Statistiques paramétriques et non paramétriques
Traitements sur des variables catégorisées

1 Travail sur des variables catégorisées avec Excel

1.1 Quelques commandes d'Excel utiles pour la saisie de données

Une configuration préalable indispensable avec Excel 2003 :

Menu : Affichage - Barres d'outils - Personnaliser. Onglet : Options. Boîte à cocher : Toujours afficher les menus dans leur intégralité.

Saisie "assistée" ou non : utilisez le menu : Outils - Options. Onglet Modifications. Boîte à cocher : Saisie semi-automatique des valeurs de cellules.

Saisissez par exemple, une dizaine de lignes d'un tableau tel que le suivant :

Identifiant	Conso	SD residence	SD age categorise	SD sexe	SD etudes	SD CSP
Q1	15,5375	Centre	- de 50 ans	Femme	Secondaire	Cadre supérieur
Q2	7,7605	Centre	50 ans à 60 ans	Homme	Primaire	Inactif
Q3	1,2815	Centre	60 ans à 65 ans	Homme	Secondaire	Retraité
Q4	7,9387	Sud	65 ans et plus	Femme	Supérieur	Retraité
Q5	38,4395	Centre	50 ans à 60 ans	Homme	CAP/BEP	Retraité
Q6	4,7260	Sud	50 ans à 60 ans	Homme	Secondaire	Cadre supérieur
Q7	6,9745	Sud	65 ans et plus	Homme	Primaire	Retraité
Q8	11,1575	Nord	- de 50 ans	Femme	Secondaire	Employé
Q9	0,1614	Autre	50 ans à 60 ans	Femme	Supérieur	Cadre supérieur
Q10	7,2769	Nord	65 ans et plus	Homme	Secondaire	Retraité

Remarques.
1) On peut utiliser les possibilités de recopie incrémentée pour générer les identifiants.
2) Les modalités des différentes variables catégorisées sont ici saisies sous forme de texte : très peu de traitements sont alors disponibles sous Excel sans recodage des données.

1.2 Tri à plat. Représentation graphique d'une variable catégorisée

Ouvrez le fichier Conso-Crustaces.xls. Les données qui y sont enregistrées correspondent à la description suivante :

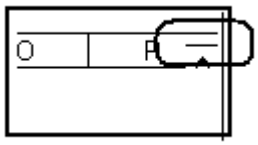
Une étude de consommation a été menée auprès des pêcheurs à pied en Bretagne (Cyndie Picot, thèse en cours). 510 pêcheurs à pied fréquentant les plages bretonnes ont rempli un questionnaire relatif à leurs habitudes de consommation. A partir des données récoltées, on a évalué leur consommation alimentaire de crustacés, et on souhaite étudier la variation du niveau de consommation selon différents facteurs socio-économiques.

On s'intéresse ici aux variables suivantes :

- Conso : consommation de crustacés en grammes par personne et par jour ;
- SD Résidence : catégorisé de 1 à 4
  - 1 : nord de la zone étudiée
  - 2 : centre de la zone
  - 3 : sud de la zone
  - 4 : autre zone
- SD Age Catégorisé : catégorie d'âge, avec la catégorisation suivante :
  - 1 : moins de 50 ans
  - 2 : de 50 ans à moins de 60 ans
  - 3 : de 60 ans à moins de 65 ans
  - 4 : 65 ans et plus
- SD sexe : sexe de la personne interrogée
- SD Etudes : niveau d'études, catégorisé de 1 à 4
  - 1 : primaire ou sans diplôme
  - 2 : CAP/BEP
  - 3 : Etudes secondaires, Bac
  - 4 : études supérieures
- SD CSP : catégorie socio-professionnelle
  - 1 : agriculteur
  - 2 : petit patron
  - 3 : cadre supérieur
  - 4 : profession intermédiaire
  - 5 : employé
  - 6 : ouvrier qualifié
  - 7 : retraité
  - 8 : inactif

1.2.1 Consultation des données

Le fichier comporte 512 observations. Il peut être intéressant de couper la fenêtre pour afficher deux sous-fenêtres. Pour cela, déplacez la barre située au-dessus de l'ascenseur vertical.



Exploration des données : placez la sélection dans l'une des cellules contenant les intitulés des variables et utilisez le menu Données - Filtrer - Filtre automatique :

	A	B	C	D	E	F
1	Conso	SD residen	SD age categoris	SD sexe	SD etude	SD CSP
2	5,2668	1	4	1	1	7

Par exemple, affichez les lignes pour lesquelles SD sexe = 1 et SD residence = 1.

Pour réafficher l'ensemble des données ou supprimer le filtre automatique, utilisez de nouveau le menu Données - Filtrer.

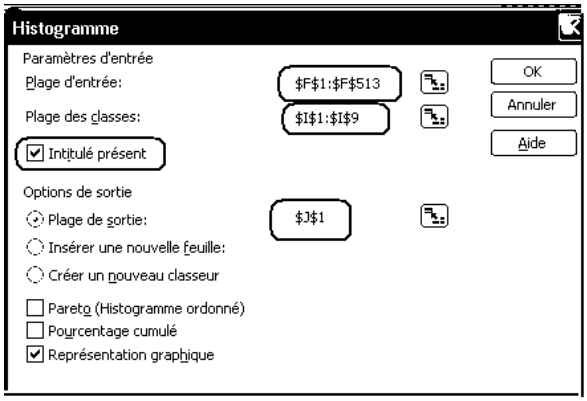
Tri à plat : on veut par exemple, obtenir un tri à plat selon les modalités (codées de 1 à 8) de la variable SD CSP.

Première solution : on utilise l'utilitaire d'analyse.

Pour activer l'utilitaire d'analyse, utilisez le menu Outils - Macro complémentaires et cochez l'item "Utilitaire d'analyse".

Complétez la feuille de données en indiquant les modalités de la variable CSP dans la plage de cellules I1 à I9, par exemple.

Utilisez ensuite le menu Outils - Utilitaire d'analyse, puis l'item Histogramme. La fenêtre de dialogue pourra par exemple être complétée comme suit :



Deuxième solution : la fonction NB.SI :

Par exemple, entrez en cellule L2 : `=NB.SI($F$1:$F$513; "=1")`
La fonction NB.SI compte le nombre de cellules de la plage indiquée qui satisfont la condition placée en deuxième paramètre.

Troisième solution : la fonction FREQUENCE, que nous verrons ultérieurement.

