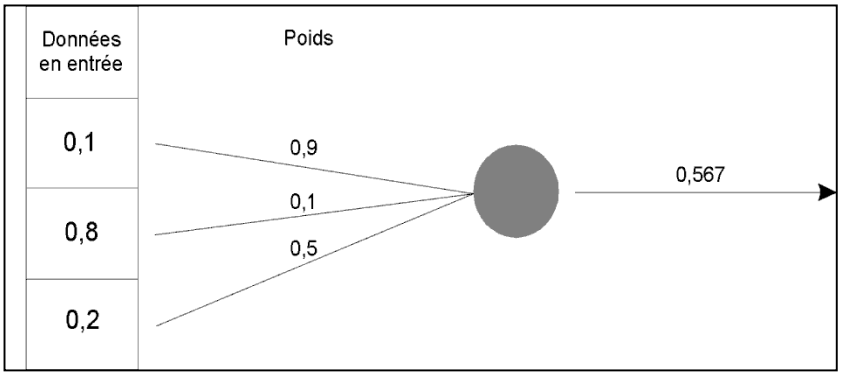
### *Le neurone : Presque comme un neurone biologique.*

Le neurone 'électronique' comme le neurone biologique, comporte plusieurs entrées et une seule sortie. Chaque entrée est pondérée par un poids. La somme pondérée obtenue est ensuite modifiée par une fonction d'activation, la plus utilisée est la fonction logistique : 1/(1+e-x).




0.1 x 0.9 + 0.8 x 0.1 + 0.2 x 0.5 = 0.27

$$sortie = \frac{1}{1+e^{-0.27}} = 0.567$$




On multiplie chaque entrée par le poids correspondant et on fait la somme totale (0.27). La fonction d'activation est appliquée sur ce résultat pour obtenir le résultat final (0.567) qui sera transmis au neurone suivant.

### *Les couches de neurones : La couche cachée identifie les 'patterns'*

Les neurones sont organisés en couches. Chaque couche contient un certain nombre de neurones. Tous les neurones d'une couche sont 'connectés' avec ceux de la couche suivante. Pour l'analyse en composantes neuronales, la régression neuronale et l'analyse discriminante neuronale, on a 3 couches : une couche d'entrée, une couche cachée, une couche de sortie.

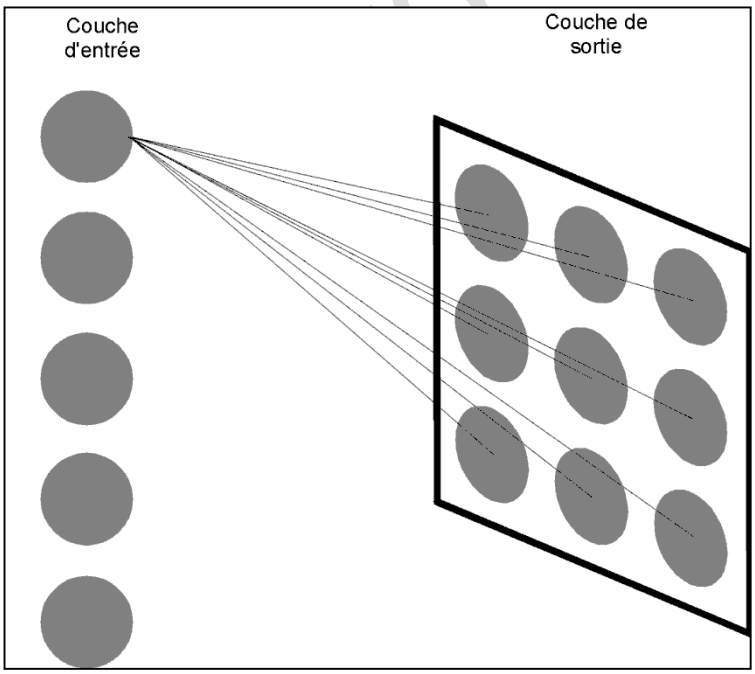
La couche d'entrée contient autant de neurones que de variables en entrée.

La couche cachée contient un nombre plus restreint de neurones par rapport aux neurones d'entrée. Une règle informelle consiste à estimer le nombre de neurones dans la couche cachée égale à la racine carrée du nombre de neurones en entrée.

Le nombre de neurones dans la couche de sortie dépend de la méthode d'analyse envisagée : 1 neurone pour la régression, le nombre de groupes pour l'analyse discriminante et le nombre de données en entrée pour l'analyse en composantes neuronales.

Pour la classification, on a seulement 2 couches :

- Une couche d'entrée contenant autant de neurones que de variables en entrée.
- Une couche de sortie contenant une matrice de neurones. Chaque neurone dans cette matrice peut représenter un groupe. Cette matrice dans StatBox est pour la plus petite de 2x2 jusqu'à 7x7, soit de 4 groupes à 49 groupes potentiels.

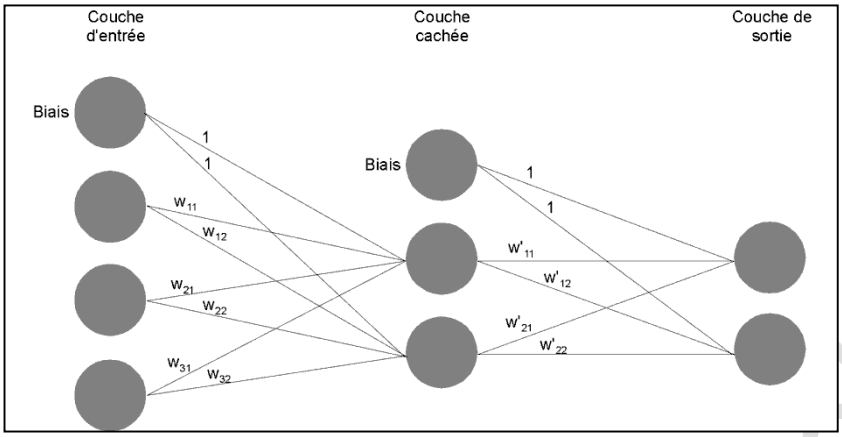



Dans le cas de la classification, les données d'entrée sont directement propagées vers la couche de sortie.

Le modèle d'apprentissage supervisé : la rétropropagation (backpropagation)

Les réseaux à rétropropagation sont ceux qui ont suscité le plus grand nombre d'applications. Ils sont utilisés dans StatBox dans le modèle de régression, en analyse discriminante et en analyse en composantes neuronales.

Au début des calculs, les poids des neurones sont définis aléatoirement. Les informations en entrée sont propagées vers la couche cachée puis vers la sortie. Les couches sont liées entre elles par des poids. Toutes les données sont présentées successivement en entrée, la somme pondérée est effectuée et modifiée grâce à la fonction logistique d'activation. Les résultats obtenus au niveau de la couche cachée sont ensuite propagés vers la couche de sortie.




Le modèle à rétropropagation va évaluer l'erreur, c'est-à-dire l'écart entre les résultats obtenus et ceux que l'on devrait obtenir. Il faut donc à chaque jeu d'entrée, un jeu de données à obtenir. Cette différence est 'rétropropagée' vers la couche cachée puis vers la couche d'entrée et les poids sont modifiés légèrement dans le sens de la réponse que l'on doit obtenir. Cette modification est effectuée à la fin d'une itération. A chaque itération, toutes les données sont présentées et l'erreur est calculée. Comme la modification des poids va vers la réduction de l'erreur, la courbe d'apprentissage doit baisser régulièrement jusqu'à se stabiliser horizontalement : alors la solution optimale est obtenue.

Le taux d'apprentissage est la part de l'erreur qui est affectée à la modification des poids.
Wt = Wt-1 + taux d'apprentissage x erreur + momentum (Wt-1 – Wt-2)
Wt : poids à l'itération t
Wt-1 : poids à l'itération t-1
Wt-2 : poids à l'itération t-2

Pour éviter des oscillations, on 'lisse' la modification du poids en ajoutant à la formule une part (momentum) de la dernière modification des poids.

L'erreur est 'rétropropagée' pendant l'apprentissage. L'ajustement des poids est un processus itératif. De la couche d'entrée vers la couche cachée puis vers la couche de sortie. Le taux d'apprentissage permet de moduler l'amplitude de la correction des poids.

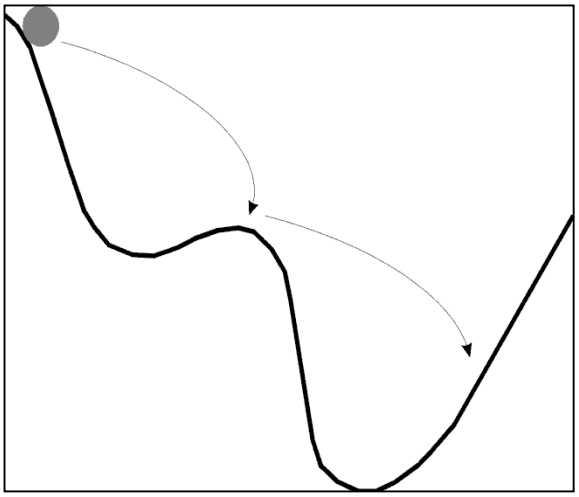
Cet ajustement est fait après chaque itération.

Le nombre d'itération suffisant varie entre 100 à 1 000 voire dans certains cas particuliers de 5 à 10 000.

L'important est d'avoir un taux d'apprentissage suffisamment petit pour que le processus de convergence s'effectue : c'est-à-dire que les modifications successives des poids réduisent l'erreur d'une part et que, d'autre part, ce taux d'apprentissage ne soit pas trop petit pour qu'à la fin des itérations on obtienne la valeur optimale des poids.

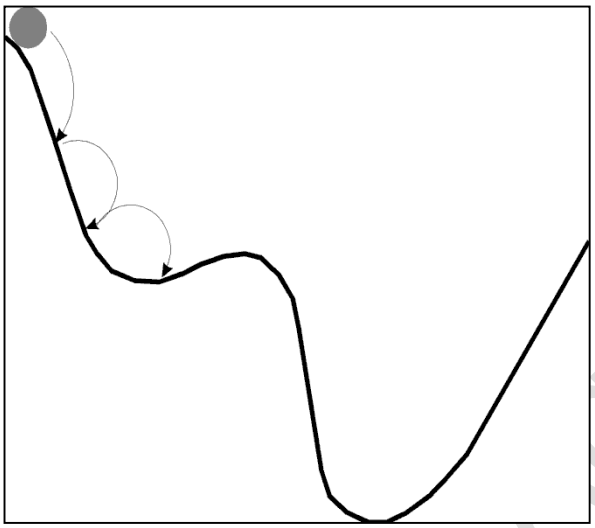
Un taux d'apprentissage élevé permet au réseau d'apprendre rapidement mais on risque de ne pas obtenir la meilleure solution. La courbe d'apprentissage oscille et n'arrive pas à se stabiliser.

On peut dans une certaine mesure, représenter l'apprentissage comme une balle qui saute le long d'une pente. Cette dernière descend dans la vallée puis remonte de l'autre côté. La longueur d'un bond symbolise le taux d'apprentissage. Si ce taux est élevé, la balle fait de grands sauts, va rebondir de l'autre côté de la pente et aura du mal à atteindre le fond de la vallée (Graphe ❶).



Graphe ❶

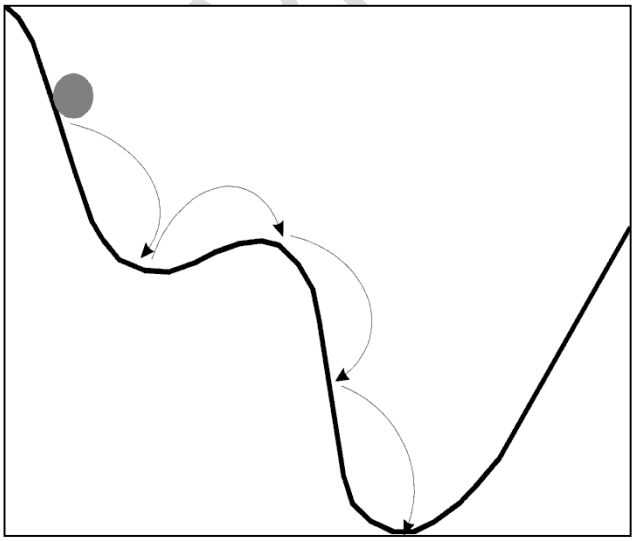
La pente présente des bosses qui peuvent bloquer la balle et l'empêcher de descendre. C'est le cas si le taux d'apprentissage est trop petit (Graphe ❷).



Graphe ❷

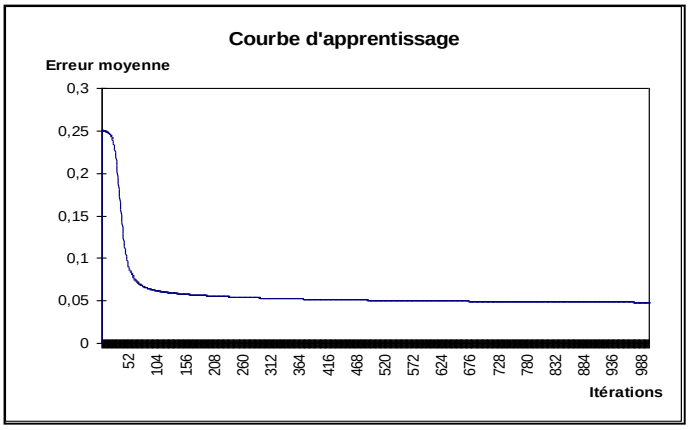
Les petits bonds conduiront la balle au fond mais s'ils sont trop petits et qu'une bosse se présente, la balle risque d'être bloquée, il s'agit en d'autres termes d'un optima local.

Un taux d'apprentissage adéquat nous permet d'atteindre le fond de la vallée (Graphe❸).



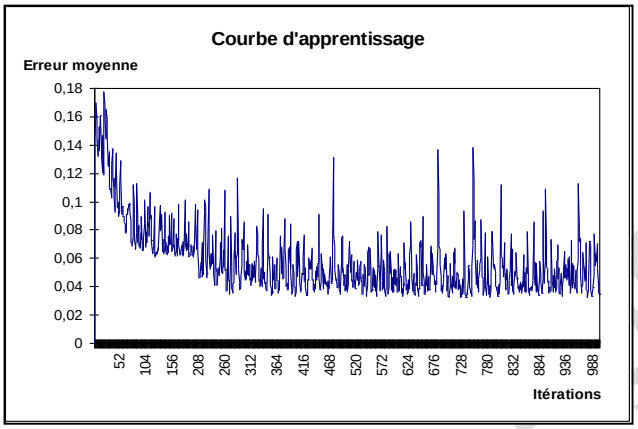
Graphe ❸

Le taux d'apprentissage permet à chaque itération de réduire l'erreur. La courbe d'apprentissage présente l'erreur en fonction du nombre d'itérations. Nous verrons maintenant 3 courbes d'apprentissage correspondant à trois taux différents :



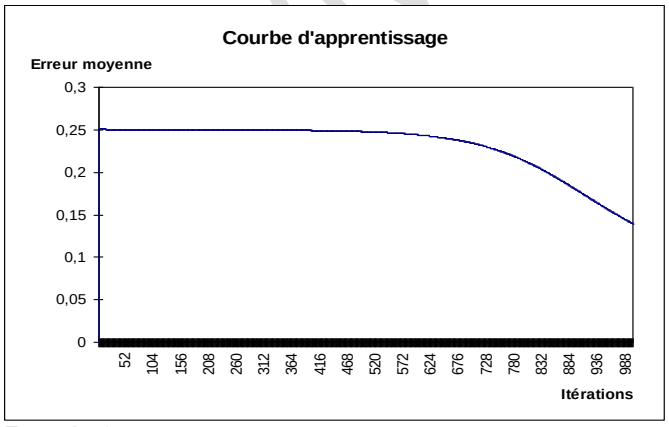

Exemple ❶

❶ Dans ce premier exemple, le taux d'apprentissage est ajusté correctement (la valeur du taux d'apprentissage est de 0.1 et le nombre maximum d'itérations est de 1000). On remarque que le réseau apprend vite, l'erreur moyenne baisse rapidement. Au-dessus de 100 itérations l'erreur se stabilise autour de 0.05. On a atteint la solution optimale.




Exemple ❷

❷ Ici le taux d'apprentissage est trop grand pour que le réseau converge. Valeur du taux : 0.9 et nombre d'itérations : 1000. Globalement la courbe baisse mais on observe de nombreuses oscillations. Dans ce cas il faut baisser le taux d'apprentissage, le diviser au moins par 2, voire plus.

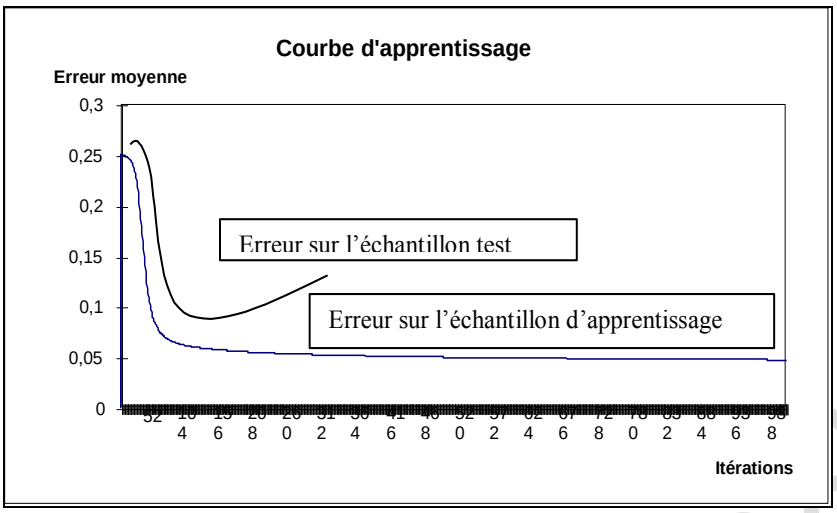



Exemple ❸

❸ Dans ce dernier exemple le taux d'apprentissage est trop petit pour atteindre la solution optimale (Valeur du taux : 0.01 et nombre d'itérations : 1000). On voit qu'à la dernière itération, dans notre cas la millième, la courbe continue à baisser si on prolonge les itérations au-delà de mille. Il faut soit augmenter le nombre d'itérations, soit plus probablement augmenter le taux d'apprentissage.

### *Le sur-apprentissage*

Lorsque la courbe d'apprentissage commence à se stabiliser horizontalement, le modèle risque d'apprendre les spécificités du jeu de données et peut perdre sa capacité à généraliser ou, en d'autres termes, à 'interpoler'. Il faudrait arrêter l'apprentissage au moment où la courbe devient horizontale. En effet, si on prend un nouveau jeu de données et qu'on applique les poids obtenus sur l'échantillon test à chaque itération, on remarque que l'erreur sur l'échantillon test va d'abord baisser puis de nouveau augmenter. Dans le graphique suivant, ce phénomène s'observe à partir de la centième itération. On devrait donc arrêter l'apprentissage à ce moment-là.




## La phase d'apprentissage et la phase de test

### *La preuve d'un bon résultat*

Contrairement à la régression multiple, il n'est pas possible de faire un test de significativité du modèle. Une solution consiste à diviser aléatoirement l'échantillon initial en deux sous-échantillons. On estime le modèle sur l'un des sous-échantillons, c'est la phase d'apprentissage. L'erreur moyenne doit être la plus petite possible. La deuxième phase consiste à tester le modèle sur l'autre sous-échantillon. Sur cet échantillon, on connaît la valeur de la variable étudiée. Si la valeur estimée n'est pas trop différente de la valeur observée, le modèle est probablement opérationnel. On pourra ensuite présenter au modèle des observations ou individus dont on ne connaît pas la valeur de la variable étudiée.

Il est intéressant de faire d'abord une analyse statistique classique et ensuite une analyse neuronale. Cette première analyse donne un point de comparaison intéressant.

### *Nombre de neurones dans la couche cachée : L'équivalent des facteurs*

Le nombre de neurones de la couche cachée correspond approximativement au nombre de facteurs en analyse factorielle. On introduit dans la couche cachée un nombre inférieur de neurones. En analyse en composantes neuronales, les neurones de la couche cachée jouent un rôle de compression des données ou de réduction du bruit.

Si on définit un trop grand nombre de neurones dans la couche cachée en régression ou en analyse discriminante, le modèle risque d'apprendre 'par cœur' les données présentées en entrée et ne saura pas généraliser sur un jeu de données inconnu. Une règle approximative consiste à prendre la racine carrée du nombre de neurones en entrée. Mais il faut également prendre en compte la manière dont les données ont été codées au départ.

### *Nature des données en entrée*

Il est possible a priori de soumettre à un réseau de neurones des données nominales ou quantitatives. L'expérience montre qu'une variable nominale transformée en plusieurs variables Oui/Non (ou en d'autres termes en variables binaires disjonctives) donne de meilleurs résultats. Il est dans ce cas conseillé de transformer les variables numériques en classes.

Les variables numériques en entrée qui ont une distribution très grande (supérieure à plus ou moins 3 écart-types par rapport à la moyenne) donnent de moins bons résultats. Dans ce cas StatBox borne les données à plus ou

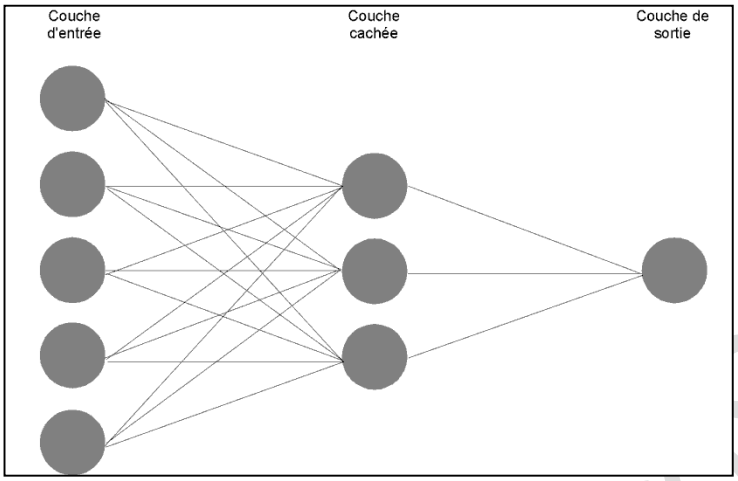
moins 3 écart-types en entrée pendant la phase d'apprentissage. D'autre part pour éviter l'effet des unités de mesure, StatBox réduit l'amplitude des données à l'intervalle 0 et 1.

L'amplitude initiale est ensuite reconstituée pour les données en sortie.

## La régression neuronale

### *Les principes*

La régression neuronale permet d'établir un lien entre une variable numérique et plusieurs autres variables numériques ou non. Elle est comparable à la régression linéaire multiple. On utilise l'algorithme de rétropropagation avec un réseau à 3 couches. La première couche contient un nombre de neurones égal au nombre de variables en entrée. La couche cachée contient un nombre plus petits de neurones. Et enfin la couche de sortie ne contient qu'un seul neurone.




Dans un premier temps, le modèle est obtenu sur un échantillon d'apprentissage. Dans un deuxième temps, on valide le modèle sur un échantillon test. Et enfin, on estime la valeur de la variable étudiée pour de nouvelles observations.

L'intérêt de la régression neuronale réside dans son algorithme non linéaire. La 'droite' de régression est en fait une courbe dans le cas d'une seule variable explicative. En conséquence, le modèle ne fournit pas une formule du type : $y = ax + b$

Pour estimer y en fonction d'un nouveau jeu de données, il suffit de lui appliquer les poids obtenus pendant la phase d'apprentissage.

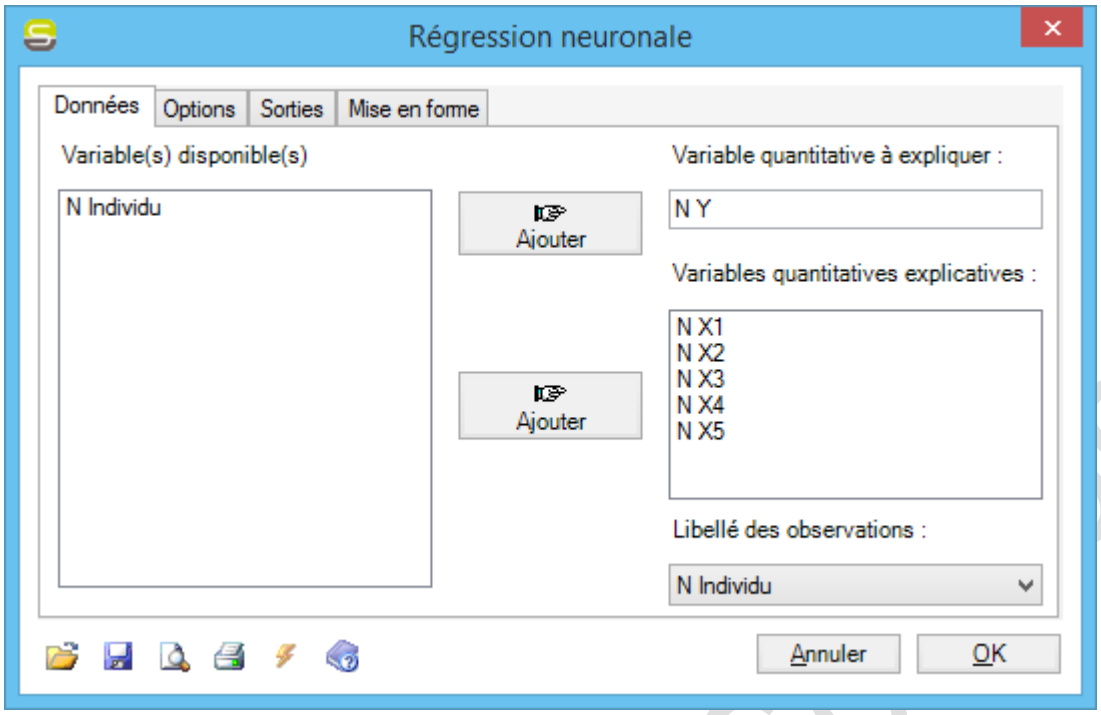
Le nombre de neurones cachés ne doit pas être trop important. En effet, l'ajustement sera meilleur avec un grand nombre de neurones cachés mais la généralisation sur de nouvelles données se fera difficilement. Le taux d'apprentissage est de 0.2 par défaut (20% de l'erreur est répercuté pour la correction des poids).

La courbe d'apprentissage représente horizontalement les itérations successives et verticalement l'erreur moyenne. Au début, l'erreur est élevée. Elle doit rapidement baisser.

Si on observe que la courbe d'apprentissage ne baisse pas et ne se stabilise pas horizontalement, il faut probablement réduire ce taux. Divisez-le par deux et relancez le modèle. Il est parfois nécessaire de le réduire encore tant que le modèle continue à osciller. Si le taux d'apprentissage est trop petit et que le nombre d'itérations n'est pas très élevé, il est probable que le modèle ne va pas atteindre la solution optimale. Augmentez dans ce cas le taux d'apprentissage et éventuellement le nombre d'itérations.

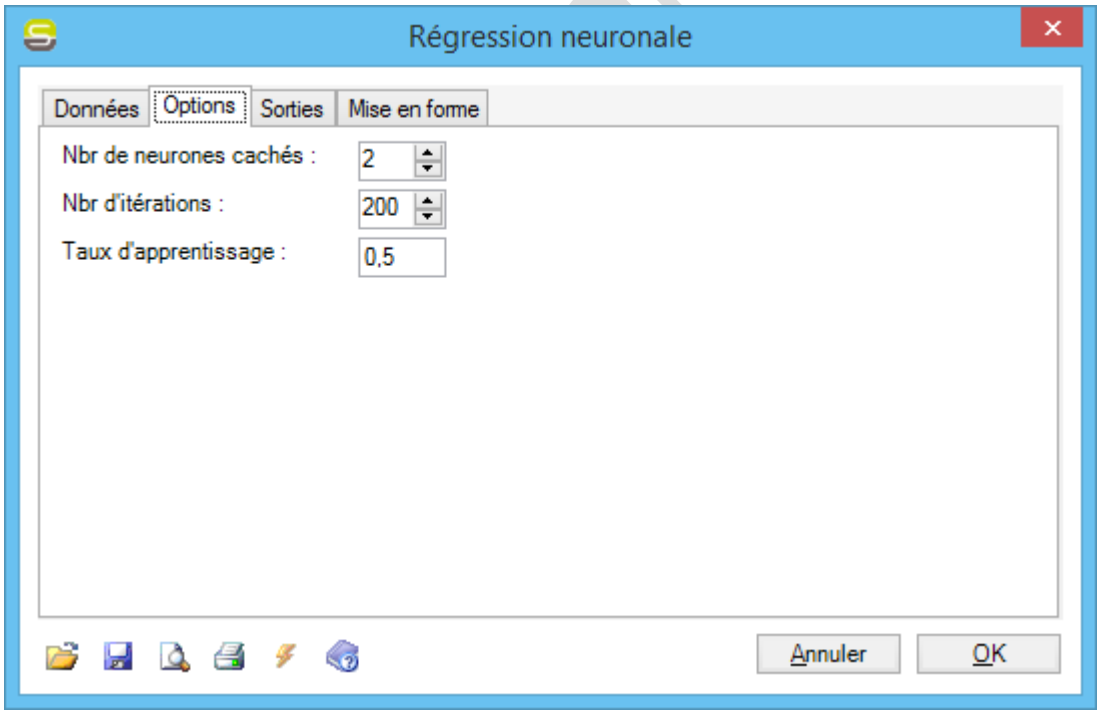
## Mise en œuvre

### *Onglet « Données »*



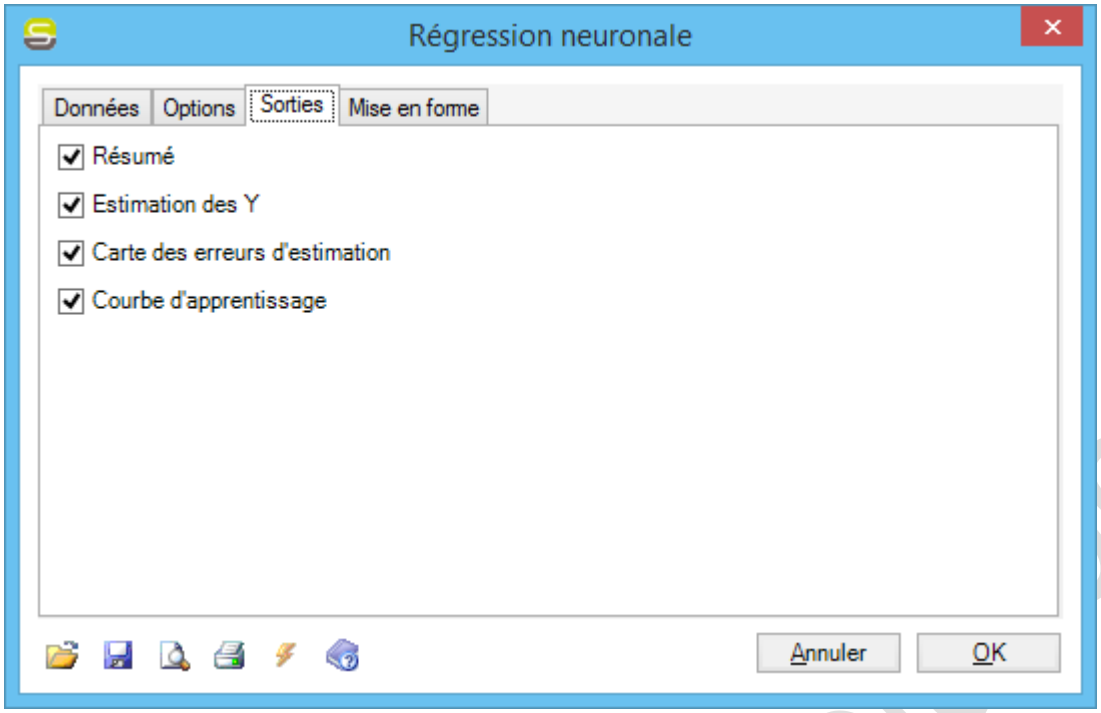

- *Variable quantitative à expliquer :* sélectionnez la variable à expliquer (Y).
- *Variables quantitatives explicatives :* sélectionnez les variables explicatives (X).
- *Libellé des observations :* sélectionnez la variable identifiant les observations.

### *Onglet « Options »*




- *Nombre de Neurones cachés :* introduisez le nombre de neurones à prendre en compte dans la couche cachée. Ce nombre doit être inférieur aux nombres de variables en entrée.
- *Nombre d'itérations :* entrez le nombre d'itérations qui sera effectué lors de l'analyse.
- *Taux d'apprentissage :* entrez le taux d'apprentissage.

*Onglet « Sorties »*

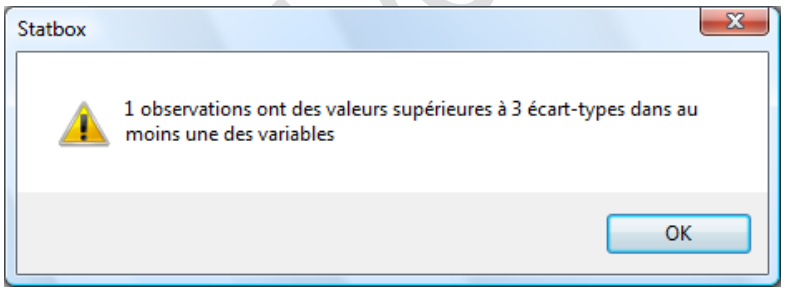



- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Estimation des Y* : sélectionnez cette option si vous êtes en mode test. Dans ce cas, l'activation des différentes options (nombre de neurones cachés, coefficient d'apprentissage) n'est pas possible. Dans cette phase de test vous devez sélectionner les mêmes variables que celles sélectionnées pendant la phase d'apprentissage. Cliquez sur Ok pour lancer la Régression Neuronale.
- *Carte des erreurs d'estimation :* affiche la carte des erreurs d'estimation associée au modèle retenu.
- *Courbe d'apprentissage :* affiche la courbe d'apprentissage de l'estimation.

## Exemple

Cet exemple est effectué sur la feuille « Régression » du fichier d'exemple « Data.xls »

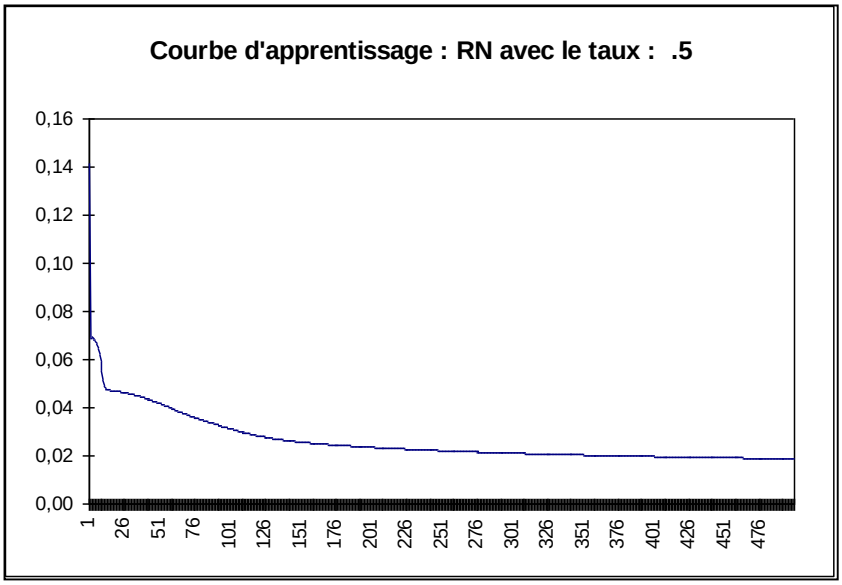
Lancez la boite de dialogue de régression neuronale. La boîte de dialogue suivante apparaît :




Ce message apparaît lorsqu'une des variables a des valeurs extrêmes dépassant 3 écart-types par rapport à la moyenne. L'algorithme de rétropropagation fonctionne mieux lorsque la distribution des données n'est pas trop importante. Toutes les valeurs qui dépassent l'intervalle seront modifiées et bornées.

À la fin du traitement, entrez un nom de fichier de sauvegarde des poids de votre réseau de neurones. Ce fichier vous permettra de relancer l'analyse en mode estimation sur un échantillon test.

La feuille apprentissage comporte la valeur de l'erreur à chaque itération. Le graphique associé montre l'évolution de l'erreur.

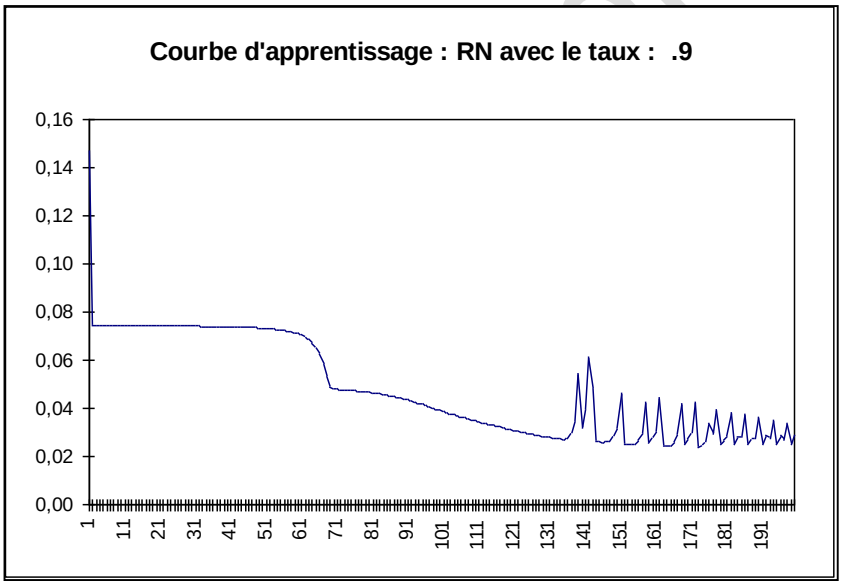



La courbe d'apprentissage baisse rapidement et ne montre pas d'oscillations. Le taux d'apprentissage est égal à 0,5 et semble être adapté au jeu de données.

Maintenant, il est conseillé de relancer l'analyse en choisissant un nombre d'itérations plus petit (150 par exemple) correspondant au début de stabilisation horizontale de la courbe afin d'éviter le phénomène de sur-apprentissage.

Les réseaux de neurones apprennent à chaque itération un peu plus les données en entrée. Si le nombre d'itérations est trop important, le réseau de neurones perd sa capacité de généraliser sur un échantillon test (voir le paragraphe sur le sur-apprentissage)

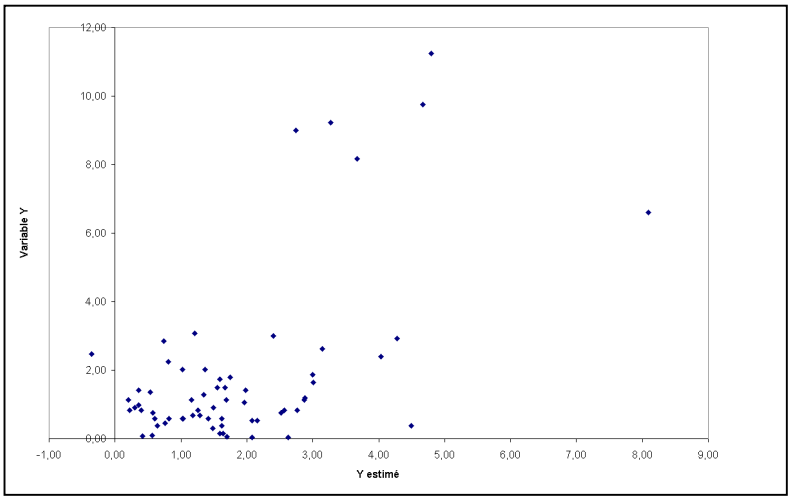
Nous aurions pu également choisir un taux plus important : 0.9 par exemple avec 200 itérations. Dans ce cas, l'évolution de l'erreur n'est pas stable et ne nous permet pas de conclure à un bon apprentissage même si la courbe converge en définitif. Le résultat risque d'être « moins bon ». Le graphique de l'apprentissage serait ici :




La feuille Y calculé contient les erreurs calculées entre la variable d'origine et la variable calculée. Les erreurs sont mises au carré afin d'éliminer l'influence du signe.

| Observation | Variable initiale | Variable calculée | Erreur | Erreur au carré |
|---|---|---|---|---|
| 1 | 1,3500 | 0,5413 | -0,8087 | 0,6539 |
| 2 | 1,8000 | 0,8520 | -0,9480 | 0,8986 |
| 3 | 2,6280 | 1,5202 | -1,1078 | 1,2273 |
| ... | ... | ... | ... | ... |
| 63 | 0,9720 | 0,7925 | -0,1795 | 0,0322 |

| | | | | |
|---|---|---|---|---|
| 64 | 0,1530 | 1,4102 | 1,2572 | 1,5804 |
| | | | Erreur totale | 151,3948 |




**Remarque :** la première analyse n'est pas souvent la bonne. Il faut en effet trouver le bon taux d'apprentissage et le bon nombre d'itérations.

Pour le bon taux d'apprentissage, la courbe doit baisser régulièrement pour se stabiliser horizontalement. Si elle oscille, réduisez le taux d'apprentissage.

Une fois la bonne courbe obtenue, refaites une analyse en limitant le nombre d'itérations. Choisissez celui qui correspond au début de la stabilisation horizontale de la courbe d'apprentissage. Le réseau de neurones aura une meilleure capacité à généraliser sur des nouveaux jeux de données.

## MULTIDIMENSIONAL SCALING (MDS)

Utilisez le multidimensional scaling (ou *positionnement multidimensionnel*) pour représenter dans un espace de faible dimension des observations pour lesquels seule une matrice de similarité ou de dissimilarité est disponible.

### Description

Le multidimensional scaling (MDS) est une méthode d'analyse d'une matrice de proximité (similarité ou dissimilarité) établie sur un ensemble d'observations. Le MDS a pour objectif de modéliser les proximités entre les observations de façon à pouvoir les représenter le plus fidèlement possible dans un espace de faible dimension (généralement 2 dimensions). Il existe différents algorithmes de MDS : StatBox utilise l'algorithme SMACOF (*Scaling by MAjorizing a COnvex Function*). Par ailleurs, il existe plusieurs modèles de MDS (ou fonctions de représentation), c'est-à-dire plusieurs façon de transformer les dissimilarités en disparités (*disparities*). Les disparités sont des distances décrivant la représentation optimale des observations. La mesure de l'écart entre les disparités et les distances mesurées sur la représentation obtenue par le MDS se nomme le *stress* : plus le *stress* est faible, meilleure est la représentation des observations.

Lorsque la fonction de représentation se contente de respecter les relations d'ordre, on parle de MDS ordinal ou non métrique (*ordinal MDS*, *nonmetric MDS*). Lorsque la transformation des dissimilarités en disparités s'effectue au moyen d'une fonction paramétrique spécifique, on parle de MDS métrique (*metric MDS*). Les modèles proposés dans la version actuelle de StatBox sont les suivants :

#### *MDS métrique*

- absolu (*absolute MDS*) : chaque dissimilarité $d_{ij}$ doit correspondre exactement à la distance entre les points *i* et *j* dans l'espace de représentation.
- rapport (*ratio MDS*) : le rapport de tout couple de distances dans l'espace de représentation doit correspondre au rapport des dissimilarités correspondantes.

- intervalle (*interval MDS*) : le rapport des différences entre distances dans l'espace de représentation doit correspondre au rapport des différences des dissimilarités correspondantes.

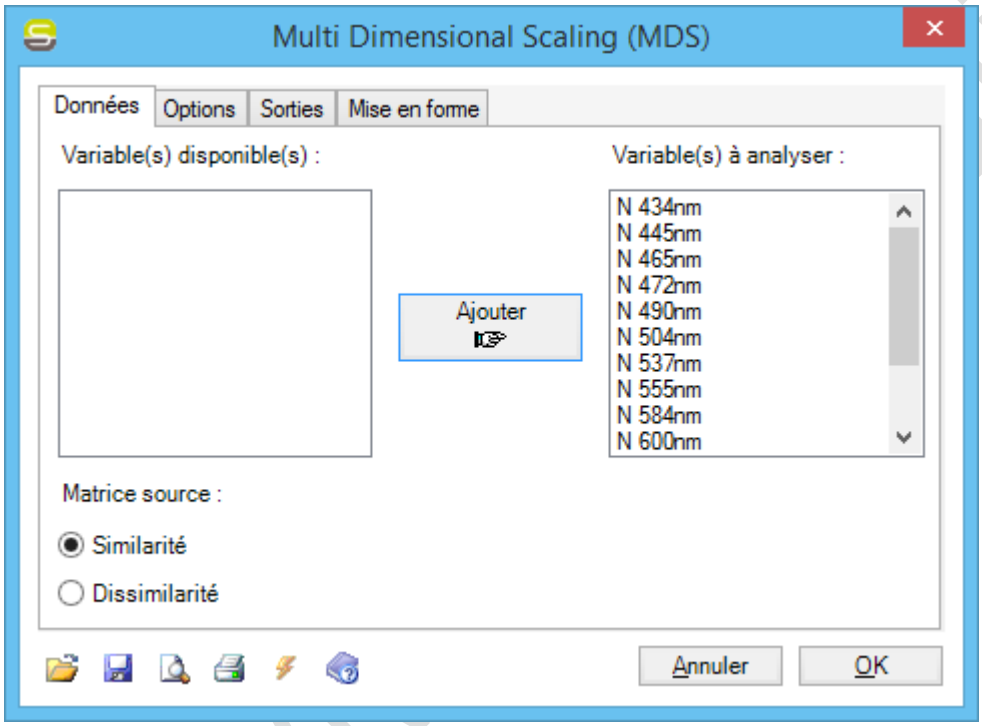
**Remarque :** StatBox ne gère pas les disparités négatives qui peuvent survenir lorsqu'on utilise le modèle « intervalle ». Si un message d'erreur est affiché à ce sujet, vous devez alors utiliser un autre modèle pour traiter vos données.