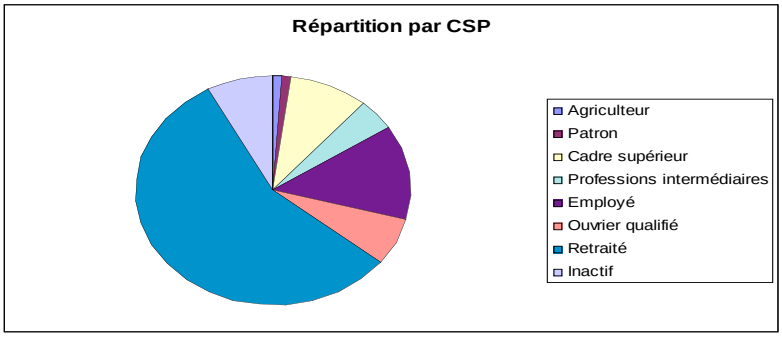
1.2.2 Représenter la distribution à l'aide d'un diagramme circulaire

Il faut partir du tri à plat de la variable à représenter. Mais le graphique sera plus explicite si les modalités sont indiquées par leurs libellés.

Constituer un tableau tel que le suivant :

Agriculteur	5
Patron	7
Cadre supérieur	46
Professions intermédiaires	23
Employé	69
Ouvrier qualifié	33
Retraité	289
Inactif	40

Utilisez ensuite le menu Insertion - Graphique pour obtenir le diagramme circulaire suivant :

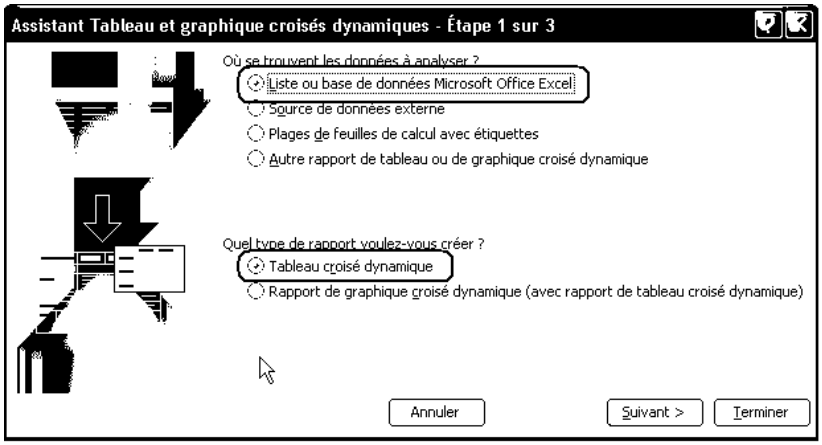


1.2.3 Effectuer un tri croisé sur deux variables catégorisées

On veut, par exemple, réaliser un tableau de contingence en croisant les variables SD Etudes et SD CSP. On va utiliser l'outil fourni par le menu Données - Rapport de tableau croisé dynamique... dont les dialogues ne sont pas très explicites, c'est le moins que l'on puisse dire !)

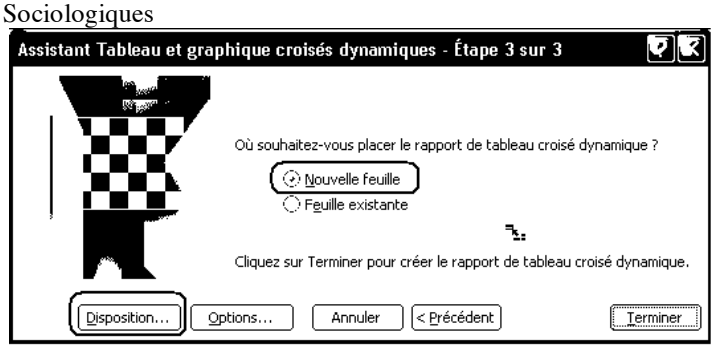
- Activez le menu Données - Rapport de tableau croisé dynamique...

Premier dialogue :

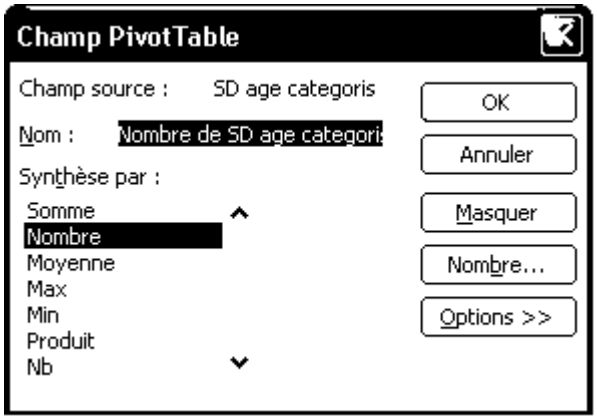


Second dialogue : sélection de la plage de données source :

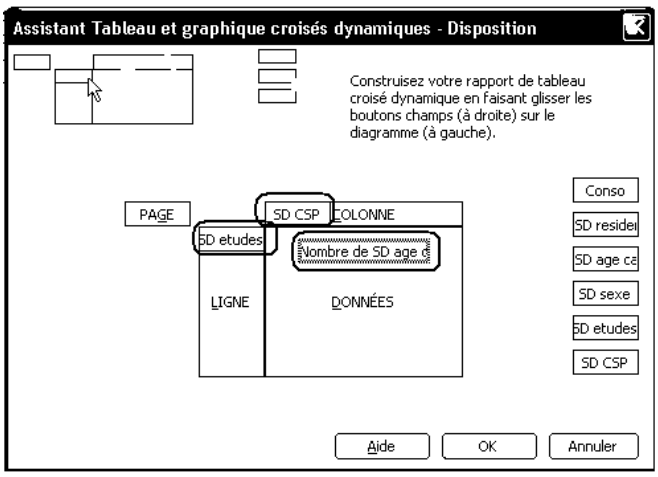
Troisième dialogue : choisissez de placer le tableau croisé dans une nouvelle feuille et *cliquez sur le bouton Disposition.*



Dans le dialogue disposition, faites glisser l'étiquette SD CSP dans la zone "colonne", l'étiquette SD études dans la zone "ligne" et l'une des autres variables dans la zone "données". Le choix a peu d'importance, mais il faut que la variable ne comporte pas de valeur manquante, faute de quoi le tableau de contingence sera incorrect. L'idéal est sans doute de prendre ici un identifiant des réponses. Double-cliquez ensuite sur l'étiquette présente dans la zone "données" et sélectionnez la fonction "Nombre" au lieu de la fonction "Somme".