### *MDS non métrique*

- ordinal (1) : la relation d'ordre entre les distances dans l'espace de représentation doit correspondre à celle des dissimilarités correspondantes. En cas de dissimilarités de même rang, aucune restriction n'est imposée sur les distances correspondantes.
- ordinal (2) : modèle identique au précédent mais en cas de dissimilarités de même rang, les distances correspondantes doivent être égales.

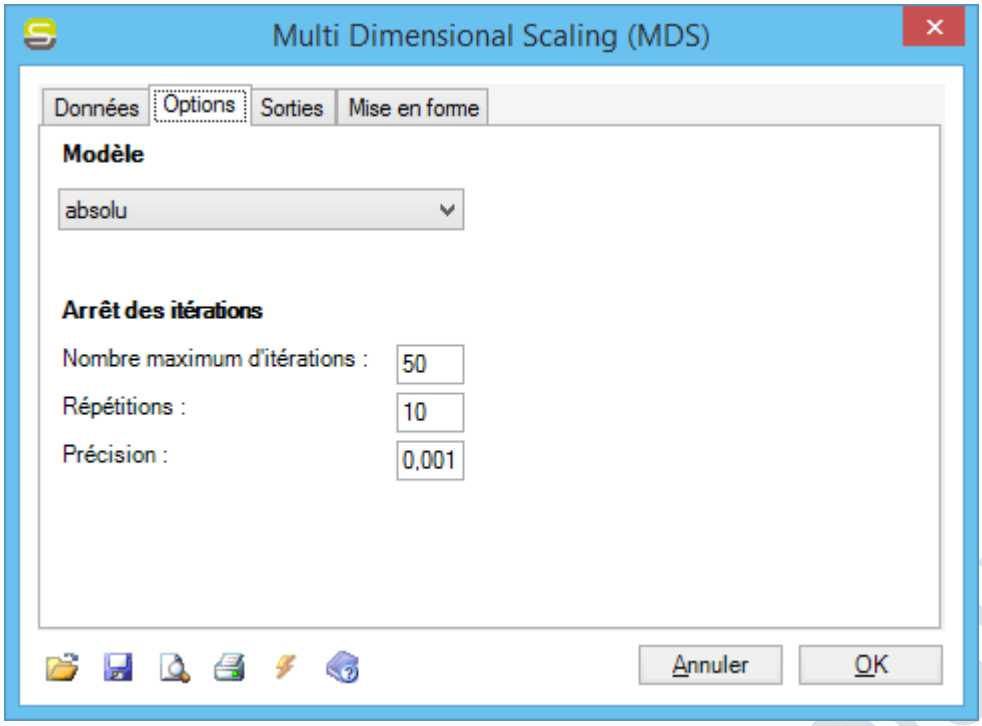
## Mise en œuvre

### *Onglet « Données »*



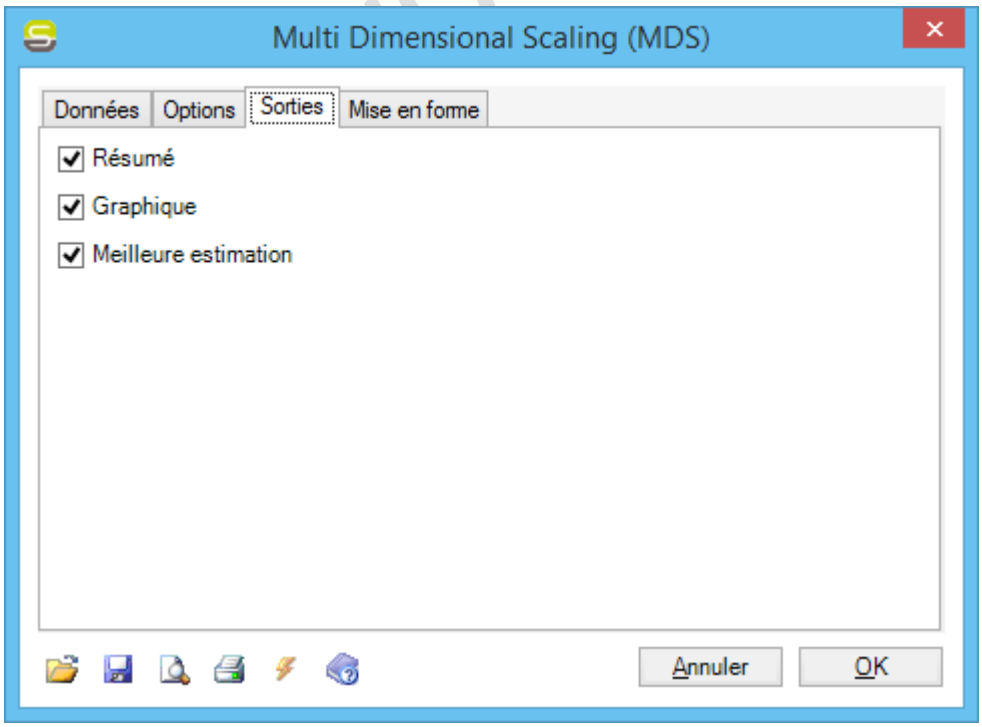

- *"Similarité" / "Dissimilarité" :* choisissez la nature des données, soit une matrice de similarité, soit une matrice de dissimilarité. StatBox travaille exclusivement avec des dissimilarités, de sorte qu'une matrice de similarité doit nécessairement être transformée en matrice de dissimilarité.
- *Variable(s) à analyser :* saisissez les variables correspondant à une matrice de proximité (similarité ou dissimilarité). Les données manquantes sont autorisées jusqu'à ce que la quantité d'information disponible soit insuffisante. Les données manquantes sont équivalentes de données dont le poids est nul.

*Onglet « Options »*




- *Modèle :* choisissez le modèle à utiliser comme fonction de représentation des dissimilarités.
- *Nombre maximal d'itérations :* entrez le nombre maximal d'itérations autorisé pour la minimisation du stress. Même si la convergence du stress n'est pas encore atteinte, l'optimisation itérative sera arrêtée au-delà du nombre maximal d'itérations spécifié. Valeur par défaut : 50.
- *Répétitions :* dans le cas d'une configuration de départ aléatoire, saisissez le nombre de répétitions de l'algorithme. Plusieurs répétitions permettent d'obtenir plusieurs configurations finales et de retenir la meilleure d'entre elles. Valeur par défaut : 10.
- *Précision :* entrez le seuil de convergence entre deux valeurs successives du stress. La convergence est atteinte lorsque l'écart absolu entre deux valeurs successives est inférieur ou égal au seuil spécifié. Valeur par défaut : 0,001.

*Onglet « Sorties »*



- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Graphique :* affiche un diagramme de Shepard. Ce graphique permet de comparer les disparités et les distances aux dissimilarités
- *Meilleure estimation :* affiche un bilan des différentes répétitions et la meilleure estimation retenue.

### Références

**Borg I. & P. Groenen (1997)**. Modern multidimensional Scaling. Theory and applications. Springer Verlag, New York.

**Dillon W.R. & M. Goldstein (1984)**. Multivariate analysis. Methods and applications. John Wiley & Sons, New York, pp. 107-156.

**Jobson J.D. (1992)**. Applied multivariate data analysis. Volume II: categorical and multivariate methods. Springer-Verlag, New York, pp. 568-605.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 172-173.

## CLASSIFICATION PAR PARTITIONNEMENT (*K*-MEANS)

Utilisez la méthode des K-means (ou *méthode des centres mobiles*) pour partitionner des observations en classes homogènes, sur la base de leur description par un ensemble de variables quantitatives.

**Remarque** : dans le cas de variables qualitatives, il est nécessaire d'effectuer au préalable une analyse des correspondances multiples (ACM) et de considérer les coordonnées des observations sur les axes factoriels obtenus comme de nouvelles variables.

### Description

L'algorithme des nuées dynamiques - analogue à l'algorithme des *k-means* - consiste à améliorer de façon itérative une partition initiale en minimisant l'inertie intra-classe. À chaque itération, l'algorithme calcule les barycentres des classes de la partition courante, puis affecte chaque observation au barycentre le plus proche afin de former une nouvelle partition dont l'inertie intra-classe est plus faible que la précédente. La variante utilisée par StatBox garantit qu'aucune classe ne peut se vider complètement de ses observations.

Cette méthode ne garantit pas que la solution obtenue à la convergence soit la solution optimale, c'est-à-dire la meilleure solution parmi toutes les solutions possibles. En ce sens, cet algorithme doit être vu comme une *heuristique*, permettant seulement d'obtenir une bonne solution, la résolution exacte du problème d'optimisation combinatoire sous-jacent n'étant généralement pas envisageable, sauf pour de très petits jeux de données. La meilleure stratégie pour obtenir une très bonne solution en un temps de calcul raisonnable consiste à exécuter l'algorithme des nuées dynamiques à partir de plusieurs partitions initiales différentes, puis de conserver la meilleure partition finale parmi toutes celles obtenues.

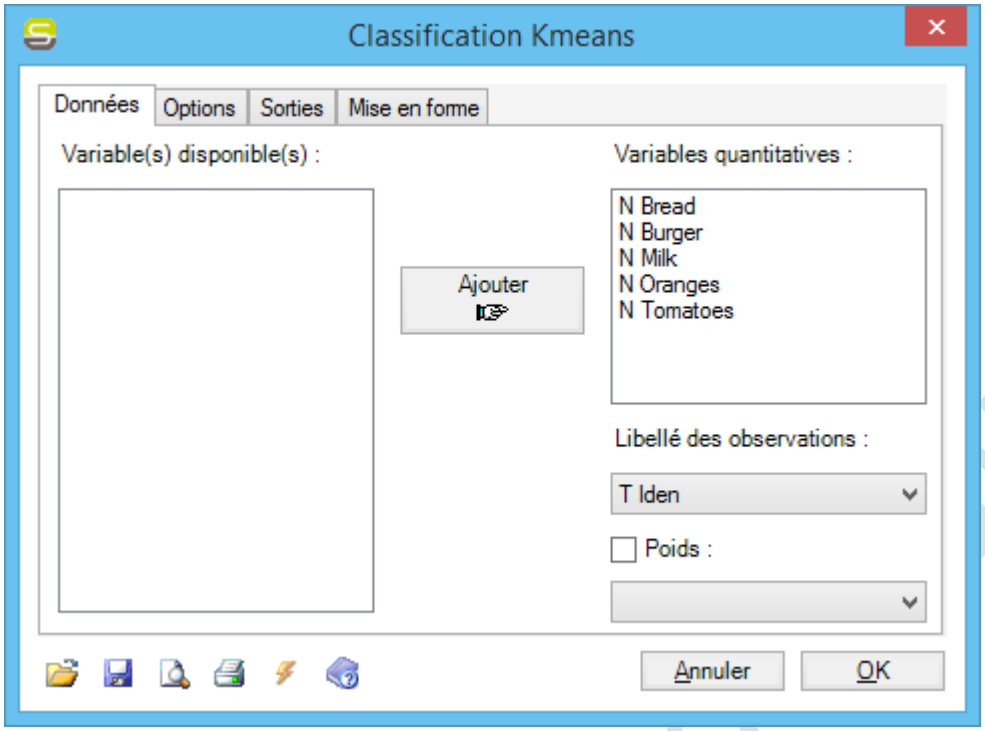
Lorsque plusieurs répétitions de la méthode sont effectuées à partir de partitions initiales différentes, StatBox identifie les formes fortes, c'est-à-dire les groupes d'observations qui ont toujours été classés ensemble. Les formes fortes représentent des groupes stables qui correspondent à l'intersection de toutes les partitions considérées. Les observations qui n'appartiennent à aucune forme forte sont affectés tantôt à une classe, tantôt à une autre, selon la partition initiale utilisée. Ces observations se trouvent généralement dans des régions intermédiaires situées entre les formes fortes. Pour identifier les formes fortes, StatBox considère au maximum les 10 meilleures partitions différentes obtenues lors des exécutions répétées de l'algorithme.

**Remarque** : l'utilisation de l'inertie intra-classe comme critère à minimiser conduit à la formation de classes compactes. Par exemple, dans un espace à deux dimensions, l'algorithme des nuées dynamiques tend à proposer des classes les plus circulaires possible. De ce fait, n'utilisez pas cette méthode si vous savez *a priori* que la forme

des classes naturelles sous-jacentes à vos données n'est pas compacte mais plutôt allongée (par exemple), le critère optimisé étant alors inadapté.

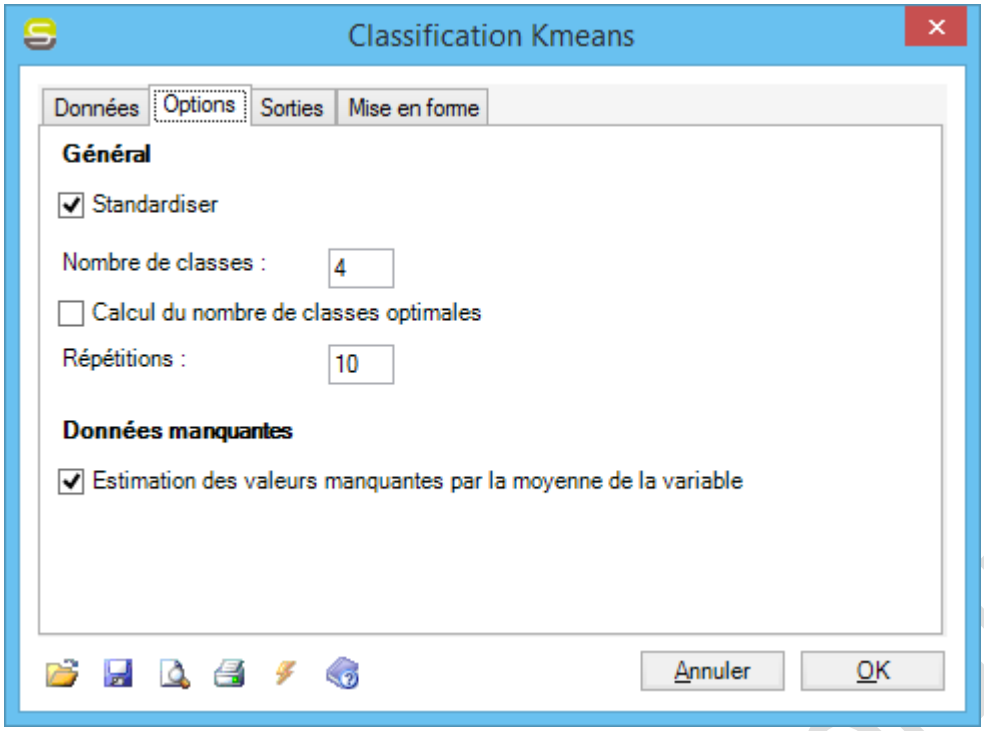
## Mise en œuvre

### *Onglet « Données »*



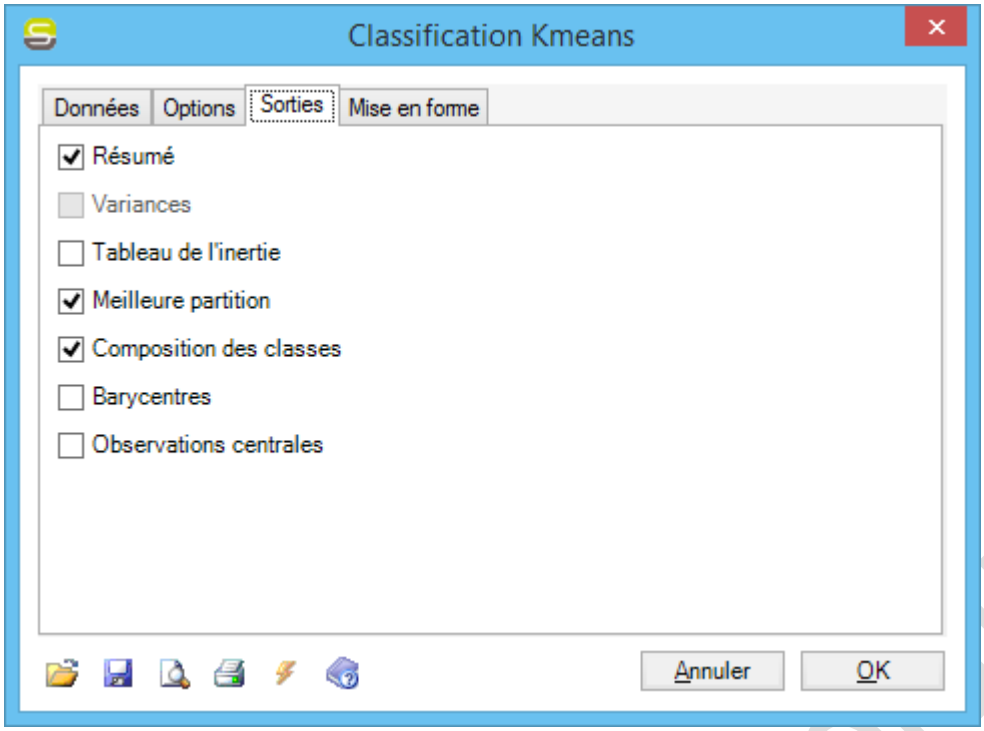

- *Données :* sélectionnez les variables correspondant à un tableau rectangulaire observations/variables. Lorsqu'il y a des valeurs manquantes, StatBox propose tout d'abord d'ignorer les lignes concernées. En cas de refus, StatBox propose alors d'estimer les valeurs manquantes de chaque variable par la moyenne (*cf.* l'option « *Estimation des données manquantes* »), sinon le traitement est abandonné.
- *Libellés des observations :* choisissez la variable contenant les libellés qui correspondent aux lignes du tableau de données.
- *Poids :* choisissez la variable contenant le poids des observations. Les valeurs manquantes dans les poids sont cumulées avec les valeurs manquantes dans les données : StatBox propose d'ignorer les lignes correspondantes ou d'estimer les valeurs manquantes par la moyenne des poids (*cf.* l'option « *Estimation des données manquantes* »), calculée sans tenir compte des éventuels poids nuls.

*Onglet « Options »*




- *Standardiser :* standardise les variables, c'est-à-dire diviser les valeurs par l'écart-type de la variable correspondante afin de supprimer des différences d'unités.
- *Nombre de classes :* entrez le nombre de classes de la partition à obtenir.
- *Calcul du nombre de classes optimales :* le logiciel calcule automatiquement le nombre de classes traduisant la meilleure partition.
- *Répétitions :* dans le cas d'une partition initiale automatique, saisissez le nombre de répétitions de l'algorithme. Plusieurs répétitions permettent d'obtenir plusieurs partitions finales et de retenir la meilleure d'entre elles. Valeur par défaut : 10.
- *Estimation des valeurs manquantes par la moyenne de la variable :* estime automatiquement les données manquantes par la moyenne de la variable considérée. Si cette option n'est pas cochée le logiciel vous demandera si vous désirez effectuer cette estimation.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Variances :* affiche l'évolution de la variance en fonction du nombre de classes.
- *Tableau de l'inertie :* affiche la table de décomposition de la variance intra-classe, inter-class et totale.
- *Meilleure partition :* affiche l'appartenance des différentes observations aux différentes classes sur la meilleure partition obtenue.
- *Composition des classes :* affiche les compositions des différentes classes
- *Barycentres :* affiche dans un tableau les coordonnées des barycentres des classes pour les différentes variables.
- *Observations centrales :* affiche, pour chaque classe, les coordonnées de l'objet le plus proche du barycentre de la classe.

## Références

**Diday E. (1971)**. Une nouvelle méthode en classification automatique et reconnaissance des formes, la méthode des nuées dynamiques. *Revue de Statistique Appliquée*, **19** 19-33.

**Diday E., J. Lemaire, J. Pouget & F. Testu (1982)**. Éléments d'analyse de données. Dunod, Paris, pp. 116-129.

**Jobson J.D. (1992)**. Applied multivariate data analysis. Volume II: categorical and multivariate methods. Springer-Verlag, New York, pp. 560-562.

**Johnson R.A. & D.W. Wichern (1992)**. Applied multivariate statistical analysis. Prentice-Hall, Englewood Cliffs, pp. 596-602.

**Lebart L., A. Morineau & M. Piron (1997)**. Statistique exploratoire multidimensionnelle. 2ème édition. Dunod, Paris, pp. 148-154.

**Roux M. (1985)**. Algorithmes de classification. Masson, Paris, pp. 61-75.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 159-165.

## CLASSIFICATION ASCENDANTE HIÉRARCHIQUE (CAH)

Utilisez la classification ascendante hiérarchique pour constituer des groupes d'observations similaires (classes) sur la base de leur description par un ensemble de variables quantitatives, ou éventuellement de tous types.

**Remarque** : pour les variables qualitatives non binaires il est préférable d'effectuer au préalable une analyse des correspondances multiples (ACM) et de considérer les coordonnées des observations sur les axes factoriels comme de nouvelles variables.

### Description

La classification ascendante hiérarchique (CAH) consiste à agréger progressivement les observations selon leur ressemblance, mesurée à l'aide d'un indice de similarité ou de dissimilarité. L'algorithme commence par rassembler les couples d'observations les plus ressemblants, puis à agréger progressivement les autres observations ou groupes d'observations en fonction de leur ressemblance, jusqu'à ce que la totalité des observations ne forme plus qu'un seul groupe. La CAH produit un arbre binaire de classification (*dendrogramme*), dont la racine correspond à la classe regroupant l'ensemble des observations. Ce dendrogramme représente une hiérarchie de partitions, une partition étant obtenue par troncature du dendrogramme à un certain niveau de ressemblance. La partition comporte alors d'autant moins de classes que la troncature s'effectue en haut du dendrogramme (c'est-à-dire vers la racine). A la limite, une troncature effectuée en dessous du premier nœud de l'arbre conduit à ce que chaque classe ne contienne qu'une observation (cette partition est l'*assise* du dendrogramme), et une troncature effectuée au-delà du niveau de la racine du dendrogramme conduit à une seule classe contenant tous les observations.

Il existe de nombreuses mesures de ressemblances (similarités ou dissimilarités), et plusieurs méthodes pour recalculer la ressemblance lorsque l'algorithme forme des groupes (critères d'agrégations). StatBox propose des indices et des critères sélectionnés en fonction de leurs propriétés mathématiques et de leur intérêt pratique ou pédagogique.

Liste des similarités/dissimilarités

StatBox propose plusieurs similarités/dissimilarités qui sont adaptées à un type de données particulier.

Pour les données quantitatives :

| Similarité | Dissimilarité |
|---|---|
| Corrélation de Pearson | Distance euclidienne |
| Corrélation de Spearman | Distance du khi² |
| Corrélation de Kendall | Distance de Manhattan |
| | Dissimilarité de Pearson |
| | Dissimilarité de Spearman |
| | Dissimilarité de Kendall |

**Remarque** : afin de traiter différents types de variables (quantitatives et qualitatives), il est possible d'utiliser une similarité/dissimilarité générale qui traite toutes les variables au niveau algébrique le plus faible, c'est-à-dire celui des variables nominales. Ceci s'accompagne nécessairement d'une perte d'information. Il peut s'avérer plus intéressant de discrétiser les variables quantitatives à l'aide du module "codage en classes », puis de les analyser conjointement aux variables qualitatives à l'aide d'une analyse des correspondances multiples (ACM), afin d'utiliser les coordonnées factorielles des observations comme nouvelles variables.

Liste des critères d'agrégation disponibles :

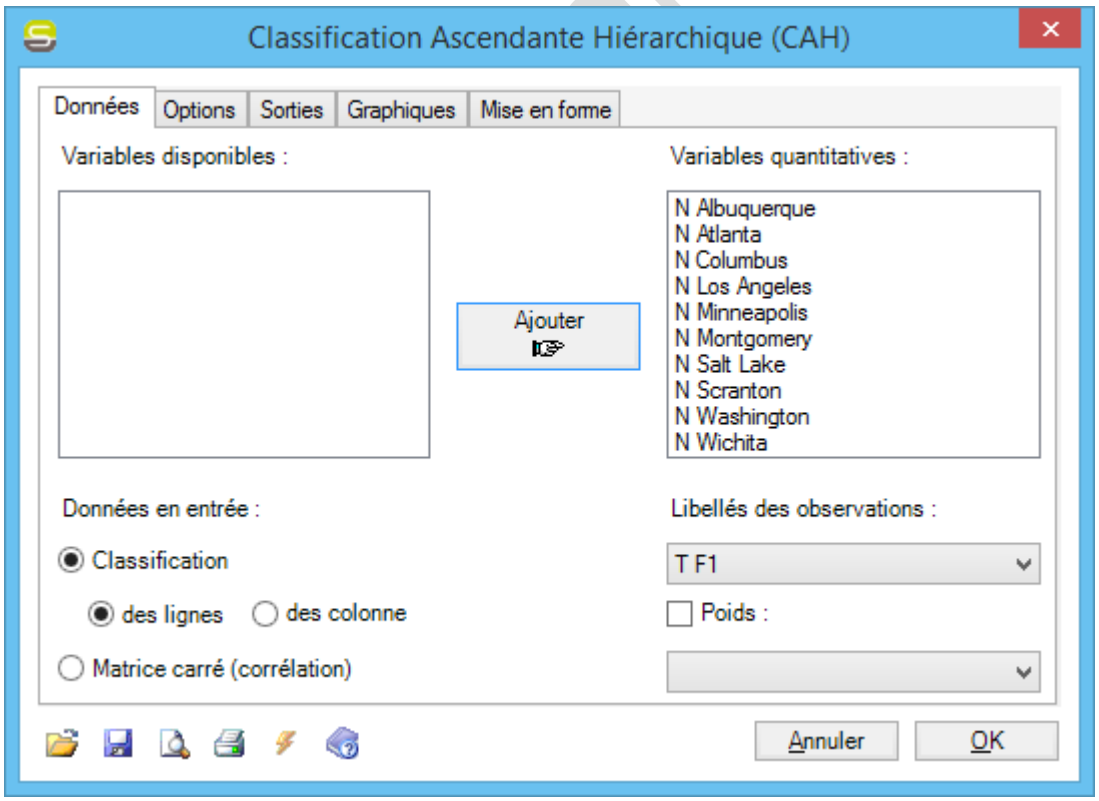
- La dissimilarité entre deux groupes d'objets A et B peut être calculée selon différentes méthodes nommées critères d'agrégation, chaque critère conditionnant la structure de la hiérarchie binaire produite par l'algorithme de CAH. Parmi les critères existants, StatBox en propose sept : liens simples, complet, moyen, proportionnel, flexible, fort, ainsi que le critère de Ward fondé sur l'augmentation de l'inertie.

- Lien simple : la dissimilarité entre A et B est la dissimilarité entre l'objet de A et l'objet de B les plus ressemblants. L'agrégation par le lien simple a tendance à contracter l'espace des données et à écraser les niveaux des paliers du dendrogramme. Comme la dissimilarité entre deux éléments de A et de B suffit à relier A et B, ce critère peut conduire à relier des classes très allongées (effet de chaînage).
- Lien complet : la dissimilarité entre A et B est la plus grande dissimilarité entre un objet de A et un objet de B. L'agrégation par le lien complet a tendance à dilater l'espace des données et produit des classes compactes.
- Lien moyen : la dissimilarité entre A et B est la moyenne des dissimilarités entre les objets de A et les objets de B. L'agrégation selon le lien moyen constitue un bon compromis entre les deux extrêmes précédents et respecte assez bien les propriétés de l'espace des données.
- Lien proportionnel : la dissimilarité moyenne entre les objets de A et de B est calculée comme une somme de dissimilarités pondérée de telle sorte qu'un poids égal soit attribué aux deux groupes. Comme le lien moyen, ce critère respecte assez bien les propriétés de l'espace des données.
- Lien fort : ce critère fait intervenir à la fois la moyenne des distances à l'intérieur de chaque groupe et la moyenne des distances entre les groupes. Son utilisation conduit à la formation de classes très compactes.
- Augmentation de l'inertie (Ward) : on agrège deux groupes de sorte que l'augmentation de l'inertie intraclasse soit la plus petite possible, afin que les classes restent homogènes. Ce critère, proposé notamment par Ward (1963), ne peut s'utiliser que dans le cas des distances quadratiques, c'est-à-dire ici, dans le cas de la distance euclidienne et de la distance du khi².

**Remarque** : par défaut, StatBox propose d'utiliser le critère d'agrégation de l'augmentation d'inertie pour les distances quadratiques (distances euclidienne et du khi²) et le critère du lien moyen dans tous les autres cas. Le choix d'un autre critère doit s'effectuer en connaissance de cause.

## Mise en œuvre

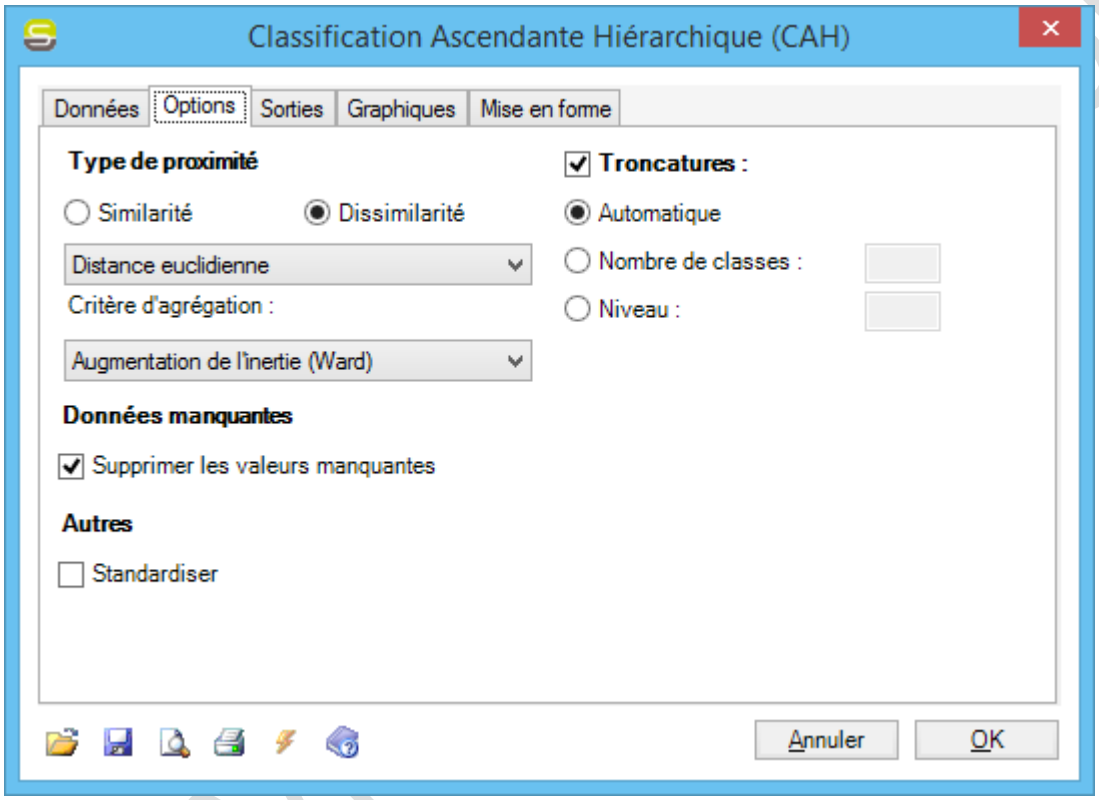
### *Onglet « Données »*




➢ *Variables quantitatives :* sélectionnez les variables correspondant à un tableau rectangulaire observations/variables ou à une matrice de similarité/dissimilarité.

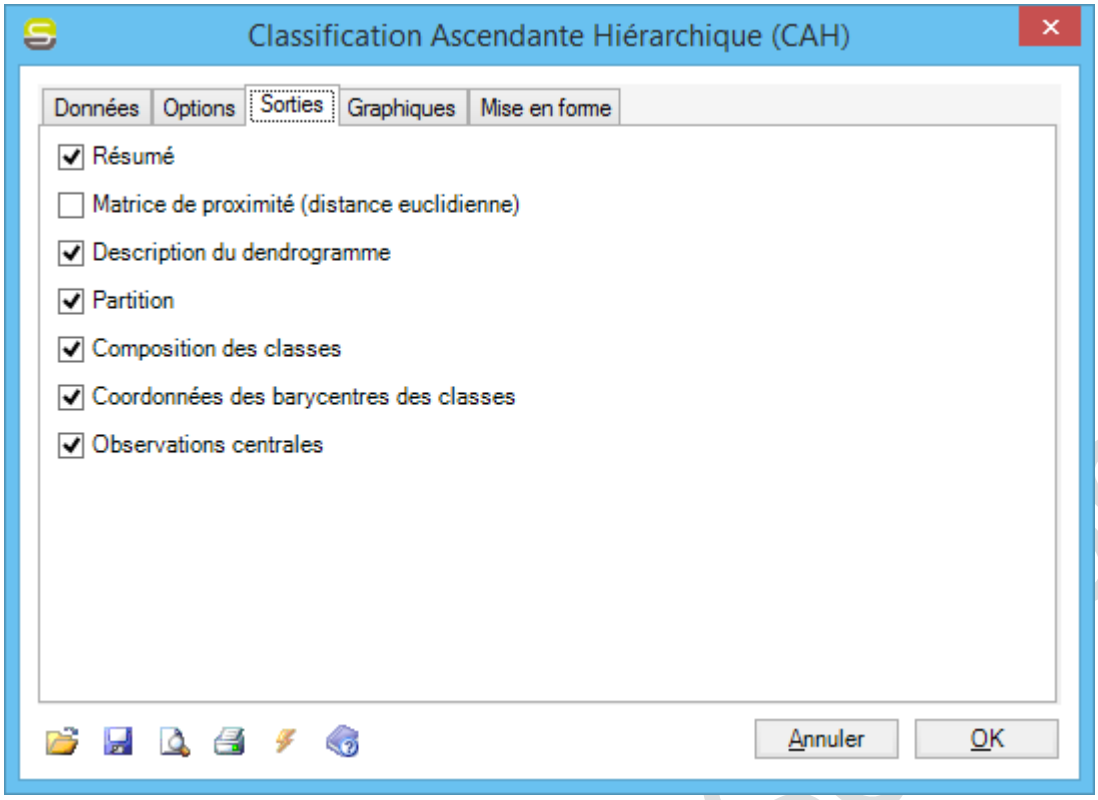
- *« Tableau » / « Matrice» :* choisissez la nature des données, tableau observations/variables ou matrice de similarité/dissimilarité. Dans le cas d'un tableau, lorsqu'il y a des valeurs manquantes StatBox propose d'ignorer les lignes concernées, sinon StatBox indique qu'il est possible d'utiliser toute l'information disponible (*pairwise deletion*) grâce au « Matrice de similarité / dissimilarité », puis la boîte de dialogue est fermée et le traitement est abandonné. Dans le cas d'une matrice de similarité/dissimilarité, les valeurs manquantes ne sont pas autorisées.
- *Classification « des lignes » / « des colonnes » :* dans le cas d'un tableau observations/variables, choisissez si la matrice de similarité/dissimilarité doit croiser les lignes du tableau de données, ou bien les colonnes.
- *Libellés des observations :* dans le cas d'un tableau observations/variables, saisissez la plage de la colonne de libellés qui correspondent aux lignes du tableau de données.
- *Poids :* dans le cas d'un tableau observations/variables, sélectionnez la variable poids des colonnes du tableau (lorsque ce sont les lignes qui sont analysées) ou des lignes du tableau (lorsque ce sont les colonnes qui sont analysées).

*Onglet « Options »*



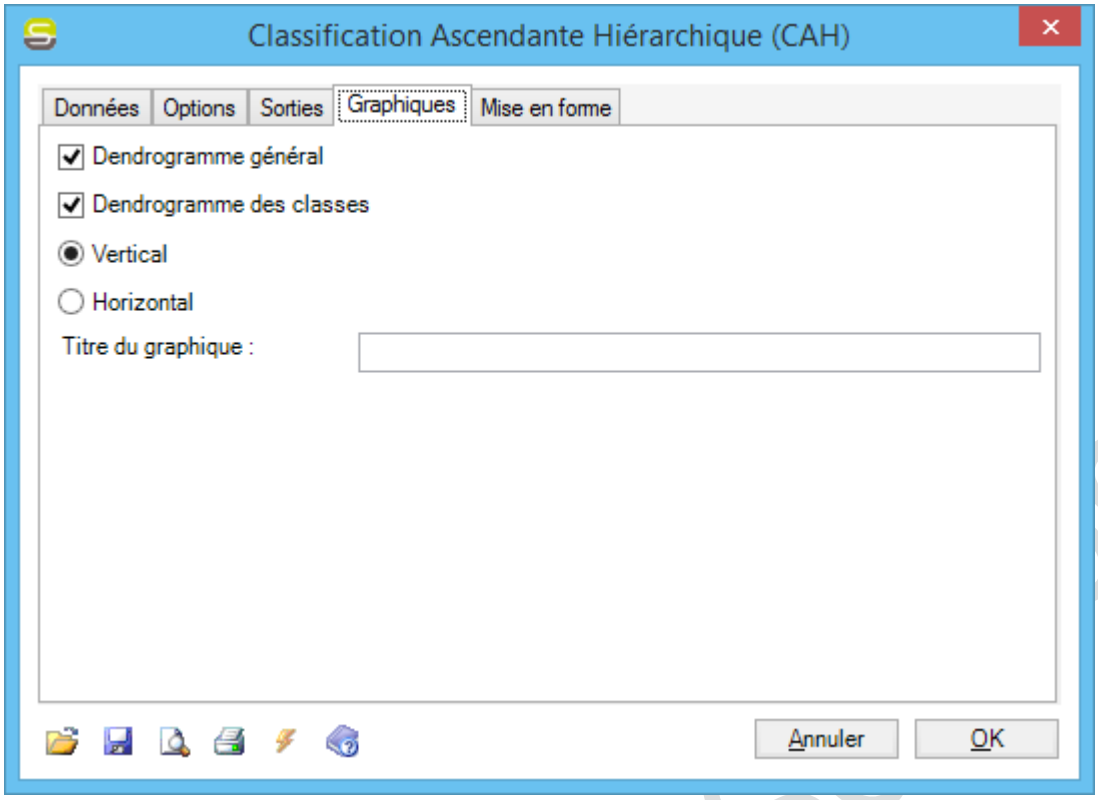

- *« Similarité » / « Dissimilarité » :* choisissez si les valeurs calculées à partir du tableau observations/variables ou les valeurs contenues dans la matrice sont des similarités, ou bien des dissimilarités. Le choix du type de mesure conditionne la liste des critères d'agrégation qui sont proposés ainsi que le traitement des données.
- *Standardiser :* dans le cas d'un tableau observations/variables contenant des données quantitatives, cochez cette case pour standardiser les variables, c'est-à-dire diviser les valeurs par l'écart-type de la variable correspondant, afin de supprimer l'effet des différences d'unités.
- *Troncature :* cochez cette case pour effectuer une troncature du dendrogramme et obtenir une partition.
- *Automatique :* le niveau de troncature du dendrogramme - et par conséquent le nombre de classes de la partition - est déterminé automatiquement par StatBox en fonction de la structure de l'histogramme des niveaux des paliers.
- *Nombre de classes :* entrez le nombre de classes de la partition à obtenir.
- *Niveau :* entrez le niveau de troncature. Une première exécution du module est généralement nécessaire afin de pouvoir décider d'un niveau de troncature correct.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Matrice de proximité :* dans le cas d'un tableau observations/variables, affiche la matrice de proximité calculée par StatBox avant d'effectuer la CAH.
- Description du dendrogramme : affiche le dendrogramme
- *Partition :* affiche la partition retenue.
- *Composition des classes :* affiche la composition des classes.
- *Coordonnées des barycentres des classes :* affiche la table des distances euclidiennes entre les barycentres des classes pour les différentes variables.
- *Observations centrales :* affiche dans une table pour chaque classe les coordonnées de l'objet le plus proche du barycentre de la classe.

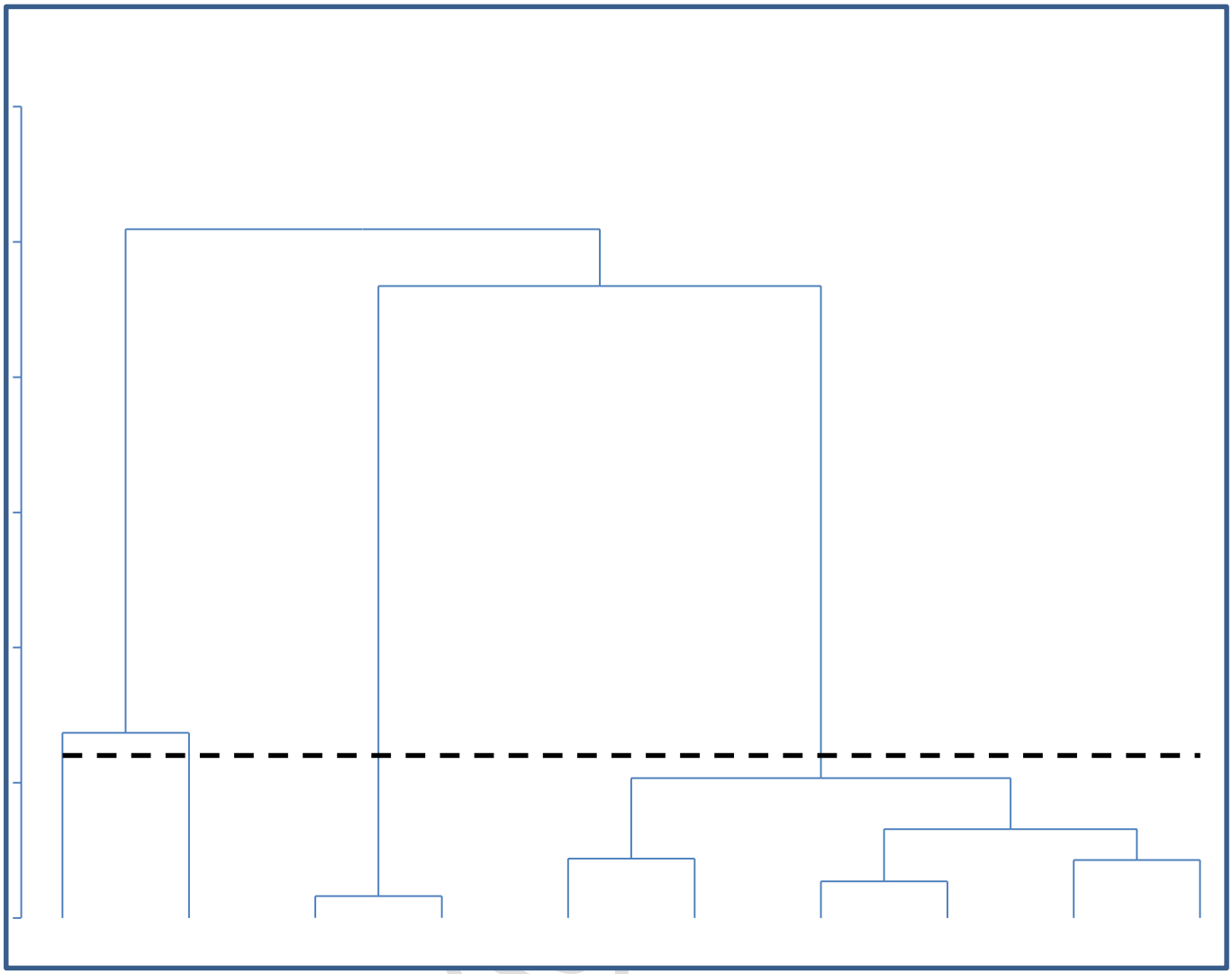
*Onglet « Graphiques »*




- *Dendrogramme général :* affiche le diagramme des niveaux permettant d'observer l'impact des regroupements successifs.
- *Dendrogramme des classes :* affiche le dendrogramme de découpage des classes.
- *« Vertical » / « Horizontal» :* choisissez « Vertical » pour que la racine du dendrogramme figure en haut du graphique, ou bien « Horizontal » pour que la racine du dendrogramme figure à droite du graphique.
- *Titre du graphique :* entrez un titre spécifique pour le graphique (facultatif).

## Exemple

Feuille " CAH " du classeur " Data.xls " (Jobson 1992, table 10.11, p. 536).



## Références


**Benzécri J.P. (1984)**. L'analyse des données. 1. La taxinomie. Quatrième édition. Dunod, Paris.

**Diday E., J. Lemaire, J. Pouget & F. Testu (1982)**. Eléments d'analyse de données. Dunod, Paris, pp. 46-116.

**Dillon W.R. & M. Goldstein (1984)**. Multivariate analysis. Methods and applications. John Wiley & Sons, New York, pp. 157-186.

**Jambu M. (1978)**. Classification automatique pour l'analyse des données. 1 - méthodes et algorithmes. Dunod, Paris.

**Jobson J.D. (1992)**. Applied multivariate data analysis. Volume II: categorical and multivariate methods. Springer-Verlag, New York, pp. 483-568.

**Johnson R.A. & D.W. Wichern (1992)**. Applied multivariate statistical analysis. Prentice-Hall, Englewood Cliffs, pp. 584-602.

**Lebart L., A. Morineau & M. Piron (1997)**. Statistique exploratoire multidimensionnelle. 2ème édition. Dunod, Paris, pp. 155-206.

**Roux M. (1985)**. Algorithmes de classification. Masson, Paris.

**Saporta G. (1990)**. Probabilités, analyse des données et statistique. Technip, Paris, pp. 251-260.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 166-174.

**Ward J.H. (1963)**. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58** : 238-244.

## ARBRES DE SEGMENTATION

La segmentation, au sens large, consiste à créer des groupes d'observations homogènes. On peut tout simplement créer des sous-populations à partir de quelques variables pour constituer ces groupes. Ces requêtes ne font intervenir qu'un nombre limité de variables (les hommes de moins de 35 ans). StatBox avec les arbres de segmentation permet de faire intervenir un ensemble complexe de variables.

Il existe plusieurs méthodes pour créer ces groupes. Soit on cherche à maximiser ou à minimiser la valeur d'une variable dans chacun de ces groupes, soit on cherche à obtenir des groupes homogènes sur un ensemble de variables. Dans ce dernier cas, on est dans le domaine de la classification.

La segmentation par arbre de décisions fait partie du premier cas. On cherche par exemple à identifier le sous-groupe d'observations en termes d'âges, de catégories sociales, etc. dans lequel se trouve le plus d'acheteurs. Ici la segmentation se fait en fonction d'une variable à expliquer : le taux d'achat.

Le taux de réponse à un mailing est généralement très faible. Il serait utile d'identifier les variables explicatives les plus importantes, les plus pertinentes. Parmi les variables dont on dispose, est-ce l'âge, la catégorie sociale, le type d'habitat, etc., qui est le plus lié, corrélé avec le taux de réponse ? La segmentation par arbre de décisions va nous permettre d'identifier les différentes variables explicatives du taux de réponse. On pourra isoler le ou les segments dont le taux de réponse est le plus élevé. On pourra également découvrir les segments dont le taux est le plus faible. L'identification de ces segments va nous permettre de réduire considérablement les coûts de nos mailings.

### La méthode CHAID

Avec la segmentation, il faut donc distinguer deux types de variables : la variable que l'on essaie d'expliquer et les variables explicatives.

Une base de données, par exemple sur des prêts contient des informations comme l'âge, le salaire, le type de logement, la profession, le nombre d'enfant etc. On dispose également d'un champ indiquant si le remboursement du crédit a été effectué avec succès ou non.

En fonction des informations disponibles, il s'agit de savoir quels sont les groupes à risque. Quels sont les variables, les attributs qui donnent le plus d'informations sur ces groupes à risque. Est-ce en priorité le salaire, la profession, l'âge qui identifiera le mieux nos groupes ?

Le logiciel va évaluer successivement toutes les variables que vous avez sélectionnées. Si, par exemple, la première variable dans la liste est la catégorie socioprofessionnelle – qui comporte les modalités agriculteurs, artisans commerçants, cadres, employés, ouvriers –, le programme va chercher le groupement des professions en deux catégories les plus différentes possibles. L'indice mesurant cette différence est le Khi-deux.

Le premier tableau sur lequel l'indice sera calculé est :

| | Agriculteurs | Artisans commerçants, cadres, employés, ouvriers |
|---|---|---|
| Risqué oui | | |
| Risqué non | | |

Puis le tableau suivant :

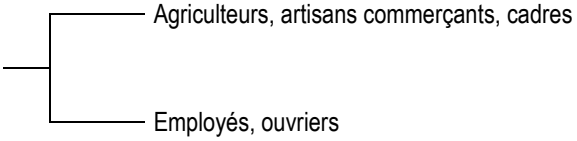
| | Agriculteurs, artisans commerçants | cadres, employés, ouvriers |
|---|---|---|
| Risqué oui | | |
| Risqué non | | |

Puis le tableau suivant :

| | Agriculteurs, artisans commerçants, cadres | employés, ouvriers |
|---|---|---|
| Risqué oui | | |
| Risqué non | | |

Toutes les combinaisons de modalités des professions sont calculées. Le programme retiendra la combinaison qui donne l'indice le plus élevé. À l'issu de ces calculs, la première catégorie de Profession sera associée à la modalité « Risqué oui » et la seconde catégorie à « Risqué non ».

Toutes les variables sont évaluées les unes après les autres et, pour chaque variable, on dispose de l'indice symbolisant l'association entre le risque et les 2 catégories obtenues. Le programme retient la variable et la combinaison de modalités ayant l'indice le plus élevé.




Ce résultat correspond au premier nœud de l'arbre. Chacune des deux branches correspond à une sous-population distincte. La même procédure est appliquée à ces deux sous-ensembles.

Pour obtenir suffisamment d'effectifs d'une part et pour simplifier les analyses d'autre part, StatBox divise la population en deux à chaque nœud. La division en plus de 2 catégories risque d'une part d'éparpiller trop vite la population initiale et, d'autre part, de rendre plus complexe l'analyse de résultats. De plus, le nombre de branches est lié à un seuil de probabilité que l'on se fixe a priori. En modifiant ce seuil, les branches changent. Il est donc difficile de connaître la bonne valeur de ce seuil. La division en deux branches a l'avantage de la clarté et a largement prouvé son efficacité.

La probabilité associée à un nœud permet d'identifier la significativité du découpage en 2 sous-populations. On admet généralement que si cette probabilité est inférieure à 0.05, on peut considérer que le découpage est significatif. Si cette probabilité est supérieure à 0.05 et inférieure à 0.10, le découpage montre une tendance. La part de hasard est ici trop importante pour en tirer des conclusions franches.

Pour effectuer une segmentation, il est conseillé de disposer d'un nombre suffisant d'observations. Certains auteurs suggèrent une taille de plus de 500 personnes. Il faut enfin souligner que la taille des segments obtenus n'est exploitable que si on obtient au moins 20 à 30 observations au niveau des feuilles de l'arbre.

La taille de l'échantillon ou du tableau de données a une certaine importance. Sur un petit tableau de moins de 500 observations, on peut moins facilement estimer la stabilité des résultats de la segmentation. L'élagage par validation croisée apporte une solution à ce problème. On effectue sur plusieurs sous-échantillons la segmentation et on compare les résultats. S'ils sont presque équivalents, on pourra dire que la segmentation est stable. S'ils sont très différents, il faudra être prudent quant aux conclusions de l'étude. C'est pour cette raison qu'il est préférable de disposer d'une population suffisante. D'un point de vue technique, on peut dire que la segmentation n'est pas vraiment une méthode multivariée dans la mesure où elle ne prend pas en compte l'ensemble des variables en même temps. Les traitements ne se font que sur 2 variables à la fois, contrairement aux régressions ou aux analyses discriminantes qui prennent en compte l'ensemble des variables explicatives dans leurs calculs. Sur des jeux de données qui comportent un certain flou, on peut se trouver dans cette situation d'instabilité.

Nous avons vu que le programme calcule des tableaux de contingences ou en d'autres termes des tris croisés et qu'il essaie successivement de créer un tableau plus petit ne comportant que 2 colonnes. Les variables explicatives ont des modalités disjointes. On dit qu'elles sont nominales ou non-numériques. Lorsqu'on est en présence de variables numériques, le programme va constituer automatiquement des classes à effectifs égaux. Le nombre de classes est déterminé par l'utilisateur. Plus le nombre de classes est grand et plus on a de chance que le découpage soit pertinent. Mais le nombre de classes est limité par la taille du tableau à analyser. Les classes obtenues sont

ordonnées. StatBox donne la possibilité de garder cet ordre dans les regroupements de ces classes. Par exemple les classes d'âges extrêmes (jeunes et vieux) ne peuvent pas être regroupées ensemble. Cette conservation de l'ordre est généralement utile pour les classes des variables numériques. Dans certains cas il est intéressant de pouvoir considérer les classes d'âge par exemple comme non ordonnées. Dans le domaine des loisirs on remarque que les classes extrêmes les plus jeunes et les plus de 55 ans ont un comportement similaire parce qu'ils disposent de plus de temps.

## La méthode CART

Bien que donnant des résultats à peu près similaires le principe de l'algorithme CART est un peu différent de celui employé dans CHAID.

Avant de présenter l'algorithme de séparation d'un nœud employé par CART, il faut d'abord parler de la notion d'impureté. L'impureté permet de mesurer l'homogénéité d'une population. Plus une population est homogène et plus on trouvera la présence d'une seule des modalités de la variable à expliquer.

Dans l'exemple suivant, la variable à expliquer est le pourcentage de satisfaction :

Supposons que la sous-population 1 soit composé de 11 personnes satisfaites et de 9 personnes insatisfaites, soit en pourcentage, 55% de satisfaits et 45% d'insatisfaits, les deux groupes de personnes sont presque autant représentés. Cette sous-population n'est donc pas homogène.

Une sous-population 2 est quant à elle est composée de 15 personnes satisfaites et de 5 personnes insatisfaites soit de 75% et 25%. On voit nettement que cette sous-population est majoritairement composée de personnes satisfaites, elle est donc plus homogène que la population précédente. L'impureté de la sous-population 1 est donc plus grande que celle de la sous-population 2.

Pour calculer l'impureté, plusieurs méthodes peuvent être utilisées. On utilise généralement la formule de Gini pour calculer cette impureté.

I= Impureté
P(X=R) étant la proportion de la modalité R dans notre population.
P(X<>R) étant la proportion de modalités différentes de R dans notre population.
La formule de Gini : $I=\Sigma\ P(X=R)*P(X<>R)$
soit : $I=\Sigma\ P(X=R)*(1- P(X=R))$

Ainsi, si on reprend l'exemple précédent l'impureté de la sous-population 1 :

Il n'y a que deux modalités dans la variable à expliquer, donc R ne peut prendre que deux valeurs : satisfait et insatisfait.

P(X=satisfait) = 11/20 = 0,55

P(X=insatisfait) = 9/20 = 0,45

I1= P(X=satisfait) * (1 - P(X=satisfait)) + P(X=insatisfait) * (1 - P(X=insatisfait))

I1 = 0,55 * 0,45 + 0,45 * 0,55 =0,495.

Pour la sous-population 2 :

P(X=satisfait) = 15/20 = 0,75

P(X= insatisfait) = 5/20 = 0,25

I2 = 0,75*0,25 + 0,25*0,75 = 0,375

Comme prévu I1 > I2.

*Évolution de l'impureté lors d'une séparation*

Lors d'une séparation d'une population P d'effectif E en deux sous-populations P1 et P2 d'effectifs E1 et E2, l'impureté suit la loi suivante : E * I(P) > E1 * I(P1) + E2 * I(P2)

En d'autres termes la somme pondérée par les effectifs des impuretés des fils d'un nœud est forcément inférieure à l'impureté du nœud père.

En procédant à des divisions successives, l'impureté globale diminue et la population de chaque nœud tend à devenir homogène au fur et à mesure des divisions.

On appelle la baisse de l'impureté le nombre :

$\Delta I = E * I(P) - ( E1 * I(P1) + E2 * I(P2) )$

(La différence entre les deux parties de la propriété précédente.)

On reprend la sous-population 1 et on introduit une variable éventuellement explicative de la satisfaction (homme, femme) :

Chez les hommes on obtient, 7 satisfaits et 3 insatisfaits soit un total de 10 hommes. Chez les femmes : 4 satisfaites et 6 insatisfaites soit un total de 10 femmes

Rappel de l'impureté de la population de départ (Calculé précédemment) :

I(1) = 0,495

On peut maintenant calculer les impuretés des deux sous-populations obtenues :

Population 3 composée d'hommes:

Satisfaits : 7/10 soit 70%

Insatisfait : 3/10 soit 30%

I3 = 0,7 * 0,3 + 0,3 * 0,7= 0,42

Pour la population 4 composée de femmes :

Satisfaits : 4/10 soit 40%

Insatisfaits : 6/10 soit 60%

I4 = 0,4 * 0,6 + 0,6 * 0,4 = 0,48

On remarque que l'impureté a diminué dans chacune des sous-populations par rapport au nœud initial. De plus, l'impureté suit la propriété exposée précédemment est vérifiée :

20 * I1 > 10 * I3 + 10 * I4

9,9 > 4,8 + 4,2

9,9 > 9

Ainsi l'impureté globale de la population de départ a diminué lors de la séparation.

La baisse d'impureté est donc de :

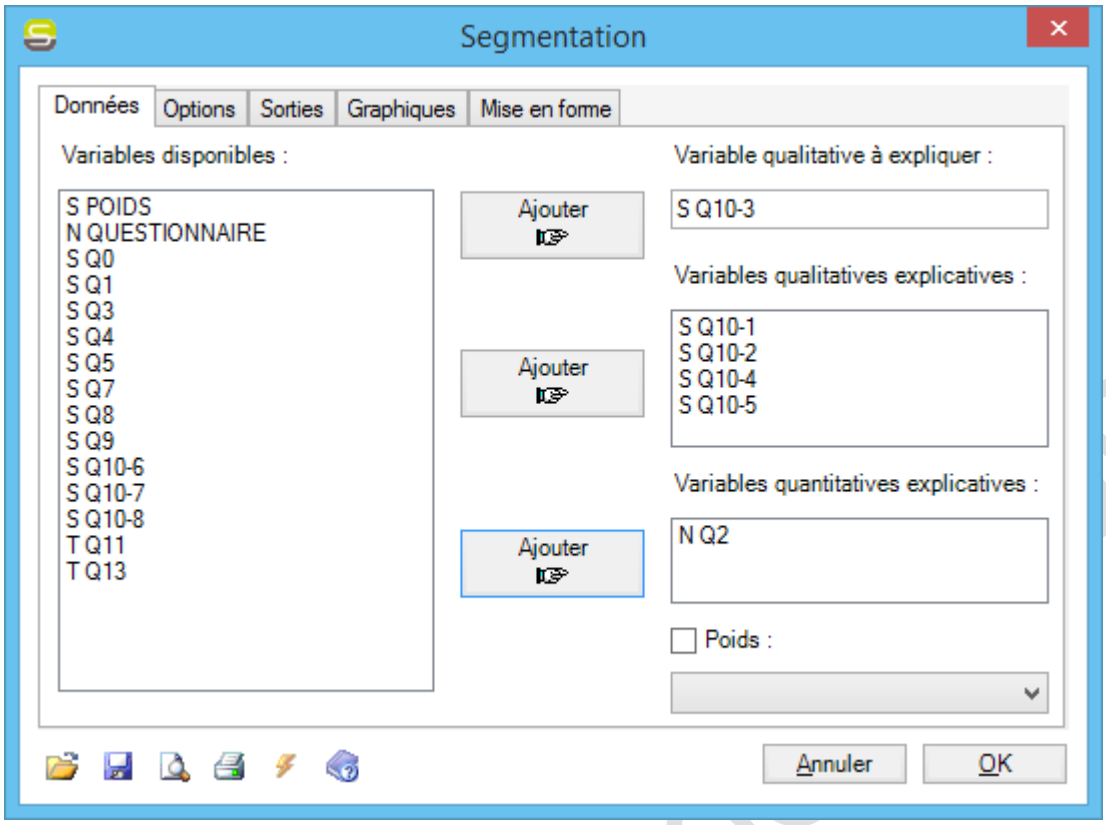
$\Delta I = 9,9 - 9 = 0,9$

Le principe de la méthode CART consiste à trouver la variable et le regroupement en 2 catégories de modalités qui donne la séparation qui diminue le plus l'impureté globale donc qui rend la baisse d'impureté maximale. On sépare ainsi successivement les populations. On obtient ainsi un arbre que l'on peut ainsi étudier de la même façon que CHAID, dans lequel chaque nœud tend à devenir homogène par rapport à une modalité de la variable à expliquer.

La différence essentielle entre les deux méthodes réside dans l'indice utilisé, khi-deux d'une part et impureté d'autre part. Les résultats quant à eux sont à peu près semblables. A noter que CART par rapport à CHAID évite les séparations qui créeraient deux populations d'effectifs complètement inégaux. Par exemple, 1 observation d'un coté et 500 de l'autre.

## Mise en œuvre

*Onglet « Données »*

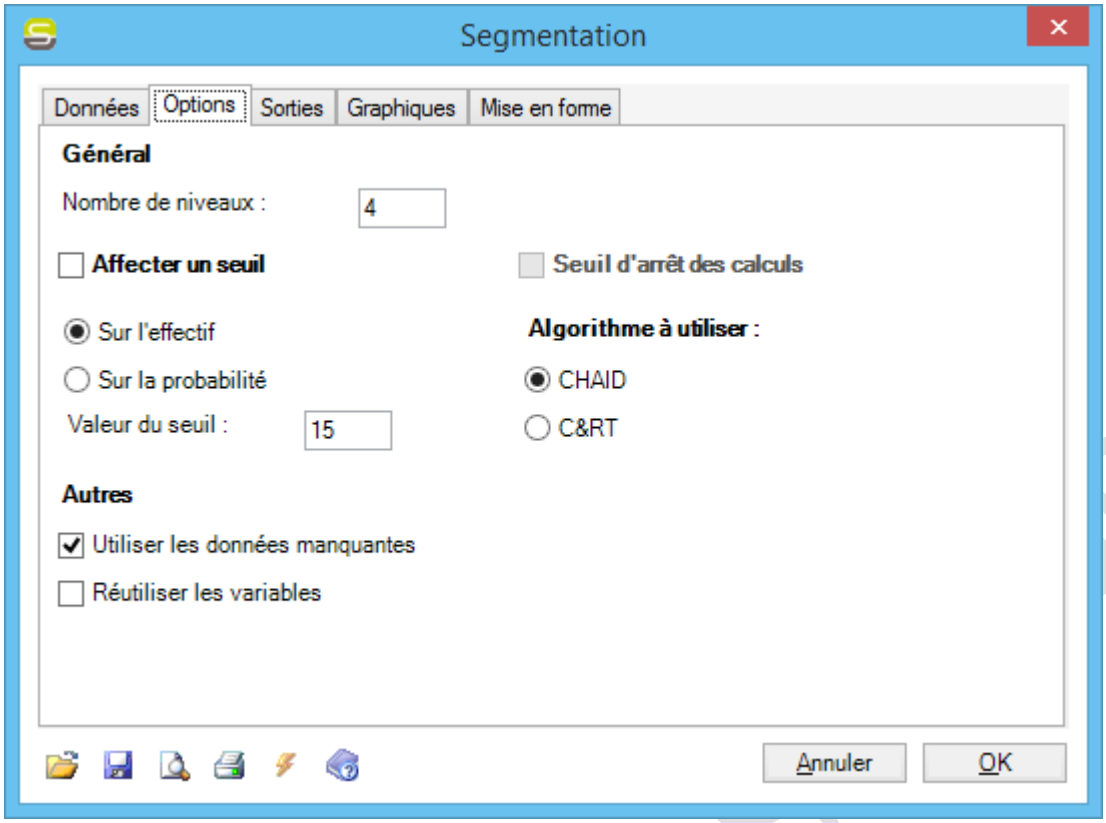



- *Variable qualitative à expliquer :* Sélectionnez la variable qualitative à expliquer. Si vous désirez expliquer une variable quantitative, transformez-la en classes.

  Pour cette variable à expliquer, le nombre de modalités doit être le plus petit possible. L'idéal est 2 modalités. Si vous en avez davantage, l'interprétation des nœuds est plus difficile. Si vous cliquez sur « *Utiliser les manquants* », la modalité 'vide' ou 'espace' est ajoutée aux autres. Cette option est intéressante lorsque les non-réponses ont une signification dans votre étude.

- *Variable(s) explicative(s) qualitative(s) :* sélectionnez dans la/les variable(s) explicative(s) qualitative(s) (ou nominale(s)). Vérifiez que le nombre de modalités de vos variables explicatives ne soit pas trop nombreux. Utilisez dans ce cas le module de regroupement des modalités.
- *Variable(s) explicative(s) quantitative(s) :* Vous sélectionnez dans cette liste les variables explicatives quantitatives. Toutes ces variables sont transformées en classes.
- *Poids :* cochez cette option pour pondérer vos observations, puis sélectionnez la variable contenant les poids.

*Onglet « Options »*



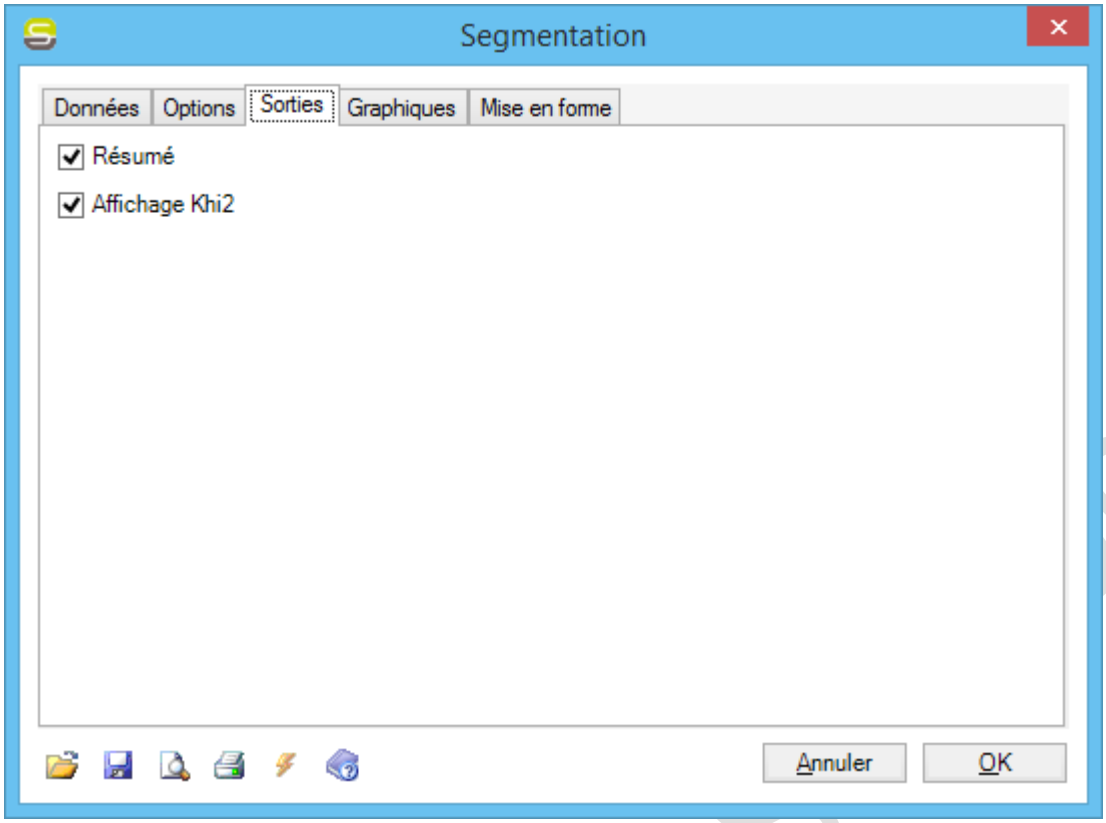

- *Nombre de Niveaux :* Le nombre de niveaux est fixé par défaut à 4. Plus vous avez une taille importante d'observations et plus vous pouvez augmenter le nombre de niveaux de votre arbre. Une fois l'arbre construit, vous pouvez le modifier en utilisant les options suivantes : « Développer un niveau », « Imposer une variable », « Supprimer une séparation ».
- *Affecter un seuil sur effectif :* dans certains cas, vous pouvez obtenir des feuilles avec un nombre très faible d'observations (1 à 5 par exemple). Ces divisions ne sont pas très intéressantes. Pour éviter de développer de telle branche, vous pouvez définir un seuil en dessous duquel la séparation ne se fait plus.
- *Affecter un seuil sur l'effectif / sur la probabilité :* lorsque la probabilité associée à une séparation d'un nœud est supérieure à 0,05, on peut considérer que le hasard peut avoir joué un rôle dans les résultats. Il faut dans ce cas être prudent dans l'interprétation de cette branche. Pour simplifier l'arbre, StatBox ne divise plus le nœud si cette probabilité dépasse un seuil.

  La valeur du seuil dépend de l'option choisie : s'il s'agit de l'effectif, tapez un seuil de 20 ou 30 par exemple, s'il s'agit d'une probabilité, tapez un seuil de 0.05 par exemple.

  Dans le cas de l'algorithme de CART, au lieu d'être une probabilité, c'est le niveau d'impureté qui est pris en compte.
- *Seuil d'arrêt des calculs :* cette option ne peut être activée que si le seuil correspond à une fréquence. Si cette option n'est pas cochée, au lieu d'arrêter le développement de l'arbre, StatBox va choisir la prochaine variable dans la liste décroissante des variables en fonction de leur Khi² ou de l'indice d'impureté.
- *Utiliser les données manquantes :* cochez cette option pour inclure les non-réponses ou les données manquantes dans l'analyse. Une donnée manquante correspond à une cellule vide dans la feuille des données. La donnée manquante est ajoutée à la liste des modalités possibles. Dans les graphiques elle est représentée par < >.
- *Réutiliser les variables :* Permet d'utiliser la même variable dans différents niveaux de l'arbre. Par défaut, lorsqu'une variable est utilisée dans un nœud, elle ne peut pas être à nouveau utilisée.

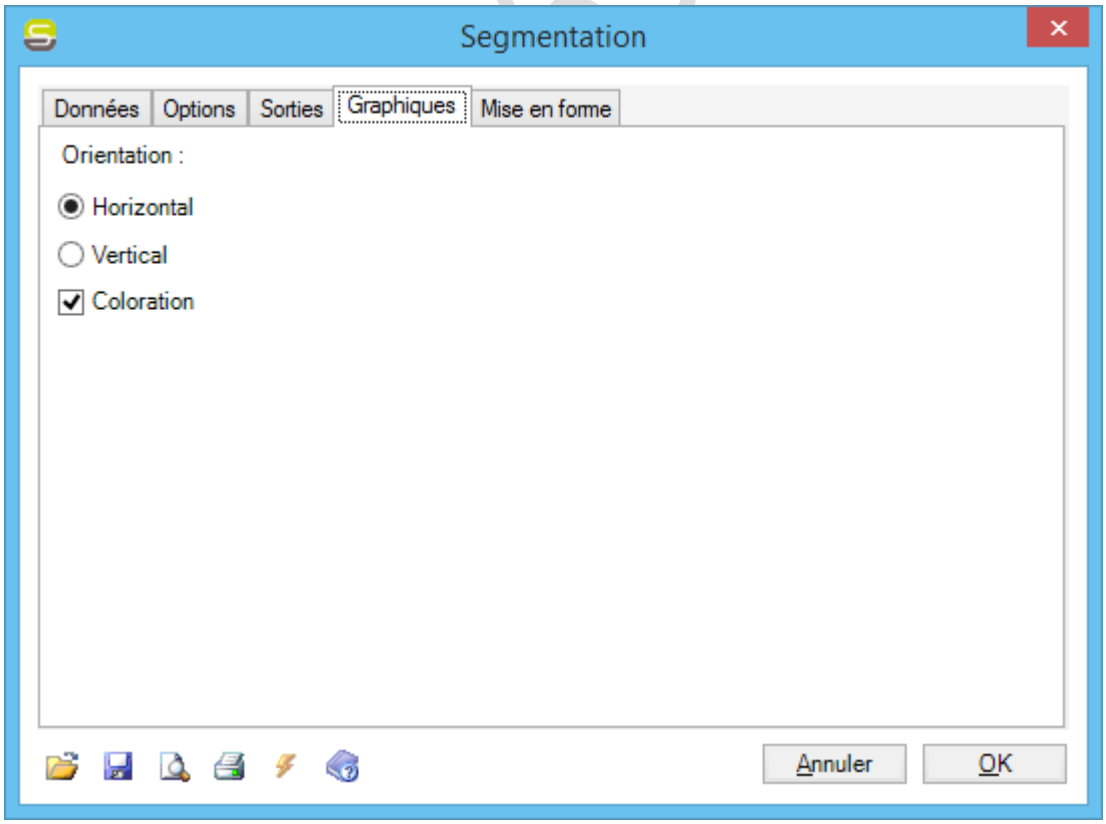
Si vous avez défini par exemple 4 niveaux et que vous cliquez dans l'arbre sur le nœud racine puis sur Développer d'un niveau, vous obtenez le même résultat que si vous aviez défini 5 niveaux.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Affichage Khi² :* cochez cette option pour que le Khi² ainsi que la probabilité associée apparaissent dans l'arbre avec l'algorithme CHAID, avec CART, c'est l'indice de baisse de l'impureté.

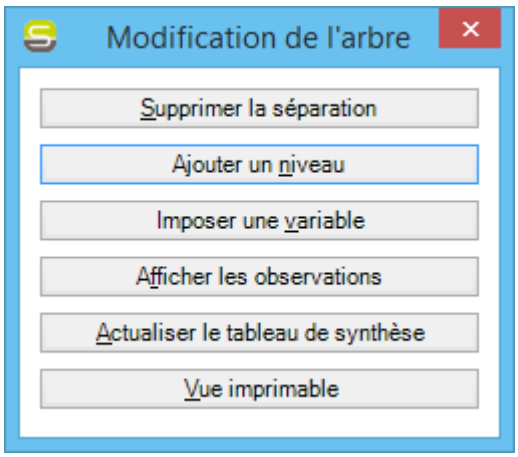
*Onglet « Graphiques »*



- *« Horizontal » / « Vertical » :* Cette option permet de changer la présentation de l'arbre. Soit le nœud racine se trouve à gauche du graphique avec l'option Horizontal, soit ce nœud se trouve en haut du graphique avec l'option Vertical.
- *Coloration :* cochez cette option pour associer chacune des modalités des variables explicatives à une couleur pour faciliter la lecture de l'arbre.

### *Modification de l'arbre en cours*

Une fois la procédure terminée, une boite de dialogue apparaît proposant plusieurs fonctions afin d'optimiser la structure ou l'affichage de l'arbre. Ce sont :




- *Supprimer la séparation :* Cette option sert à enlever une branche de votre arbre. Si une branche ne vous paraît pas pertinente, vous pouvez l'enlever pour n'imprimer que la partie intéressante de l'arbre. Sélectionnez préalablement le nœud à supprimer.
- *Ajouter un niveau :* cette option sert à ajouter une branche dans l'arbre afin de développer l'arbre selon une nouvelle variable.
- *Imposer une variable :* Cette option est très utile pour développer l'arbre en fonction de vos préférences. Supposons qu'au niveau d'un nœud, le logiciel a trouvé que la variable A est la plus pertinente pour séparer la population du nœud. En utilisant sur ce nœud l'option « Imposer une variable », StatBox affiche la liste de toutes les variables possibles en ordre décroissant de pertinence. Si A obtient un Khi-deux de 12 et que la variable D arrive en second avec un Khi-deux de 10.5, on peut dire que cette seconde variable joue également un rôle important. Pour vous cette dernière variable peut être plus facile à utiliser d'un point de vue opérationnel. Dans ce cas sélectionnez-la, ce n'est peut-être pas la meilleure séparation mais la plus adaptée à vos possibilités d'action.
- *Afficher les observations :* Cette option permet de lister tous les observations appartenant à un ou plusieurs nœuds ou feuilles. Elle est utile pour effectuer des croisements entre variables pour mieux analyser cette ou ces sous-populations.
- *Actualiser le tableau de synthèse :* Cette option permet de mettre à jour le tableau de synthèse après une modification de l'arbre.
- *Vue imprimable :* Cette fonction permet l'affichage de l'arbre dans la fenêtre de Prévisualisation d'Excel et éventuellement de procéder à l'impression.

Remarques :

- Une fois les nœuds affichés, vous pouvez les déplacer mais vous ne pouvez pas les renommer.
- Chaque nœud est affecté à une des modalités de la variable à expliquer. On peut suivre dans le graphique cette affectation grâce à la couleur du nœud.
- Si l'arbre est très grand, utilisez le zoom d'Excel pour afficher l'ensemble du graphique. Vous pouvez dans ce cas cliquer sur un nœud et à nouveau changer le zoom pour voir en détail cette partie de l'arbre.

## Exemple


## Références


Data Mining, techniques appliquées au marketing, à la vente et aux services clients

Michael J.A. Berry, Gordon Linoff, InterEditions, Masson, Paris,1997

Data Mining with neural networks, Solving Business Problems- Application development to decision support, Joseph P. Bigus, McGraw-Hill,1996

Data Mining, Pieter Adriaans Dolf Zantinge, Addison-Wesley, 1996

Discovering Data Mining from concept to implementation, Cabena Hadjinian, Stadler, Verhees, Zanasi, Prentice Hall PTR 1998

Analyse discriminante sur variables qualitatives, Gilles Celeux, Jean-Pierre Nakache, Polytechnica, Paris 1994


# ANOVA (MODÈLE LINÉAIRE GÉNÉRAL)

## Description

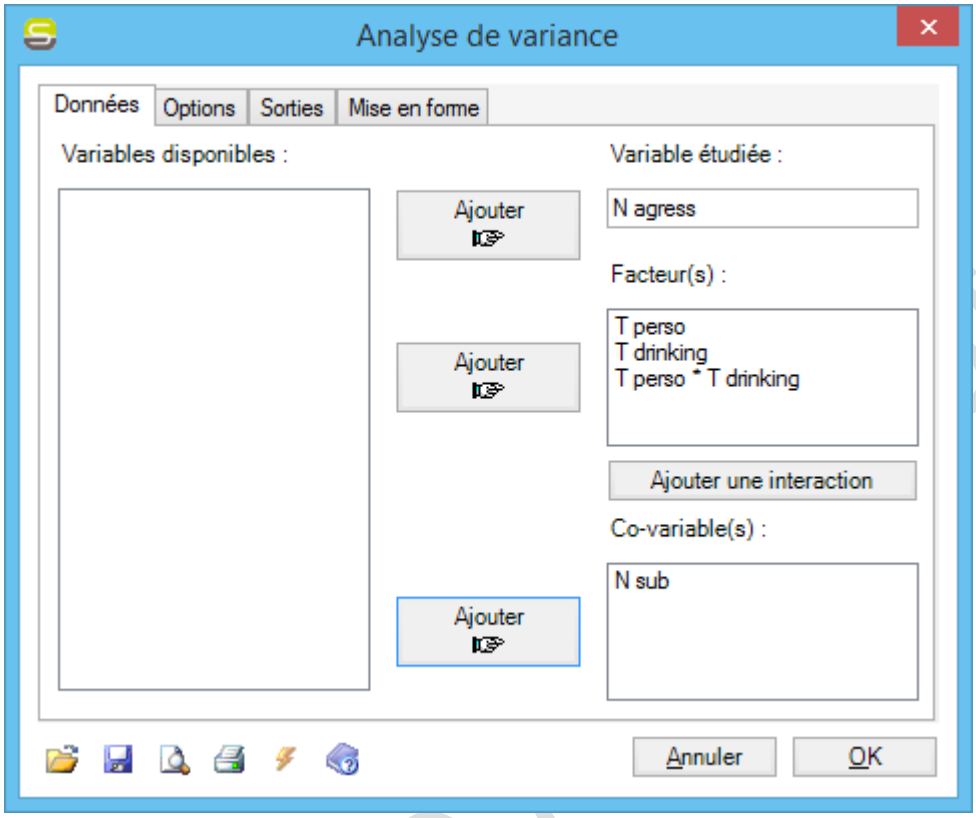
La variable à analyser est continue et la ou les variables explicatives sont nominales. Il possible d'ajouter des variables explicatives continues : les co-variables.

Ce module de StatBox permet de traiter un grand nombre de plans d'expériences :

- de 1 à n facteurs
- les différentes interactions d'ordre 2 et 3
- mesures répétées
- les plans déséquilibrés comportant un nombre différents d'observations par cellule

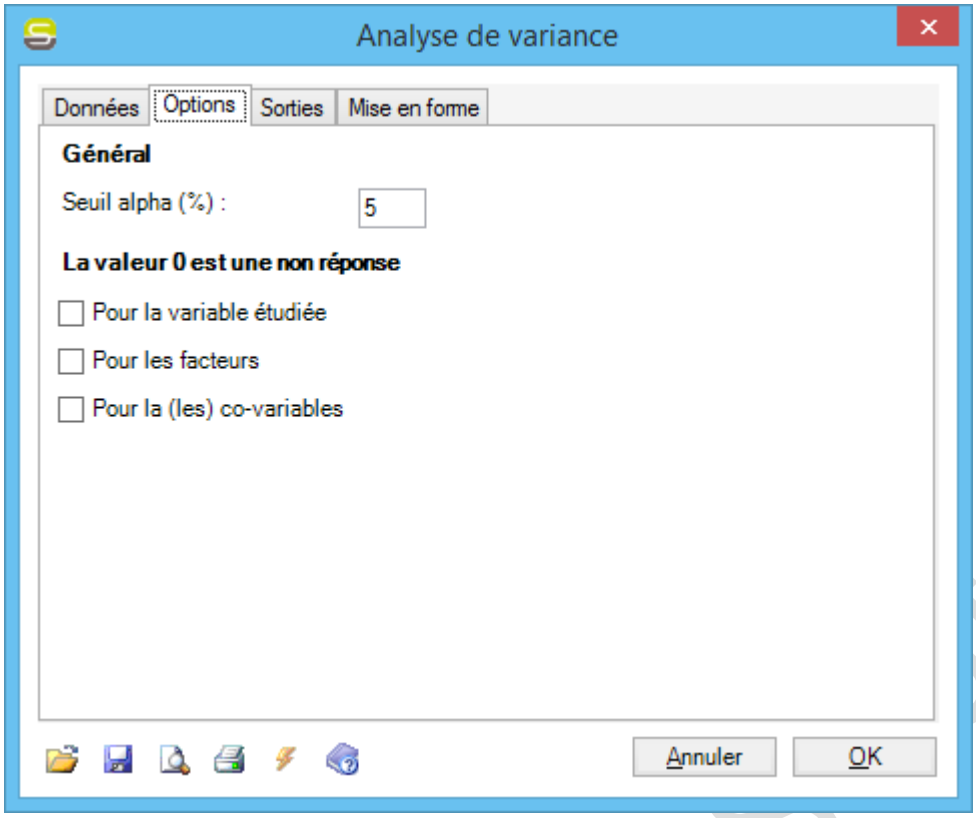
## Mise en œuvre

### *Onglet « Données »*



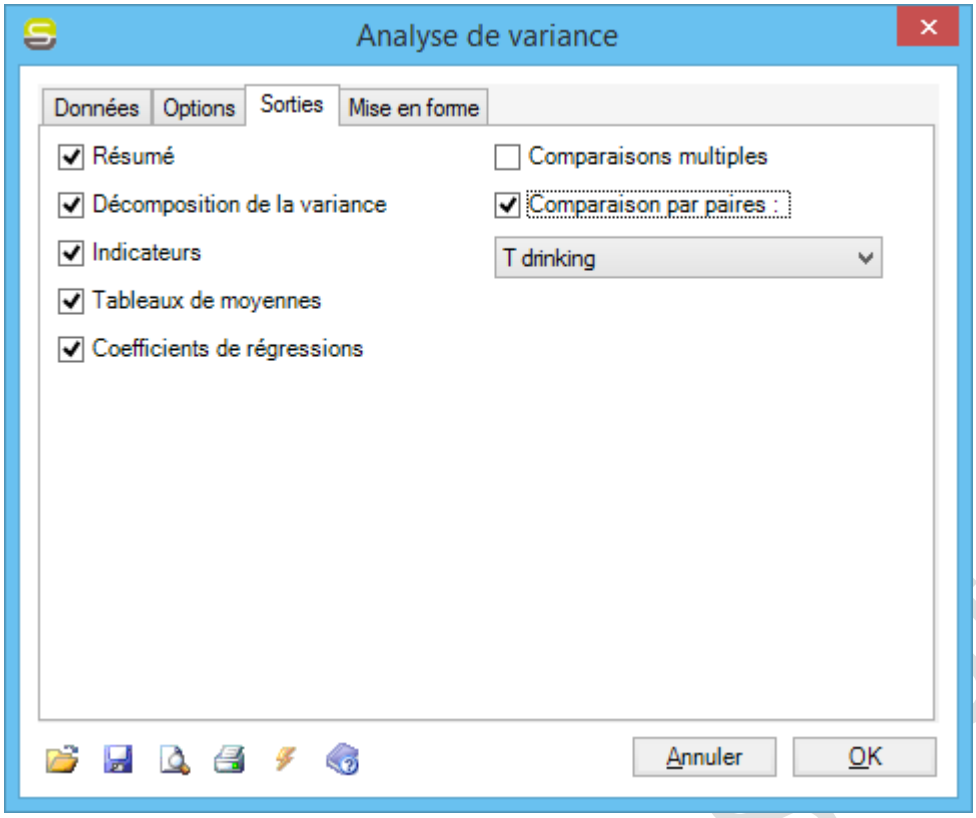

- *Variable étudiée :* sélectionnez la variable à étudier.
- *Facteur(s) :* sélectionnez la/les variable(s) qualitatives correspondant aux facteurs étudiés.
- Les facteurs peuvent comporter soit des codes (1 pour le premier niveau, 2 pour le second, etc), ou des noms de niveaux en clairs. Dans le cas de code, ne commencez pas par 0 votre numérotation. Ainsi, si vous avez 2 niveaux, ne les identifiez pas par 0 et 1, mais par 1 et 2.
- Pour ajouter une interaction, il suffit de sélectionner dans la liste des facteurs sélectionnés 2 facteurs et de cliquer sur le bouton **Ajouter une interaction**. Pour ajouter une interaction d'ordre 3, il suffit de sélectionner une interaction d'ordre 2 et un facteur.
- *Co-variable(s) :* sélectionnez la/les variable(s) numérique(s) explicative(s).

*Onglet « Options »*




- *Seuil alpha (%) :* entrez la valeur du risque de première espèce pour les tests de comparaison de moyenne.
- *Pour la variable étudiée :* cochez cette option si les observations ayant une valeur nulle pour la variable étudiée doivent être ignorées.
- *Pour les facteurs :* cochez cette option si les observations ayant une valeur nulle pour la/les facteur(s) doivent être ignorées.
- *Pour la (les) co-variables :* cochez cette option si les observations ayant une valeur nulle pour la variable étudiée doivent être ignorées.

*Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Décomposition de la variance :* affiche la table de décomposition de la variance pour les facteurs étudiés et les niveaux d'interactions.
- *Indicateurs :* affiche des indicateurs de base sur la variable étudiée (moyenne, écart-type, % de variation)
- *Tableaux de moyennes :* affiche les tables de moyennes pour les facteurs étudiés et les interactions.
- *Coefficients de régressions :* affiche les tables des coefficients de régression pour les co-variables.
- *Comparaisons multiples :* effectue des comparaisons multiples de moyennes
- *Comparaison par paires :* effectue des comparaisons par paires. Sélectionnez alors le facteur ou l'interaction pour lequel les moyennes doivent être comparées.

## Exemple

Le classeur Data.xls contient des données qui ont été proposées par S.A. Glantz, B K Slinker dans leur ouvrage Primer of Applied regression & analysis of variance. Ces exemples y ont été traités à l'aide de la procédure General Linear Model de SAS.