L'affichage final doit se présenter ainsi :



On obtient le résultat suivant :

Nombre de SD age categorise	SD CSP								
SD etudes	1	2	3	4	5	6	7	8	Total
1	1	5	2	2	14	5	101	14	144
2		1	3	3	24	22	83	11	147
3	2	1	12	5	21	5	57	5	108
4	2		29	13	10	1	48	10	113

Total	5	7	46	23	69	33	289	40	512

Remarques.

1) Les palettes d'outils flottantes permettent éventuellement de modifier le tableau : changement de champ ligne, de champ colonne ou de champ pivot, changement de fonction à appliquer au champ pivot, etc.
2) On peut également indiquer un champ "page" (par exemple, la variable SD sexe), de manière à afficher alternativement les tableaux de contingence obtenus sur les sous-populations correspondant aux deux modalités de cette variable.

3) Comment Excel traite-t-il les valeurs manquantes ?
Supprimez les valeurs de SD études ou de SD CSP dans une ou deux lignes du fichier, faites recalculer le tableau croisé et observez le résultat. De même, explorez l'effet d'une valeur manquante dans le champ pivot.

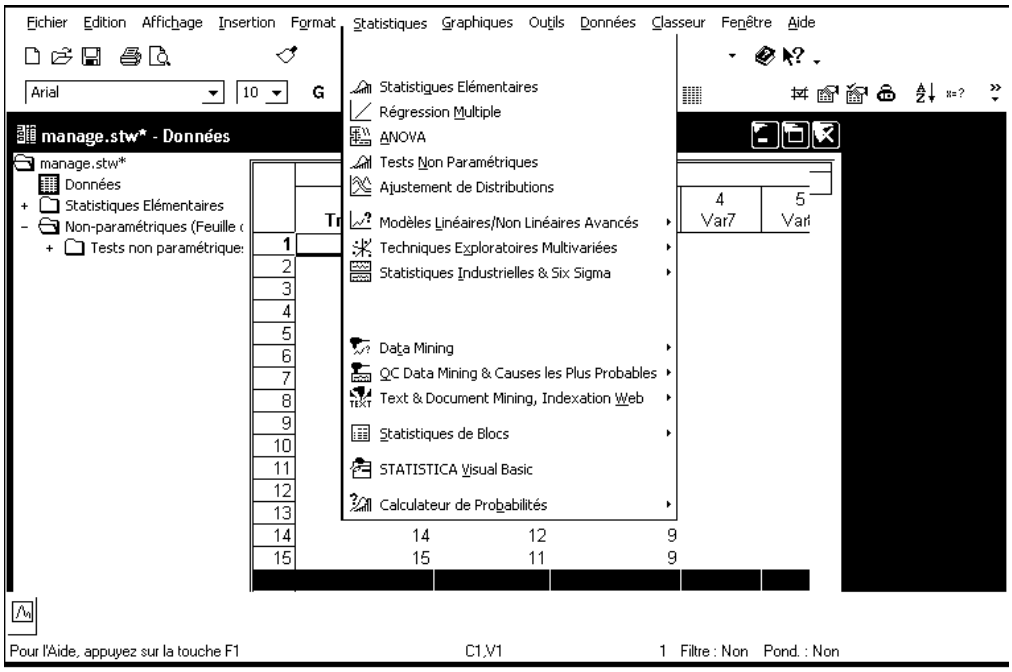
2 Travail sur des variables catégorisées avec Statistica

2.1 Présentation de Statistica

Statistica : l'interface utilisateur

L'écran de travail

Statistica 7.1 est un logiciel dédié aux traitements statistiques. C'est également la "brique" de base des logiciels proposés par Statsoft, et ses possibilités d'interaction avec d'autres logiciels (tableurs, systèmes de gestion de bases de données, traitements de textes, ...) sont nombreuses. En revanche, l'interface utilisateur pourra sembler un peu déconcertante au premier abord.



Les objets manipulés par Statistica

La **feuille de données** est organisée en variables et observations. Les colonnes sont les variables. Chaque ligne représente un individu statistique, appelé observation.

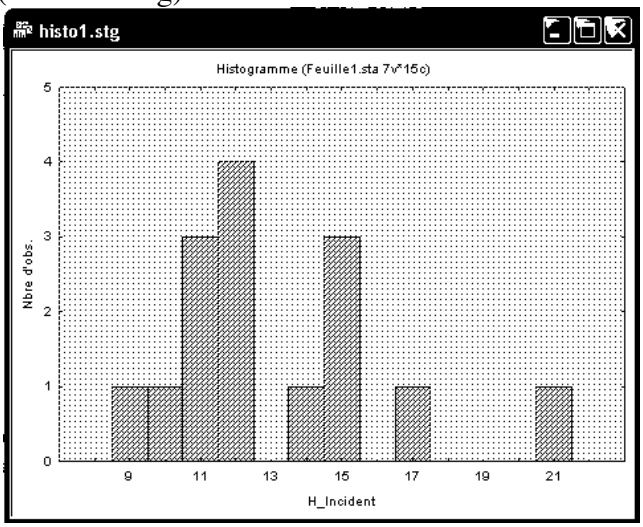
Données : Feuille1.sta (7v par 15c)

	1 Trimestre	2 H_Incident	3 D_Incident	4 Var7
1	1	11	8	
2	2	11	13	
3	3	14	12	
4	4	21	17	
5	5	12	14	
6	6	10	9	
7	7	15	10	
8	8	15	12	
9	9	17	13	
10	10	9	10	
11	11	12	8	
12	12	12	13	
13	13	15	12	
14	14	12	9	
15	15	11	9	

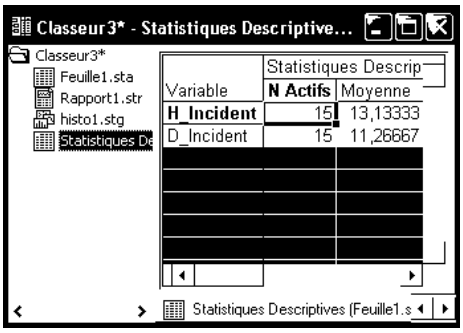
Les feuilles de données peuvent être enregistrées comme fichiers autonomes (fichiers *.sta). Elles contiennent les données d'entrée sur lesquelles s'effectuent les traitements statistiques. Les résultats de ces traitements s'affichent dans un document de sortie. Plusieurs possibilités sont offertes.

Fenêtre de rapport : C'est la méthode traditionnelle pour gérer les résultats produits par le logiciel. Un rapport se comporte plus ou moins comme un document produit par un traitement de textes. On peut insérer des commentaires, modifier la mise en forme, spécifier la mise en page, la numérotation des pages, l'en-tête et le pied de page en vue de l'impression. Les rapports peuvent être enregistrés comme fichiers autonomes (fichiers *.str).

Les résultats de sortie peuvent également être dirigés vers des fenêtres individuelles. Les résultats numériques sont alors affichés dans des fenêtres de données. Les graphiques sont affichés dans des **fenêtres de graphiques** (fichiers *.stg).

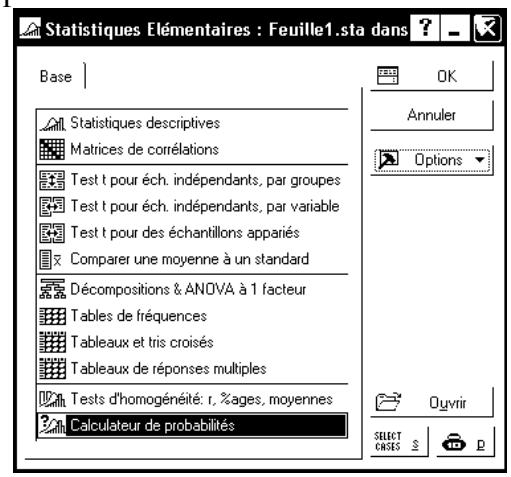


Les classeurs : les données d'entrée et de sortie peuvent également être stockées comme onglets dans un classeur. Un classeur est un "container" accueillant d'autres objets, organisés sous forme hiérarchique. Ils correspondent aux fichiers de type *.stw.



Traitements statistiques

Statistica est organisé en modules, accessibles à partir du menu Statistiques. Chaque module contient un groupe de procédures statistiques reliées entre elles. Par exemple, le module "Statistiques élémentaires" se présente comme suit :



2.2 Paramétrage de Statistica

Statistica est un logiciel assez stable et fiable, à condition de respecter quelques règles élémentaires. La première d'entre elles est :

> Pour travailler avec Statistica, ne pas utiliser un compte tel que le compte "etudiant" dont le profil est verrouillé.

Ouvrez une session *avec votre login sous XP*, puis chargez le logiciel Statistica. La configuration par défaut du logiciel n'est pas vraiment satisfaisante. Nous allons donc commencer par régler la configuration à nos besoins.

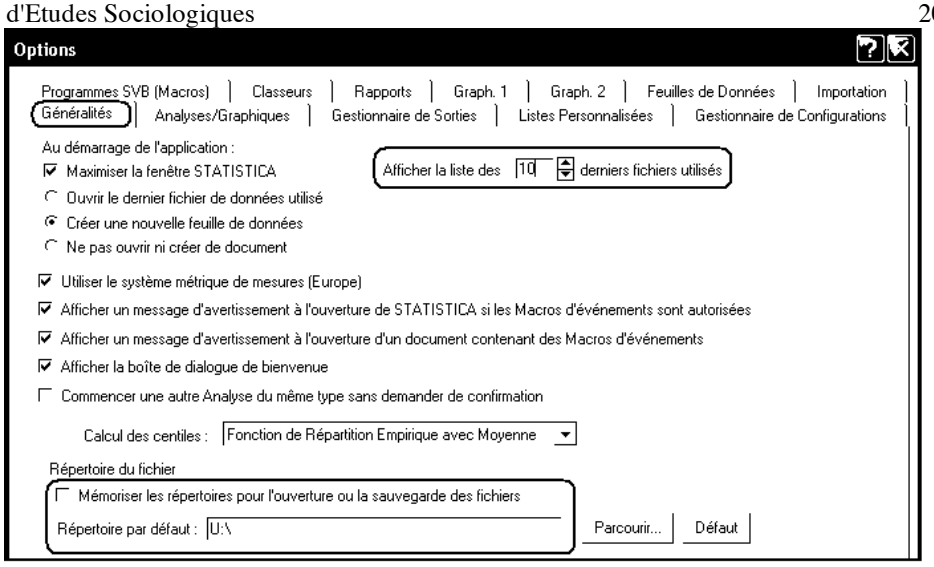
2.2.1 Le menu Outils - Options

Le menu Outils - Options contient de nombreuses possibilités de paramétrage de Statistica. Heureusement, seules quelques-unes d'entre elles méritent d'être retouchées.
Ouvrez la fenêtre de dialogue accessible par le menu Outils-Options et explorez les différents onglets qui y sont rassemblés.

N.B. Les options ainsi choisies sont enregistrées dans le profil de l'utilisateur lorsque l'on quitte le logiciel. *Il n'y a aucun enregistrement si le compte est verrouillé ou si Statistica se plante en cours de travail.*

2.2.1.1 Spécifier le répertoire d'enregistrement par défaut

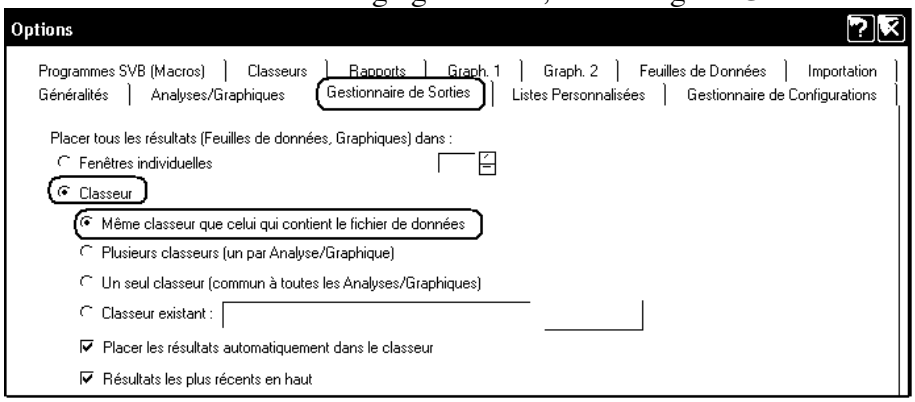
Par défaut, Statistica propose le répertoire "Mes Documents" pour l'enregistrement des nouveaux documents. On peut modifier ce comportement en utilisant l'onglet Généralités :