Exemple feuille **anova1** : Analyse de variance à 2 facteurs effectifs équilibrés, avec interaction dans Primer of Applied regression & analysis of variance, S.A. Glantz, B K Slinker, page 328

| | ddl | S.C.E | CM | F | Proba |
|---|---|---|---|---|---|
| F1 | 1 | 2838,811 | 2838,811 | 22,640 | 0,000 |
| F2 | 1 | 1782,045 | 1782,045 | 14,212 | 0,001 |
| S F1 * S F2 | 1 | 108,045 | 108,045 | 0,862 | 0,361 |
| Var.résiduelle | 28 | 3510,908 | 125,390 | | |
| Total | 31 | 8239,809 | | | |

Exemple feuille **anova4** : Analyse de variance à 2 facteurs, observations manquantes et appariées, avec interaction dans Primer of Applied regression & analysis of variance, S.A. Glantz, B K Slinker, page 488

La colonne Sub (subjects) devient un facteur.

| | ddl | S.C.E | CM | F | Proba |
|---|---|---|---|---|---|
| SUB | 7 | 6,917 | 0,988 | 12,776 | 0,000 |
| GUM | 1 | 0,947 | 0,947 | 12,238 | 0,004 |
| TIME | 2 | 13,458 | 6,729 | 87,001 | 0,000 |
| S SUB * T GUM | 7 | 0,116 | 0,017 | 0,213 | 0,975 |
| S SUB * T TIME | 14 | 1,997 | 0,143 | 1,845 | 0,147 |
| T GUM * T TIME | 2 | 2,402 | 1,201 | 15,528 | 0,000 |
| Var.résiduelle | 12 | 0,928 | 0,077 | | |
| Total | 45 | 29,199 | | | |

Exemple feuille **anova6** : Analyse de variance à 1 facteurs, avec la covariable Apolipoprotein dans Primer of Applied regression & analysis of variance, S.A. Glantz, B K Slinker, page 488

| | ddl | S.C.E | CM | F | Proba |
|---|---|---|---|---|---|
| pregnancy | 1 | 2695,235 | 2695,235 | 22,514 | 0,000 |
| Apolipoprotein | 1 | 1084,535 | 1084,535 | 9,059 | 0,008 |
| Var.résiduelle | 17 | 2035,114 | 119,713 | | |
| Total | 19 | 7356,610 | | | |

# TESTS PARAMÉTRIQUES

## COMPARAISON DES PARAMÈTRES DE 2 ÉCHANTILLONS

Utilisez ce module de tests paramétriques lorsque vous êtes en présence de 2 échantillons, pour déterminer si les échantillons proviennent de populations :

- qui ont même variance (test F de Fisher),
- dont les espérances (moyennes théoriques) diffèrent d'une quantité *D* donnée (test *t* de Student, test *z*).

**Remarque :** les échantillons peuvent être indépendants pour tous les tests, et éventuellement appariés dans le cas des tests portant sur les moyennes. En revanche, le test *F* de Fisher requiert des échantillons indépendants.

### Description du test F de Fisher

Le *F* de Fisher est le rapport des estimations des variances des populations 1 et 2. StatBox divise toujours la plus grande variance $\sigma^2_{max}$ par la plus petite $\sigma^2_{min}$. La valeur de la statistique est testée par rapport à la loi de Fisher de degrés de libertés $n_{max}-1$ et $n_{min}-1$, avec $n_{max}$ la taille de l'échantillon ayant la plus grande variance et $n_{min}$ la taille de l'échantillon ayant la plus petite variance. Le test effectué est unilatéral à droite, les hypothèses nulle ($H_0$) et alternative ($H_1$) étant les suivantes :

$$H_0 : \sigma^2_{max} / \sigma^2_{min} = 1$$
$$H_1 : \sigma^2_{max} / \sigma^2_{min} > 1$$

### Description du test t de Student pour échantillons indépendants

Les échantillons 1 et 2 sont prélevés respectivement dans deux populations d'espérances $\mu_1$ et $\mu_2$. Le test bilatéral correspond au test de la différence entre $\mu_1$ - $\mu_2$ et *D*, et les hypothèses nulle ($H_0$) et alternative ($H_1$) sont les suivantes :

$$H_0 : \mu_1 - \mu_2 = D$$
$$H_1 : \mu_1 - \mu_2 \neq D$$

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche (ou inférieur) et le test unilatéral à droite (ou supérieur).

Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

$$H_0 : \mu_1 - \mu_2 = D$$
$$H_1 : \mu_1 - \mu_2 < D$$

Dans le test unilatéral à droite les hypothèses sont les suivantes :

$$H_0 : \mu_1 - \mu_2 = D$$
$$H_1 : \mu_1 - \mu_2 > D$$

Ce test a été développé en considérant que :

- les deux échantillons sont des échantillons aléatoires tirés de leurs populations respectives, distribuées selon des lois normales de même variance,
- en plus de l'indépendance au sein de chaque échantillon, il y a indépendance mutuelle entre les deux échantillons,
- les données sont quantitatives.

**Remarque** : le test suppose en principe l'égalité des variances théoriques des deux populations. Toutefois, StatBox permet d'effectuer ce test même si l'égalité des variances n'est pas satisfaite, en utilisant une combinaison linéaire de valeurs critiques de *t*.

## Description du test t de Student pour échantillons appariés

Notons $\delta$ l'espérance des différences $d_i = x_{i2} - x_{i1}$, avec $x_{i2}$ la $i^{ème}$ valeur pour l'échantillon 2 et $x_{i1}$ la $i^{ème}$ valeur pour l'échantillon 1. Le test bilatéral correspond au test de la différence entre $\delta$ et $D$, et les hypothèses nulle ($H_0$) et alternative ($H_1$) sont les suivantes :

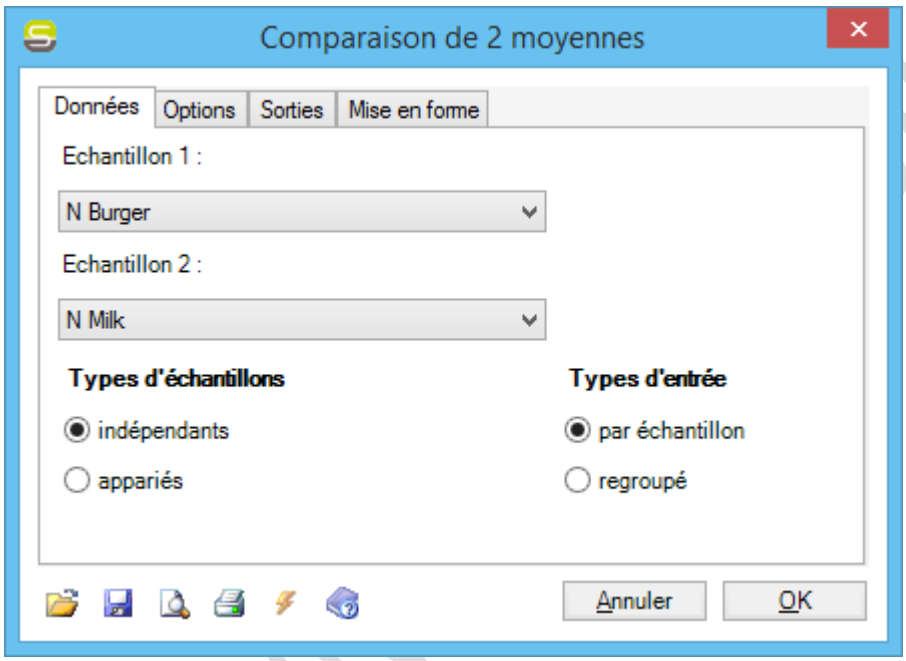
$$H_0 : \delta = D$$
$$H_1 : \delta \neq D$$

Ce test a été développé en considérant que :

- les deux échantillons sont des échantillons aléatoires tirés de leurs populations respectives,
- les deux échantillons sont appariés,
- la différence est distribuée selon une loi normale, ce qui constitue une condition moins restrictive que la normalité des deux populations d'origine,
- les données sont quantitatives.

## Mise en œuvre

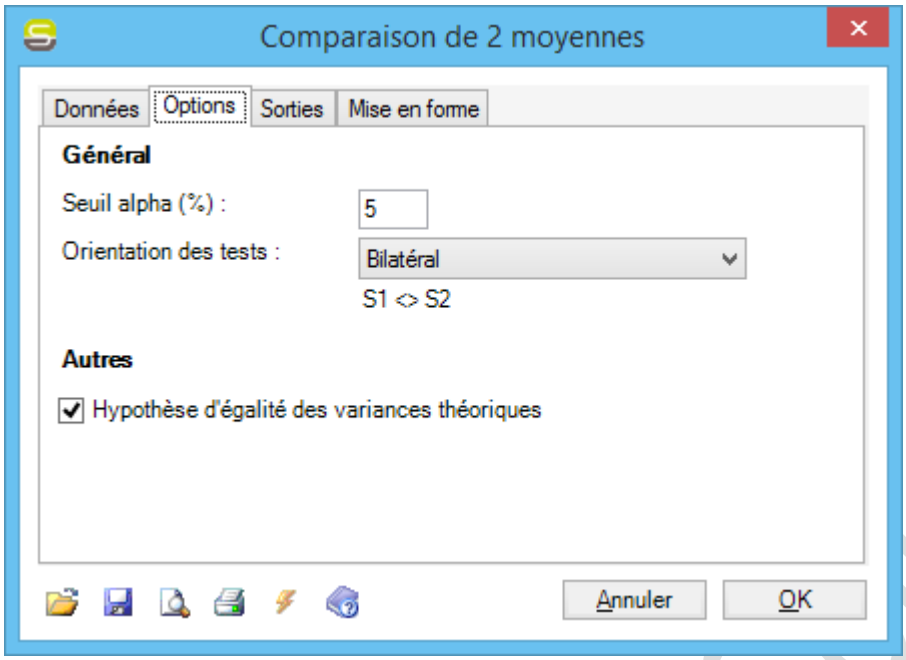
### *Onglet « Données »*




- *« Par échantillon » / « Regroupées » :* si les échantillons figurent dans des colonnes différentes, sélectionnez les échantillons 1 et 2, la taille des colonnes pouvant être différente. Si les données sont regroupées, la variable des données correspond à une colonne de valeurs, l'appartenance aux échantillons étant indiquée par un *descripteur d'échantillon.*
- Pour des données par échantillon
- *Échantillon 1 :* sélectionnez la variable correspondant au premier échantillon. Les valeurs manquantes ne sont pas autorisées.
- *Échantillon 2 :* sélectionnez la variable correspondant au deuxième échantillon. Les valeurs manquantes ne sont pas autorisées.
- Pour des données regroupées
- *Données :* dans le cas des données regroupées, sélectionnez la variable correspondant aux valeurs des deux échantillons. Les valeurs manquantes ne sont pas autorisées.
- *Descripteur d'échantillon :* dans le cas des données regroupées, sélectionnez la variable correspondant à une variable qualitative indiquant l'échantillon d'appartenance de chaque valeur. Les valeurs manquantes ne sont pas autorisées.

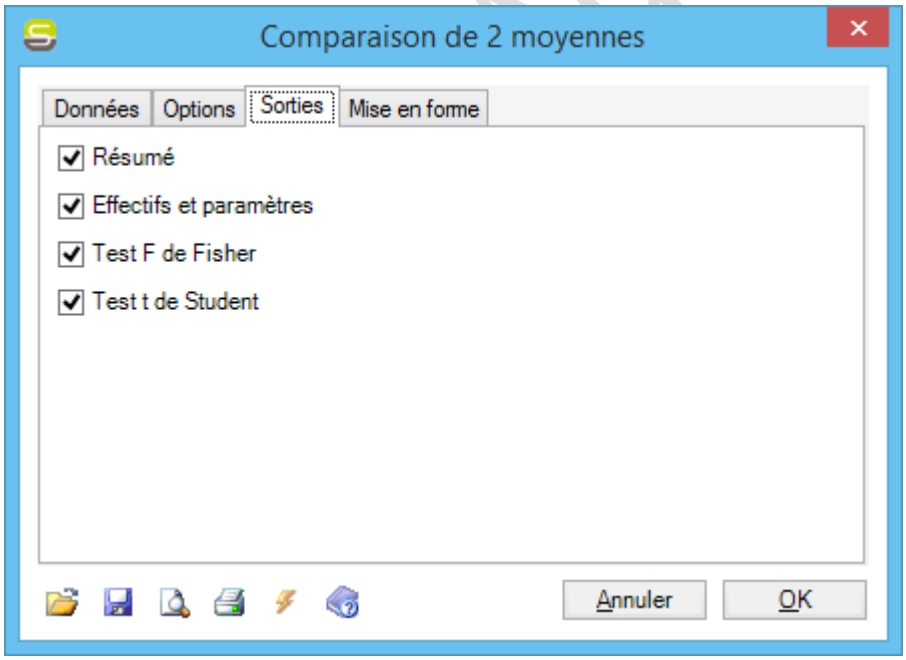
- *« Indépendants » / «Appariés » :* choisissez la nature de la relation entre les deux échantillons. Les échantillons appariés peuvent correspondre par exemple à deux traitements portant sur un même ensemble de sujets expérimentaux.

### *Onglet « Options »*




- *Seuil alpha (%) :* entrez la valeur du risque de première espèce des tests.
- *Orientation du test :* choisissez le type de test à réaliser, bilatéral, unilatéral à gauche, ou unilatéral à droite
- *Hypothèse d'égalité des variances théoriques :* cochez cette case pour faire l'hypothèse que les variances théoriques sont égales.

### *Onglet « Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- Effectifs et paramètres :
- *Test F de Fisher :* effectue un test d'égalité des variances des populations, utilisant la distribution de Fisher. Cette case est automatiquement décochée dans le cas des échantillons appariés.

- *Test t de Student :* effectue un test sur les moyennes des populations, utilisant la distribution du *t* de Student. Une seconde boîte de dialogue spécifique permet de choisir l'hypothèse testée.

### Références

**Dagnelie P. (1986)**. Théorie et méthodes statistiques. Vol. 2. Les Presses Agronomiques de Gembloux, Gembloux, pp. 16-17, 21-29, 35-39, 50-53.

**Frontier S. (1981)**. Méthode statistique. Masson, Paris, pp. 119-127, 189-190.

**Manoukian E.B. (1986)**. Guide de statistique appliquée. Hermann, Paris, pp. 125-132, 135-136.

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 184-190, 223-227.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 70-72.

## COMPARAISON DE DEUX PROPORTIONS

Utilisez ce module pour comparer deux proportions.

### Description

L'effectif *n* des observations qui vérifient une certaine propriété, parmi un total de *N* observations examinés, suit une loi binomiale de paramètres *N* (nombre d'essais) et *p* (probabilité de succès). Lorsque *N* est assez grand, et que *p* n'est ni trop proche de 0, ni trop proche de 1, la loi binomiale peut être approximée par une loi normale d'espérance *Np* et de variance *Np*(1-*p*). La proportion *n*/*N* suit approximativement une loi normale de moyenne *p* et de variance *p*(1-*p*)/*N*. StatBox réalise un test *z* adapté au cas de deux proportions en utilisant l'approximation de la loi binomiale par la loi normale.

Le test bilatéral correspond au test de la différence entre $p_1 - p_2$ et $D$, et les hypothèses nulle ($H_0$) et alternative ($H_1$) sont les suivantes :

$$H_0 : p_1 - p_2 = D$$
$$H_1 : p_1 - p_2 \neq D$$

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche (ou inférieur) et le test unilatéral à droite (ou supérieur). Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

$$H_0 : p_1 - p_2 = D$$
$$H_1 : p_1 - p_2 < D$$

Dans le test unilatéral à droite les hypothèses sont les suivantes :

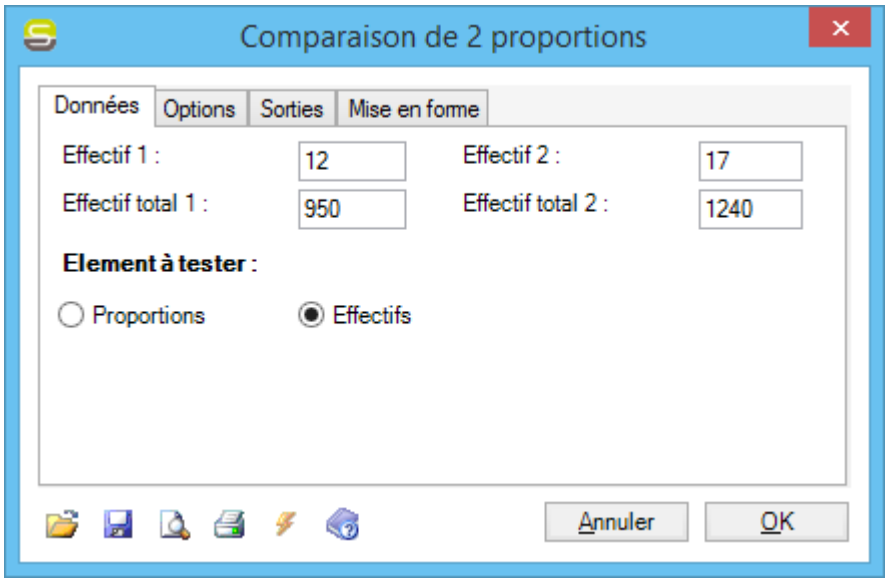
$$H_0 : p_1 - p_2 = D$$
$$H_1 : p_1 - p_2 > D$$

Ce test a été développé en considérant que :

- les observations sont mutuellement indépendantes,
- la probabilité p de posséder la propriété considérée est la même pour toutes les observations,
- les effectifs sont assez grands, et p n'est ni trop proche de 0, ni trop proche de 1.

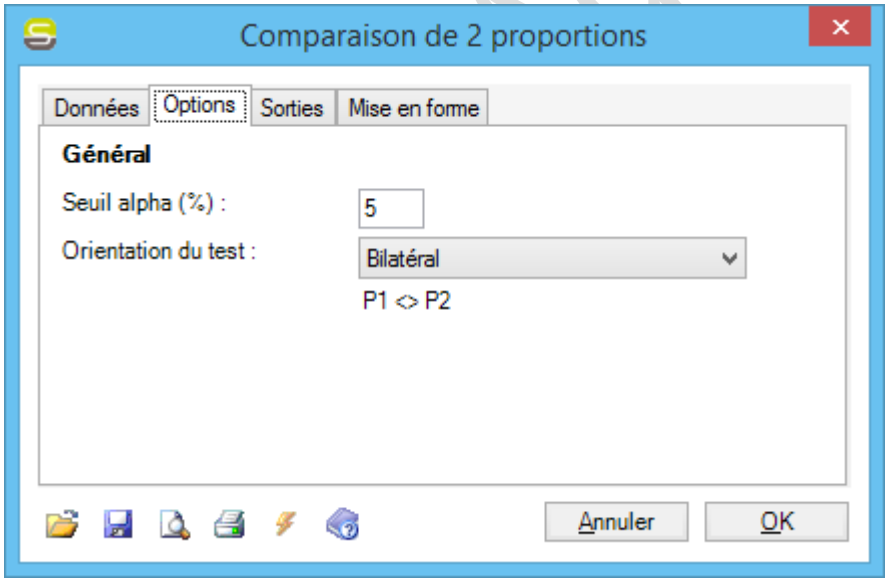
## Mise en œuvre

### *Onglet « Données »*



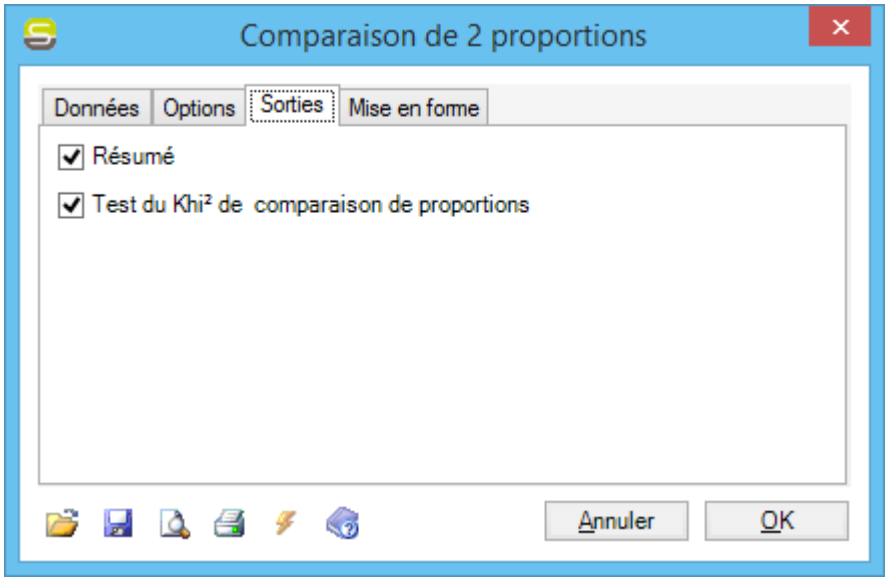

- *« Proportions » / « Effectifs » :* choisissez la nature des données, soit des proportions (valeurs comprises entre 0 et 1), soit des effectifs (valeurs inférieures ou égales aux effectifs totaux respectifs).
- *Proportion 1 / Effectif 1 :* entrez la proportion ou l'effectif des observations possédant la propriété $C_1$ dans le groupe 1.
- *Effectif total 1 :* entrez l'effectif total du groupe 1.
- *Proportion 2 / Effectif 2 :* entrez la proportion ou l'effectif des observations possédant la propriété $C_2$ dans le groupe 2.
- *Effectif total 2 :* entrez l'effectif total du groupe 2.

### *Onglet « Options »*




- *Seuil alpha (%) :* entrez la valeur du risque de première espèce du test du Khi².
- *Orientation du test :* choisissez le type de test à réaliser, bilatéral, unilatéral à gauche ou unilatéral à droite

*Onglet Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Test du Khi² de comparaison de proportions :* affiche un test du Khi² de comparaison de proportions pour les proportions/effectifs entrés.

## Références

**Dagnelie P. (1986)**. Théorie et méthodes statistiques. Vol. 2. Les Presses Agronomiques de Gembloux, Gembloux, pp. 90-96.

**Fleiss J.L. (1981)**. Statistical methods for rates and proportions. John Wiley & Sons, New York.

**Frontier S. (1981)**. Méthode statistique. Masson, Paris, pp. 128-134.

**Manoukian E.B. (1986)**. Guide de statistique appliquée. Hermann, Paris, pp. 133-134.

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, pp. 686-687.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, p. 70.

# TESTS NON PARAMÉTRIQUES

## COMPARAISON DE 2 ÉCHANTILLONS INDÉPENDANTS

Utilisez ce module de tests non paramétriques lorsque vous êtes en présence de 2 échantillons indépendants, afin de déterminer si les échantillons proviennent de la même population ou de 2 populations différentes. StatBox propose deux tests :

- le test de Kolmogorov-Smirnov,
- le test de Mann-Whitney.

**Remarque** : l'utilisation du test de Mann-Whitney constitue une alternative non paramétrique au test *t* de Student (équivalent à l'analyse de variance à 1 facteur dans le cas de deux échantillons). Comme pour le test *t* de Student, les échantillons peuvent être de tailles différentes.

### Description du test de Kolmogorov-Smirnov

L'objectif du test de Kolmogorov-Smirnov est de déterminer si les fonctions de répartition des populations à l'origine des échantillons sont différentes. StatBox réalise un test bilatéral.

Soient $F(x)$ et $G(x)$ les fonctions de répartition des deux populations d'où sont tirés les deux échantillons. Le test bilatéral correspond au test de la différence entre les deux populations, et les hypothèses nulle ($H_0$) et alternative ($H_1$) sont les suivantes :

$$H_0 : F(x) = G(x) \text{ pour tout } x$$
$$H_1 : F(x) \neq G(x) \text{ pour au moins une valeur de } x$$

### Description du test de Mann-Whitney

L'objectif du test de Mann-Whitney est de déterminer si les échantillons proviennent d'une même population ou de deux populations différentes. StatBox peut réaliser un test bilatéral ou unilatéral.

Soient deux populations *A* et *B* dont sont prélevés les échantillons comportant des valeurs *a* et *b*. Le test bilatéral correspond au test de la différence entre *A* et *B*, et les hypothèses nulle ($H_0$) et alternative ($H_1$) sont les suivantes :

$$H_0 : P(a < b) = 1/2$$
$$H_1 : P(a < b) \neq 1/2$$

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche (ou inférieur) et le test unilatéral à droite (ou supérieur). Dans le test unilatéral à gauche, l'hypothèse alternative indique que la population *A* admet en général des valeurs inférieures à celles de la population *B* :

$$H_0 : P(a < b) \leq 1/2$$
$$H_1 : P(a < b) > 1/2$$

Dans le test unilatéral à droite, l'hypothèse alternative indique que la population *A* admet en général des valeurs supérieures à celles de la population *B* :

$$H_0 : P(a < b) \geq 1/2$$
$$H_1 : P(a < b) < 1/2$$

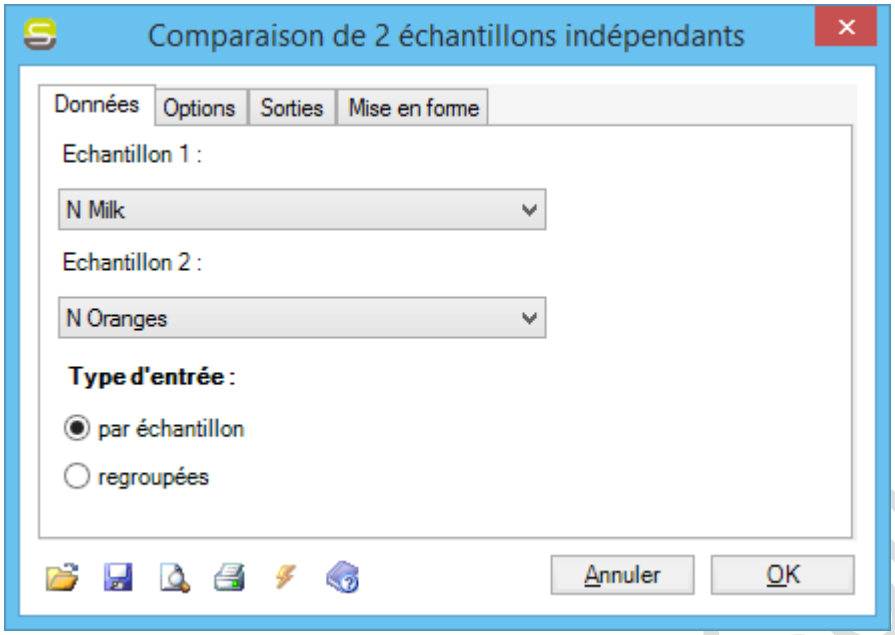
Ce test a été développé en considérant que :

- les deux échantillons sont des échantillons aléatoires tirés de leurs populations respectives,
- en plus de l'indépendance au sein de chaque échantillon, il y a indépendance mutuelle entre les deux échantillons,
- les données sont au moins des données ordinales.

**Remarque** : la statistique de Mann-Whitney est reliée à la statistique de Wilcoxon, de sorte que le test de Wilcoxon (non signé) est équivalent au test de Mann-Whitney.

## Mise en œuvre

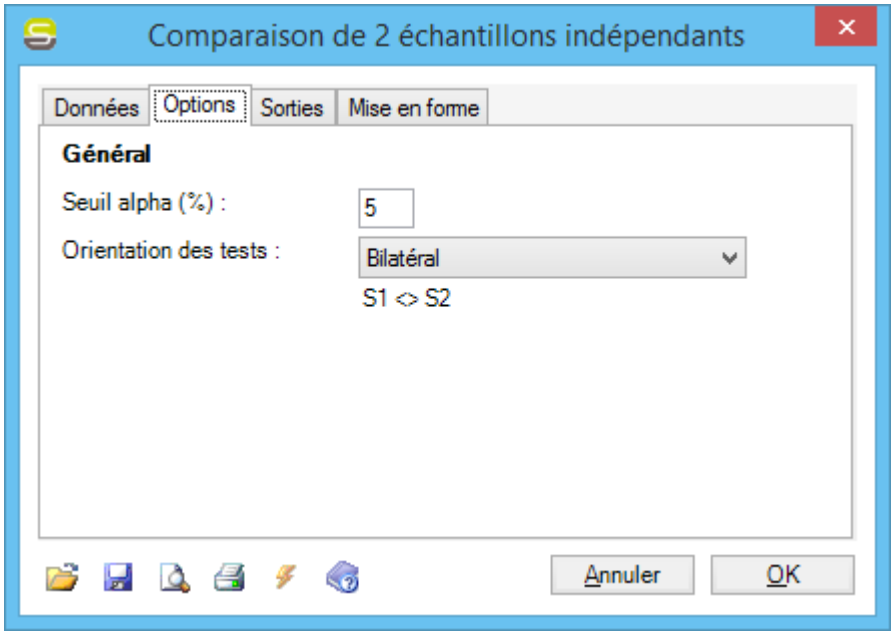
*Onglet « Données »*




- *« Par échantillon » / « Regroupées » :* si les échantillons figurent dans des colonnes différentes, sélectionnez l'option « Par échantillon ». Si les données sont « Regroupées », la variable des données correspond à une colonne de valeurs, l'appartenance aux échantillons étant indiquée par un *descripteur d'échantillon.*
- Pour des données par échantillons
- *Échantillon 1 :* sélectionnez la variable correspondant au premier échantillon. Les valeurs manquantes ne sont pas autorisées.
- *Échantillon 2 :* sélectionnez la variable correspondant au deuxième échantillon. Les valeurs manquantes ne sont pas autorisées.
- Pour des données regroupées
- *Données :* dans le cas des données regroupées, sélectionnez la variable correspondant aux valeurs des deux échantillons. Les valeurs manquantes ne sont pas autorisées.
- *Descripteur d'échantillon :* dans le cas des données regroupées, sélectionnez la variable correspondant à une variable qualitative indiquant l'échantillon d'appartenance de chaque valeur. Les valeurs manquantes ne sont pas autorisées.

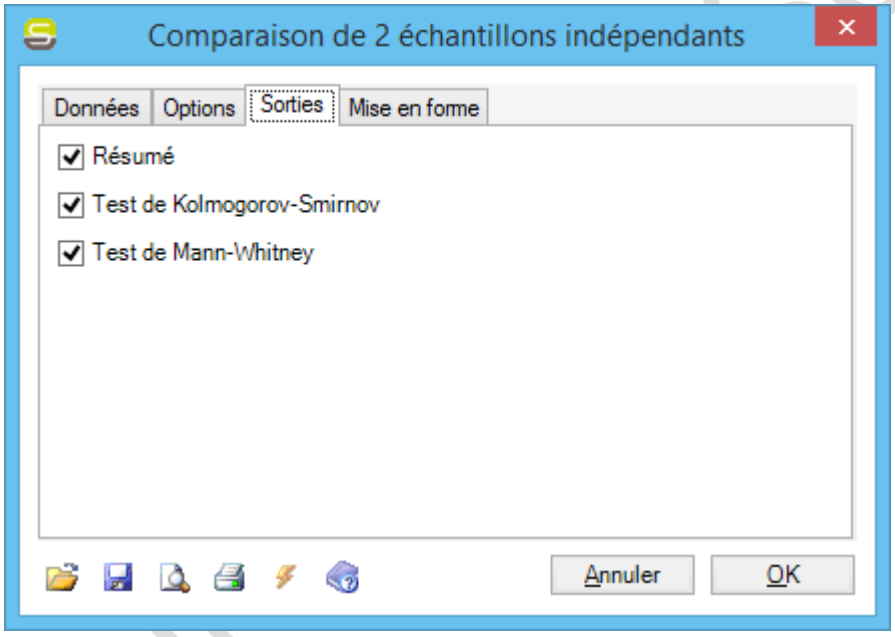
**Remarque** : dans le cas de l'option « Par échantillon » la taille des colonnes peut être différente.

*Onglet « Options »*




- *Seuil alpha (%) :* entrez la valeur du risque de première espèce du test.
- *Orientation du test :* choisissez le type de test à réaliser, bilatéral, unilatéral à gauche ou unilatéral à droite

*Onglet Sorties »*




- *Résumé* : cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- Test de Kolmogorov-Smirnov :
- Test de Mann-Withney :

## Références

**Dagnelie P. (1986)**. Théorie et méthodes statistiques. Vol. 2. Les Presses Agronomiques de Gembloux, Gembloux, pp. 381-385.

**Lehmann E.L. & H.J.M. D'Abrera (1975)**. Nonparametrics. Statistical methods based on ranks. Holden-Day, San Francisco, pp. 5-31.

**Manoukian E.B. (1986)**. Guide de statistique appliquée. Hermann, Paris, pp. 139-140, 146.

**Siegel S. (1956)**. Nonparametric statistics for the behavioral sciences. McGraw-Hill Kogakusha, Tokyo, Japan, pp. 116-136.

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 427-439.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 216-220.

## COMPARAISON DE 2 ÉCHANTILLONS APPARIÉS

Utilisez ce module de tests non paramétriques lorsque vous êtes en présence de 2 échantillons appariés, afin de déterminer si les échantillons proviennent de la même population ou de 2 populations différentes. StatBox propose deux tests:

- le test de Wilcoxon signé,
- le test du signe.

**Remarques** : l'utilisation de ces tests constitue une alternative non paramétrique au test *t* de Student pour données appariées. Les échantillons étant appariés, ils doivent nécessairement comporter le même nombre d'observations.

### Description du test de Wilcoxon signé

L'objectif du test de Wilcoxon signé est de déterminer si les échantillons proviennent d'une même population ou de deux populations différentes. StatBox peut réaliser un test bilatéral ou unilatéral.

Soient deux populations *A* et *B* dont sont prélevés les échantillons comportant des valeurs *a* et *b*. Notons *d* la médiane des différences $d_{b-a} = b-a$ pour tous les couples de données appariées. Le test bilatéral correspond au test de la différence entre *A* et *B*, et les hypothèses nulle ($H_0$) et alternative ($H_1$) sont les suivantes :

$$H_0 : d = 0$$
$$H_1 : d \neq 0$$

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche (ou inférieur) et le test unilatéral à droite (ou supérieur). Dans le test unilatéral à gauche, l'hypothèse alternative indique que la population *A* admet en général des valeurs inférieures à celles de la population *B* :

$$H_0 : d \leq 0$$
$$H_1 : d > 0$$

Dans le test unilatéral à droite, l'hypothèse alternative indique que la population *A* admet en général des valeurs supérieures à celles de la population *B* :

$$H_0 : d \geq 0$$
$$H_1 : d < 0$$

Ce test a été développé en considérant que :

- la distribution des db-a est symétrique,
- les db-a sont indépendants,
- les $db_{-a}$ se mesurent en valeurs réelles.

### Description du test du signe

L'objectif du test du signe est de déterminer si les échantillons proviennent d'une même population ou de deux populations différentes. StatBox peut réaliser un test bilatéral ou unilatéral.

Soient deux populations *A* et *B* dont sont prélevés les échantillons comportant des valeurs *a* et *b*. Le test du signe considère le nombre de différences *b-a* de signe positif. Le test bilatéral correspond au test de la différence entre *A* et *B*, et les hypothèses nulle ($H_0$) et alternative ($H_1$) sont les suivantes :

$$H_0 : P(a < b) = P(a > b)$$
$$H_1 : P(a < b) \neq P(a > b)$$

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche (ou inférieur) et le test unilatéral à droite (ou supérieur). Dans le test unilatéral à gauche, l'hypothèse alternative indique que la population *A* admet en général des valeurs inférieures à celles de la population *B* :

$$H_0 : P(a < b) \leq P(a > b)$$
$$H_1 : P(a < b) > P(a > b)$$

Dans le test unilatéral à droite, l'hypothèse alternative indique que la population *A* admet en général des valeurs supérieures à celles de la population *B* :

$$H_0 : P(a < b) \geq P(a > b)$$
$$H_1 : P(a < b) < P(a > b)$$

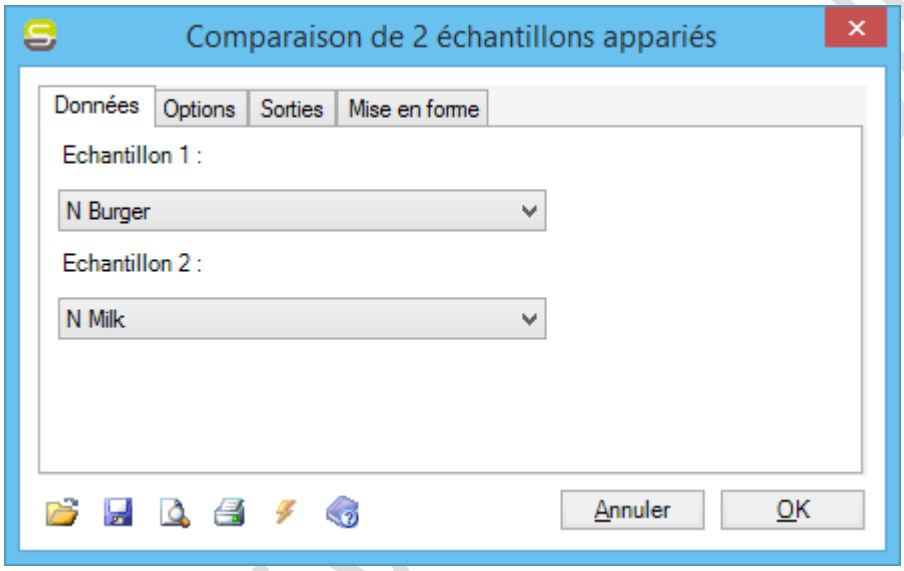
Ce test a été développé en considérant que :

- les couples de données appariées sont indépendants,
- les données sont au moins des données ordinales.

**Remarque** : pour calculer la *p*-value associée au nombre de différences positives, StatBox utilise la loi binomiale dans tous les cas, et pas l'approximation de la loi binomiale par la loi normale.

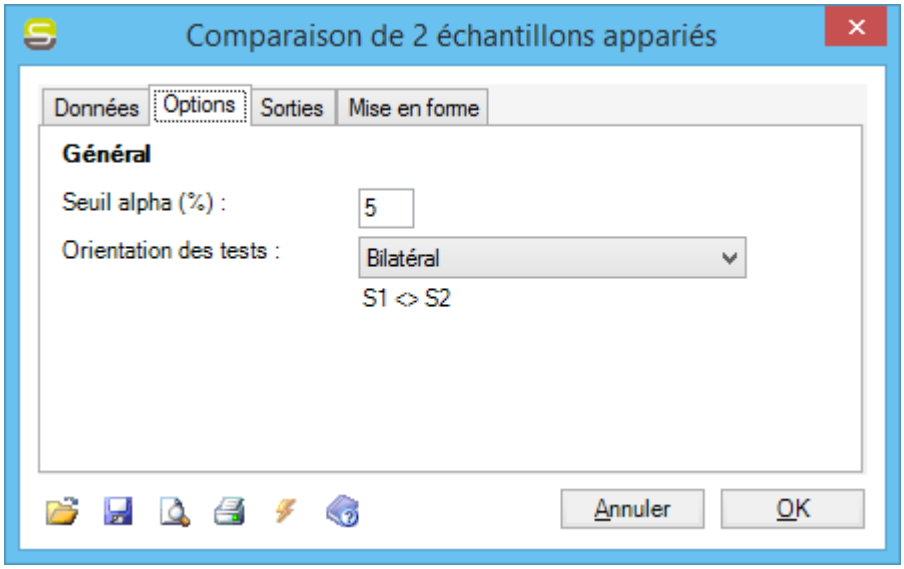
## Mise en œuvre

### *Onglet « Données »*



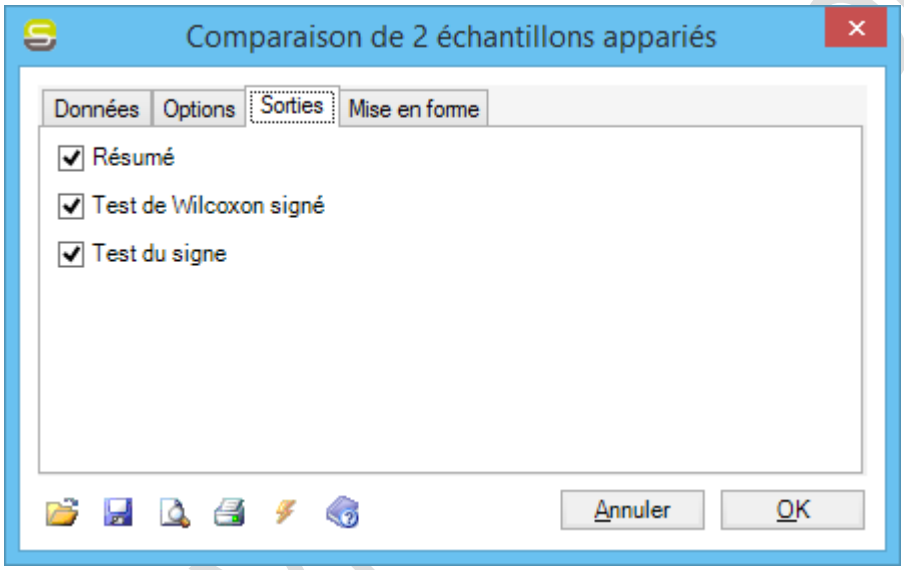

- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Échantillon 1 :* sélectionnez la variable correspondant au premier échantillon. Les valeurs manquantes ne sont pas autorisées.
- *Échantillon 2 :* sélectionnez la variable correspondant au second échantillon. Les valeurs manquantes ne sont pas autorisées.

*Onglet « Options »*




- *Seuil alpha (%) :* entrez la valeur du risque de première espèce du test.
- *Orientation du test :* choisissez le type de test à réaliser, bilatéral, unilatéral à gauche ou unilatéral à droite

*Onglet « Sorties »*




- *Test de Wilcoxon signé :* effectue le test de Wilcoxon signé.
- *Test du signe :* effectue le test du signe.

## Références

**Dagnelie P. (1986)**. Théorie et méthodes statistiques. Vol. 2. Les Presses Agronomiques de Gembloux, Gembloux, pp. 385-389.

**Lehmann E.L. & H.J.M. D'Abrera (1975)**. Nonparametrics. Statistical methods based on ranks. Holden-Day, San Francisco, pp. 120-132.

**Siegel S. (1956)**. Nonparametric statistics for the behavioral sciences. McGraw-Hill Kogakusha, Tokyo, Japan, pp. 68-83.

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 440-444.

## COMPARAISON DE K ÉCHANTILLONS INDÉPENDANTS (TEST DE KRUSKAL-WALLIS)

Utilisez ce test non paramétrique lorsque vous êtes en présence de *k* échantillons indépendants, afin de déterminer si les échantillons proviennent d'une même population ou si au moins un échantillon provient d'une population différente des autres.

**Remarque** : l'utilisation du test de Kruskal-Wallis constitue une alternative non paramétrique à l'utilisation de l'analyse de variance à 1 facteur (ANOVA 1). Comme dans l'ANOVA 1, les échantillons peuvent être de tailles différentes.

### Description

L'objectif du test de Kruskal-Wallis est de déterminer si les échantillons proviennent d'une même population ou si au moins un échantillon provient d'une population différente des autres. Les hypothèses nulle ($H_0$) et alternative ($H_1$) du test sont donc les suivantes :

$H_0$ : les *k* échantillons proviennent de la même population

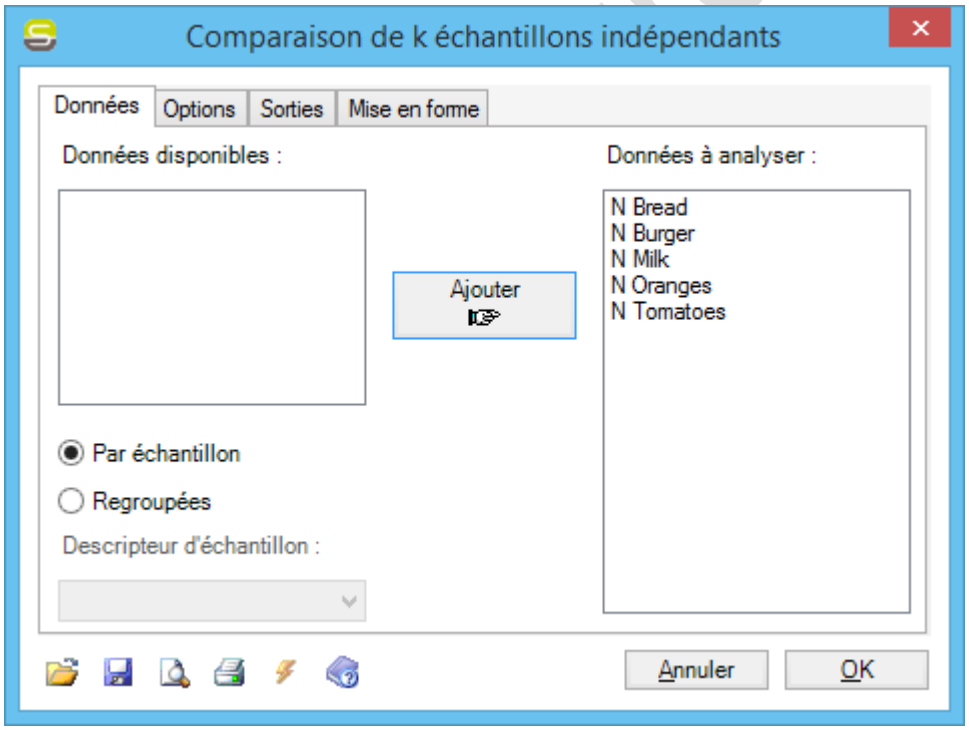
$H_1$ : au moins un des échantillons provient d'une population différente des autres

Ce test a été développé en considérant que :

- tous les échantillons sont des échantillons aléatoires tirés de leurs populations respectives,
- en plus de l'indépendance au sein de chaque échantillon, il y a indépendance mutuelle entre les échantillons,
- les données sont au moins des données ordinales.