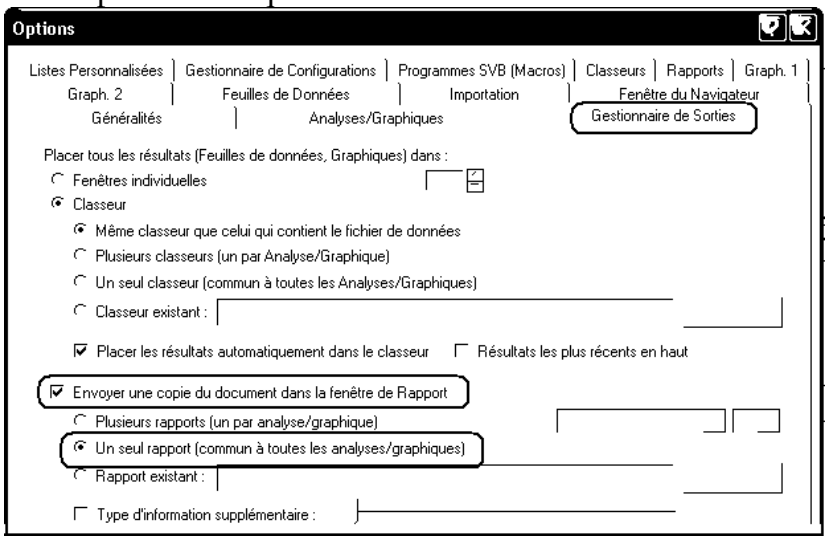
On peut aussi, sans inconvénient, réduire la longueur de la liste des derniers fichiers utilisés (indiquer 10 au lieu de 16 par exemple).

2.2.1.2 Gérer les sorties

La manière la plus commode de gérer nos documents avec Statistica consiste à rassembler dans un même classeur la ou les feuilles de données et les résultats de traitements concernant ces données. Ce comportement sera obtenu à l'aide du réglage suivant, sous l'onglet "Gestionnaire de sorties" :



Il peut également être commode de demander à Statistica de placer une copie des résultats dans un rapport, en utilisant des options telles que :



En effet un rapport peut être enregistré au format .rtf pour être relu sur une autre machine par un logiciel de traitement de textes, même si Statistica n'est pas installé sur l'appareil. Cependant :

- Les rapports produisent rapidement des fichiers très volumineux. Un rapport, ou un classeur contenant un ou des rapports devra être compressé (zippé) avant d'être envoyé par mail. Et par ailleurs, un rapport trop volumineux semble provoquer des plantages du logiciel dans certains cas.

- Si plusieurs séances de travail sont nécessaires pour réaliser le traitement, un nouveau rapport sera créé à chaque séance, ce qui est assez peu pratique.

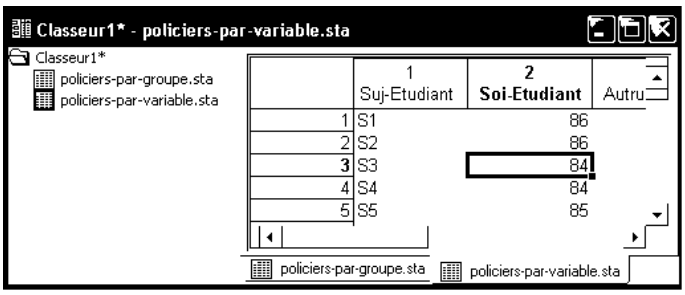
- Si le rapport produit au cours d'une séance de travail est inséré dans le classeur (par exemple à l'aide du menu local Insérer - Toutes les fenêtres du classeur), Statistica se plante de façon systématique, lorsque l'on quitte le logiciel.

Ne pas insérer dans le classeur courant le rapport produit au cours d'une séance de travail à l'aide du menu local "Insérer - Toutes les fenêtres", faute de quoi Statistica se plante lorsque l'on quitte le logiciel.

En revanche, il semble que l'on évite le plantage en utilisant le menu : Insérer - Document Statistica - Créer à partir d'une fenêtre.

2.2.1.3 La feuille de données active

Les traitements demandés via les menus s'appliquent à la fenêtre de données **active**. Dans le cas de données rassemblées dans plusieurs fenêtres indépendantes, la feuille active est celle qui se trouve au premier plan sur l'écran. Dans le cas d'un classeur, la feuille active est repérée par un liseré rouge :



Dans le classeur ci-dessus, la feuille active est "policiers-par-variable.sta"

On peut rendre active une feuille, ou changer de feuille active :

- soit en cliquant sur l'icône de la feuille et en utilisant le menu : Classeur - Feuille de données active;

- soit en cliquant avec le bouton droit sur l'icône de la feuille et en utilisant l'item "Feuille de données active" du menu local.

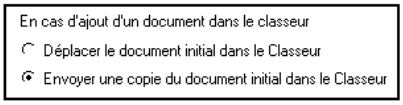
2.2.2 Manipulations de base sur un classeur

2.2.2.1 Copier - coller entre classeurs, entre un classeur et un objet Statistica

Pour déplacer un objet d'un classeur à un autre, il suffit de déplacer son icône depuis le volet gauche du premier classeur dans le volet gauche du second. On peut également utiliser les menus locaux Copier et Coller obtenus à l'aide d'un clic droit dans le volet gauche de chaque classeur.

Le menu local "Insérer" du volet gauche d'un classeur permet également d'insérer dans ce classeur un document contenu dans une fenêtre indépendante. Il suffit de choisir les options : Document Statistica - Créer à partir d'une fenêtre.

L'opération faite par Statistica est soit une copie (l'original de l'objet est conservé) soit un déplacement (l'original de l'objet n'est pas conservé) selon le paramétrage choisi dans le menu Outils - Options - Onglet Classeurs - Item "En cas d'ajout d'un document dans le classeur".



2.2.2.2 Supprimer un objet d'un classeur

Il est également possible de supprimer un objet d'un classeur, à l'aide d'un clic droit et de l'item de menu Supprimer. Cela permet notamment de ne garder, pour un traitement donné, que le résultat le plus abouti. Attention cependant : lorsque l'on supprime un objet qui n'est pas une feuille de la hiérarchie, on supprime en même temps tous les objets qui en dépendent.