### Mise en œuvre

*Onglet « Données »*

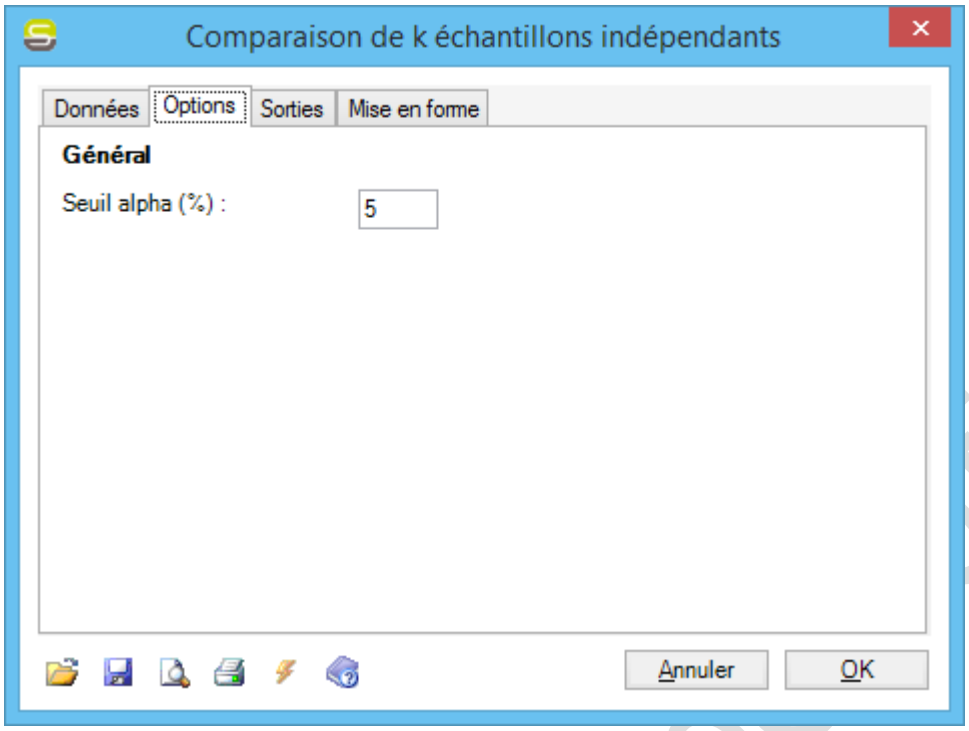



- *« Par échantillon » / « Regroupées » :* si les échantillons figurent dans des colonnes différentes, la plage des données correspond à un tableau avec les observations en lignes et les échantillons en colonnes. Les valeurs manquantes ne sont pas autorisées. Si les données sont regroupées, la plage correspond à une colonne de valeurs, l'appartenance aux échantillons étant indiquée par un *descripteur d'échantillon.*
- *Descripteur d'échantillon :* dans le cas des données regroupées, sélectionnez la variable qualitative indiquant l'échantillon d'appartenance de chaque valeur. Les valeurs manquantes ne sont pas autorisées.

- *Données à analyser :* sélectionnez la/les variable(s) correspondant aux données. Les valeurs manquantes ne sont pas autorisées.

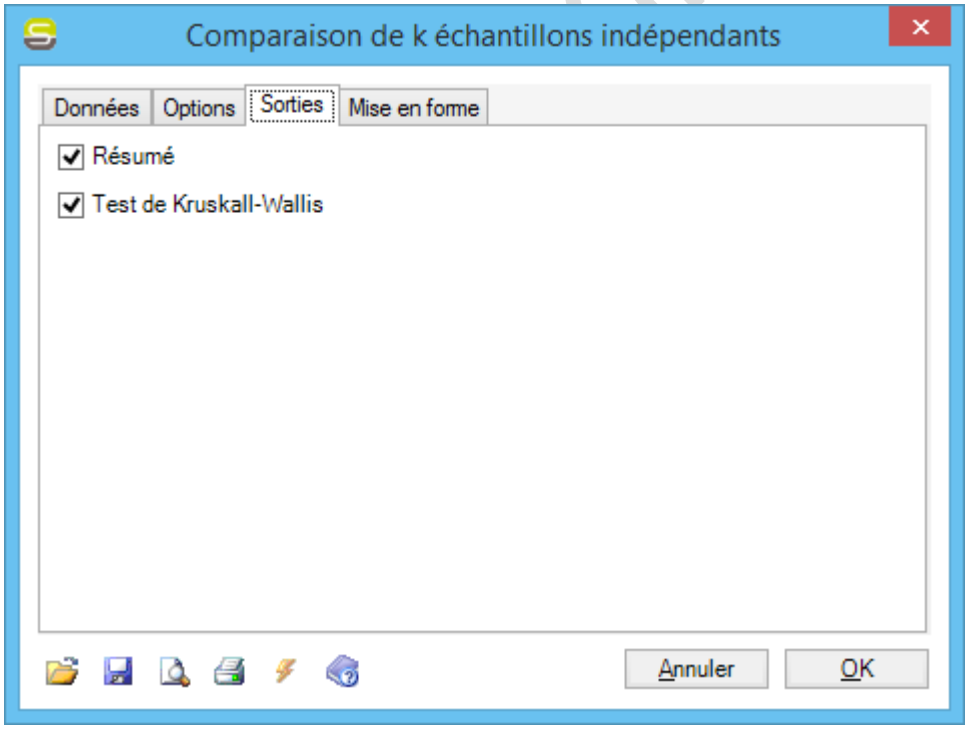
**Remarque** : dans le cas de l'option « Par échantillon » la taille des colonnes peut être différente.

### *Onglet « Options »*




- *Seuil alpha (%) :* entrez la valeur du risque de première espèce pour le test de Kruskal-Wallis.

### *Onglet Sorties »*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Test de Kruskal-Wallis* : effectue un test de Kruskal-Wallis.

## Références

**Dagnelie P. (1986)**. Théorie et méthodes statistiques. Vol. 2. Les Presses Agronomiques de Gembloux, Gembloux, pp. 390-392.

**Lehmann E.L. & H.J.M. D'Abrera (1975)**. Nonparametrics. Statistical methods based on ranks. Holden-Day, San Francisco, pp. 204-210.

**Manoukian E.B. (1986)**. Guide de statistique appliquée. Hermann, Paris, pp. 181-182.

**Siegel S. (1956)**. Nonparametric statistics for the behavioral sciences. McGraw-Hill Kogakusha, Tokyo, Japan, pp. 184-194.

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 423-427.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 240-241.


# COMPARAISON DE K ÉCHANTILLONS APPARIÉS (TEST DE FRIEDMAN)

Utilisez ce test non paramétrique lorsque vous êtes en présence de *k* échantillons appariés correspondant à *k* traitements portant sur les mêmes blocs, afin de mettre en évidence une différence entre les traitements.

**Remarque** : l'utilisation du test de Friedman constitue une alternative non paramétrique à l'utilisation de l'analyse de variance à 2 facteurs contrôlés (ANOVA 2). Les termes « traitement » et « bloc » doivent être pris dans un sens très général. En effet, il peut s'agir par exemple :

- de *k* traitements médicaux, les blocs étant des sujets volontaires,
- des appréciations sensorielles émises par un panel de consommateurs au sujet de k produits alimentaires, les blocs étant les consommateurs et les traitements étant les produits alimentaires,
- d'une cotation d'abondance d'espèces biologiques dans *k* zones géographiques différentes, les blocs étant les espèces et les traitements étant les zones géographiques et les conditions écologiques qui y règnent.

Les échantillons étant appariés, ils doivent nécessairement comporter le même nombre de blocs.

## Description

L'objectif du test de Friedman est de déterminer si tous les traitements donnent le même résultat ou si au moins un de traitements diffère des autres. Les hypothèses nulle ($H_0$) et alternative ($H_1$) du test sont donc les suivantes :

$H_0$ : les *k* échantillons ont été prélevés dans une même population

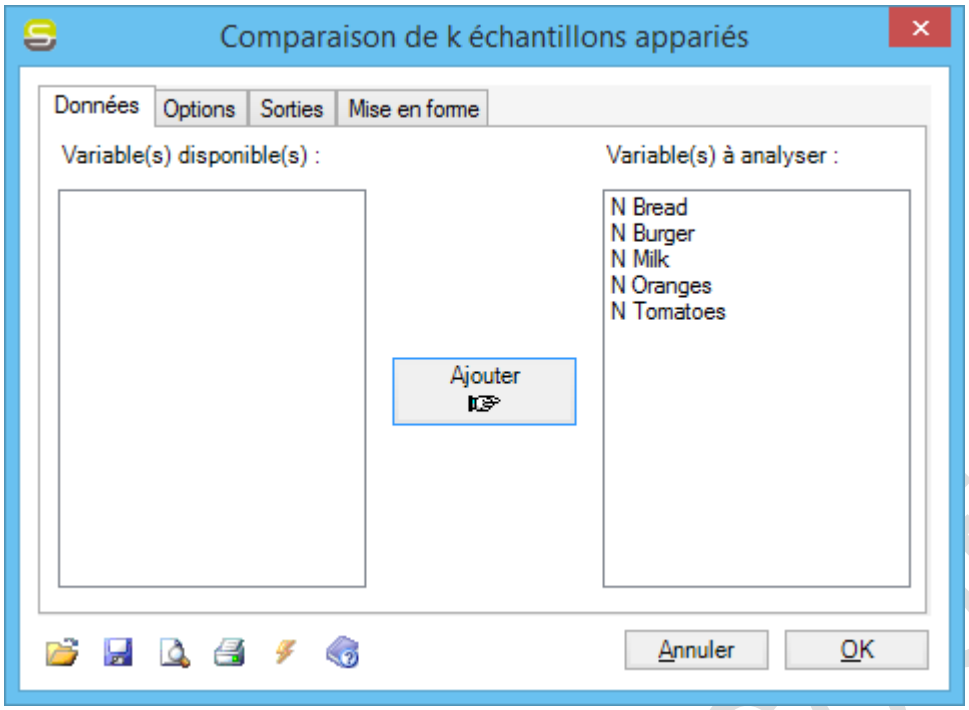
$H_1$ : au moins un des échantillons provient d'une population différente des autres

Ce test a été développé en considérant que :

- les blocs sont randomisés,
- les échantillons sont appariés,
- les données sont au moins des données ordinales.

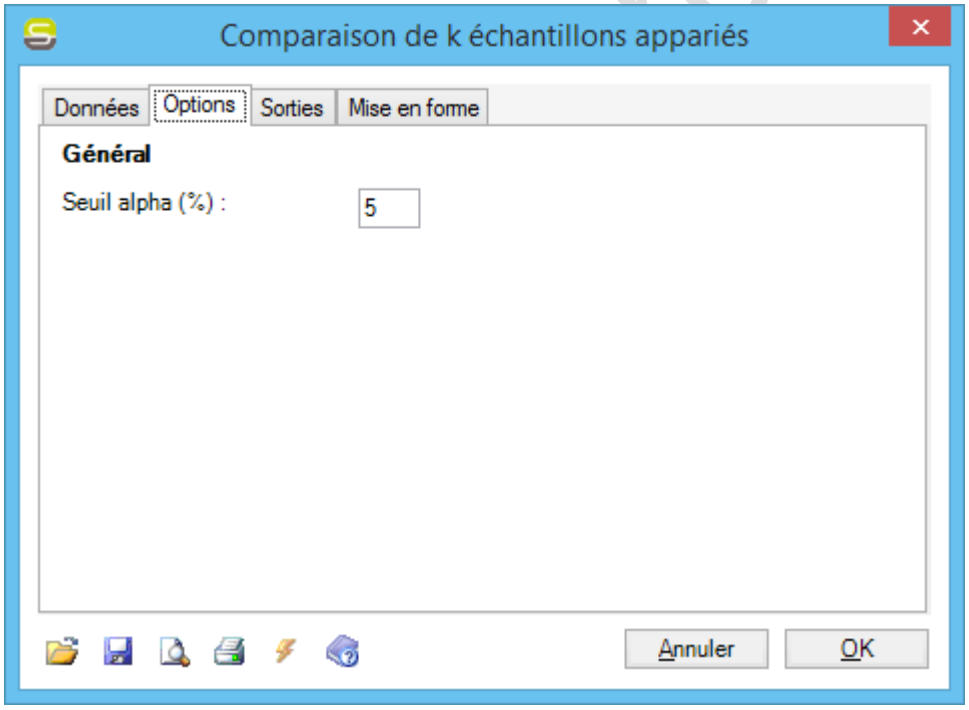
## Mise en œuvre

### *Onglet « Données »*



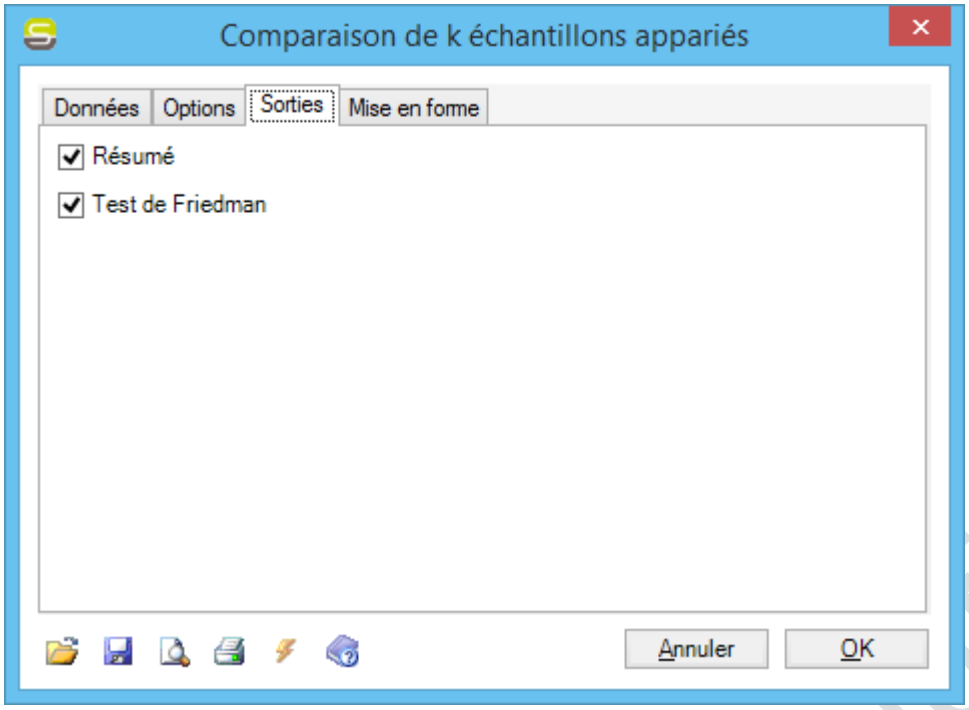

- *Variable(s) à analyser :* sélectionnez les variables correspondant à un tableau avec les blocs en lignes et les traitements en colonnes. Les valeurs manquantes ne sont pas autorisées.

### *Onglet « Options »*




- *Seuil alpha (%) :* entrez la valeur du risque de première espèce du test.

*Onglet Sorties*




- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport.
- *Test de Friedman :* effectue un test de Friedman.

## Références

**Dagnelie P. (1986)**. Théorie et méthodes statistiques. Vol. 2. Les Presses Agronomiques de Gembloux, Gembloux, pp. 393-394.

**Lehmann E.L. & H.J.M. D'Abrera (1975)**. Nonparametrics. Statistical methods based on ranks. Holden-Day, San Francisco, pp. 262-270.

**Manoukian E.B. (1986)**. Guide de statistique appliquée. Hermann, Paris, pp. 183-184.

**Siegel S. (1956)**. Nonparametric statistics for the behavioral sciences. McGraw-Hill Kogakusha, Tokyo, Japan, pp. 166-173.

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 440-442.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 242-243.

# LES ESSAIS EN AGRICULTURE

## INTRODUCTION

Sous le terme « Analyse de Variance », le programme fournit notamment :

- un test de l'interaction traitements * blocs (test de Tukey) pour vérifier la validité du modèle retenu (pour certains dispositifs expérimentaux seulement),
- un histogramme des résidus pour vérifier leur Normalité,
- les écarts-types des résidus intra-traitements (et intra-blocs) pour vérifier que dans tous les traitements (et tous les blocs) les résidus ont la même dispersion,
- une cartographie des résidus sur le plan réel de l'essai pour vérifier l'indépendance des erreurs,
- le tableau proprement dit d'analyse de variance qui permet de conserver ou de rejeter l'hypothèse d'homogénéité de l'ensemble des moyennes,
- la puissance de l'essai, utile pour en apprécier les chances de réussite,
- des tests de comparaisons multiples de moyennes.

Ce programme permet de réaliser l'analyse de variance de tous les plans d'expérience orthogonaux et équilibrés, comprenant de 1 à 4 facteurs (étudiés ou contrôlés) selon des modèles croisés. Ces plans sont :

1 facteur étudié

- randomisation totale avec répétitions
- bloc
- carré latin
- alpha-plan

2 facteurs étudiés

- factoriel 2 facteurs en randomisation totale avec ou sans répétitions
- factoriel 2 facteurs en blocs
- factoriel 2 facteurs en carré latin
- split-plot
- criss-cross

3 facteurs étudiés

- factoriel 3 facteurs en randomisation totale avec ou sans répétitions
- factoriel 3 facteurs en blocs
- factoriel 3 facteurs en carré latin
- split-plot 3 étages 1/2/3
- split-plot factoriel 1/(2*3)
- factoriel split-plot (1*2)/3
- criss-cross factoriel 1 # (2*3)
- criss-cross split-plot 1 # (2/3)

4 facteurs étudiés

- factoriel 4 facteurs en randomisation totale avec ou sans répétitions

### Lexique :

Répétition : on entendra par répétitions les différentes observations recevant le même traitement.

Niveau : on entendra par niveaux le nombre de modalités pour un facteur.

## TRAITEMENT DES DONNÉES NULLES

Pour la variable analysée, il est possible que certaines valeurs soient nulles.

Deux cas sont alors possibles:

1 - Ces valeurs nulles correspondent à la réalité observée (par exemple, on effectuait des comptages de pucerons sur des épis de blé... et il n'y en avait pas !). Ces informations doivent être retenues dans l'analyse et il suffit de répondre ensuite que ce n'est pas une donnée manquante.

2 - Ces valeurs nulles représentent des données manquantes, non enregistrées (ou mises à zéro par l'expérimentateur, car leur relevé était totalement aberrant). Dans ce cas, supprimez complètement les données nulles afin que le logiciel les détecte comme manquantes. Plusieurs solutions peuvent alors être envisagées.

- Si une ou deux valeurs sont manquantes, et à condition qu'elles n'appartiennent ni à un même bloc, ni à un même traitement, le programme pourra les estimer par la méthode de **Yates**. Celle-ci consiste tout simplement à « boucher le trou » avec une valeur telle que son résidu soit nul dans le modèle additif correspondant au plan de l'essai. Mais attention ! Cette méthode peut conduire à un malentendu, car, une fois le « trou » bouché, tout semble se passer comme s'il n'était rien arrivé .Cette impression est tristement fausse ! Il y a toujours perte d'information pour estimer les paramètres et calculer les tests : on perd autant de degrés de liberté à la variance résiduelle qu'il y a de données estimées.
- Si plus de deux valeurs sont manquantes, ou si deux ou plusieurs données manquent dans un même bloc, ou un même traitement, on pourra supprimer le (ou les) bloc(s) ou traitement(s) de l'analyse. **Attention** ! Ce(s) bloc(s) ou traitement(s) sera aussi éliminé pour toutes les autres variables analysées en même temps dans l'essai, sauf si on réalise l'analyse variable par variable.
- Si la variable enregistrée est pleine de « trous »… ne vaut-il pas mieux la supprimer de l'analyse ?... « Les maladies désespérées demandent des remèdes désespérés ! » (S.C. PEARCE attribue cette phrase à Guy FAWKES lorsqu'il tenta de faire sauter le Parlement Anglais).

## LE DISPOSITIF

### Création

Vous devez connaître quel dispositif vous allez mettre en place avant de lancer la création du dispositif.

Cliquez sur le menu « Nouveau » une boite de dialogue apparait vous permettant de sélectionner le type de dispositif et le nombre de facteurs correspondant au dispositif souhaité. Ce programme permet de générer (ou de saisir) tous les plans d'expérience orthogonaux et équilibrés, comprenant de 1 à 4 facteurs (étudiés ou contrôlés selon des modèles croisés).

Une fois le type de dispositif sélectionné, validez en cliquant sur OK ou double-cliquez sur celui-ci.

Un nouveau classeur est créé contenant une seule feuille nommée « Dispositif » destinée à recueillir toutes les informations relatives à l'essai et aux facteurs étudiés ou contrôlés.

Certaines zones de saisies sont facultatives, ce sont : le titre de l'essai, le protocole, l'année, le numéro d'essai, le code essai. D'autres sont indispensables au bon fonctionnement du classeur. Ce sont, selon le type de dispositif :

- Le nombre de répétitions / blocs (compris entre 2 et 300)
- Le nombre de sous-blocs (uniquement en alpha-plan)
- La taille des sous-blocs (uniquement en alpha-plan)

Et pour tous les classeurs :

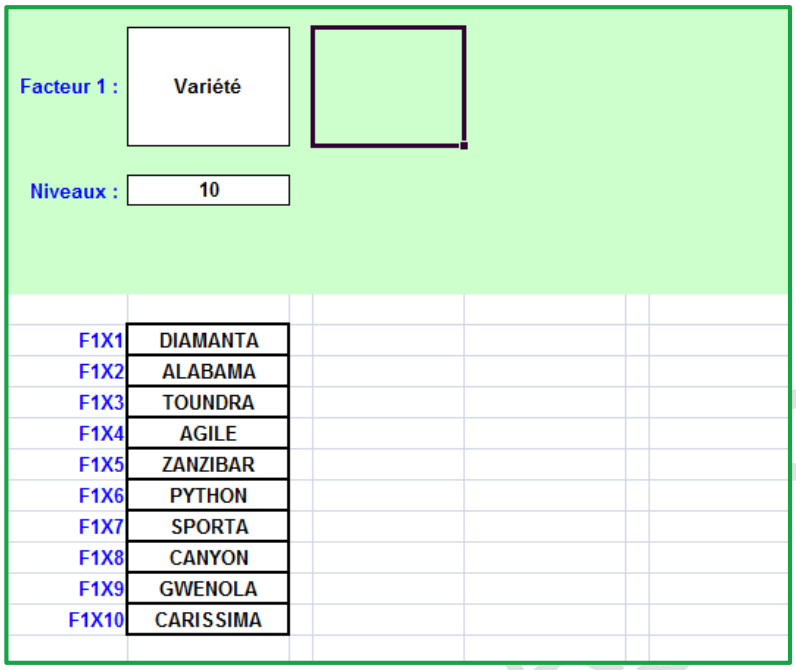
- Le nom de chacun des facteurs étudiés
- Pour chacun des facteurs le nombre de niveaux étudiés (compris entre 2 et 300)
- Pour chacun des facteurs le libellé de chaque niveau (plusieurs niveaux peuvent porter le même nom)

Remarque :

- **Ne jamais donner un numéro comme nom de niveau**

Afin de faciliter la saisie des libellés des niveaux de facteurs, vous pouvez générer automatiquement des listes de niveaux avec des noms par défaut. Pour cela, une fois que le nombre de niveaux étudiés par facteur est renseigné, cliquez sur « Générer les modalités ». Personnalisez ensuite les noms des niveaux.

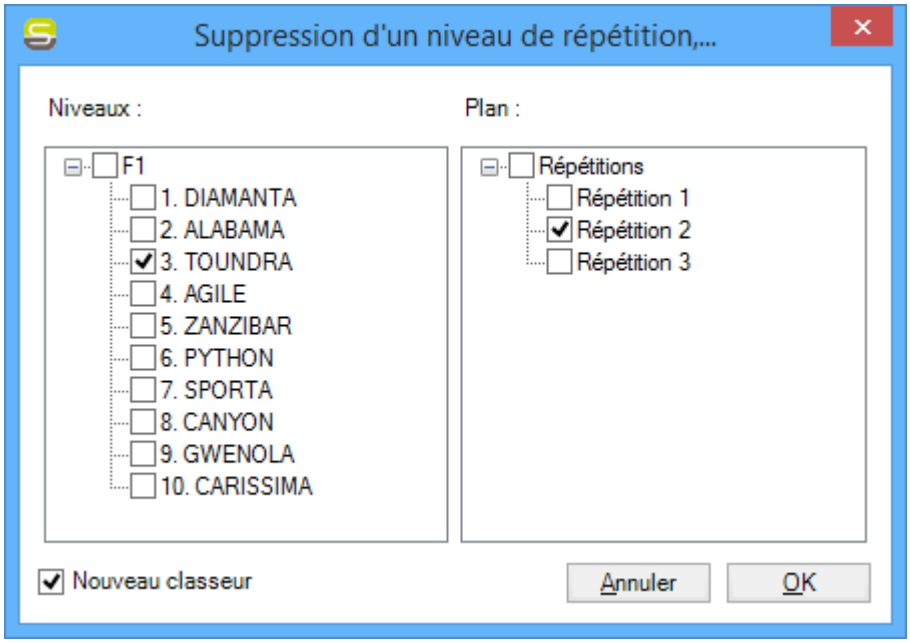
La saisie des libellés de niveau doit se faire de la manière suivante (exemple à 1 facteur en alpha-plan) :




## Supprimer : niveau, bloc, …

Vous pouvez à tout moment éliminer un niveau, une répétition ou un bloc, pour obtenir un classeur contenant une feuille de saisie réduite ce qui vous permet de réaliser des analyses sur une partie des données.

Pour cela, cliquez sur « Supprimer : niveau, bloc, … ». La boite de dialogue suivante s'affiche :




- *Niveaux :* sélectionnez les niveaux **à supprimer** pour chaque facteur.
- *Plan :* sélectionnez les répétitions/blocs/essais **à supprimer**
- *Nouveau classeur :* cochez cette option pour que le dispositif issu de la réduction de niveau soit affiché dans un nouveau classeur. Si cette option est décochée, la réduction de niveau s'effectuera sur le classeur en cours et des données seront donc **définitivement perdues**.

Validez en cliquant sur « OK ».

**Remarque :** cette fonction n'est pas disponible pour les carrés latins.

### Dupliquer un dispositif

Vous pouvez créer un nouveau dispositif à partir d'un dispositif existant. Cela est utile, par exemple, dans le cas ou vous souhaitez reconduire un même essai sur plusieurs lieux différents, vous n'avez ainsi pas à ressaisir la totalité de l'information concernant les facteurs.

Cliquez sur « Dupliquer le dispositif ». Dans le cas où le plan du classeur d'origine a déjà été généré, il vous est demandé si vous souhaitez conserver le plan existant ou si vous souhaitez générer un nouveau plan. Dans ce deuxième cas, la boite de dialogue de génération de plan s'affiche.

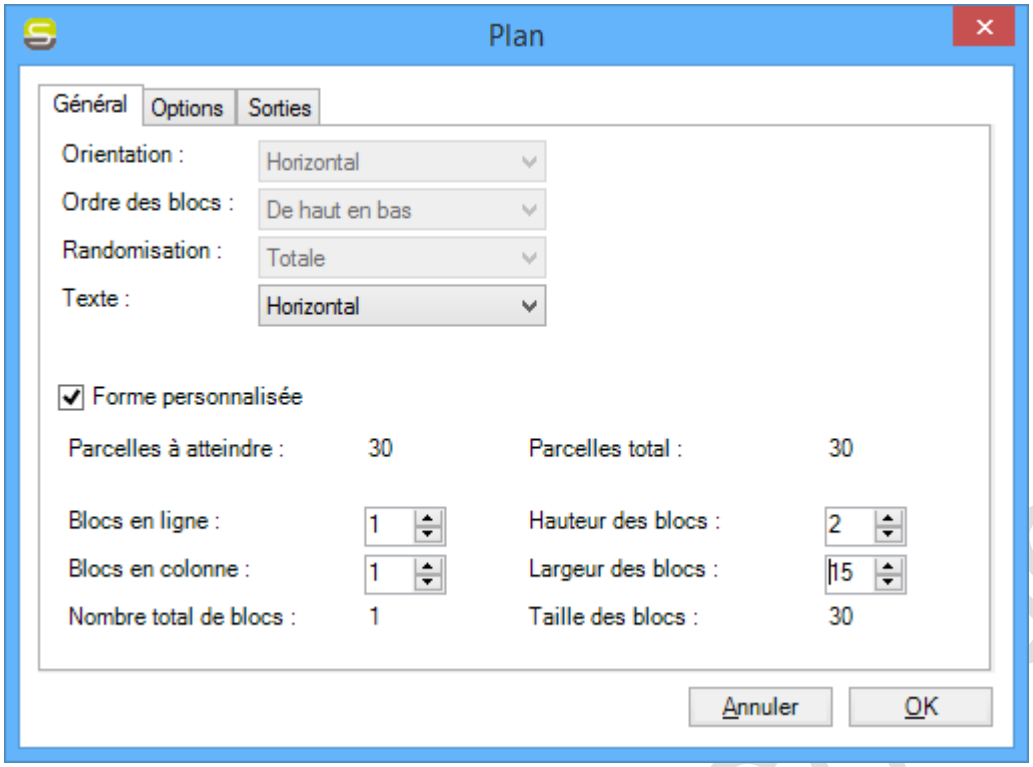
## LE PLAN

### Génération du plan

Une fois toutes les informations indispensables au dispositif renseignées, vous devez déterminer la répartition des traitements étudiés sur le terrain d'expérimentation. Pour cela, lancez la génération du plan d'expérience en cliquant sur « Générer ».

Une boite de dialogue apparait vous proposant de nombreuses options de génération :

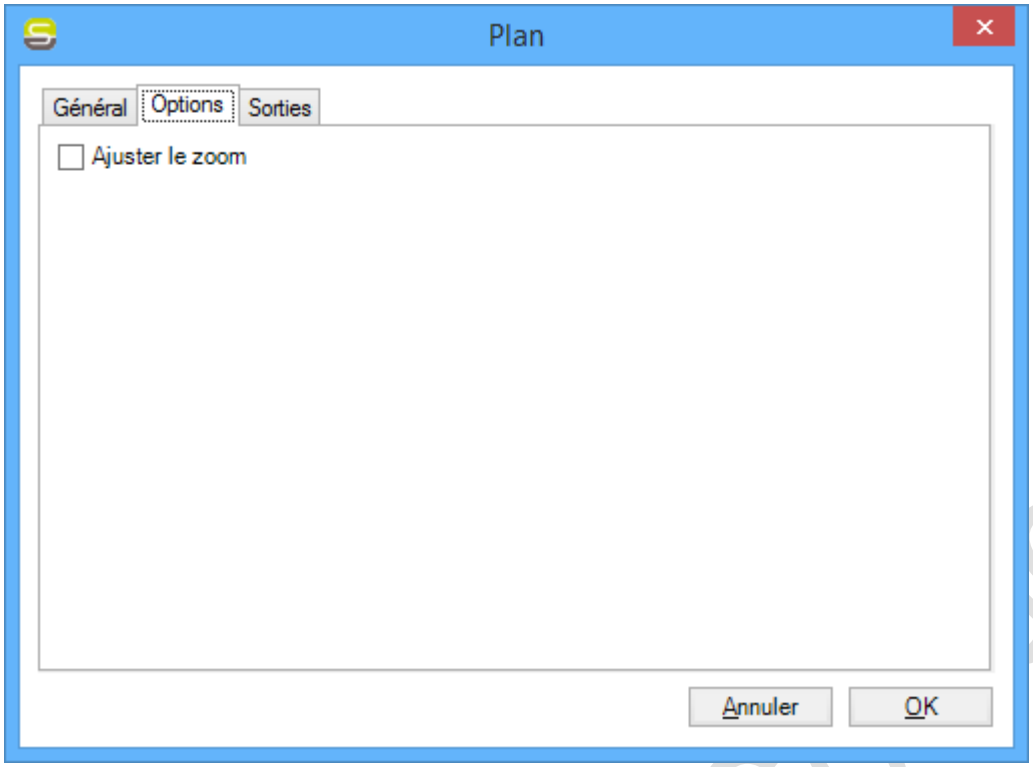
*Onglet « Options »*




- *Orientation :* sélectionnez l'orientation des répétitions/blocs.
- *Ordre des blocs :* sélectionnez l'ordre de numérotation des blocs.
- *Randomisation :* choisissez si la randomisation doit être totale, exclure 1 ou 2 blocs ou sans aucune randomisation. Ces deux dernières options peuvent être utiles si vous souhaitez par exemple conserver sur un bloc les modalités dans l'ordre saisie sur la feuille dispositif afin de les identifier plus rapidement sur le terrain.
- *Texte :* sélectionnez l'orientation du texte sur la feuille de plan.
- *Plan Traité/non traité* (disponible sur les essais Vegetal seulement) : choisissez cette option si vous souhaiter un plan traité/non traité. Dans StatBox Vegetal les plans traités/non traité gèrent le bloc 1 non traité et les autres blocs traités. Lorsque cette option est cochée, les traitements se font automatiquement sur les blocs traités.
- *Forme personnalisée :* cochez cette option si vous souhaitez donner une forme particulière au plan en terme de disposition des blocs et de taille de ceux-ci. Vous pouvez ainsi sélectionner le nombre de blocs en ligne et en colonne ainsi que la hauteur et la largeur des blocs. Si les informations saisies ne permettent pas de tirer le plan, le nombre de parcelle totale est indiqué en rouge.

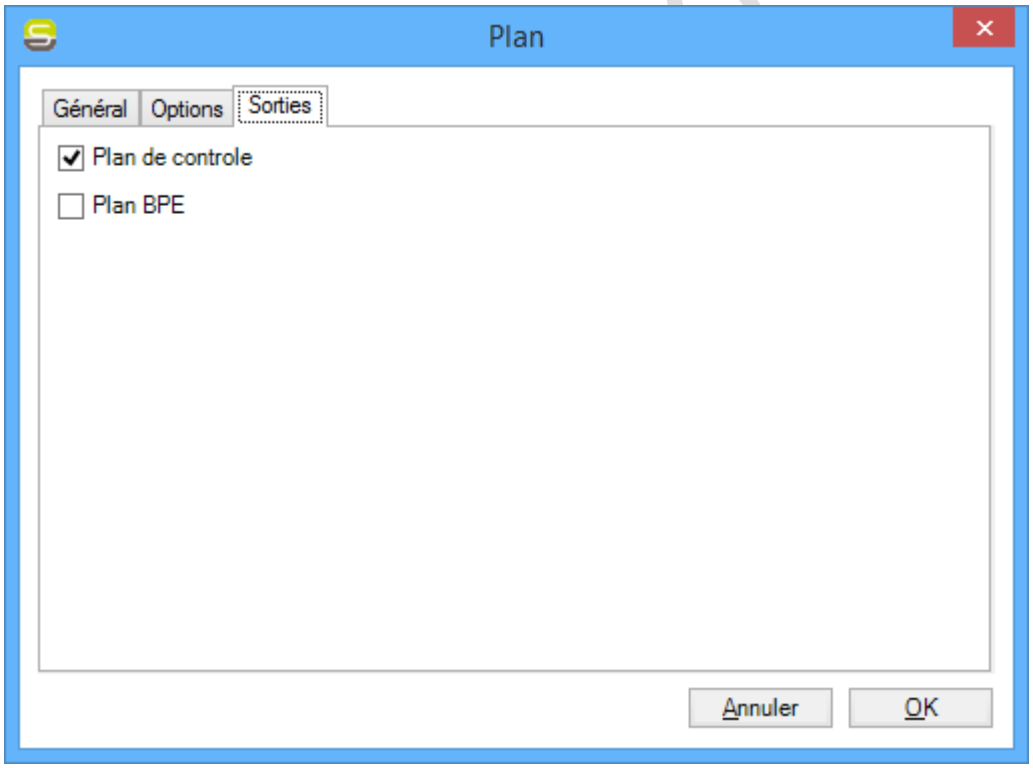
**Remarques** : selon le type de dispositif certaines options de génération peuvent ne pas être disponibles.

*Onglet « Options »*




*Ajuster le zoom :* ajuste le zoom de la feuille Excel afin de rendre visible la totalité du plan.

*Onglet « Sorties »*




- *Plan de contrôle :* génère en parallèle un plan colorant chaque modalité différemment pour étudier rapidement leur répartition.
- *Plan BPE :* génère un plan identifiant chaque parcelle par un code Ligne/colonne qui ne traduira pas l'appartenance à tel ou tel niveau de modalité étudiée.

### *Les alpha-plans*

Les alpha-plans sont utilisés lorsque le nombre de modalités ou de niveaux du facteur est trop important pour assurer une certaine homogénéité à l'intérieur d'un bloc.

Les alpha-plans sont constitués de sous-blocs ne comportant qu'un sous ensemble des niveaux du facteur. Ces sous-blocs de petite taille permettent de mieux contrôler l'homogénéité à l'intérieur des blocs.

Il n'est pas possible des générer des alpha-plans pour toutes les tailles de dispositif. Ainsi il est impossible de générer un alpha-plan pour moins de 10 niveaux. D'autres combinaisons Nombre de niveaux * Nombre de répétitions seront également impossible à générer, dans ce cas un message vous alertera au lancement de la génération du plan.

2 informations supplémentaires sont indispensables à la génération d'un alpha plan : la taille des sous-blocs et le nombre de sous-blocs. Ces informations doivent être renseignées sur la feuille de dispositif. Au cours du tirage, si cela est possible, ces paramètres seront optimisés afin de tirer le plan le plus « parfait » possible et la feuille de dispositif sera alors mise à jour, dans le cas contraire, ce sont les paramètres saisis par l'utilisateur qui seront pris en compte.

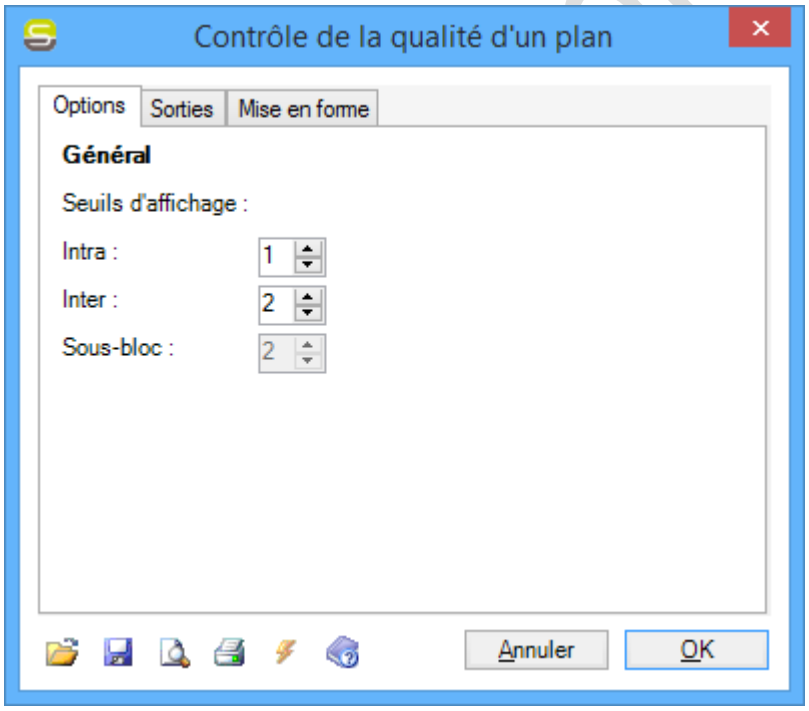
## Contrôle de la qualité du plan

La qualité d'un plan est notamment liée à la qualité de la répartition des différents niveaux étudiés à l'intérieur de celui-ci. Ainsi, si toutes les répétitions d'un même niveau se retrouvent côte à côte (on parle de concomitance), il est sans doute préférable de régénérer le plan.

StatBox propose d'établir rapidement les tables de dénombrement des concomitances intra-traitement (répétitions d'un même niveau côte à côte), des concomitances inter-traitements (2 niveaux sont plusieurs fois côte à côte) ou intra sous-blocs (2 niveaux sont plusieurs fois côte à côte dans les différents sous-blocs).

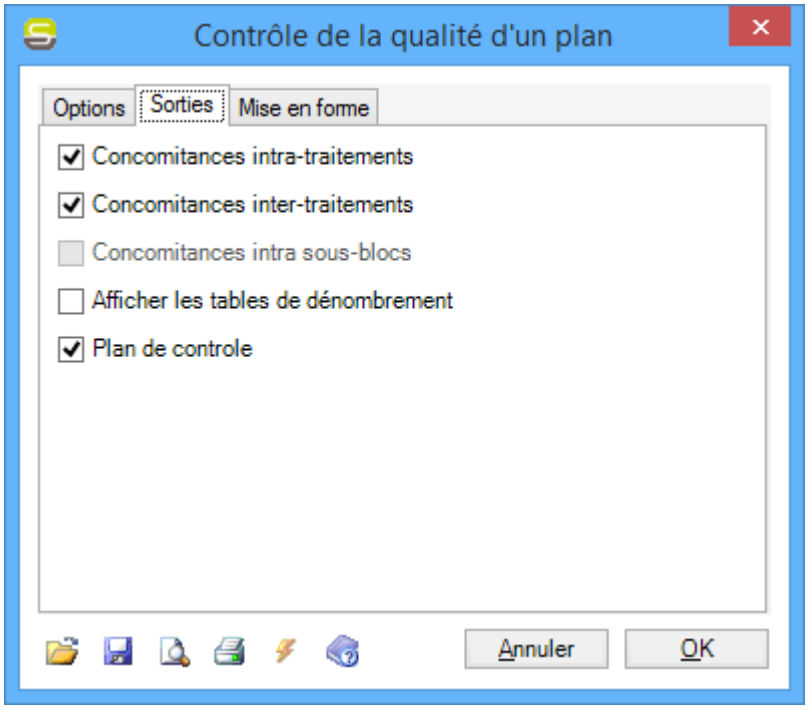
Pour effectuer un contrôle de qualité, cliquez sur « Contrôle de qualité », la boite de dialogue suivante apparait :

### *Onglet « Options »*




- *Seuils d'affichage :* Sélectionnez pour chacun des types de concomitance le seuil à partir duquel les concomitances doivent être signalées dans les résultats.

*Onglet « Sorties »*




- *Concomitances intra-traitements :* affiche les résultats concernant les concomitances intra-traitements
- *Concomitances inter-traitements :* affiche les résultats concernant les concomitances inter-traitements
- *Concomitances intra sous-blocs :* affiche les résultats concernant les concomitances intra sous-blocs
- *Afficher les tables de dénombrement :* affiche des tables pour chaque type de concomitance permettant de déterminer les traitements ou croisements de traitements qui ont le nombre de concomitances les plus élevés. Si vous souhaitez par la suite effectuer une personnalisation manuelle du plan, il sera ainsi préférable de modifier en priorité la localisation de ces traitements.
- *Plan de contrôle :* génère en parallèle un plan colorant chaque modalité différemment pour étudier rapidement leur répartition.

## Personnalisation de la position des parcelles dans le plan de l'essai

Dans le cas où le plan généré ne correspondrait pas à votre plan réel (présence d'un arbre, zones inutilisables, …), vous pouvez le modifier.

Cliquez sur « Personnaliser le plan », une nouvelle feuille nommée « PlanPS » s'ajoute au classeur reprenant dans la partie supérieure le plan actuel de l'essai et proposant dans la partie basse un plan vierge.

Pour créer le plan personnalisé, déplacer les parcelles du plan situé en haut vers le plan situé en dessous en faisant, soit un Couper/Coller, soit en faisant glisser les parcelles vers le plan vide du dessous :

- Couper/Coller : **Ctrl + X** de la ou les parcelle(s) au point d'origine puis **Ctrl + V** à la destination.
- Glisser : Sélectionnez la ou les parcelle(s), placez le pointeur sur le bord de la cellule de sorte à obtenir une croix noire et déplacez la sélection à l'emplacement souhaité en continuant d'appuyer sur le bouton de la souris.

Une fois que le nouveau plan est satisfaisant cliquez sur « Actualiser » (sous « Personnaliser le plan ») afin que la feuille de plan reprenne le plan personnalisé. Il vous est alors proposé d'effectuer un contrôle de qualité du nouveau plan et la feuille de plan personnalisé est détruite.

Remarques :

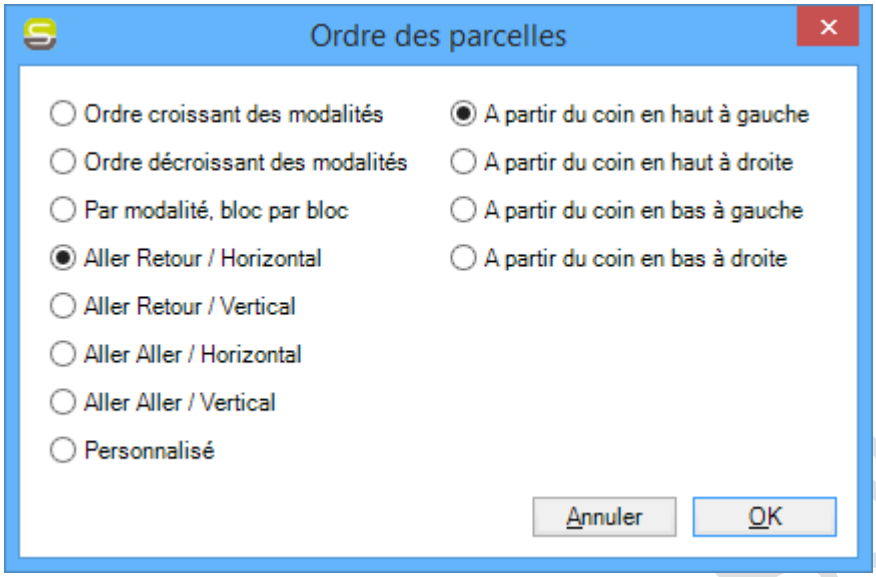
- Ne faites pas de copier/coller (**Ctrl + C** / **Ctrl + V**)
- Ne superposez pas les parcelles
- Toutes les parcelles du plan d'origine doivent être placées sur le nouveau plan

- Il est possible de lancer cette procédure autant de fois que nécessaire.

### Gestion de l'ordre de saisie

Il est possible de faire varier l'ordre des parcelles sur la feuille de saisie afin de refléter l'ordre réel de saisie lors des notations sur le terrain. Pour cela utilisez un plan de saisie, qui attribue sur le plan réel de l'essai un numéro d'ordre à chaque parcelle correspondant à l'ordre de saisie.

Cliquez sur « Générer un plan de saisie », la boite de dialogue suivante apparait :




Sélectionnez dans la liste de gauche l'ordre de navigation dans le plan et dans la liste de droite le point de départ de la saisie, puis validez en cliquant sur OK.

L'option « Personnalisé » permet de partir d'un plan de saisie vide et de renseigner manuellement l'ordre de saisie en saisissant les valeurs des rangs dans chacune des parcelles. Dans ce cas, un rang doit être fourni pour chaque parcelle et il ne doit pas y avoir de discontinuité dans les rangs.

Il est également possible de modifier manuellement une partie des rangs généré automatiquement sur la feuille de plan de saisie. Corriger les valeurs des rangs dans les parcelles, il ne doit pas y avoir de discontinuité ni de doublons.

Une fois l'ordre de saisie satisfaisant, cliquez sur « Actualiser » sous « Générer un plan de saisie » afin que la feuille de saisie de base reprenne l'ordre de saisie. Toutes les feuilles de saisie créée par la suite reprendront cet ordre.

**Remarque** : La feuille de plan de saisie existe tant que vous désirez la conserver vous pouvez donc à tout moment modifier tout ou partie de l'ordre de saisie et actualiser la feuille de saisie.

## LES SAISIES

### Gestion des feuilles de saisie

Vous pouvez créer autant de feuilles de saisie que nécessaire en plus de la feuille de saisie créée par défaut.

Pour ajouter une nouvelle feuille de saisie, cliquez sur « Nouvelle feuille de saisie ». Une copie de la feuille de saisie initiale est alors créée avec un nom incrémentiel. L'ordre de saisie de la nouvelle feuille est ainsi identique à celui en cours sur la feuille de saisie de base.

Ne jamais supprimer la feuille de saisie de base.

Pour les analyses, seules les variables présentes sur la feuille de saisie de base seront proposées. Vous pouvez fusionner les variables de toutes les feuilles de saisie en cliquant sur « Fusionner » dans le menu saisie. La feuille de saisie de base reprend alors toutes les variables disponibles et les autres feuilles de saisies sont détruites.

Remarques :

- Il est possible de fusionner des feuilles de saisie reprenant des ordres de saisie différents
- Il n'est pas possible de fusionner lorsque 2 variables ont le même nom sur l'une ou l'autre des feuilles

### Affichage sur la feuille de saisie

Par défaut, les codes sont affichés dans la feuille de Saisie. Il est cependant possible de modifier l'affichage sur cette feuille pour afficher selon les cas les codes ou les libellés des niveaux.

Pour afficher les codes, cliquez sur « Afficher les identifiants » et pour afficher les libellés, cliquez sur « Afficher les libellés ».

## L'ANALYSE DE VARIANCE

### Description

L'analyse de variance est une méthode statistique qui permet de tester l'hypothèse d'homogénéité d'un ensemble de k moyennes.

Pour tester cette hypothèse, le choix d'un modèle est nécessaire. Par exemple, lorsque vous comparez des traitements selon un dispositif en "blocs", le modèle que vous retenez (peut-être sans le savoir !...) est le suivant :

| rendement « espéré » d'une parcelle de l'essai | = | effet moyen du champ d'essai | + | effet du traitement reçu par cette parcelle | + | effet du bloc auquel appartient cette parcelle |
|---|---|---|---|---|---|---|

$$\hat{Y}ij = \mu + \alpha i + \beta j$$

Ensuite, réaliser l'analyse de variance, c'est tester si les effets des traitements sont identiques ou non. En termes statistiques, c'est rechercher si l'effet "traitements" est "significatif" ou non (bien sûr, avec un certain risque d'erreur).

Dans le cas où l'effet "traitements" est globalement significatif, vous voulez évidemment connaître les traitements qui ont des effets différents. Il faudra alors poursuivre l'analyse en choisissant le test de comparaison de moyenne adapté à l'objectif de votre essai.

Dans le cas où l'effet "traitements" n'est pas significatif, un calcul de puissance vous sera utile pour savoir si votre essai avait toutes les chances ou non de mettre en évidence les différences entre traitements que vous jugiez intéressantes à déceler.

Maintenant, il ne faut pas oublier que vous avez choisi un modèle a priori : il est honnête de vérifier son bien-fondé. Que peut-on en dire ?

Le modèle est par construction additif : $\hat{Y}ij = \mu + \alpha i + \beta j$

Il conviendra de vérifier qu'il y a bien additivité des effets "traitements" et des effets "blocs", c'est à dire qu'il n'y a pas d'interaction traitements * blocs.

Le modèle est bien sûr théorique ; dans la réalité, il y a un écart, appelé "résidu", entre le rendement que vous mesurez sur la parcelle et le rendement obtenu par le modèle. Ce résidu est la traduction de différents types d'erreurs indissociables : mauvais choix de modèle, erreurs de mesures, erreurs aléatoires.

3 conditions importantes doivent être remplies par ces résidus. Ils doivent :

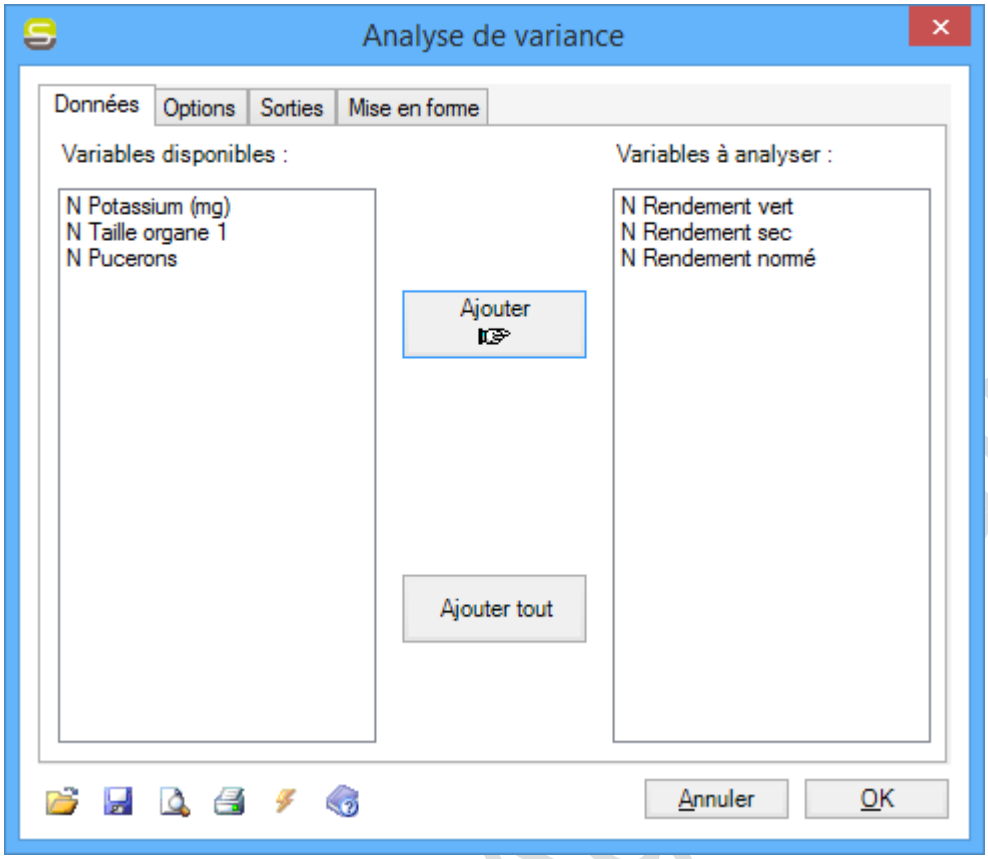
- être distribués normalement,
- avoir une variance constante (la même pour tous les traitements),
- être indépendants.

Il conviendra de vérifier ces conditions d'application.

### Mise en œuvre

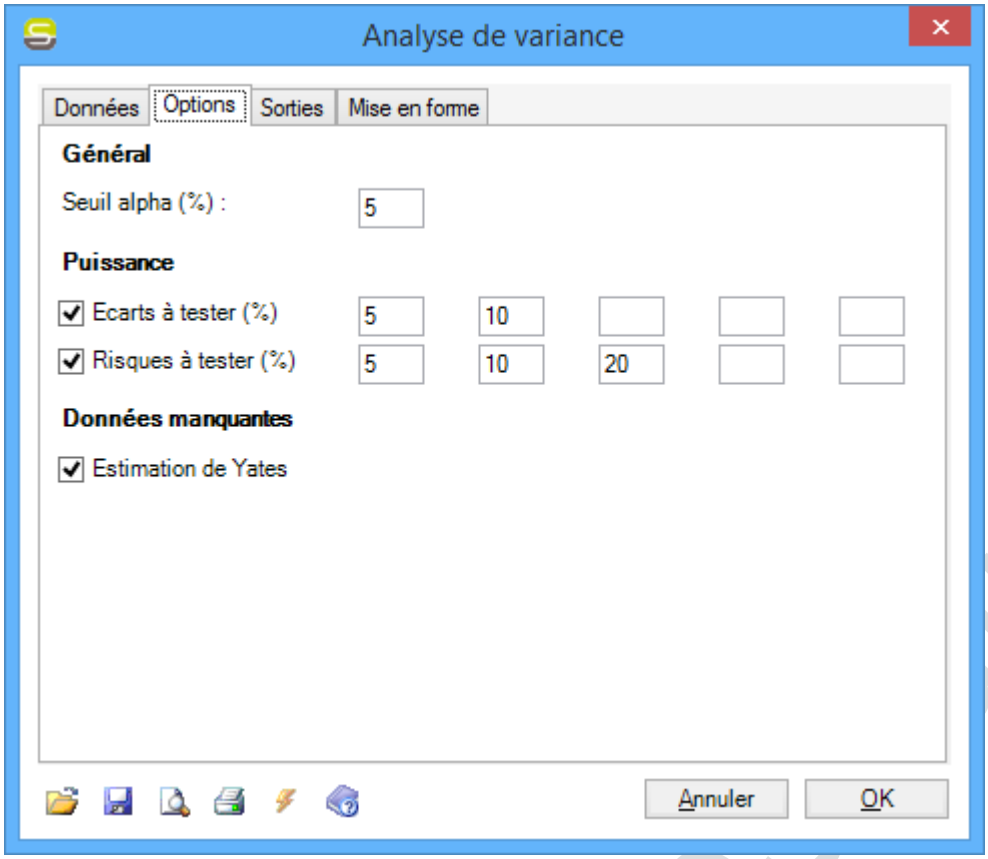
Pour lancez une analyse cliquez sur « Analyse ».

*Onglet « Données »*



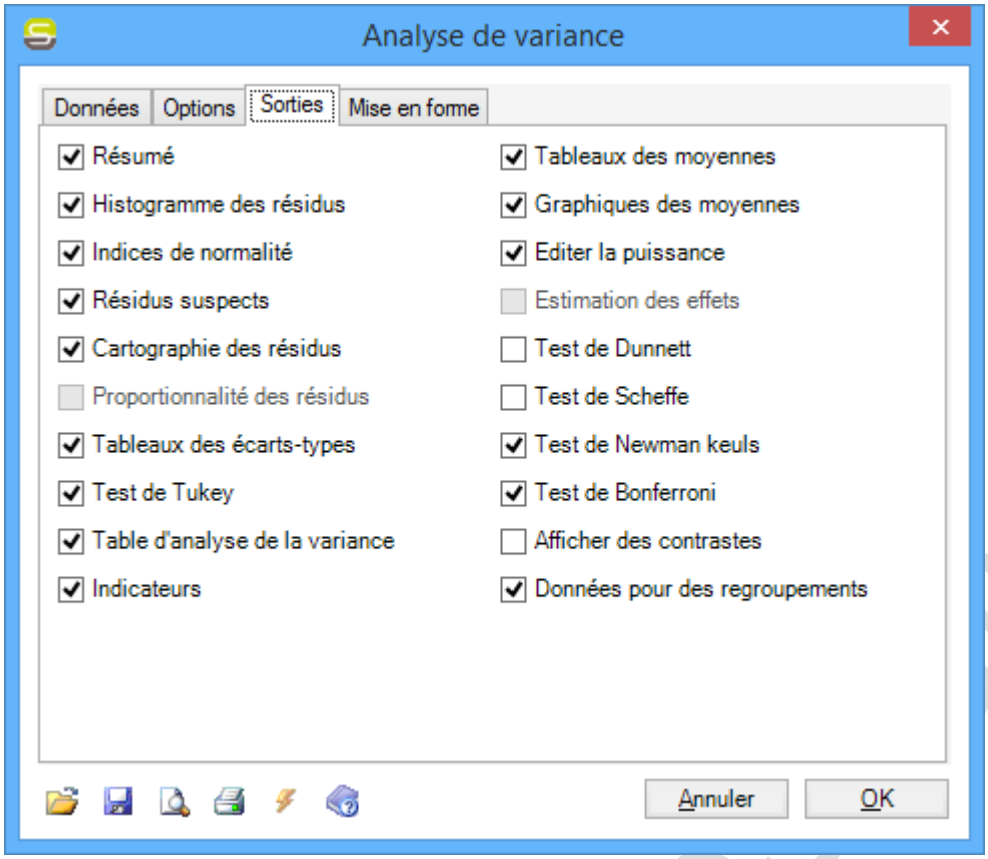

- *Variable(s) à analyser :* sélectionnez les variables à analyser en les faisant passer dans la liste de droite. Vous pouvez sélectionner toutes les variables disponibles en cliquant sur « Ajouter tout ».

*Onglet « Options »*




- *Seuil alpha (%) :* entrez la valeur du risque de première espèce pour les tests de comparaisons de moyennes. Ce risque doit être de 1 ou 5.
- *Écarts à tester :* entrez les valeurs des écarts à tester pour les tests de puissance. Il s'agit des écarts que vous cherchez à montrer sur les variables mesurées, par exemple un gain de 5 quintaux sur des variétés présentant un rendement moyen de 100 quintaux correspond à un écart de 5%
- *Risques à tester :* entrez les valeurs des risques à tester pour les tests de puissance. Vous pouvez ainsi tester un gain de rendement de 5% à la fois pour un risque d'erreur de 5 ou de 10 %
- *Estimation de Yates :* Si vous avez des données manquantes, le logiciel vous propose, dans la mesure du possible d'estimer ces données manquantes. En cochant cette option, le logiciel calcule automatiquement les données manquantes et continu le traitement.

*Onglet « Sorties »*



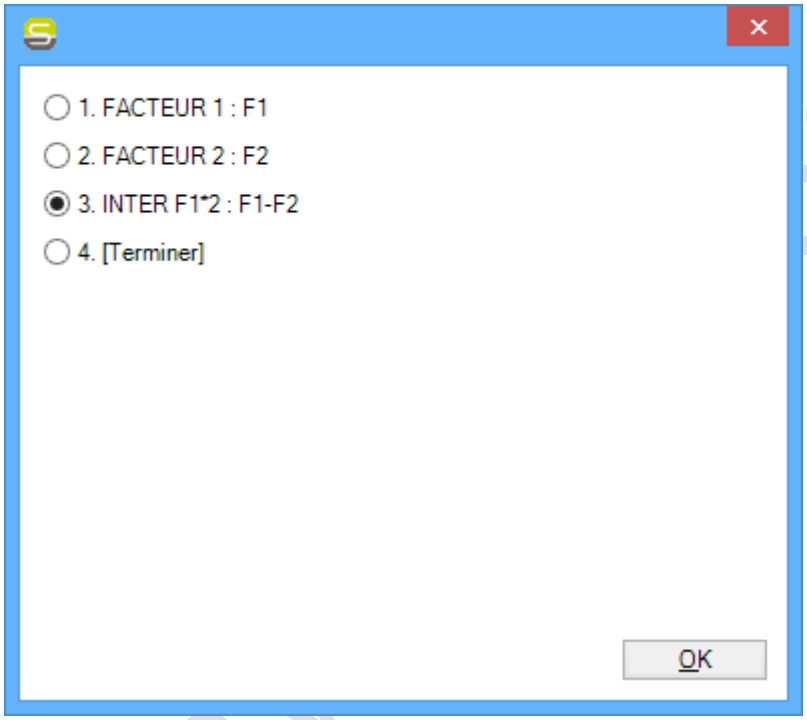

- *Résumé :* cochez cette option pour obtenir une brève synthèse des données et des options sélectionnées pour le rapport
- *Histogramme des résidus* : affiche un histogramme de répartition des résidus afin d'étudier de manière visuelle la normalité de leur distribution
- *Indices de normalité :* affiche les indices de symétrie et d'aplatissement de Pearson associés à l'histogramme des résidus.
- *Résidus suspect :* effectue une détection des résidus suspects par la méthode de Grubbs
- *Cartographie des résidus :* affiche la répartition des résidus sur le plan réel de l'essai, chaque parcelle est colorée selon un gradient de couleur traduisant la valeur de son résidu (création du gradient de couleur par la méthode des quartiles sur l'intervalle de variation des résidus). Cela permet d'apprécier visuellement l'indépendance des résidus entre eux.
- *Proportionnalité des résidus :* vérifie l'indépendance des résidus par rapport à la valeur de la variable étudiée.
- *Tableau des écarts-types :* affiche les tables d'écarts types pour chacun des facteurs étudiés, contrôlés ou des niveaux d'interactions.
- *Test de TUKEY :* vérifie si l'interaction traitement*bloc est significative. Cette option n'est active que si votre dispositif comporte des blocs.
- *Table d'analyse de la variance :* affiche la table de décomposition de la variance pour les facteurs étudiés, contrôlés et les interactions.
- *Indicateurs :* affiche une table d'indicateurs sur la variable analysée : moyenne, écart type résiduel et coefficient de variation.
- *Tableaux des moyennes :* affiche les tables de moyennes pour chacun des facteurs étudiés, contrôlés et les niveaux d'interactions.
- *Graphiques des moyennes :* affiche des histogrammes des moyennes pour chacun des facteurs étudiés, contrôlés et les niveaux d'interactions.
- *Éditer la puissance :* permet d'étudier les risques $\alpha$ de 1ère espèce ainsi que les risques $\beta$ de 2ème espèce. Vous pouvez saisir jusqu'à 5 écarts à tester en % ainsi que 5 risques à tester en %. Les valeurs doivent être comprises entre 0.1 et 99.

- *Test de Dunnet :* effectue un test de comparaison de moyennes avec présence de témoins
- *Test de Scheffe :* effectue un test de Newman Keuls pour les traitements supérieur aux témoins.
- *Test de Newman-keuls :* constitue des groupes homogènes de traitements par comparaison de moyennes.
- *Test de Bonferroni :* effectue des comparaisons de moyennes 2 à 2.
- *Afficher des contrastes :* effectue des comparaisons particulières entre les facteurs.
- *Données pour les regroupements :* édite une table de synthèse des résultats de l'essai. Ces résultats permettent la constitution des essais en regroupement.

Si le dispositif le permet, au lancement de l'analyse, le message suivant apparaît : «L'analyse se fait sur toutes les modalités et sur toutes les répétitions ? ». Le logiciel vous propose d'effectuer l'analyse sur un nombre réduit de niveaux ou de Répétitions/Blocs/Essais. Cela est utile dans le cas ou vous avez de nombreuses données manquantes ou lorsque vous décidez de supprimer les résidus suspects de l'analyse. Pour supprimer un niveau, cliquez sur « Non », la boite de dialogue de suppression de niveau apparaît alors. Sélectionnez les niveaux **à supprimer** de l'analyse et validez. Dans ce cas, la suppression de niveau est toujours temporaire et n'affecte jamais les données initiales du classeur.

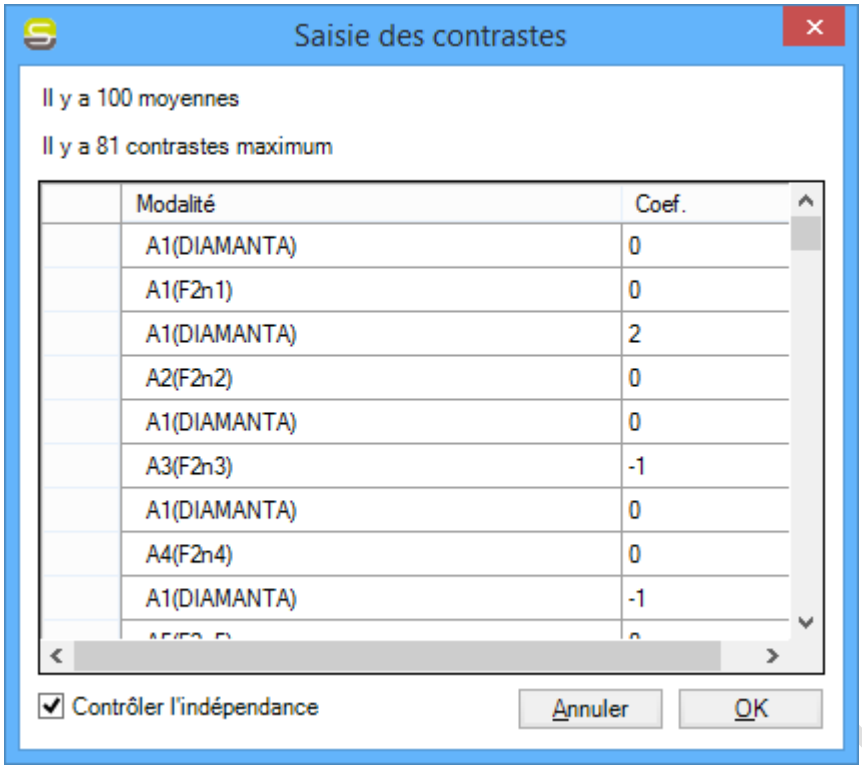
Si vous avez choisi d'éditer un test de Dunnet, le programme vous demande de paramétrer le nombre total de témoin et de déterminer les niveaux concernés.

Si vous choisissez d'éditer des contrastes, sur les dispositifs à 2 facteurs et plus, la boite (n°1) suivante apparait :




Sélectionnez le facteur ou le niveau d'interaction pour lequel vous souhaitez éditer un contraste et validez en cliquant « OK »

La boite (n°2) suivante apparait :




Saisissez dans la colonne de droite les coefficients des contrastes pour chacune des moyennes

Décochez au besoin l'option de contrôle de l'indépendance des contrastes.

Validez en cliquant sur « OK ». Si cela est possible, le programme vous demande si vous souhaitez éditer un contraste supplémentaire pour ce facteur / interaction. Si vous choisissez « Oui », la boite n°2 réapparait pour la saisie des coefficients du second contraste. Si vous choisissez « Non », la boite n°1 réapparait pour la saisie de contraste sur un autre facteur/interaction.

Lorsque tous les contrastes ont été saisis, cochez « Terminer » sur la boite n°1 et validez.

Remarques :

- Lorsque le logiciel détecte des résidus suspects pour une variable, il vous propose d'arrêter l'analyse pour cette variable.
- Il est possible d'analyser des plans d'expérience non généré par StatBox. Pour cela il suffit de remplir une feuille dispositif correspondant au type d'essai désiré et de reconstituer une feuille de saisie, ayant strictement la même structure qu'une feuille de saisie généré par StatBox, pour ce type de dispositif. Il est par exemple possible d'analyser des alpha-plans sortant des bornes de génération d'alpha plan dans StatBox.
- La présence d'une feuille de plan n'est jamais nécessaire pour réaliser une analyse.

## REGROUPEMENTS D'ESSAIS

### Pourquoi des regroupements ?

En expérimentation, les différences de classement des traitements sont généralement plus importantes d'un lieu à un autre, qu'à l'intérieur d'un même lieu (entre les blocs ou les répétitions d'un essai). Il est donc nécessaire de travailler en « réseau » d'essais et il vaut alors mieux augmenter le nombre d'essais, quitte à diminuer le nombre de blocs (de répétitions) pour chaque essai individuel.

Il faut donc considérer l'analyse de variance d'un essai comme une analyse critique des résultats, une validation de ceux-ci : l'examen des résidus, des erreurs (histogramme, cartographie, écart-type intra-traitement) et de l'interaction traitement*bloc, sont donc particulièrement important.

## Mise en œuvre

Si vous devez effectuer un regroupement d'essais, il faut lancer d'abord les différentes analyses de variance en cochant dans la boîte de dialogue du traitement, la case « Données pour les regroupement d'essais ». Les moyennes et les variances résiduelles apparaissent ainsi à la fin des résultats. Ces données devront être introduites dans le calcul final.

Choisissez un type de classeur correspondant au nombre de facteurs étudiés (de 1 à 3 facteurs) dans le menu « Nouveau » et validez. Dans le nouveau classeur, renseignez les informations indispensables au dispositif :

- Nombre de lieux d'expérimentation
- Libellé des facteurs et nombre de niveaux étudiés par facteur
- Noms des différents niveaux

Dans le menu regroupement, cliquez sur générer les feuilles de saisie. 2 feuilles sont alors créées : une feuille « Résiduelle » servant à l'introduction des variances résiduelles, des nombres de degré de liberté et nombres de blocs pour chaque essai et une feuille « Saisie » servant à l'introduction des moyennes. Renseignez correctement ces 2 feuilles.

Le fonctionnement du classeur est ensuite identique à celui des autres classeurs. Le déroulement des analyses est notamment semblable.

Si vous désirez faire des transformations, utilisez dans le menu Codage, l'option Transformation.

## Références

**PHILIPPEAU G. (1983).** Une exploitation des principaux paramètres statistiques élaborés lors de l'analyse des essais de variétés de céréales à l'ITCF en 1980, 1981 et 1982, PUBLICATION ITCF.

**GOUET J.P - PHILIPPEAU G. (1986).** Comment interpréter les résultats d'une analyse de variance ? PUBLICATION ITCF.

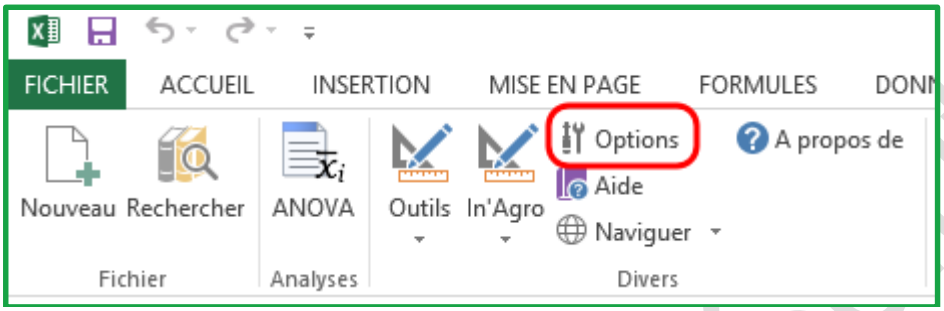
# STATBOX VEGETAL – PRISE EN MAIN

StatBox Vegetal est une extension des essais en agriculture permettant de gérer, de manière plus simple et ergonomique, les essais à 1 facteur.

## PREMIERS PARAMÉTRAGES

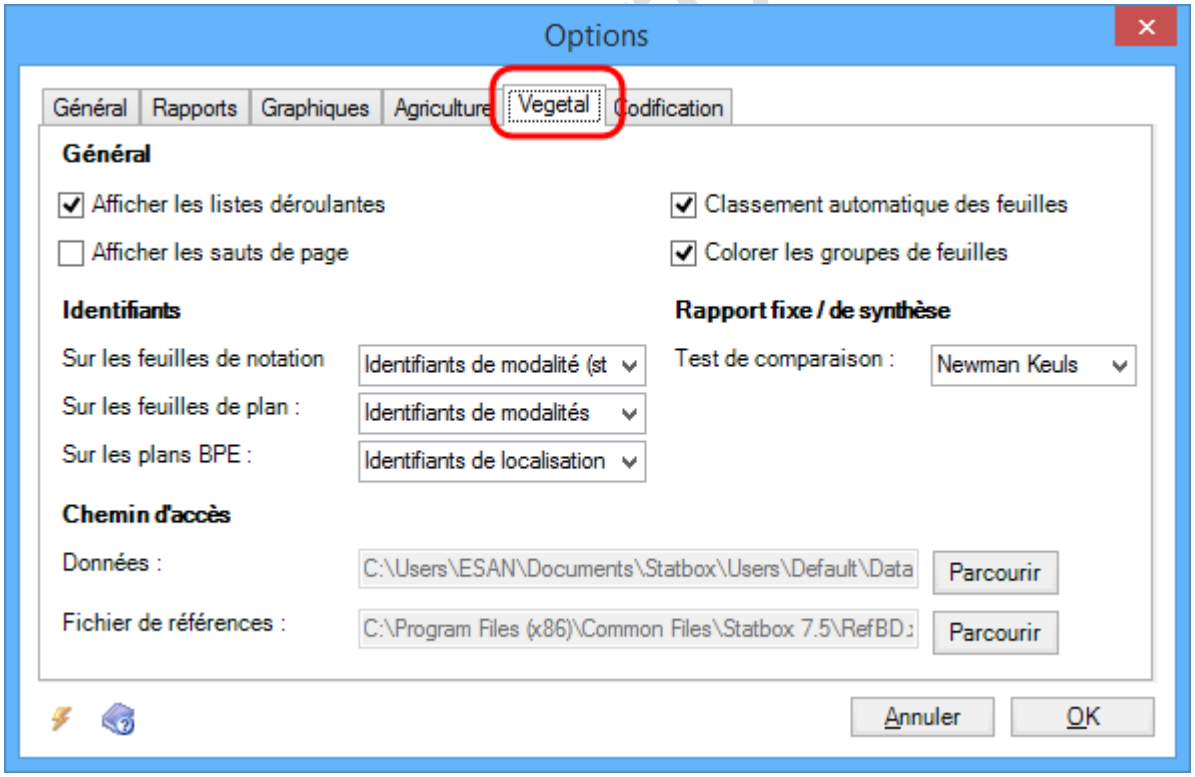
Avant la première utilisation, vous devez paramétrer le logiciel selon vos préférences.

Cliquez sur le menu StatBox Vegetal, et sur Options.




*Onglet « vegetal »*

Cliquez sur l'onglet Vegetal : les options apparaissent.




Vous avez le choix, entre-autre, de paramétrer avant l'utilisation :

- Les paramétrages de lecture (partie « Général »).
- Le mode d'identification des parcelles

- Le chemin de sauvegarde : si vous désirez utiliser un autre emplacement pour l'enregistrement des essais, dans la partie Chemins d'accès, cliquez sur le bouton Parcourir de la ligne Données et enregistrer votre nouveau chemin d'accès.

**Identification des parcelles dans le plan**

Dans les options, pour identifier les parcelles dans le plan, vous avez le choix entre 2 identifications différentes :

1. Localisation de la parcelle
2. Identification de modalités

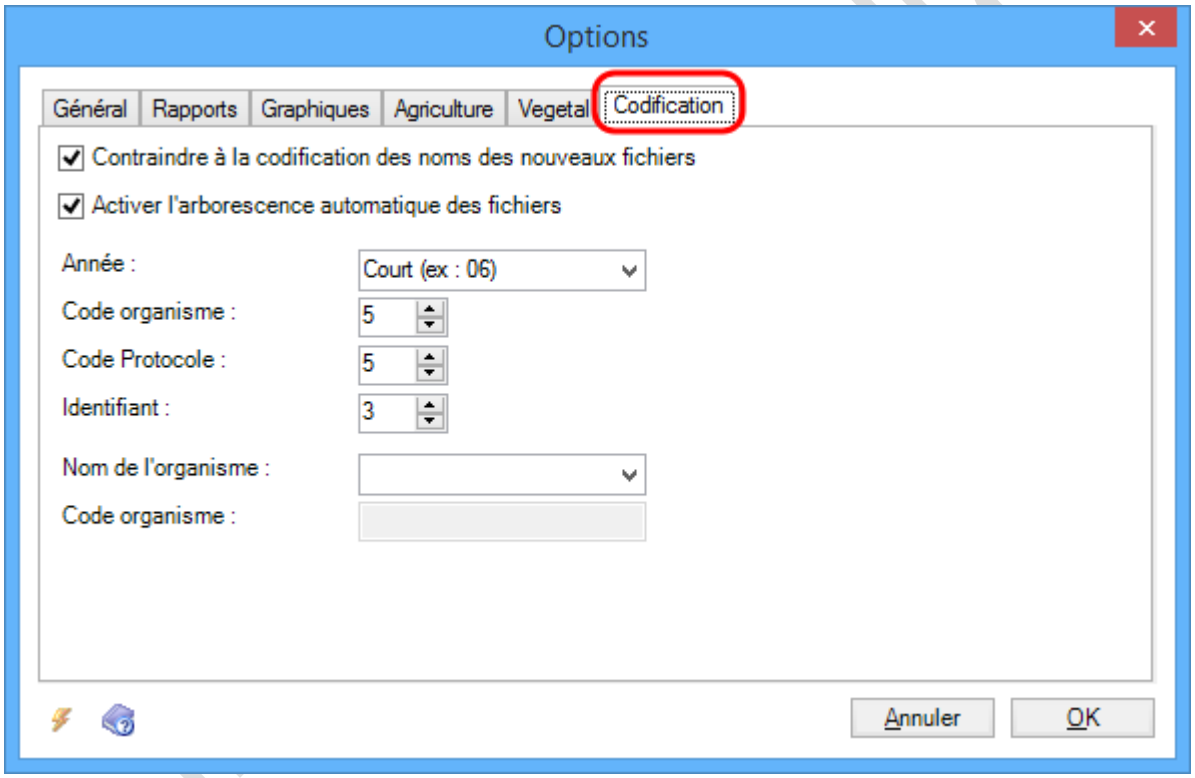
La localisation de parcelle permet de connaitre la position de la parcelle dans le plan. La parcelle 205 représentant la 5e colonne dans le bloc 2

L'identification de modalité permet de connaitre la modalité présente dans le plan. La parcelle 205 représentant la 5e modalité dans le bloc 2

Cette option n'est applicable que pour les essais issus des modèles Vegetal. Elle n'est pas accessible sur l'édition Vision 4.

En ce qui concerne les essais Simples ou les essais en regroupement, l'option utilisée est toujours la localisation de la parcelle quelle que soit l'option définie dans cet onglet.

*Onglet « Codification »*



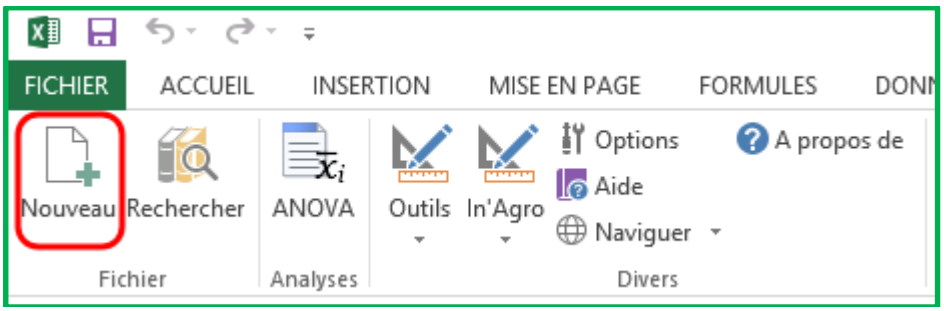

- *Contraindre la codification des noms des nouveaux fichiers* : activer la codification des fichiers, notamment en sélectionnant le nom de l'organisme. Le code de votre coopérative se génère automatiquement.

L'ensemble des options sera retenu pour toutes les utilisations de StatBox Vegetal.

## CRÉATION D'UN CLASSEUR

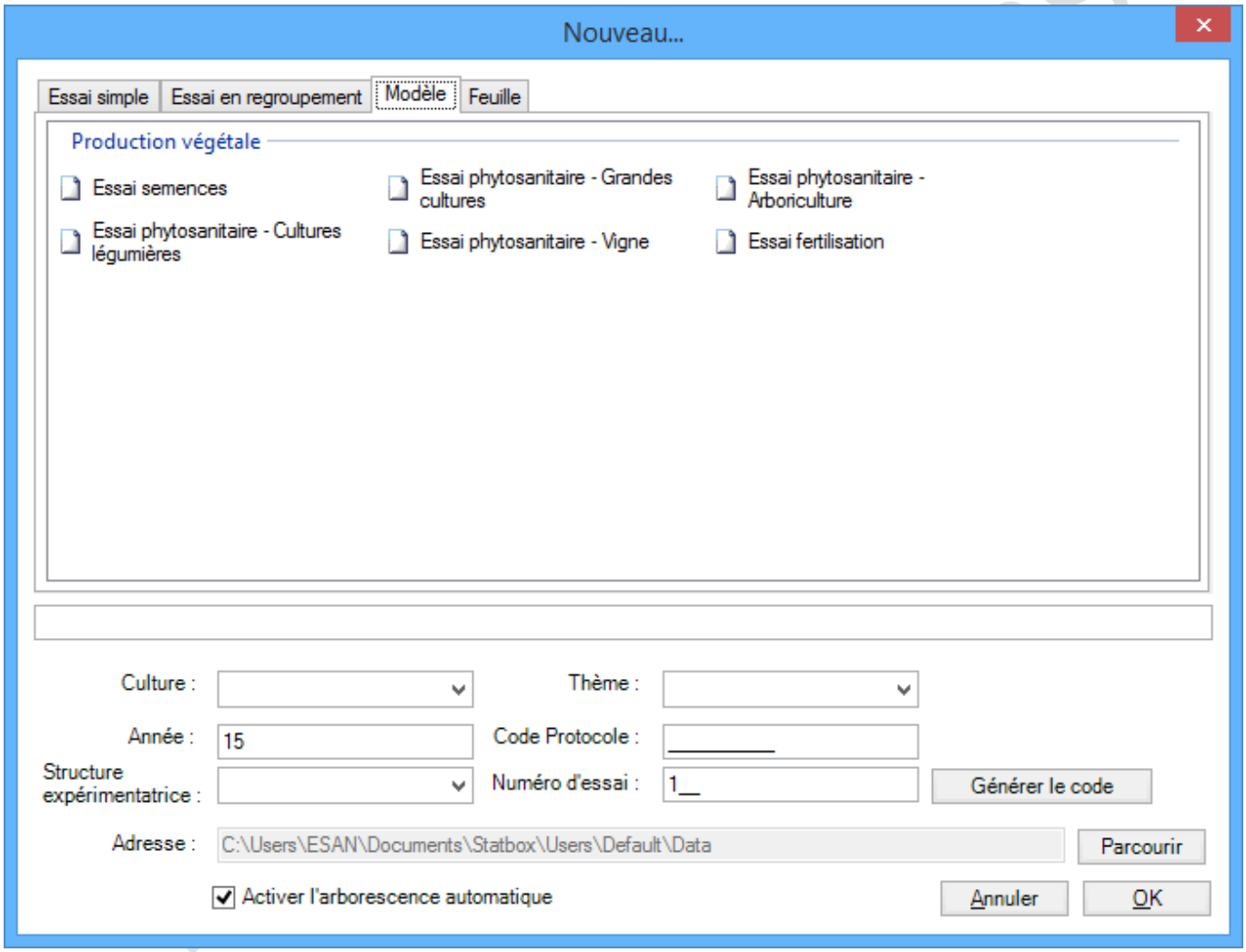
Ouvrez Microsoft Excel : le menu StatBox Vegetal s'affiche automatiquement dans la barre de menu.

Lancez Vegetal en cliquant sur le menu StatBox Vegetal puis Nouveau




Vous avez le choix entre 6 modèles de classeurs : Semences, phytosanitaire Grandes Cultures, phytosanitaire Vigne, phytosanitaire Arboriculture, phytosanitaire Cultures Légumières, et Fertilisation.

Chacun des classeurs a été adapté aux différents types de cultures et aux différentes thématiques : veillez donc bien à choisir le bon classeur.




Les classeurs que vous saisissez sont enregistrés par défaut dans le répertoire sélectionné au préalable dans Options (cf. PREMIERS PARAMETRAGES) dans le répertoire correspondant à la culture choisie. L'option « activer l'arborescence automatique » propose un classement par Année / Classeur / Culture / Thème. Il sera ensuite plus aisé de retrouver les fichiers saisis.

Si vous ne désirez pas l'arborescence automatique, décochez l'option.

**Important** : chaque dossier est constitué d'un fichier Excel (suffixe .xls ou xlsm) et d'un fichier texte (suffixe « .txt »). Veillez à les conserver ensemble.

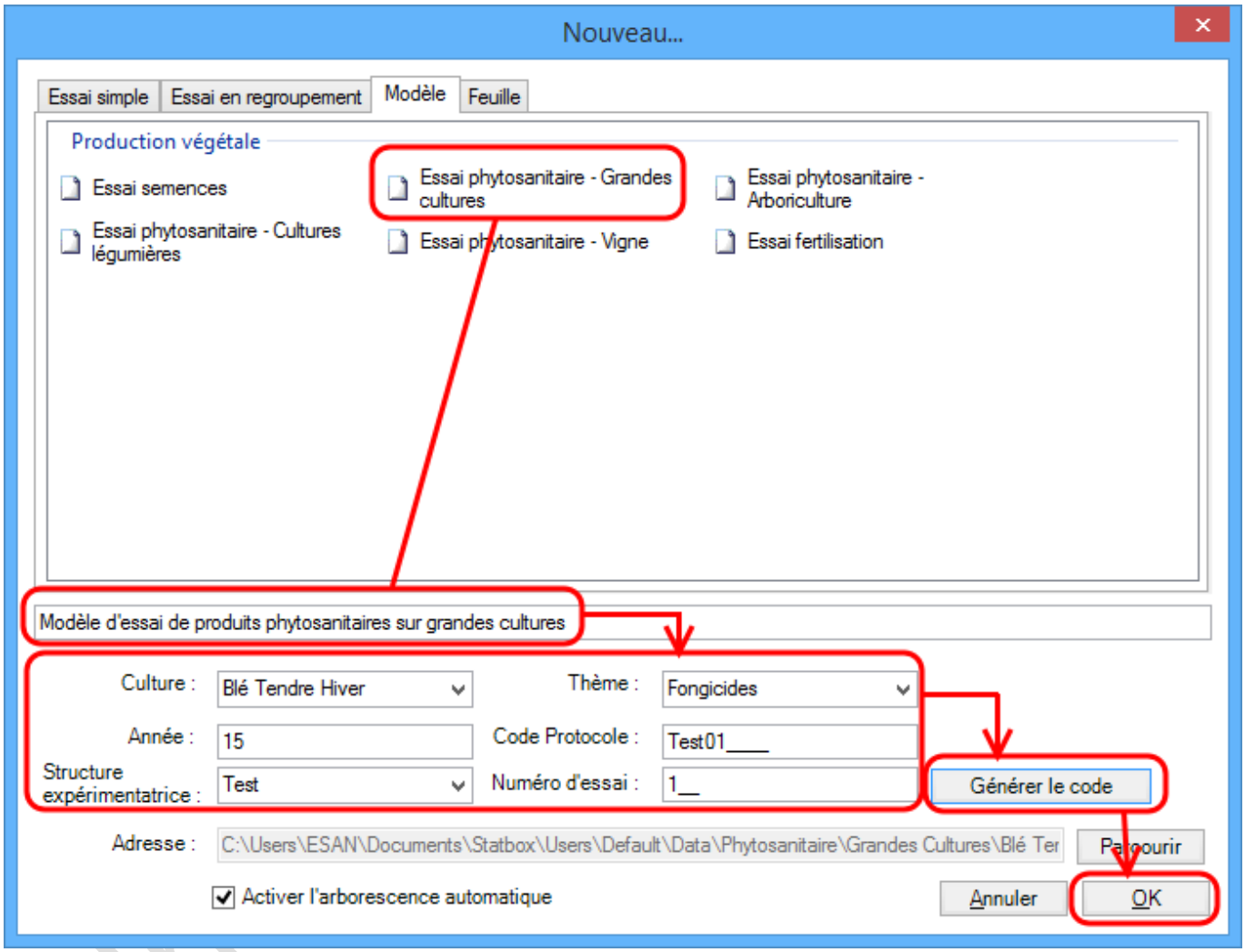
**Important** : comme tout fichier Excel, si vous sortez d'un classeur sans l'enregistrer, vous perdez le fruit de votre travail. De même, faites attention à ne pas écraser un classeur par un autre du même nom.