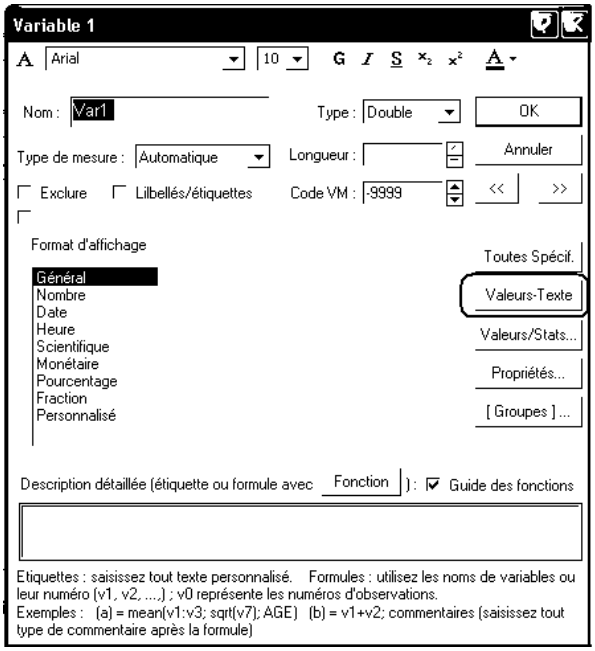
2.2.3 Saisie de données avec Statistica

Nous avons vu qu'avec Excel, il était difficile de travailler sur des variables catégorisées pour lesquelles ce sont les libellés des catégories qui ont été saisis. De ce point de vue, le comportement de Statistica est tout à fait différent :

- par défaut, les variables sont toutes de type numérique ;
- lorsque l'on saisit un texte dans une telle variable, Statistica génère une valeur numérique (101 pour le premier texte, 102 pour le second, etc) et traite le texte comme une étiquette associée à ce nombre. Ce texte est appelé "valeur texte" par le logiciel.

Exemple : définissez un nouveau classeur, avec une feuille de données vierge. Saisissez successivement Homme, Femme puis 101 dans la première colonne. Observez le résultat.

La liste des valeurs textes définies pour une variable donnée peut être consultée en double-cliquant sur la tête de la colonne considérée, puis en cliquant sur le bouton Valeurs-texte :



On peut également indiquer les valeurs texte correspondant à notre codage des données, saisir les données sous forme numérique et obtenir l'affichage sous forme de texte.

Définissez un nouveau classeur, avec une feuille de données. Indiquez les noms des variables et les valeurs texte correspondantes, en suivant la description de la page 2, puis saisissez quelques lignes. Par exemple :

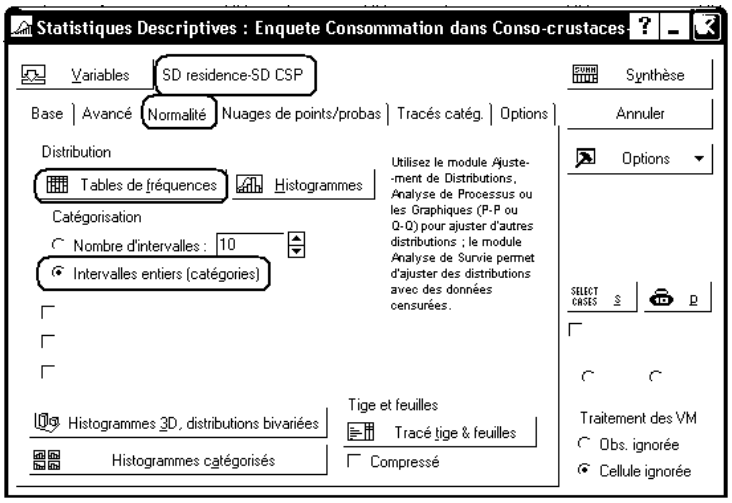
Conso	SD residence	SD age categorise	SD sexe	SD etudes	SD CSP
15,5375	2	1	2	3	3
7,7605	2	2	1	1	8
1,2815	2	3	1	3	7
7,9387	3	4	2	4	7
38,4395	2	2	1	2	7

2.2.4 Tri à plat avec Statistica

Ouvrez le classeur Statistica Conso-Crustaces.stw

Possibilité de filtres automatiques analogues à ceux d'Excel : utilisez le menu Données - Filtre automatique après avoir sélectionné les colonnes sur lesquelles vous souhaitez disposer d'un filtre.

Pour effectuer un tri à plat sur les différentes variables concernées, on peut utiliser le menu : Statistiques - Statistiques élémentaires - Statistiques descriptives, puis l'onglet Normalité et le bouton Tables de fréquences. Indiquez également une catégorisation par intervalles entiers :

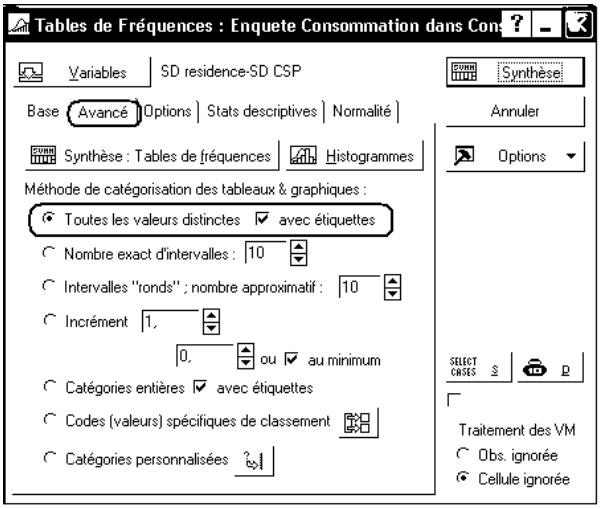


Le même menu permet également de représenter la distribution de chacune des variables sélectionnées sous forme d'histogramme.

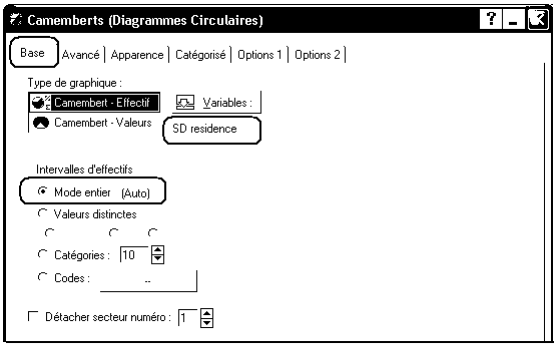
Exemple de résultat :