Par la suite, nous prendrons l'exemple d'un essai Fongicide sur blé tendre d'hiver (Essai phytosanitaire sur Grandes Cultures).

1. Lancez un nouvel Essai (depuis Excel, StatBox Vegetal, Nouveau).
2. Choisissez le type d'essai

Les menus déroulants vous permettent de sélectionner rapidement la culture et le thème.

3. Une fois le code protocole et le numéro d'essai saisis, cliquez sur **Générer le code**. Le code va se générer automatiquement. Ce code devient le nom de votre fichier il vous permettra de retrouver un essai par le seul nom du classeur.
4. Cliquez sur « OK » pour lancer la création du classeur correspondant à votre essai.




## LA SAISIE DANS LES CLASSEURS

**Important** : utilisez à chaque fois que c'est possible les listes de choix présentes dans un bon nombre de cellules. Cela permettra de limiter les erreurs de saisie (mauvaise orthographe d'un produit) et ainsi de pouvoir faire par la suite des recherches fructueuses sur certains champs.

**Important** : la saisie s'effectue uniquement dans les zones blanches et bleues (les cellules bleues devant être obligatoirement renseignées).

**Important** : ne supprimez jamais une feuille de classeur autrement que par le menu « StatBox Vegetal – Outils – Suppression de feuilles ».

## Introduction

Les classeurs que vous créez vont être conservés dans la mémoire de votre ordinateur tant que vous ne le supprimez pas. Veillez donc bien à remplir le plus complètement possible tous les renseignements que l'analyse minutieuse d'un essai agronomique peut nécessiter.

## Présentation d'un classeur

À l'ouverture, un nouveau classeur se compose de 6 feuilles (ou « onglets ») : Site expérimental, Modalités, Plan, Rapport fixe, Expertise et TNT Notations. Ne pas séparer ou supprimer ces feuilles.

### 1- Feuille « Site expérimental »

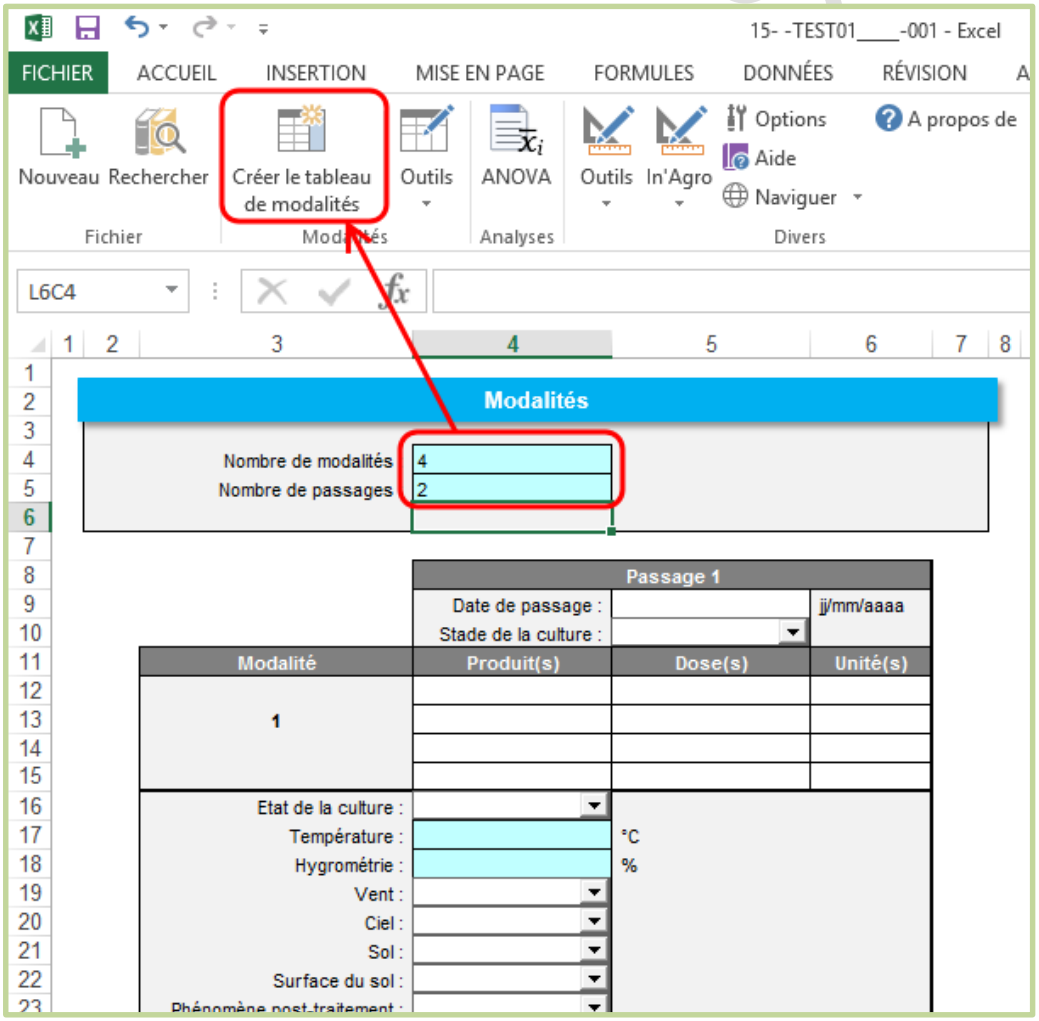
Cette feuille vous permet de remplir toutes les caractéristiques de votre essai en dehors des modalités testées et du plan.

**Important** : le type de dispositif vous permet de choisir si votre dispositif est un dispositif en Blocs, Randomisation, Carré Latin ou Alpha plan. Ce choix est déterminant pour la création de votre plan. **Vous ne devez jamais le modifier après avoir créé le Plan**.

### 2- Feuille « modalités »

Saisissez le nombre de modalités, le nombre de passages (ou dates de traitement).

Cliquez sur le menu StatBox Vegetal, Modalités, Créer le tableau de modalités.



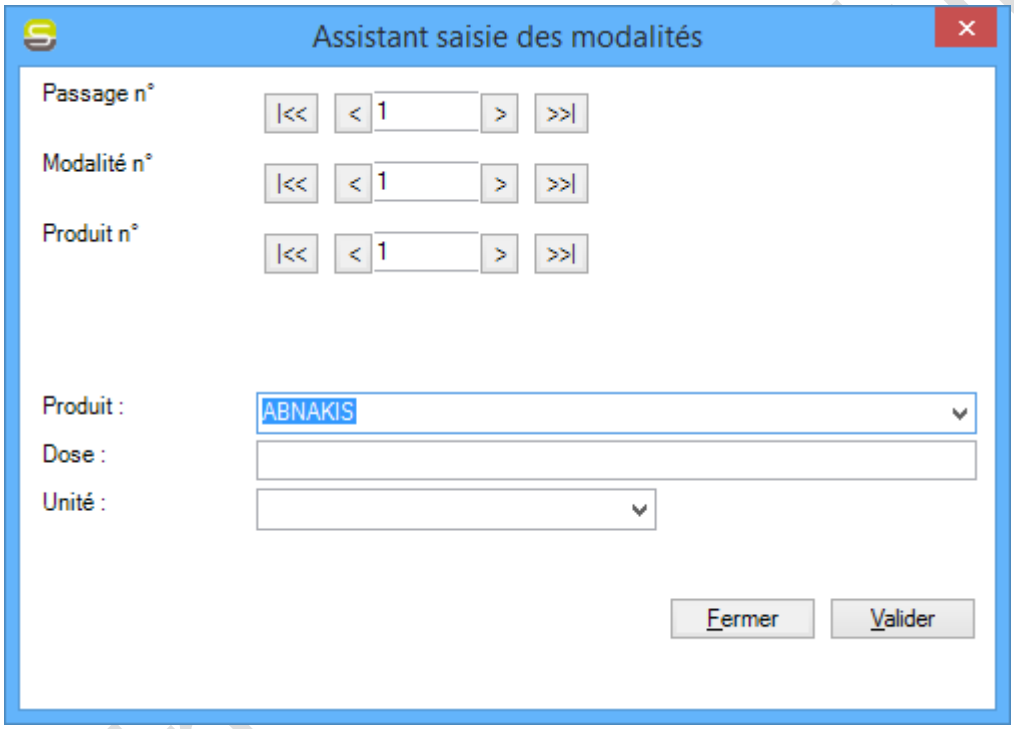
Dans le tableau de saisie :

- Chaque modalité peut être saisie sur 4 lignes (soit un mélange de 4 produits maximum)
- Pour chaque produit (ou variétés pour les Semences) vous pouvez saisir la dose (densité pour les Semences)
- Pour chaque dose (hors Semences), vous pouvez saisir l'unité.
- Pour les Semences vous pouvez saisir le TS

**Attention** : ne pas oublier de saisir les dates de passage (date de semis pour les Semences) et stade de la culture, informations indispensables pour la bonne compréhension d'un essai.

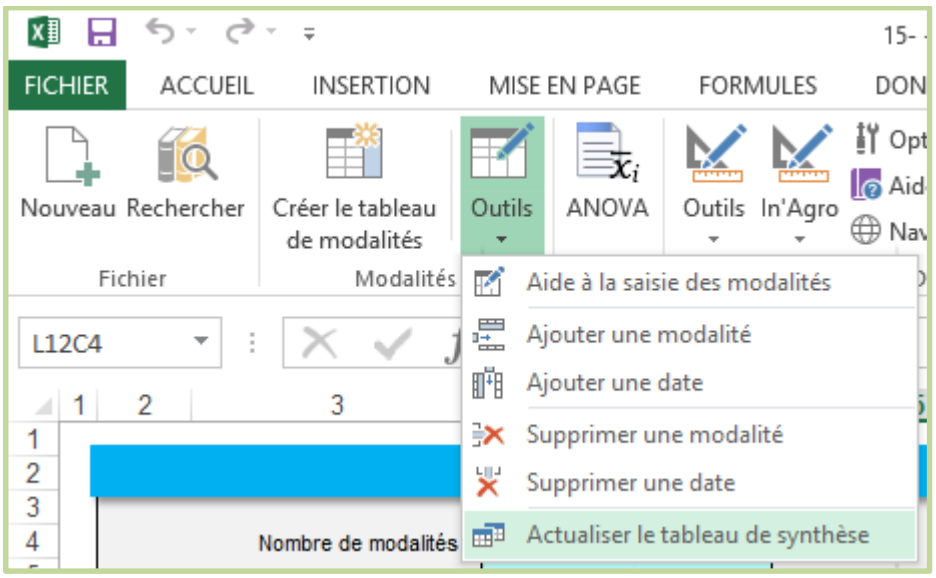
Afin de normaliser la saisie et minimiser les erreurs, un assistant « Saisie des modalités » vous est proposé dès la création du tableau de saisie :

- Les flèches vous permettent de vous déplacer
  - Dans les passages
  - Dans les modalités d'un passage
  - Dans les produits d'une modalité
- Les flèches doubles vous permettent d'aller au début ou à la fin
- Les flèches simples vous permettent de vous déplacer
- Le bouton Valider vous permet d'insérer la saisie dans la feuille

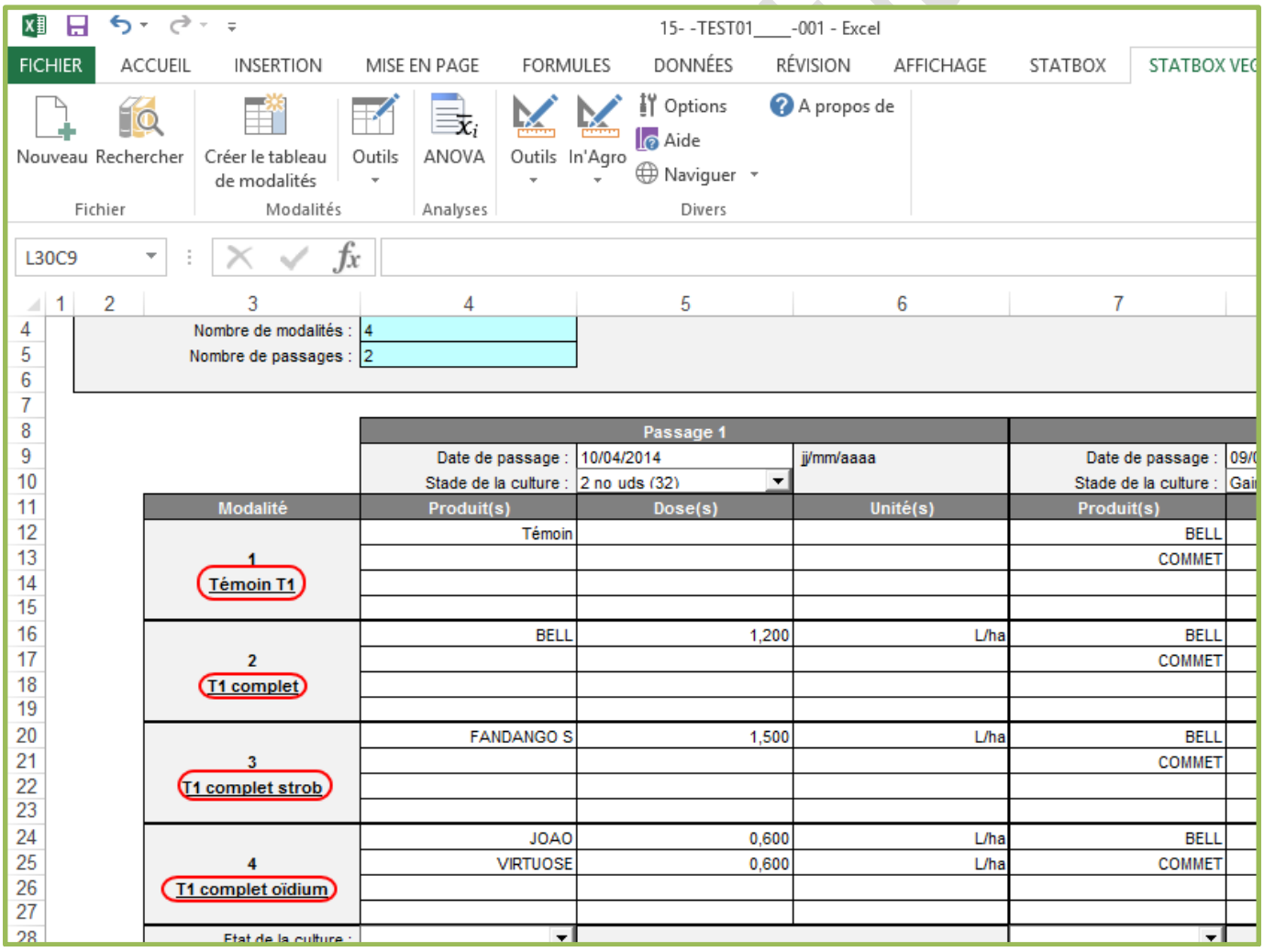



Une fois votre saisie terminée à l'aide de l'assistant, cliquer sur fermer. Une fenêtre vous demandera si vous voulez actualiser le tableau de synthèse. Répondre **oui**.

**Important** : une fois la saisie terminée, actualisez le tableau de synthèse. Si vous n'avez pas utilisé l'assistant de saisie n'oubliez pas d'actualiser le tableau de synthèse en cliquant sur le menu StatBox Vegetal, Modalités, (Outils), Actualisez le tableau de synthèse.




Sous les numéros de modalité, vous avez la possibilité de « Nommer » les modalités (**Etiquettes**). Sachez dans ce cas, que c'est ce nom (étiquette) qui figurera sur le plan et non pas le détail des produits, par contre les 2 apparaitrons sur le rapport.




**NB** : dans le cas d'une suppression d'une modalité ou d'une date de passage vous avez la possibilité de choisir le numéro de la modalité ou du passage que vous souhaitez supprimer. Attention, ce choix n'est possible que dans le cas d'une suppression, dans le cas d'un ajout celui-ci se fera toujours en dernier. Une fois un ajout (ou suppression) effectué(e), pensez à mettre à jour le tableau de synthèse.

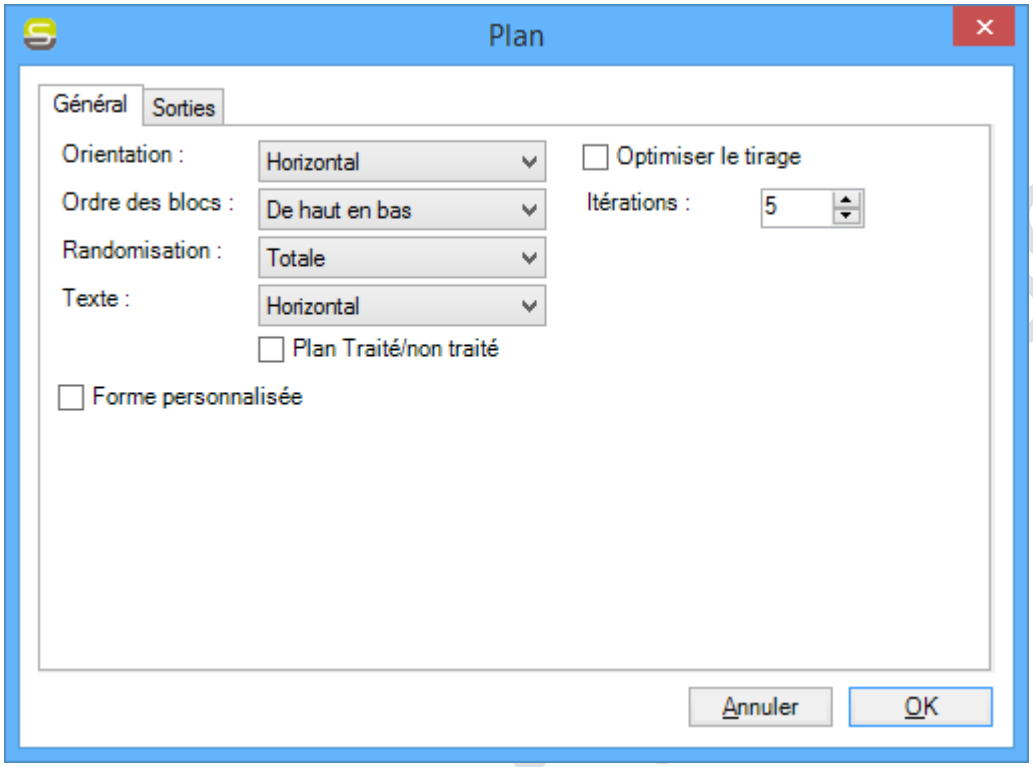
### 3- Feuille « Plan »

**Attention** : le plan doit être obligatoirement réalisé (même s'il est fictif) pour pouvoir accéder aux feuilles suivantes.

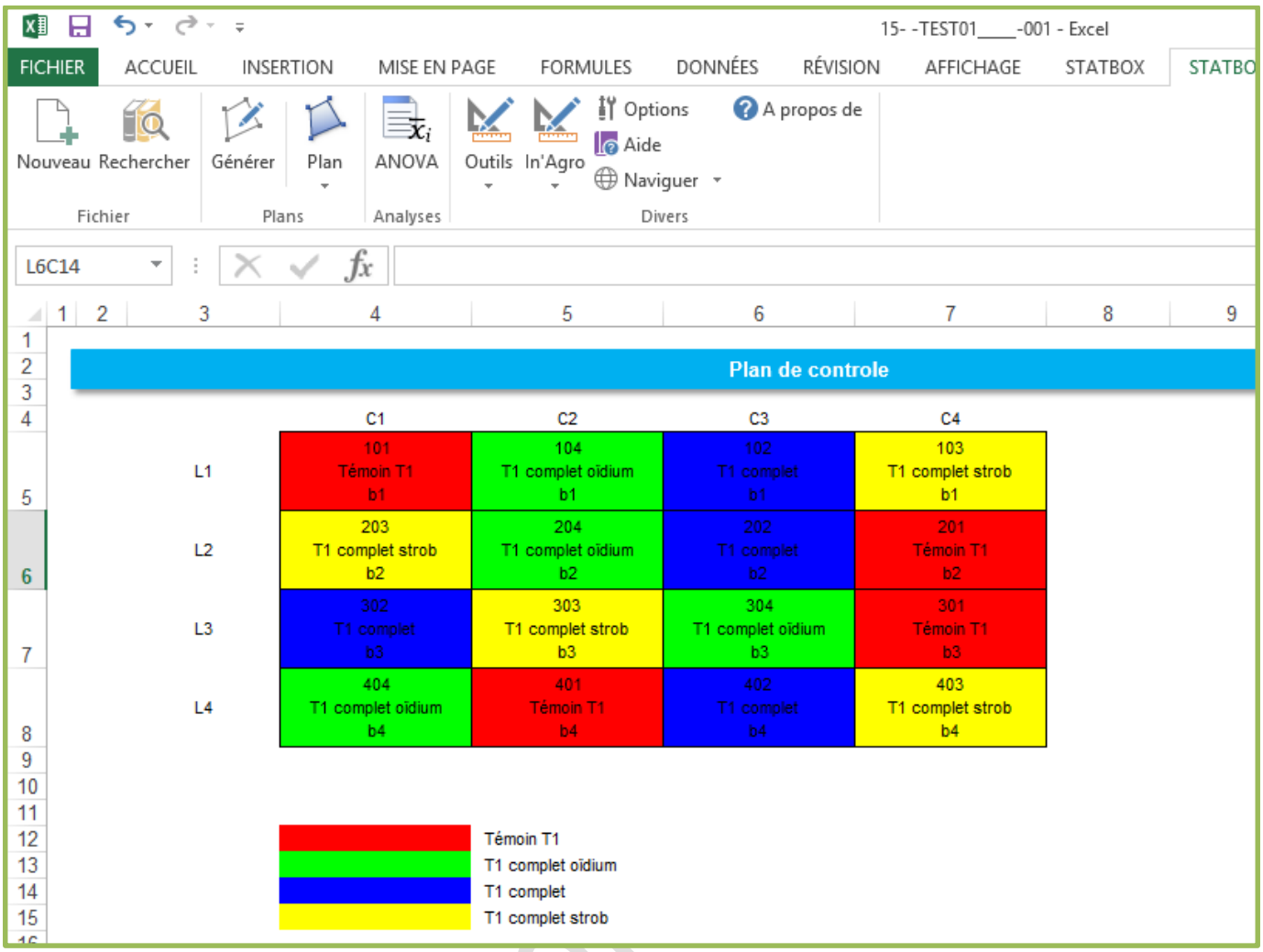
**Important** : avant de construire votre plan, il est indispensable d'avoir renseigné dans la feuille « Site expérimental », le type de dispositif sélectionné (blocs, randomisation, carré latin, alpha plan).

Positionnez-vous sur la feuille « Plan ». Renseignez le nombre de répétitions (blocs). Cliquez sur le menu « StatBox Vegetal », Plan, Générer le plan. Une fenêtre d'options va alors s'ouvrir : vous avez la possibilité de choisir l'orientation du plan, d'effectuer une randomisation totale ou partielle (etc.).

**Attention**, pour choisir l'Alpha-plan, il faut au minimum 10 modalités et 4 blocs.




Une fois le plan généré, une feuille appelé « plan de contrôle » va se générer : cette feuille vous permet de visualiser via des codes couleurs l'organisation du plan.




Si le tirage aléatoire ne vous convient pas, vous avez toujours la possibilité soit :

- de générer une nouvelle fois le plan en reproduisant la procédure préalablement décrite.
- De créer un plan personnalisé : cliquez sur le menu « StatBox Vegetal », Plan, Personnaliser le plan. Une feuille « PlanPS » va être créé. Vous pouvez créer vous-même votre plan en effectuant un « couper-coller » ou « copier-glisser » en vous positionnant sur les modalités que vous voulez déplacer. Déplacez les parcelles du plan situé en haut vers le plan vide situé en dessous. Pour que ce plan soit bien pris en compte, vous devez impérativement actualiser le plan à partir du Menu StatBox Vegetal, Plan, Actualiser.

*Pour plus d'informations sur le plan, se reporter au paragraphe* ***Personnalisation de la position des parcelles dans le plan de l'essai*** *du chapitre* ***plan*** *de la section* ***Les essais en agriculture***

### *Numérotation des parcelles :*

Les numéros des parcelles comportent 3 chiffres :

Par défaut, le premier chiffre correspond au bloc, les 2 suivants à la modalité.

Exemple : 308 → Modalité 08 du bloc 3

*Pour plus d'informations sur la numérotation des parcelles, se reporter au paragraphe* ***Identification des parcelles dans le plan*** *du chapitre* ***Premiers Paramétrages***.

## 4- Feuille « Rapport Fixe »

La feuille du rapport fixe est la feuille du classeur dans laquelle vous pouvez ajouter au fur à mesure toutes les analyses statistiques que vous jugez d'intérêt. En plus de ces résultats, le rapport fixe reprend les principales caractéristiques de l'essai (si vous les avez renseignés dans les différents onglets).

Les rapports personnalisés sont des rapports qui peuvent être transitoires et effacés si vous le souhaitez.

## 5- Feuille « Expertise »

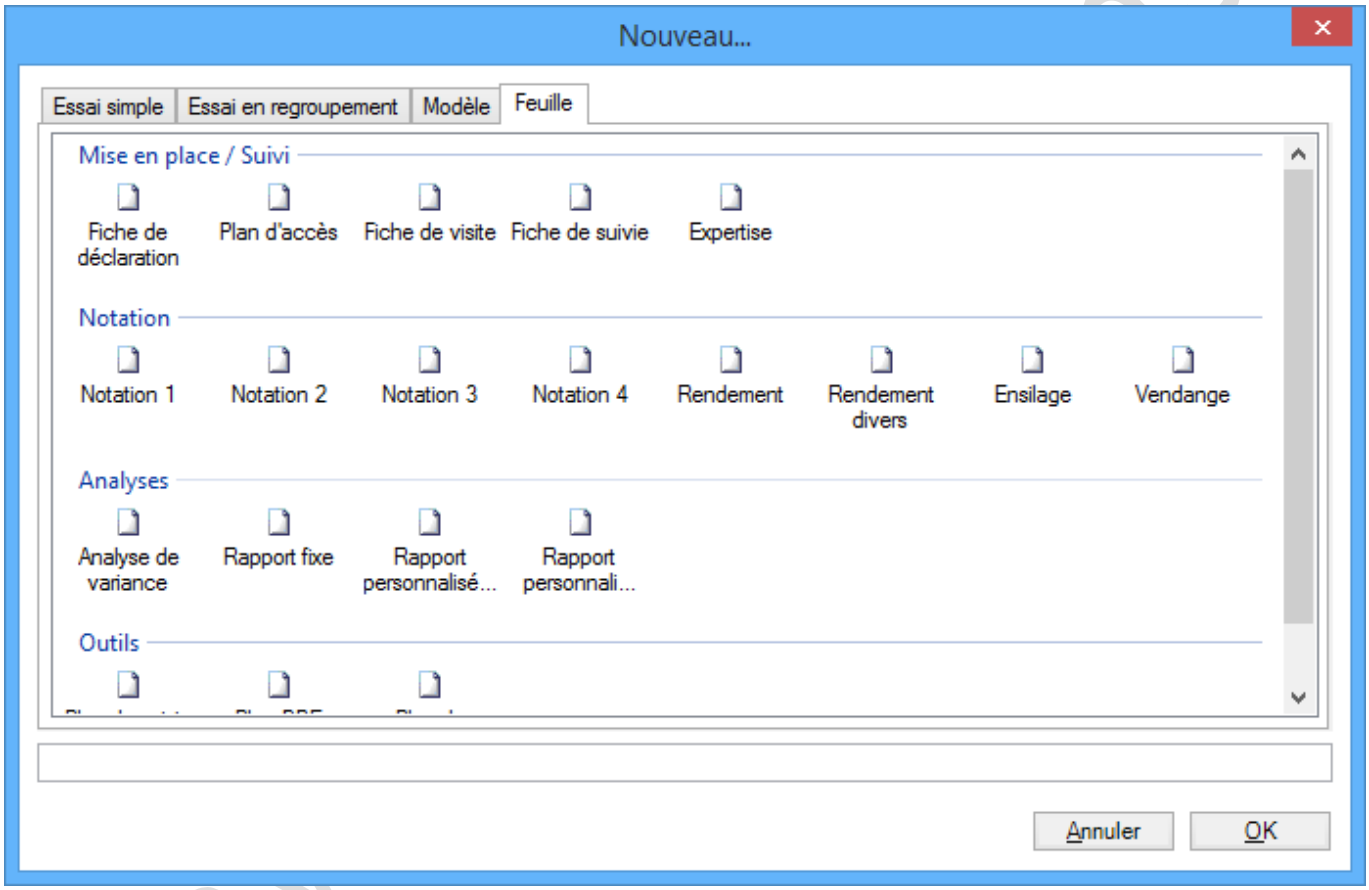
La feuille Expertise va vous permettre d'associer des commentaires à l'essai mis en place à la fois sur

- la qualité du dispositif, de l'application, de l'observation, de l'analyse statistique,
- le contexte général
- le niveau d'intérêt de l'essai

Remplir la feuille expertise est essentiel tant au niveau de l'appréciation de l'essai que de la traçabilité.

## 6- Feuilles de Notation

Les autres feuilles (mise en place, notations, etc.) sont accessibles via le menu « StatBox Vegetal », Nouveau, onglet Feuille.




### *6a- Les feuilles de notations*

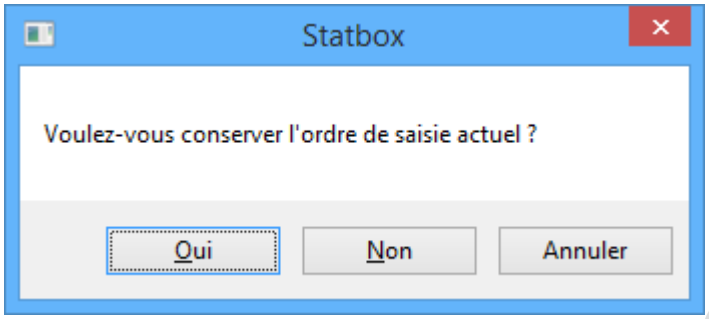
StatBox Vegetal fourni pour chacun des modèles les feuilles de notations suivantes :

- Notation 1 = 1 note par variable
  - type efficacité (ex : efficacité sur gaillet = 8 / 10)
- Notation 2 = plusieurs notes par variables
  - type infestation (ex : 20 plantes notées pour le piétin verse)
- Notation 3 = plusieurs organes par individu
  - type maladie blé (ex : notation septoriose sur 20 plantes ou individus et sur plusieurs étages foliaires F1, F2…)
- Notation 4 = 1 note par modalité
  - 1 note pour l'ensemble des répétitions (ex : analyse qualité). Les variables dans cette feuille de notation ne peuvent pas faire l'objet d'une analyse de variance.

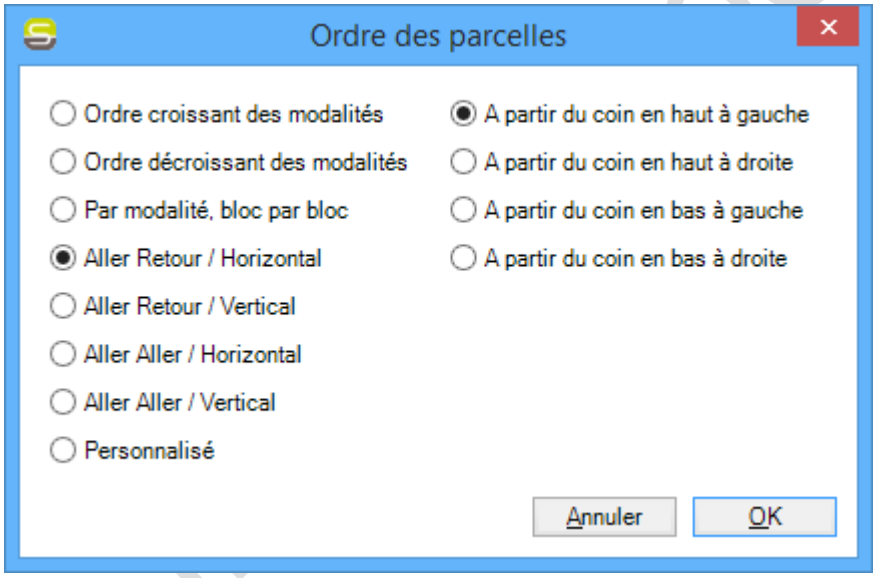
- Rendement
    - 11 variables fixées avec calculs automatiques
- Rendement divers
    - variables au choix
- Ensilage
    - 5 variables fixées avec calculs automatiques
- Vendange
    - 10 variables fixées avec calculs automatiques

Ordre de saisie

Lors de la création de toute feuille de notation, un message apparait pour vous demander si vous souhaitez garder ou changer l'ordre de saisie des parcelles.




Si vous souhaitez en changer, répondez non, une autre boite de dialogue va alors s'ouvrir avec les choix possible, à vous de cocher l'ordre qui vous convient.




*Se reporter au chapitre **Gestion de l'ordre de saisie** de la section **Essais en agriculture**.*

Création des variables et du tableau de saisie

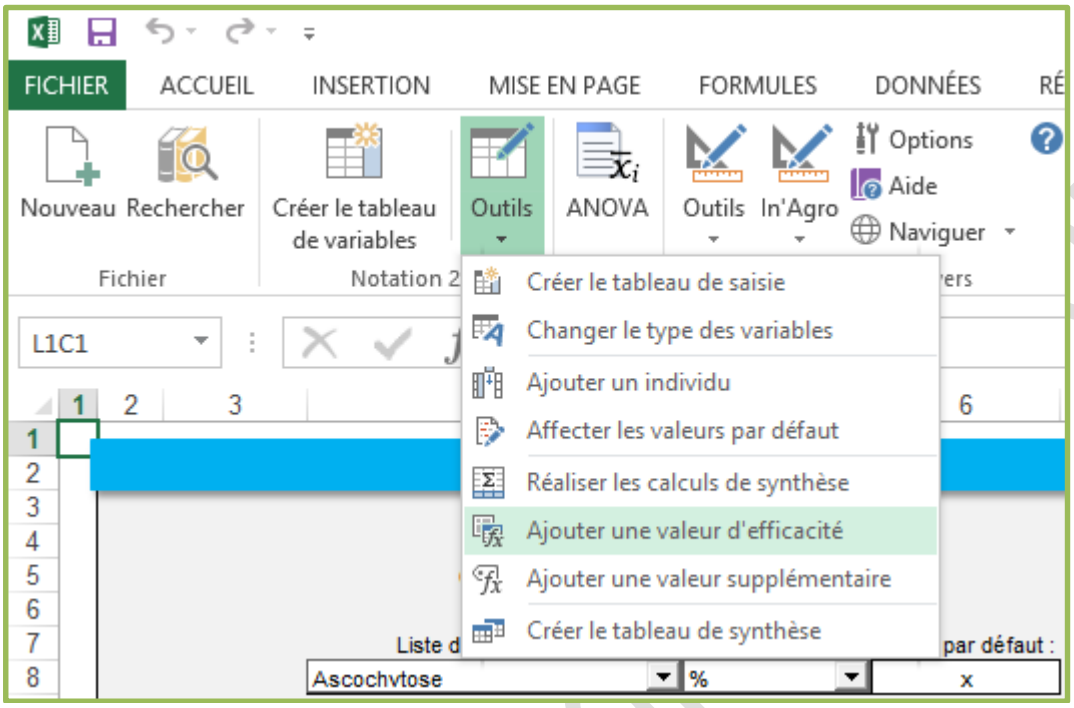
Sur les feuilles de notations, pour créer le tableau de variables, le tableau de saisie et lancer les analyses statistiques, vous devez passer par le menu StatBox Vegetal.

- Notation 1 et Notation 4 : définissez le nombre de variable puis StatBox Vegetal > Créer le tableau de saisie
- Notation 2 et Notation 3 : définissez le nombre de variable et d'observations (et d'organes pour Notation 3) puis StatBox Vegetal > Créer le tableau de variables. Saisissez les données des variables puis StatBox Vegetal > Outils > Créer le tableau de saisie
- Rendement Ensilage et Vendange : définissez les données de la parcelle puis StatBox Vegetal > Créer le tableau de saisie

- Rendement divers : définissez le nombre de variables et les données de la parcelle puis StatBox Vegetal > Créer le tableau de saisie

Calcul d'efficacité

Pour les feuilles de notation 2 et 3, vous avez la possibilité en plus de la fréquence, IOA et de l'intensité d'avoir le calcul automatique de l'efficacité, il vous suffit après avoir fait le tableau de synthèse de retourner dans le menu StatBox Vegetal et de cliquer sur « Ajouter une valeur d'efficacité » et d'entrer le numéro de la parcelle témoin, le calcul apparait au bout du tableau de saisie, ne pas oublier d'Actualiser le tableau de synthèse pour que ce calcul apparaisse




Astuce de saisie dans les tableaux de notation

Sélectionnez la zone de saisie puis :

- appuyez sur « Entrée » pour aller de bas en haut.
- appuyez sur « tabulation » (touche ↹) pour aller de gauche à droite

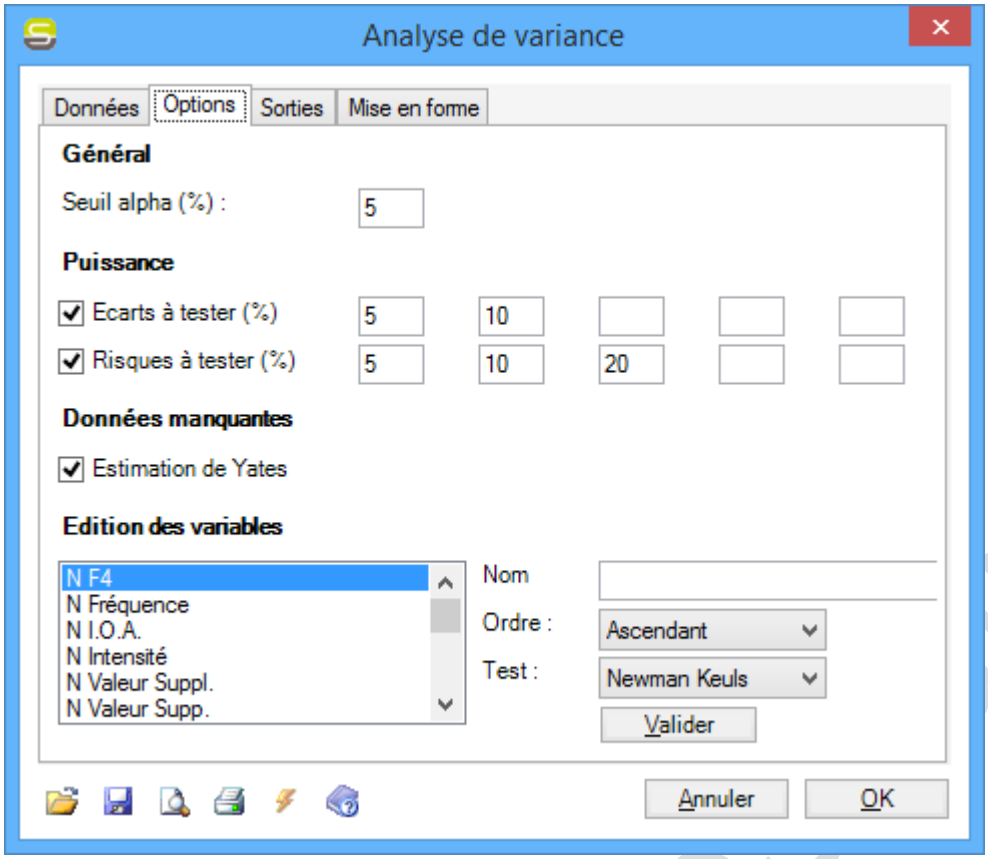
## ANALYSE STATISTIQUE

*Pour plus d'informations sur l'analyse de variance, se reporter au chapitre* ***Analyse de variance*** *de la section* ***Essais en agriculture****.*

### Estimation des variables

Lorsque vous choisissez les variables à analyser, dans la boite de dialogue l'onglet Options vous permet de choisir le test souhaité (Newman Keuls ou Bonferroni) ainsi que de choisir l'ordre de classement des classes statistiques (croissant ou descendant). Le test employé est spécifié sur le rapport.

Par défaut, les analyses sont faites avec le test de Newman & Keuls et les classes sont dans l'ordre croissant.

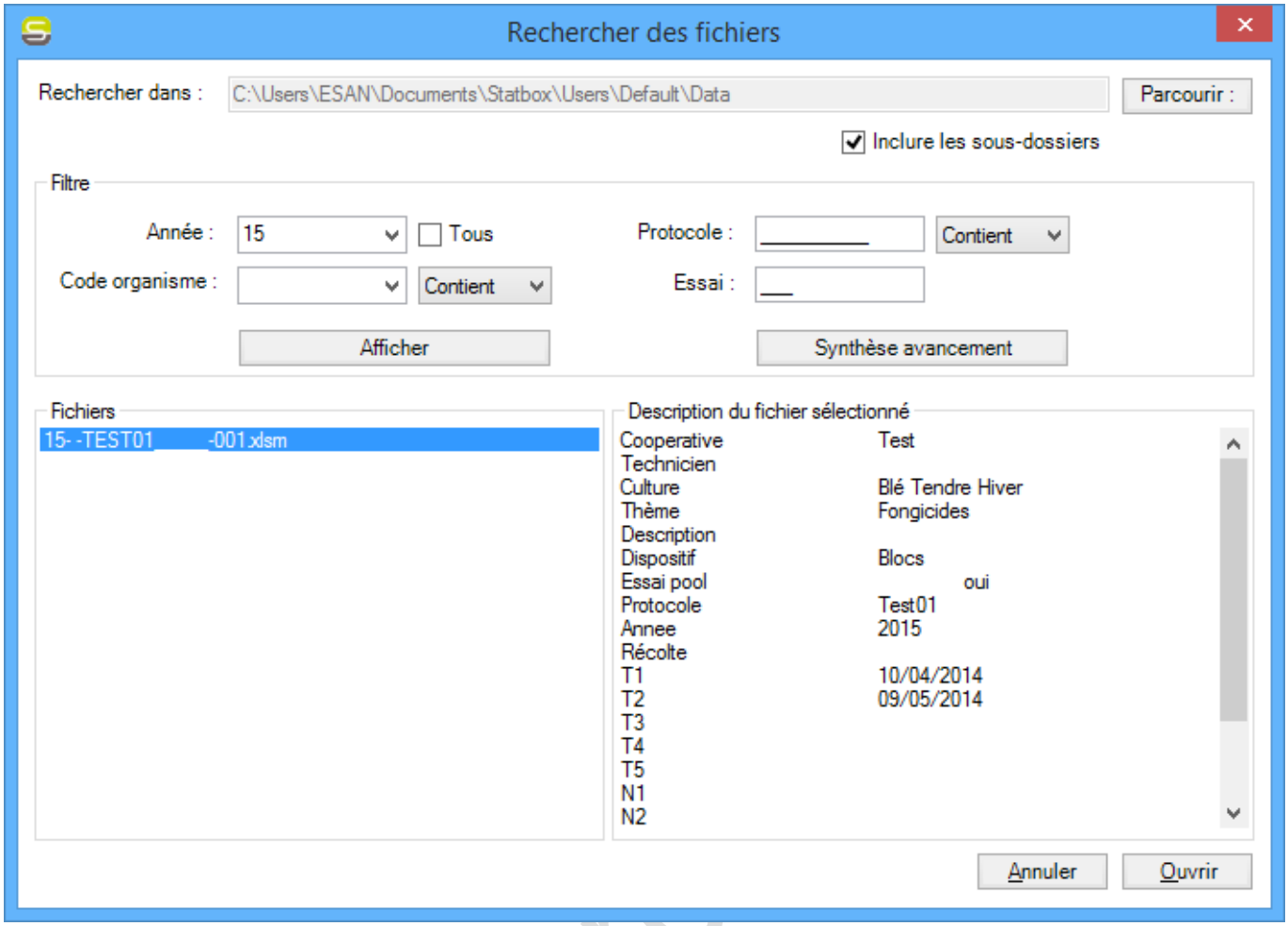



### Les rapports

Les rapports StatBox (histogrammes des résidus, cartographie…) sont enregistrés par défaut dans le classeur, il est toutefois possible de supprimer ces feuilles de façon à alléger les classeurs.