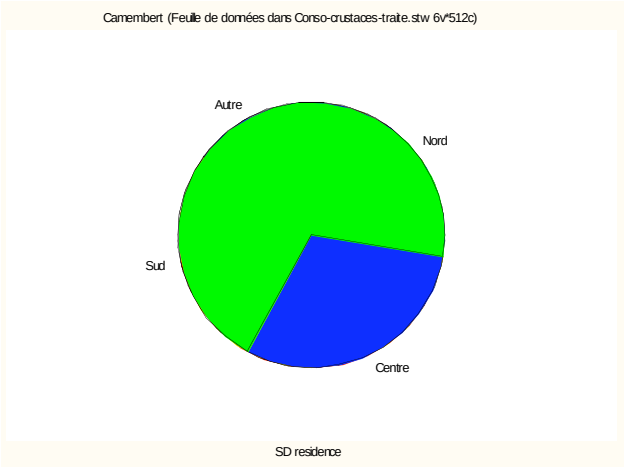
	Table de fréquences : SD sexe (Enquete Consommation dans Conso-crustaces-traite.stw)					
Catégorie	Effectifs	Effectifs Cumulés	% Indiv. Actifs	% Cumulé Ind. Act.	% toutes Observ.	% Cumulé du Total
Homme	195	195	38,08594	38,0859	38,08594	38,0859
Femme	317	512	61,91406	100,0000	61,91406	100,0000
VM	0	512	0,00000		0,00000	100,0000

On peut obtenir des résultats analogues à l'aide du menu Statistiques - Statistiques élémentaires - Tables de fréquences, puis l'onglet Avancé et le bouton Tables de fréquences



Les représentations graphiques sous forme de graphiques à barres pourront être obtenues à l'aide du bouton Histogrammes des dialogues précédents. On peut aussi obtenir des diagrammes circulaires à l'aide du menu Graphiques - Graphique en 2D - Camemberts :

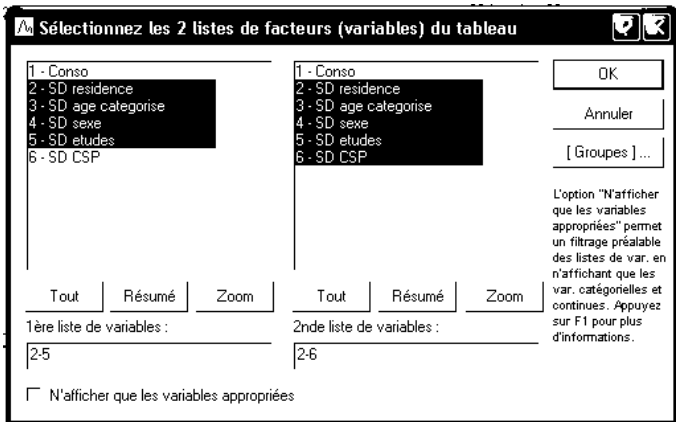




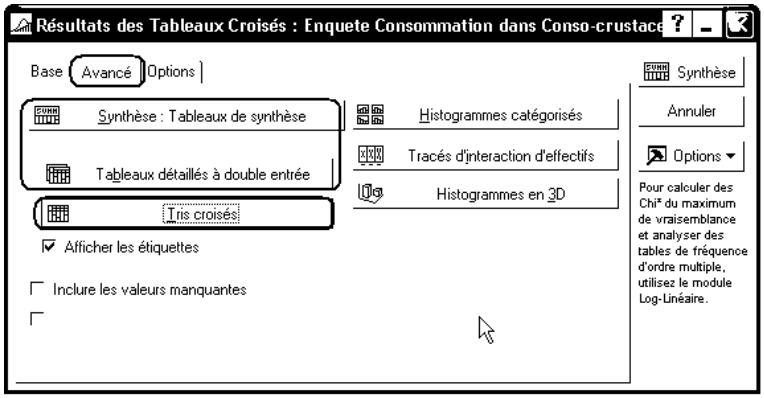
2.2.5 Tris croisés et graphiques catégorisés

On utilise le menu Statistiques - Statistiques élémentaires - Tableaux et tris croisés.

L'onglet Tris croisés permet de croiser les variables deux à deux :
On indique deux listes de variables, une même variable pouvant être présente dans chacune des listes.



On peut alors afficher les résultats sous deux formes : dans l'onglet Avancé, le bouton Synthèse et le bouton Tableaux détaillés à double entrée génèrent autant de feuilles de résultats que de couples de variables obtenus en croisant les deux listes. Le bouton Tri croisés permet d'obtenir les résultats rassemblés dans un seul tableau.



Exemple :

SD residence	Tableau de Synthèse : Effectifs Observés (Enquete Consommation dans Conso-crustaces-traite.stw) Effectifs en surbrillance > 10 SD age categorise - de 50 ans	SD age categorise 50 ans à 60 ans	SD age categorise 60 ans à 65 ans	SD age categorise 65 ans et plus	Totaux Ligne
Nord	29	42	36	34	141
Centre	35	46	42	33	156
Sud	27	44	38	30	139
Autre	29	13	16	18	76
Total	120	145	132	115	512

L'onglet Tableaux croisés permet d'indiquer plus de deux listes de variables. Statistica définit alors un facteur, en combinant les modalités des variables des premières listes et croise ce facteur avec les modalités des variables de la dernière liste. On peut ainsi, par exemple, obtenir un tableau de contingence croisant les combinaisons Sexe-Age avec la variable Lieu de résidence :