## RECHERCHER UN ESSAI

Cette fonction vous permet de rechercher un essai (différents filtres) mais aussi de voir l'état d'avancement d'un essai sans avoir à ouvrir celui-ci.




## AUTRES FEUILLES

D'autres feuilles sont à votre disposition :

- Fiche de déclaration
- Plan d'accès
- Fiche de visite
- Fiche de suivie

## RAPPELS STATISTIQUES

### Seuils de significativité indiqués sur le rapport :

- 0.1 = Non significatif
- 0.005 à 0.01 = Significatif
- < 0.001 = Hautement significatif

### Les 2 tests statistiques proposés :

- Newman & Keuls : PPAS (plus petite amplitude significative)
- Bonferroni : PPDS (plus petite différence significative)

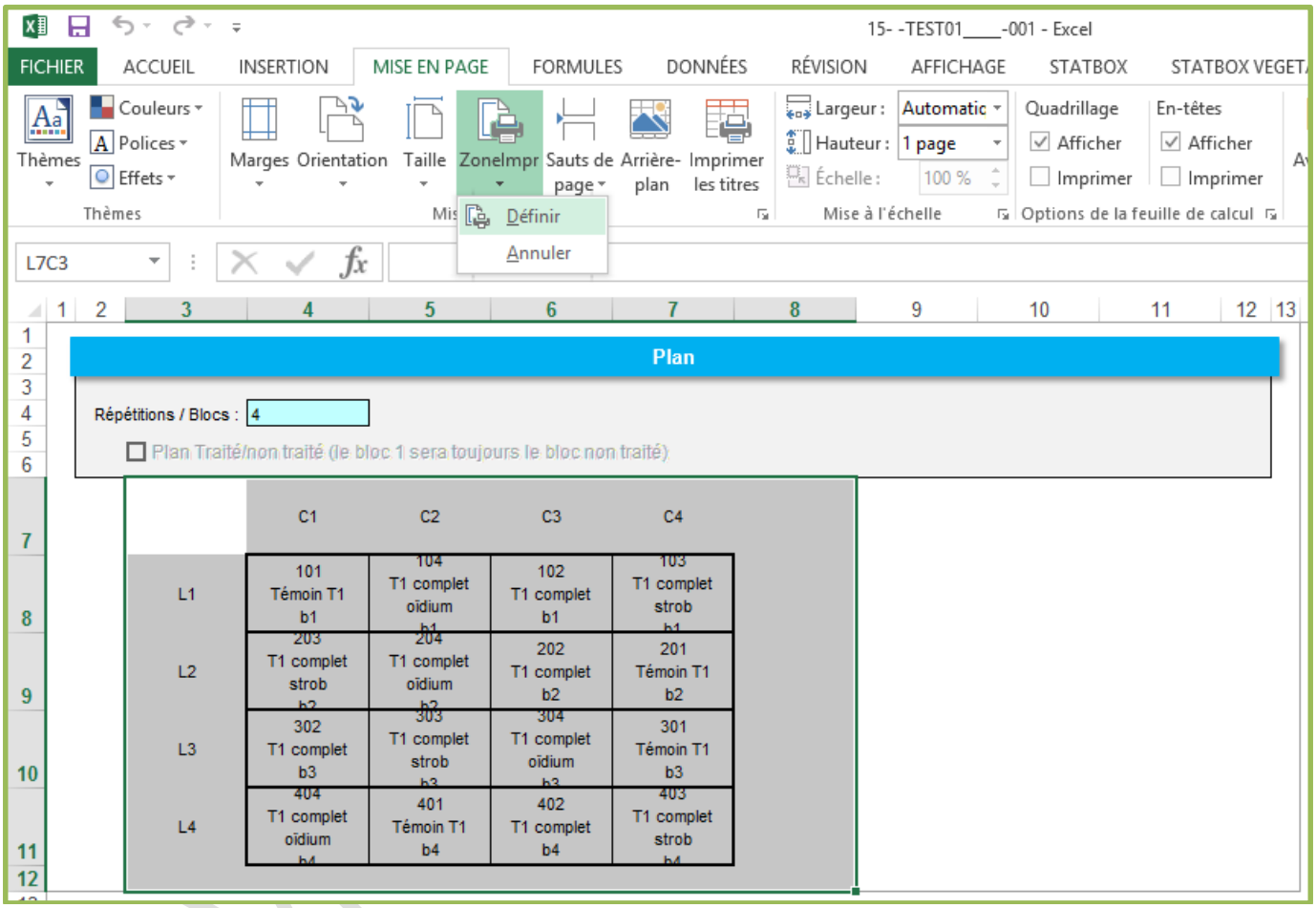
## Copier des données dans les essais

Les données extraites d'un autre fichier Excel doivent <u>toujours</u> être collées en faisant **<u>collage spécial</u>** puis Valeur

## Impression du plan (ou du Rapport fixe)

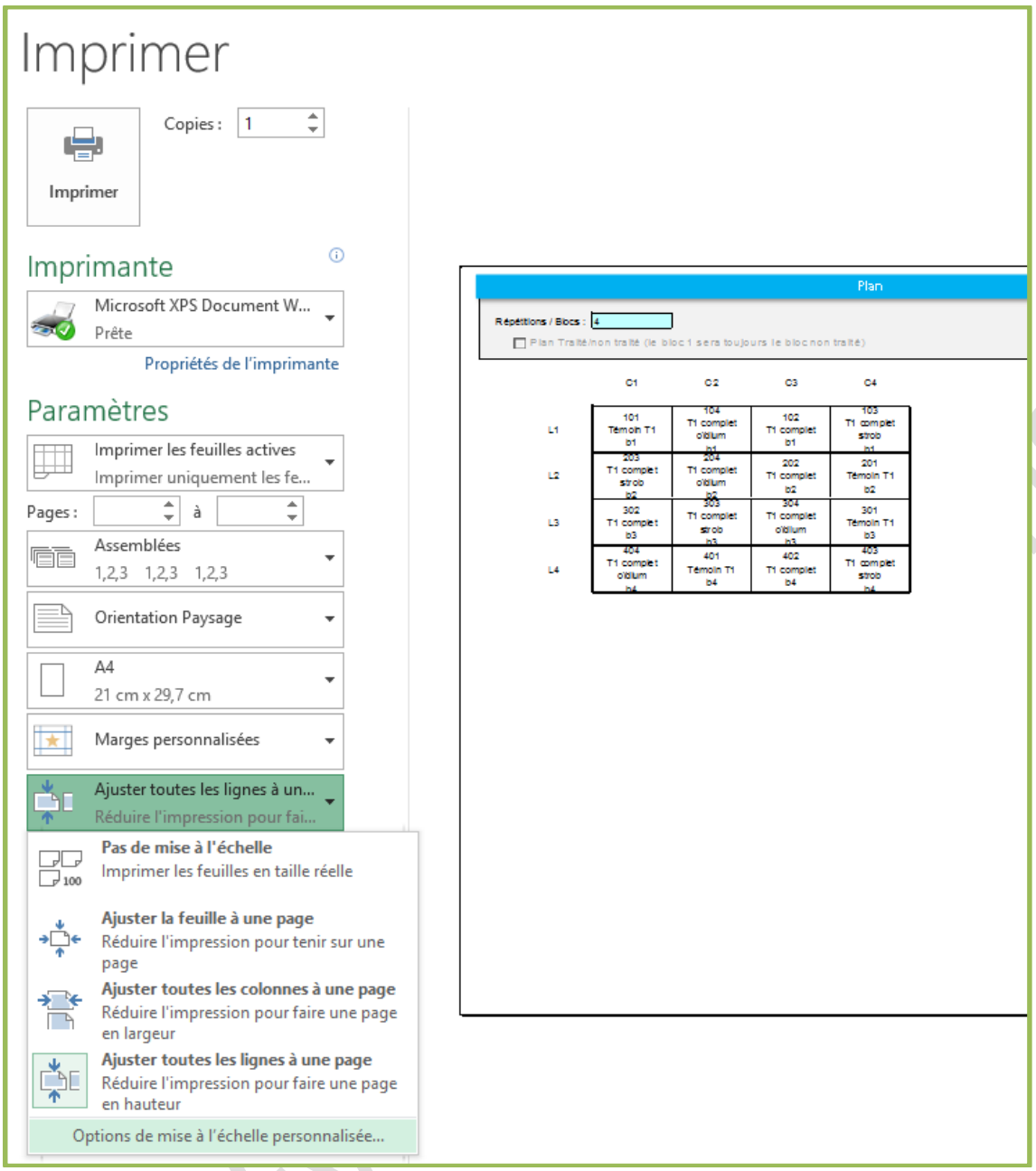
Si le plan est trop grand pour une bonne impression :

Si l'on veut exclure le bandeau lors de l'impression, sélectionner la zone de la feuille Plan qu'on souhaite imprimer puis Mise en page > ZoneImpr > Définir

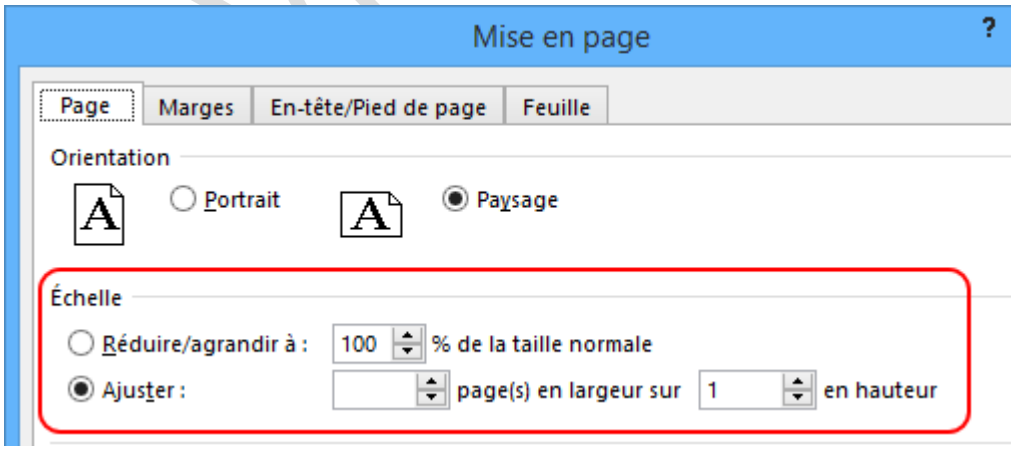



*Pour imprimer*

Sur Office 2010/2013 : menu Fichier > Imprimer > Ajuster toutes les lignes à une page > Options de mise à l'échelle personnalisée :




Puis ajustez à votre guise :




Sur Office 2007 : bouton Office > Imprimer > Aperçu avant impression > Mise en page : ajustez à votre guise

## Impression d'un rapport complet

Ces rapports ne peuvent pas être mis en page pour une impression. Comme pour le plan ou le rapport fixe, vous devez ajuster les paramètres d'impression

## Modifications des produits/variétés dans la feuille Modalité

Si vous modifiez les produits/variétés dans la feuille Modalités (ou que vous rajoutez des étiquettes sur les modalités) alors que vous avez déjà créé un Plan :

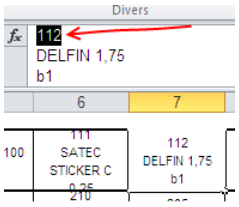
- N'oubliez pas de mettre à jour le tableau de synthèse (Outils > Actualiser le tableau de synthèse)
- Et de mettre à jour le plan (Plan > Rafraichir le plan)

Attention : ceci n'est valable que si vous ne modifiez pas le nombre de modalités. Dans le cas contraire, il vous faudra recréer le plan.

## Déplacement de certaines parcelles dans la feuille Plan

Si vous avez un nombre restreint de parcelle à repositionner dans votre plan :

Modifiez le numéro de parcelle des parcelles concernées, sans modifier ni supprimer le libellé de la micro-parcelle puis Rafraichissez le plan (Plan > Rafraichir le plan)




Attention :

- Contrôlez que toutes les micros-parcelles sont bien dans le plan
- Pour un repositionnement de beaucoup de parcelle, il est préférable d'utiliser la fonction « Personnaliser un plan ».
- Dans le cas d'un alpha-plan, vous ne devez pas utiliser les fonctions « Personnaliser un plan » et « Rafraichir le plan »

## Rajout de commentaires dans le rapport fixe

Sous le rapport, vous pouvez ajouter des commentaires libres en formatant les cellules comme vous le souhaitez.

| Modalités | | Variables | | | | | |
|---|---|---|---|---|---|---|---|
| Date : | | Variable : | Acariose bronzée | | | Cigarier | |
| Stade : | | Infestation : | | | | | |
| Temp (°C) : | | Date : | 20/02/2013 | | | 20/02/2013 | |
| H.R. (%) : | | Stade : | 8 feuilles étalées (18) | | | 8 feuilles étalées (18) | |
| | | Unité : | | | | | |
| | | Numéro : | 1 | | | 2 | |
| 1 | Témoin | | 99,000 | A | 1/12 | 97,250 | A 2/12 |
| 2 | Moda1 | | 75,000 | AB | 7/12 | 95,000 | A 3/12 |
| 3 | Moda2 | | 97,500 | A | 2/12 | 83,500 | A 4/12 |
| 4 | Moda3 | | 95,500 | A | 3/12 | 72,500 | AB 6/12 |
| 5 | Moda4 | | 78,750 | AB | 6/12 | 42,500 | AB 9/12 |
| 6 | Moda5 | | 32,500 | C | 9/12 | 75,000 | AB 5/12 |
| 7 | Moda6 | | 22,500 | CD | 10/12 | 98,689 | A 1/12 |
| 8 | Moda7 | | 94,500 | A | 4/12 | 35,000 | AB 10/12 |
| 9 | Moda8 | | 87,500 | AB | 5/12 | 20,000 | AB 11/12 |
| 10 | Moda9 | | 20,000 | CD | 11/12 | 0,000 | B 12/12 |
| 11 | Moda10 | | 64,750 | B | 8/12 | 50,000 | AB 8/12 |
| 12 | Moda11 | | 7,500 | D | 12/12 | 72,500 | AB 7/12 |
| | | Test : | | N.K. 5% | | | N.K. 5% |
| | | Sign. : | | 0,000 | | | 0,003 |
| | | C.V. : | | 19,519 | | | 55,354 |
| | | Moyenne : | | 64,583 | | | 61,828 |
| | | Ecart type : | | 12,606 | | | 34,224 |
| | | Nombre de modalités : | | 12 | | | 12 |
| | | Nombre de répétitions : | | 4 | | | 4 |
| | | Résidus suspects : | | OUI | | | NON |
| | | Distribution normale : | | NON | | | OUI |
| | | Interaction traitements / blocs : | | NON | | | NON |

# ANNEXES

## LE RISQUE $\alpha$ DE LA PREMIÈRE ESPÈCE

Y-a-t-il des différences entre les traitements (entre des variétés ou des régimes alimentaires par exemple) ?

Votre démarche pour répondre à cette question est la suivante :

- Vous supposez au départ que tous les traitements sont identiques.
- En réalité, vous constatez toujours des écarts entre les traitements.

Alors, quelle est votre conclusion ? Pouvez-vous admettre que ces écarts sont :

**A** - dus aux aléas de l'expérience = Vos traitements sont alors identiques.

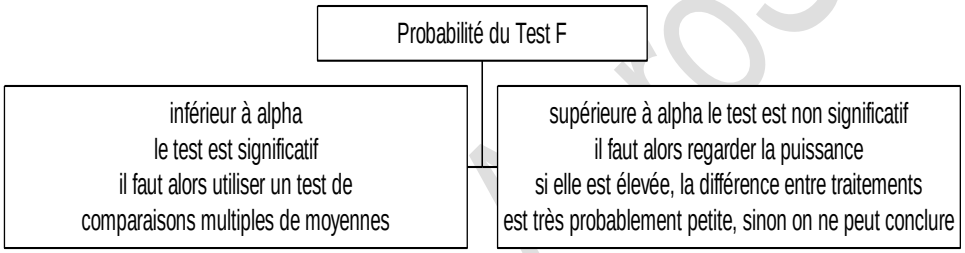
**B** - réels = Vos traitements sont alors différents.

L'analyse de variance (le test F entre traitements) vous donne la probabilité d'apparition de tels écarts.

La comparaison de cette probabilité à un seuil $\alpha$ fixé à l'avance va vous permettre de choisir entre A et B.

Ce seuil $\alpha$ est le risque de première espèce $\Rightarrow$ C'est le risque de décider que des traitements effectivement identiques sont différents.

Le choix de son niveau dépend tout simplement du coût d'une mauvaise décision.




### *Interaction Traitements * Blocs : le test de Tukey*

Vous avez réalisé un dispositif en blocs : l'analyse de variance de celui-ci suppose, entre autre, que les différents effets (traitements et blocs) sont additifs.

Une interaction traitements * blocs signifie que les écarts entre les traitements ne sont pas les mêmes dans les différents blocs.

TUKEY a mis au point une méthode qui prend un degré de liberté à la variation résiduelle pour tester l'éventuelle présence d'une interaction multiplicative entre les traitements et les blocs (l'écart entre 2 traitements sera plus élevé dans un bloc dont la valeur moyenne sera importante).

Dans le cas où cette interaction est significative $\Rightarrow$ regardez attentivement la cartographie des résidus pour en déterminer l'origine. L'interprétation des résultats est cependant très délicate.

## GRAPHIQUES DE L'ANALYSE EXPLORATOIRE

### Box plot

Un *box plot* (ou boîte à moustaches) est une représentation graphique qui donne des indications sur la tendance centrale des valeurs, leur variabilité, la symétrie de la distribution et la présence d'*outliers* (valeurs très différentes des autres). Le box plot est souvent utilisé pour comparer plusieurs ensembles de données.

Il existe plusieurs possibilités de représentation du box plot. StatBox utilise la forme suivante :

- le premier quartile Q1 correspond au bord inférieur de la boîte,
- la médiane Q2 correspond à un trait noir,
- la moyenne correspond à un trait rouge,
- le troisième quartile $Q_3$ correspond au bord supérieur de la boîte.

Deux intervalles sont définis de part et d'autre des premier et troisième quartiles :

$$I_{Q1} = [Q_1 - 1{,}5 \times (Q_3 - Q_1) \, , \, Q_1]$$
$$I_{Q3} = [Q_3 \, , \, Q_3 + 1{,}5 \times (Q_3 - Q_1)]$$

- la moustache inférieure du box plot s'étend de Q1 jusqu'à la valeur la plus proche de la borne inférieure de IQ1, en restant à l'intérieur de IQ1,
- la moustache supérieure du box plot s'étend de Q3 jusqu'à la valeur la plus proche de la borne supérieure de IQ3, en restant à l'intérieur de IQ3,
- les valeurs en deçà de la moustache inférieure et au-delà de la moustache supérieure sont représentées individuellement par des cercles. Ces cercles sont pleins lorsque les valeurs sont au-delà de 3 fois l'écart interquartile (Q3 - Q1), et vides s'ils sont situés à l'intérieure de cet intervalle,
- les valeurs minimale et maximale sont indiquées sur le box plot.

## Stem and leaf plot

Un *stem and leaf plot* (ou diagramme « tige et feuille ») est une représentation semi-graphique qui donne des indications sur la distribution de fréquence d'un ensemble de données, en utilisant les valeurs elles-mêmes. La partie *stem* (ou tige) correspond aux intervalles de classes de valeurs, et la partie *leaf* (ou feuille) correspond au nombre de données dans la classe, représenté par les différentes valeurs.

Pour construire un diagramme « tige et feuille », il faut couper chaque valeur en une partie principale (*stem*) et une partie secondaire (*leaf*), cette coupure ne s'effectuant pas nécessairement au niveau de la décimale. Les tiges sont affichées les unes en dessous des autres par ordre croissant, et les feuilles sont affichées horizontalement à droite des tiges, également par ordre croissant. StatBox détermine automatiquement l'unité qui lui semble la plus appropriée pour couper les valeurs en tige et feuille, mais vous pouvez modifier l'unité par défaut. Pour plus de clarté, StatBox affiche avant chaque diagramme l'unité utilisée en donnant la signification d'une tige et feuille élémentaire **1|1**.

## Q-Q plot et p-p plot

Le *Q-Q plot* (ou *normal probability plot*, ou graphique « quantile-quantile ») et le *p-p plot* (ou *probability-probability plot*) permettent d'apprécier visuellement si les données sont susceptibles de suivre une loi normale en comparant la distribution de fréquence cumulée des données à la fonction de répartition de la loi normale de mêmes moyenne et variance. Le Q-Q plot effectue cette comparaison du point de vue des valeurs tandis que le p-p plot se place du point de vue des probabilités. Dans les deux cas, lorsque les points s'organisent selon la première bissectrice du graphique, cela indique que la loi normale est compatible avec les données.

### *p-p plot*

Dans un p-p plot, l'axe des abscisses correspond aux fréquences relatives des valeurs et les ordonnées correspondent aux probabilités qu'auraient les valeurs si elles étaient distribuées selon une loi normale de mêmes moyenne et variance que les données.

Ainsi, chaque abscisse du p-p plot correspond à l'ordonnée de chaque valeur sur la distribution de fréquence cumulée des données, et l'ordonnée correspondante dans le p-p plot est l'ordonnée de la fonction de répartition de la loi normale de mêmes moyenne et variance que les données, pour la valeur considérée.

### *Q-Q plot*

Dans un Q-Q plot, l'axe des abscisses correspond aux valeurs observées et les ordonnées correspondent aux valeurs de la loi normale de mêmes moyenne et variance que les données, calculées pour les fréquences relatives des valeurs observées.

Ainsi, chaque abscisse du Q-Q plot correspond à l'abscisse de chaque valeur sur la distribution de fréquence cumulée des données, et l'ordonnée correspondante dans le Q-Q plot est l'abscisse de la fonction de répartition de la loi normale de mêmes moyenne et variance que les données, pour la probabilité considérée.

### Références

**Jobson J.D. (1991)**. Applied multivariate data analysis. Volume I: regression and experimental design. Springer-Verlag, New York, pp. 35-36, 45-46, 62-65.

**Johnson R.A. & D.W. Wichern (1992)**. Applied multivariate statistical analysis. Prentice-Hall, Englewood Cliffs, pp. 154-158.

**Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, pp. 28-30, 116-123, 151-152.

**Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, pp. 119-121.

## SIMILARITÉS/DISSIMILARITÉS

Il existe de nombreuses mesures de ressemblance (similarités ou dissimilarités). StatBox propose des indices sélectionnés en fonction de leurs propriétés mathématiques et de leur intérêt pratique ou pédagogique.

### Données quantitatives

- Corrélation de Pearson : covariance des deux lignes ou des deux colonnes comparées, standardisées par les variances, ou ce qui revient au même, covariance calculée sur les données centrées-réduites. Résultat dans l'intervalle [-1,+1].
- Corrélation de Spearman : coefficient de corrélation non paramétrique strictement équivalent au coefficient de corrélation de Pearson calculé sur les rangs des valeurs. Résultat dans l'intervalle [-1,+1].
- Corrélation de Kendall : coefficient de corrélation non paramétrique, c'est-à-dire calculé sur les rangs des valeurs. Résultat dans l'intervalle [-1,+1].

**Remarque :** Les coefficients de corrélation ont été créés avec l'intention de mesurer la ressemblance entre variables. Pour évaluer la ressemblance entre observations, ils devraient être employés avec circonspection.

- Distance euclidienne : métrique de l'espace euclidien (espace de la géométrie classique). La distance euclidienne vaut 0 pour deux lignes ou deux colonnes identiques, mais elle ne possède pas de borne supérieure. La distance euclidienne augmente à mesure que s'accroît le nombre de variables, et sa valeur dépend également de l'échelle de chacune des variables de sorte qu'en changeant simplement leur échelle, on peut obtenir des résultats très différents. Ce problème peut être évité en standardisant les variables.
- Distance du khi² : Pour pallier les inconvénients liés à l'utilisation de la distance euclidienne, il est possible d'utiliser la distance du khi² qui fait intervenir à la fois les sommes des colonnes et des lignes du tableau de données. Dans le cas du calcul de la distance du khi² entre deux lignes par exemple, les termes de chaque ligne sont rapportés à leur somme et une colonne contribue à la distance en raison inverse de son poids. Le calcul de la distance du khi² revient à calculer la distance euclidienne sur des données transformées selon : xij -> xij / (xi.√x.j) avec xi. la somme sur les colonnes pour la ligne i et x.j la somme sur les lignes pour la colonne j. La distance du khi² satisfait au principe d'équivalence distributionnelle c'est-à-dire que la distance ne change pas entre les lignes ou entre les colonnes en remplaçant deux colonnes ou deux lignes de même profil par leur somme. La distance du khi² est particulièrement adaptée aux tableaux homogènes d'effectifs ou de grandeurs additives (ex. tonnes, kilomètres, pourcentages) .

- Distance de Manhattan : métrique dite L1, calculée sur la base des écarts absolus au lieu des écarts quadratiques comme dans le cas de la distance euclidienne.
- Dissimilarité de Pearson : transformation de la corrélation de Pearson en une dissimilarité variant dans l'intervalle [0,1], soit r -> (1 - r) / 2.
- Dissimilarité de Spearman : transformation de la corrélation de Spearman en une dissimilarité variant dans l'intervalle [0,1], soit rS -> (1 - rS) / 2.
- Dissimilarité de Kendall : transformation de la corrélation de Spearman en une dissimilarité variant dans l'intervalle [0,1], soit $\tau$ -> (1 - $\tau$) / 2.

### Données binaires

Si *i* et *j* sont deux entrées dans le tableau (deux lignes ou deux colonnes), notons *a* le nombre de 1 communs à *i* et *j*, *b* le nombre pour 1 de *i* qui correspondent à des 0 pour *j*, *c* le nombre de 1 pour *j* qui correspondent à des 0 pour *i* et *d* le nombre de 0 communs à *i* et *j*. Les indices pour données binaires sont définis à partir de *a*, *b* et *c*, et éventuellement de *d*. Notez que les données *a*, *b*, *c* et *d* sont simplement les effectifs du tableau de contingence 2 × 2 suivant :

| *i* / *j* | 1 | 0 | |
|---|---|---|---|
| 1 | a | b | a + b |
| 0 | c | d | c + d |
| | a + c | b + d | n = a + b + c + d |

Les indices sont présentés sous la forme de similarités *S*, mais peuvent s'exprimer très facilement sous la forme de dissimilarités *D* en calculant *D* = 1 - *S* lorsque *S* varie dans l'intervalle [0,1], et en calculant *D* = (1 - *S*)/2 lorsque *S* varie dans l'intervalle [-1,+1].

- Indice de Jaccard : *a* / (a + b + c). Résultat dans l'intervalle [0,1]. Donne un poids égal aux différents termes, et ne prend pas en considération les doubles-0 (terme d).
- Indice de Dice : 2a / (2a + b + c), où a est divisé par la moyenne arithmétique des nombres de 1 pour i et j. Résultat dans l'intervalle [0,1]. Construit selon le modèle de l'indice de Jaccard, cet indice donne un poids deux fois plus élevé aux doubles-1 (terme a).
- Indice de Sokal & Sneath (2) : a / (a + 2b +2c). Résultat dans l'intervalle [0,1]. Construit selon le modèle de l'indice de Jaccard, cet indice donne un poids deux fois plus élevé aux différences figurant au dénominateur (termes b et c).

**Remarque** : les indices de Jaccard, Dice et Sokal & Sneath (2) donnent la même ordonnance, c'est-à-dire les mêmes relations d'ordre entre les observations. En conséquence, dans une classification ascendante hiérarchique on obtient des dendrogrammes qui ont la même structure (ou *topologie*).

- Indice de Sokal & Michener : (*a* + *d*) / (*a* + *b* + *c* + *d*). Résultat dans l'intervalle [0,1]. En employant cet indice, on part du principe que les doubles-1 (terme a) et les doubles-0 (terme d) jouent un rôle symétrique, ce qui implique que les deux modalités de la variable peuvent être indifféremment codées 1 ou 0.
- Indice de Rogers & Tanimoto : (a + d) / (a + 2b + 2c + d). Résultat dans l'intervalle [0,1]. Construit selon le modèle de l'indice de Sokal & Michener, cet indice donne aux différences (termes b et c) un poids deux fois plus important qu'aux concordances (termes a et d).
- Indice de Sokal & Sneath (1) : (2a + 2d) / (2a + b + c + 2d). Résultat dans l'intervalle [0,1]. Construit selon le modèle de l'indice de Sokal & Michener, cet indice donne aux concordances (termes *a* et *d*) un poids deux fois plus important qu'aux différences (termes *b* et *c*).

**Remarque** : les indices de Sokal & Michener, Rogers & Tanimoto et Sokal & Sneath (1) donnent la même ordonnance. En conséquence, dans une classification ascendante hiérarchique on obtient des dendrogrammes qui ont la même topologie.

- Phi de Pearson : (*ad* - *bc*) / $\sqrt{(a + b)(c + d)(a + c)(b + d)}$. Résultat dans l'intervalle [-1,+1]. Cet indice soustrait le produit des différences (terme bc) au produit des concordances (terme ad). Le Phi de Pearson $\varphi$ est relié

au khi² par la relation $\chi^2 = n\varphi^2$, avec n l'effectif total. Pour obtenir une dissimilarité, StatBox effectue la transformation vers l'intervalle [0,1] : φ -> (1 - φ)/2.

- Indice de Ochiai : a / √(a + b) (a + c) où a est divisé par la moyenne géométrique des nombres de 1 pour i et j. Résultat dans l'intervalle [0,1].
- Indice de Kulczinski : a(1/(a + b) + 1/(a + c))/2 où a est divisé par la moyenne harmonique des nombres de 1 pour *i* et *j*. Résultat dans l'intervalle [0,1].

**Remarque** : les indices de Ochiai et de Kulczinski sont des variantes de l'indice de Dice faisant intervenir respectivement la moyenne géométrique et la moyenne harmonique au lieu de la moyenne arithmétique. On peut donc s'attendre à ce que les valeurs de ces indices soient voisines, s'écartant les unes des autres lorsque (*a* + *b*) et (*a* + *c*) sont très différents.

### Références


**Dagnelie P. (1986)**. Théorie et méthodes statistiques. Vol. 2. Les Presses Agronomiques de Gembloux, Gembloux, pp. 88-90, 395-398.

**Dillon W.R. & M. Goldstein (1984)**. Multivariate analysis. Methods and applications. John Wiley & Sons, New York, pp. 157-167.

**Gower J.C. & P. Legendre (1986)**. Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3** : 5-48.

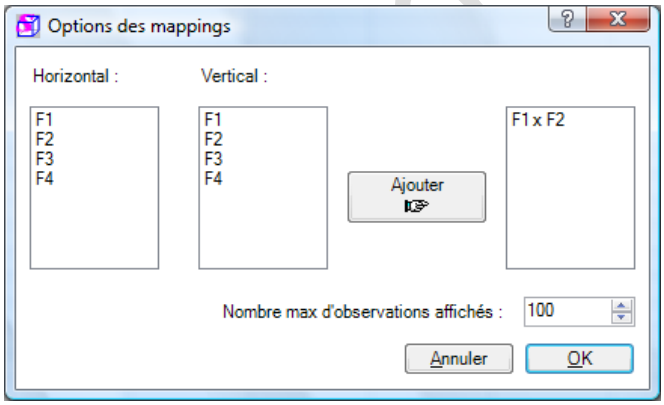
**Jambu M. (1978)**. Classification automatique pour l'analyse des données. 1 - méthodes et algorithmes. Dunod, Paris, pp. 484-518.

**Jobson J.D. (1992)**. Applied multivariate data analysis. Volume II: categorical and multivariate methods. Springer-Verlag, New York, pp. 345-388.

**Legendre L. & P. Legendre (1984)**. Ecologie numérique. Tome 2. La structure des données écologiques. Masson, Paris, pp. 5-50.

**Roux M. (1985)**. Algorithmes de classification. Masson, Paris, pp. 126-134.


## BOÎTE D'AFFICHAGE DES GRAPHIQUES




Sélectionnez les croisements de facteurs symbolisant les axes à représenter pour les graphiques de résultats sur les variables ou les observations. Pour cela sélectionnez un facteur dans la liste pour l'axe horizontal, un facteur dans la liste pour l'axe vertical puis cliquez sur « Ajouter »

*Nombre max. d'observations affichées :* entrez le nombre d'observations actives à représenter, classées par ordre décroissant des contributions (moyenne des contributions sur les deux axes définissant le plan factoriel, pondérée par le pourcentage de variance correspondant à chaque axe). Les observations supplémentaires sont forcément représentées.

## ROTATION DES FACTEURS

Il existe deux types de techniques de rotation des facteurs visant à simplifier l'analyse : la rotation orthogonale et la rotation oblique. Au contraire d'une rotation oblique, une rotation orthogonale préserve l'orientation originelle entre les facteurs de sorte qu'ils sont toujours orthogonaux (non corrélés) après rotation. StatBox propose les deux techniques de rotation orthogonale les plus communément utilisées : les rotations varimax et quartimax.

### Rotation varimax

Utilisez la rotation varimax pour simplifier l'interprétation des facteurs en minimisant le nombre de variables qui ont des contributions élevées sur chaque facteur.

L'objectif de la rotation orthogonale varimax est d'identifier une structure factorielle telle que pour chaque facteur, quelques variables aient des contributions élevées, les autres ayant des contributions très faibles. Cet objectif est atteint en maximisant, pour un facteur donné, la variance des carrés des contributions parmi les variables, sous la contrainte que la variance de chaque variable soit conservée.

### Rotation quartimax

Utilisez la rotation quartimax pour simplifier l'analyse des variables en minimisant le nombre de facteurs nécessaires pour expliquer chaque variable.

L'objectif de la rotation quartimax est d'identifier une structure factorielle telle que les variables aient des contributions élevées pour un même facteur. En outre, chaque variable doit avoir une contribution non nulle pour un autre facteur, et des contributions pratiquement nulles pour tous les facteurs restants. Cet objectif est atteint en maximisant la variance des contributions parmi les facteurs, sous la contrainte que la variance de chaque variable soit inchangée.

### Références

**Dillon W.R. & M. Goldstein (1984)**. Multivariate analysis. Methods and applications. John Wiley & Sons, New York, pp. 87-95.

**Sharma S. (1996)**. Applied multivariate techniques. John Wiley & Sons, New York, pp. 137-141.

## *P*-VALUE

Dans StatBox, chaque test statistique est accompagné d'une *p*-value. La *p*-value est définie comme la probabilité, calculée sous l'hypothèse nulle, d'obtenir une valeur de la statistique aussi extrême que celle observée pour les données (dans une direction particulière). Cette définition implique qu'une *p*-value est utile dans un test unilatéral parce que la direction utilisée pour la calculer correspond à l'hypothèse alternative du test. Par exemple, dans un test *t* de Student unilatéral à droite, la *p*-value correspond à l'aire contenue sous la loi de Student à droite de $t_{obs}$, tandis que dans le test unilatéral à gauche, la *p*-value correspond à l'aire contenue sous la loi de Student à gauche de $t_{obs}$.

Intuitivement, la *p*-value peut être vue comme la force de l'évidence contre l'hypothèse nulle. En effet, plus la *p*-value est faible, plus la probabilité d'obtenir par hasard un résultat aussi extrême que celui observé est faible, et par conséquent, plus le résultat est significatif. L'usage classique d'un risque de première espèce $\alpha$ consiste alors à accepter l'hypothèse alternative si la *p*-value est inférieure ou égale à $\alpha$. La relation entre la *p*-value et le risque de première espèce conduit à interpréter la *p*-value comme le niveau de signification le plus faible auquel la valeur observée de la statistique est significative, dans une direction particulière. La *p*-value est parfois désignée comme la *provalue* ou la *probabilité associée*.

### Références

**Berger J.O. & T. Sellke (1987)**. Testing a point null hypothesis: the irreconcilability of P values and evidence (with discussion, pp. 123-139). *Journal of the American Statistical Association*, **82 :** 112-122.

**Casella G. & R.L. Berger (1987)**. Reconciling bayesian and frequentist evidence in the one-sided testing problem (with discussion, pp. 123-139). *Journal of the American Statistical Association*, **82** : 106-111.

**Gibbons J.D. (1986)**. P values. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, John Wiley & Sons, New York, pp. 366-368.

**Yoccoz N.G. (1991)**. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*, **72** : 106-111.

## IDENTIFICATION DES OBSERVATIONS POUR L'HISTOGRAMME DES RÉSIDUS (AGRICULTURE)

L'examen de l'histogramme des résidus permet de vérifier aisément la normalité de leur distribution et de détecter d'éventuelles valeurs suspectes. De façon à pouvoir repérer facilement les observations (parcelles, animaux...) correspondant aux résidus, chaque parcelle est identifiée sur l'histogramme par son numéro.

Dans l'exemple suivant la parcelle 104 (ligne 1 colonne 4) à un résidu anormalement élevé.

| | | | | |
|---|---|---|---|---|
| 8 | 305 | | | |
| 7 | 201 | | | |
| 6 | 107 | 307 | | |
| 5 | 106 | 203 | | |
| 4 | 304 | 308 | | |
| 3 | 204 | 202 | | |
| 2 | 103 | 302 | | |
| 1 | 101 | 207 | 104 | |
| Effectifs | | | | |
| | 6 | 8 | 0 | 1 |
| Bornes | | | | |
| | -1,6 à -0,44 | -0,44 à 0,72 | 0,72 à 1,87 | 1,87 à 3,03 |

## DÉTECTION DES VALEURS ANORMALES, MÉTHODE DE GRUBBS

La comparaison d'une valeur apparemment anormale à l'ensemble des autres observations, est identique à la comparaison d'un échantillon ne comportant qu'une observation (la valeur anormale), avec la moyenne d'un échantillon formé des (n-1) autres observations. Mais cette méthode est très longue, pour explorer toute une série de données (il faut faire n comparaisons).

GRUBBS a proposé de calculer une seule fois la moyenne ( $\overline{X}$ ) et l'écart-type estimé (S) de l'échantillon de l'ensemble des n observations, puis de déterminer, pour chaque donnée, un "T observé", tel que :

$$Tobs = \frac{\left| Xi - \overline{X} \right|}{S}$$

On considère, alors, qu'une observation est anormale, lorsque :

$Tobs. \geq Tg$

La valeur de *Tg* peut être lue dans des tables données par GRUBBS, ou encore calculée à partir des distributions du t de Student, ou de la loi Normale réduite, pour un risque de première espèce de :

$$\frac{\alpha}{2n}$$

C'est cette méthode qui est programmée pour détecter les "résidus suspects" dans le module "analyse de variance".

## PUISSANCE

Le risque $\alpha$ de 1ère espèce est le risque de décider que des traitements effectivement identiques sont différents.

On peut aussi décider que des traitements effectivement différents sont identiques $\Rightarrow$ C'est le risque ß de 2ème espèce.

Mais votre problème est souvent de montrer que des traitements réellement différents sont bien différents $\Rightarrow$ Il vous faut alors apprécier la puissance de votre essai, c'est-à-dire la probabilité que vous avez de mettre en évidence une différence donnée "d" entre des traitements. C'est donc la capacité de votre essai à vous faire voir quelque chose. Cette puissance dépend :

- du risque $\alpha$ de 1ère espèce
- de la variabilité des résultats (de l'écart-type résiduel)
- de la différence "d" entre les traitements (différence intéressante techniquement ou économiquement à mettre en évidence)
- du nombre de répétitions (de blocs ou d'essais)

Son calcul permet d'aller plus loin dans l'interprétation des résultats, dans le cas où l'effet traitement est "non significatif".

$\Rightarrow$ Si la puissance est faible (par exemple 20 %)

Vous n'avez pas vu de différences entre les traitements, mais vous ne vous en étiez pas donné les moyens.(Vous n'avez qu'une chance sur 5 de voir une différence si elle existe vraiment).

$\Rightarrow$ Si la puissance est élevée (par exemple 80 %)

Vous n'avez pas vu de différences entre les traitements, mais, s'il en existe une, vous aviez les moyens de la voir $\Rightarrow$ Il y a donc de grandes chances que la différence réelle entre vos traitements soit inférieure à "d".

## LE TEST DE NEWMAN-KEULS

Ce test de comparaison de moyennes permet de constituer des groupes homogènes de traitements ; ceux appartenant à un même groupe sont considérés comme non différents au risque de 1ère espèce choisi. La constitution des groupes homogènes se fait à partir des plus petites amplitudes significatives (p.p.a.s.). Lorsque l'amplitude observée entre les moyennes extrêmes d'un groupe de k moyennes est inférieure à la p.p.a.s. pour k moyennes, on déclarera que ces k moyennes constituent un groupe homogène.

Vous pouvez utiliser ce test si tous vos traitements jouent le même rôle (il n'y a ni témoin, ni traitement de référence), comme c'est souvent le cas dans la comparaison de variétés de céréales par exemple.

## LE TEST T DE BONFERRONI

Aussi appelé « test du t corrigé », le test de Bonferroni permet de réaliser toutes les comparaisons 2 à 2 de moyennes, c'est à dire (t(t-1))/2 comparaisons avec t traitements, en respectant globalement le risque $\alpha$ de 1ère espèce choisi. Cela signifie que chacune des comparaisons est effectuée au risque

$$\frac{\alpha}{(t(t-1)/2}$$

Comme le test de Newman-Keuls, vous pouvez l'utiliser si tous vos traitements jouent le même rôle.

## LE TEST DE DUNNETT

Dans ce test de comparaison de moyennes, tout traitement dont l'écart au(x) témoin(s) est supérieur au plus petit écart significatif (p.p.e.s.) est déclaré supérieur (inférieur) au(x) témoin(s).

L'utilisation de ce test suppose donc la présence de témoin(s). Un témoin peut être, par exemple :

- une parcelle non traitée dans un essai de produits phytosanitaires
- un traitement de référence (produit de référence) dans un essai de produits phytosanitaires, une variété de référence dans un essai variétés... La référence est un traitement bien connu, parmi les plus utilisés en pratique.

## LA MÉTHODE DES CONTRASTES

Cette méthode de comparaison de moyennes a pour but de vous permettre de répondre précisément aux diverses questions posées que vous avez formulées au départ d'un essai, dans le protocole expérimental.

Vos questions peuvent être du genre :

- en moyenne, les nouveaux traitements sont-ils meilleurs que le témoin (le traitement de référence)?
- parmi les nouveaux traitements, vaut-il mieux appliquer une dose simple ou une dose double ?
- lorsque l'on utilise une dose double, y-a-t-il une interaction avec l'espèce ?

L'utilisation de ce test suppose donc que vous ayez des questions précises... et qu'elles soient formalisées. Cette méthode permet de décomposer une somme des carrés des écarts factorielle du tableau d'analyse de la variance en (t-1) sommes des carrés des écarts (si on a t traitements) indépendantes, et chacune avec 1 degré de liberté. On obtient alors (t-1) « contrastes ».

Tout « contraste » est une combinaison linéaire des moyennes comparées. Le programme va donc vous demander, pour chacun d'eux, d'affecter un coefficient aux différentes moyennes, en respectant les règles suivantes :

- pour un contraste donné, la somme des coefficients doit être nulle.
- deux contrastes seront indépendants si la somme des doubles produits des coefficients est nulle.

Exemple : Soient trois moyennes, si A est une référence, B et C deux nouveaux « traitements » ; on peut se poser, par exemple, les questions suivantes :

==> Est-ce que les nouveaux « traitements » sont meilleurs que la référence ?

==> Les 2 nouveaux traitements sont-ils différents ?

La traduction de ces questions en « contrastes » donne :

A B C

On a bien

- 1re question +2 -1 -1 (+2) + (-1) + (-1) = 0
- 2e question 0 -1 +1 (0) + (-1) + (+1) = 0

Les 2 questions sont indépendantes car :

(+2)*(0) + (-1)*(-1) + (-1)*(+1) = 0

Ces 2 contrastes ne sont pas les seuls possibles avec 3 moyennes.

Si la définition des traitements change, les questions ne sont plus les mêmes... et les contrastes doivent être modifiés.

## Références

**GOUET J.P. (1974).** LES COMPARAISONS DE MOYENNES ET DE VARIANCES. Application à l'agronomie, PUBLICATION I.T.C.F.