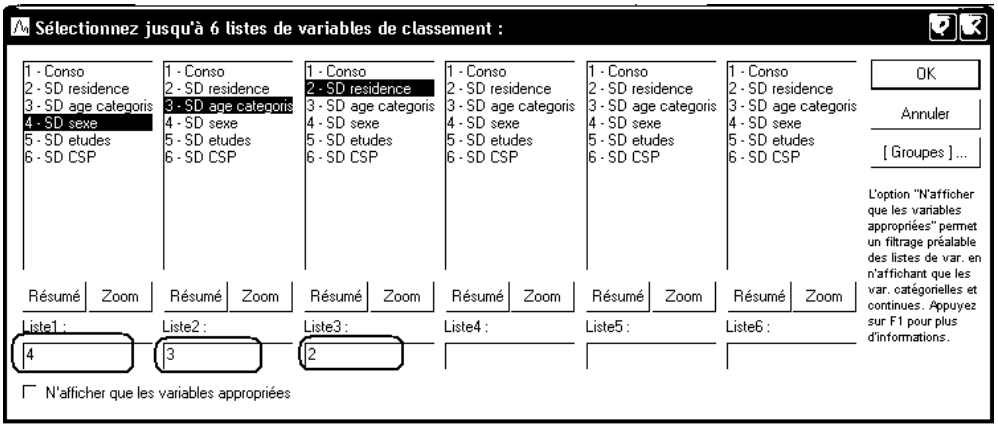
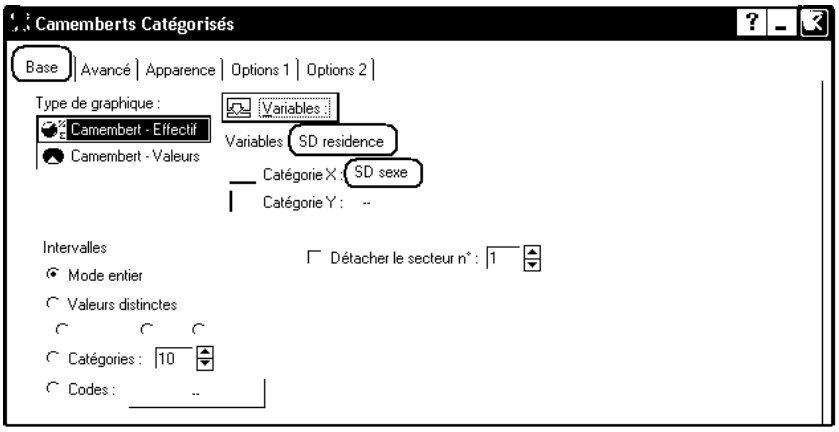
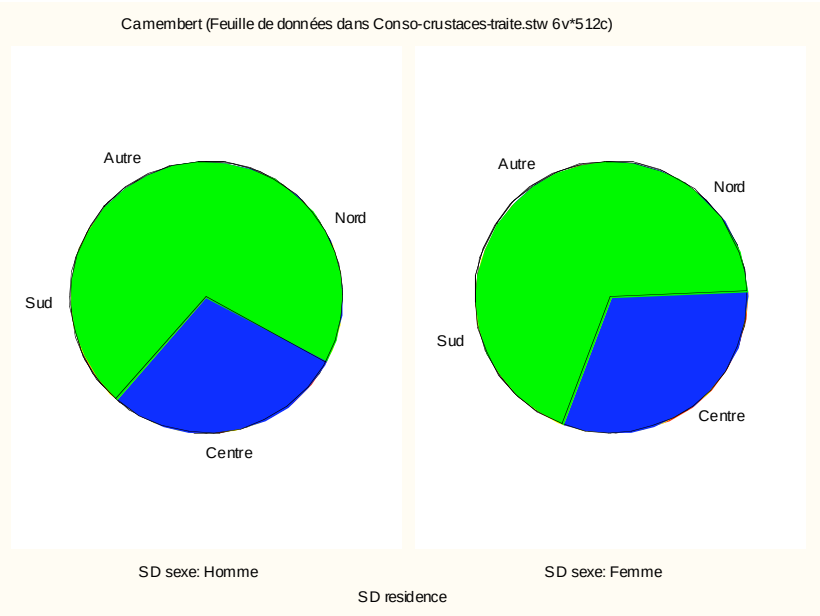


Table de Fréquences - Synthèse (Enquete Consommation dans Conso-crustaces-traite.stw)
Effectifs en surbrillance > 10
(effectifs marginaux non marqués)

SD sexe	SD age categorise	SD residence Nord	SD residence Centre	SD residence Sud	SD residence Autre	Totaux Ligne
Homme	- de 50 ans	13	14	12	11	50
Homme	50 ans à 60 ans	21	13	20	4	58
Homme	60 ans à 65 ans	17	15	8	5	45
Homme	65 ans et plus	13	14	9	6	42
Total		64	56	49	26	195
Femme	- de 50 ans	16	21	15	18	70
Femme	50 ans à 60 ans	21	33	24	9	87
Femme	60 ans à 65 ans	19	27	30	11	87
Femme	65 ans et plus	21	19	21	12	73
Total		77	100	90	50	317
Tot. Colonne		141	156	139	76	512

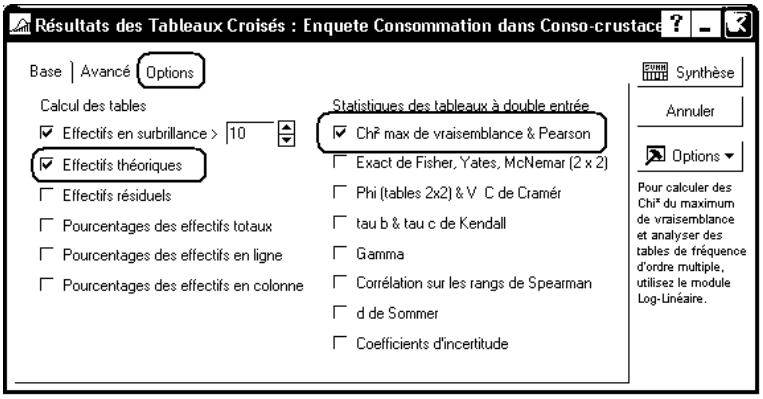
Les boutons Histogrammes catégorisés et Histogrammes en 3D permettent de représenter les données sous forme de diagrammes à bandes. On peut aussi obtenir une juxtaposition de diagrammes circulaires à l'aide du menu Graphiques - Graphiques catégorisés - Camemberts. Par exemple :





2.2.6 Test du khi-2 sur un tableau de contingence

Le module Tableaux et Tris croisés permet aussi de calculer une statistique du khi-2 pour chacun des tableaux générés. Pour cela, il suffit, dans l'onglet Options, de cocher la boîte Khi-2 Max de vraisemblance et Pearson, et de demander l'affichage de l'un ou l'autre des résultats complémentaires proposés :



Par exemple, il semble y avoir un lien significatif entre l'âge et le niveau d'études dans la population étudiée.

Synthèse : Effectifs Théoriques (Enquete Consommation dans Conso-crustaces-traite.stw)
Effectifs en surbrillance > 10
Chi? de Pearson : 58,5299, dl=9, p=,000000

SD etudes	SD age categorise - de 50 ans	SD age categorise 50 ans à 60 ans	SD age categorise 60 ans à 65 ans	SD age categorise 65 ans et plus	Totaux Ligne
Primaire	33,7500	40,7813	37,1250	32,3438	144,0000
CAP/BEP	34,4531	41,6309	37,8984	33,0176	147,0000
Secondaire	25,3125	30,5859	27,8438	24,2578	108,0000
Supérieur	26,4844	32,0020	29,1328	25,3809	113,0000
Ts Grpes	120,0000	145,0000	132,0000	115,0000	512,0000