

TRAVAUX DIRIGES DE STATISTIQUE (année P2)

ELEMENTS DE STATISTIQUE INFERENTIELLE

Ecole Nationale Vétérinaire de Toulouse

Département des Sciences biologiques et fonctionnelles

UP de Biométrie

F. Lyazrhi , D.Concordet

Objectif du TD :

Ces travaux dirigés sont l'occasion d'illustrer à travers des simulations et des exemples concrets les différentes notions de base en calcul de probabilités et statistique vues en cours. Nous insisterons sur les résultats fournis par les logiciels statistiques car ceux-ci sont de plus en plus élaborés, et les résultats qu'ils fournissent sont plus ou moins facilement interprétables.

Le logiciel que nous utiliserons tout le long de ces travaux dirigés est **SYSTAT**.

Plan

Première partie : Application du calcul des probabilités à la décision médicale

Deuxième partie : Présentation de SYSTAT

Troisième partie : Analyse statistique à l'aide de SYSTAT

Première partie : Calcul de probabilités et décision médicale**Exercice 1 :**

Le quart d'une population de chiens a été vaccinée contre une maladie contagieuse. Au cours d'une épidémie, on constate qu'il y a parmi les chiens malades un vacciné pour quatre non vaccinés. On sait, de plus, qu'au cours de cette épidémie il y avait un malade sur douze parmi les vaccinés.

Quelle était la probabilité qu'un chien soit malade alors qu'il n'est pas vacciné ? le vaccin est-il efficace ?

Exercice 2 :

Parmi 300 chiens atteints d'une maladie M , on a dénombré 200 chiens qui souffrent d'une insuffisance rénale. Par ailleurs, on a observé 30% de chiens présentant une insuffisance rénale dans un échantillon de 1500 chiens représentatif de l'ensemble de la population des chiens.

Calculer le risque relatif d'être atteint de la maladie M pour les chiens atteints d'une insuffisance rénale par rapports à ceux qui ne le sont pas.

Exercice 3 :

Dans une population de brebis deux maladies M_1 (listériose) et M_2 (mammites) sont présentes respectivement chez 10% et 20% des brebis (le nombre de celles qui souffrent des deux maladies est négligeable). On entreprend un dépistage systématique des maladies M_1 et M_2 . Pour cela on applique un test qui réagit à la maladie sur 90% des malades de M_1 , sur 70% des malades de M_2 , et sur 10% des brebis qui n'ont aucune de ces deux affections.

1. Quand on choisit au hasard une brebis de la population quelle est la probabilité pour que le test réagisse ?
2. Sachant que pour une brebis le test a réagi, donner les probabilités pour que ce soit à cause de M_1 , à cause de M_2 , sans que cette brebis ait l'une des deux maladies.
3. On décide pour les brebis dont le test est positif, d'effectuer des examens divers et éventuellement les traiter. En moyenne le coût pour une brebis atteinte de M_1 est de 1500 F, pour une brebis atteinte de M_2 il est de 1000 F, et pour une brebis non malade il est de 400 F.

Donner la moyenne de ce coût sur l'ensemble des brebis ayant un test positif.

Si on répartit le coût uniformément sur l'ensemble de la population des brebis, combien devra-t-on payer pour chaque brebis de la population ?

Exercice 4 :

Un chien de 9 ans souffre d'insuffisance cardiaque congestive, il est traité par digoxine, et il vient d'être hospitalisé. A l'entrée, sa digoxinémie (concentration plasmatique) est de 2.5 ng/ml.

On sait que parmi les chiens hospitalisés, 32 % présentent une intoxication digitale.

Dans la pratique, le test de diagnostic utilisé consiste à décider qu'un chien est malade si sa digoxinémie dépasse 1.7 ng/ml.

On donne :

La probabilité que le test soit positif sachant qu'un chien est malade : $P(T^+ / M^+) = 58\%$

La probabilité que le test soit négatif sachant qu'un chien est sain : $P(T^- / M^-) = 85\%$

Calculer la probabilité que le chien soit malade sachant que le test de diagnostic est positif.

Exercice 5 :

Diagnostic d'angine streptococcique

Environ 30% des adultes jeunes (11 à 30 ans) et souffrant de mal de gorge, présentent une pharyngite streptococcique qui peut être mise en évidence à l'aide de deux tests de diagnostic T1 et T2. Le premier test est basé sur la prise de température et le second sur l'analyse d'un prélèvement pharyngé.

	T1 : prélèvement pharyngé	T2 : Fièvre $\geq 38.3^\circ$
sensibilité	0.95	0.65
spécificité	0.70	0.85

Le tableau ci-dessus montre la sensibilité et la spécificité de chacun des tests.

1°) Calculer pour chacun des tests la probabilité qu'un patient soit malade sachant que son test est positif, c'est-à-dire $P(M+/T1+)$ et $P(M+/T2+)$.

2°) Supposons que les deux tests soient indépendants. Calculer la sensibilité et la spécificité du test T3 qui consiste à déclarer le patient malade si, et seulement si, les deux tests sont positifs, et à le déclarer sain dans le cas contraire. Soit, $P(T1+ \text{ et } T2+/M+)$ et $P(T1- \text{ ou } T2-/M-)$.

3°) Calculer $P(M+/T3+)$.

4°) Connaissant les propriétés de la maladie et de son traitement :

-Quelles sont pour le clinicien les probabilités qu'il doit interpréter ?

-Quel est "le meilleur" des trois tests que vous venez d'étudier ?

Exercice 6 :

On propose un questionnaire comprenant 10 questions qui comportent chacune deux réponses possibles l'une vraie, l'autre fausse. Pour tester si la personne interrogée essaie de deviner « au hasard », on adopte la règle de décision suivante :

-si 7 réponses, ou pus, sont bonnes, on admet que la personne interrogée n'a pas essayé de « deviner au hasard » ;

-sinon, on admet la conclusion contraire.

1. Quelle est la probabilité de rejeter l'hypothèse « la personne interrogée a essayé de deviner au hasard les réponses », quand celle-ci est vraie ?

2. Que devient cette probabilité lorsque chacune des questions posées comporte 3 réponses dont une seule est vraie ?

Rappels :

1°) - si A et B sont deux événements indépendants, alors $P(A \cap B/C) = P(A/C) \cdot P(B/C)$

- si A et B sont deux événements disjoints, alors $P(A \cup B/C) = P(A/C) + P(B/C)$

2°) Formule de Bayes :
$$P(A/B) = \frac{P(B/A) P(A)}{P(B/A) P(A) + P(B/\bar{A}) P(\bar{A})}$$

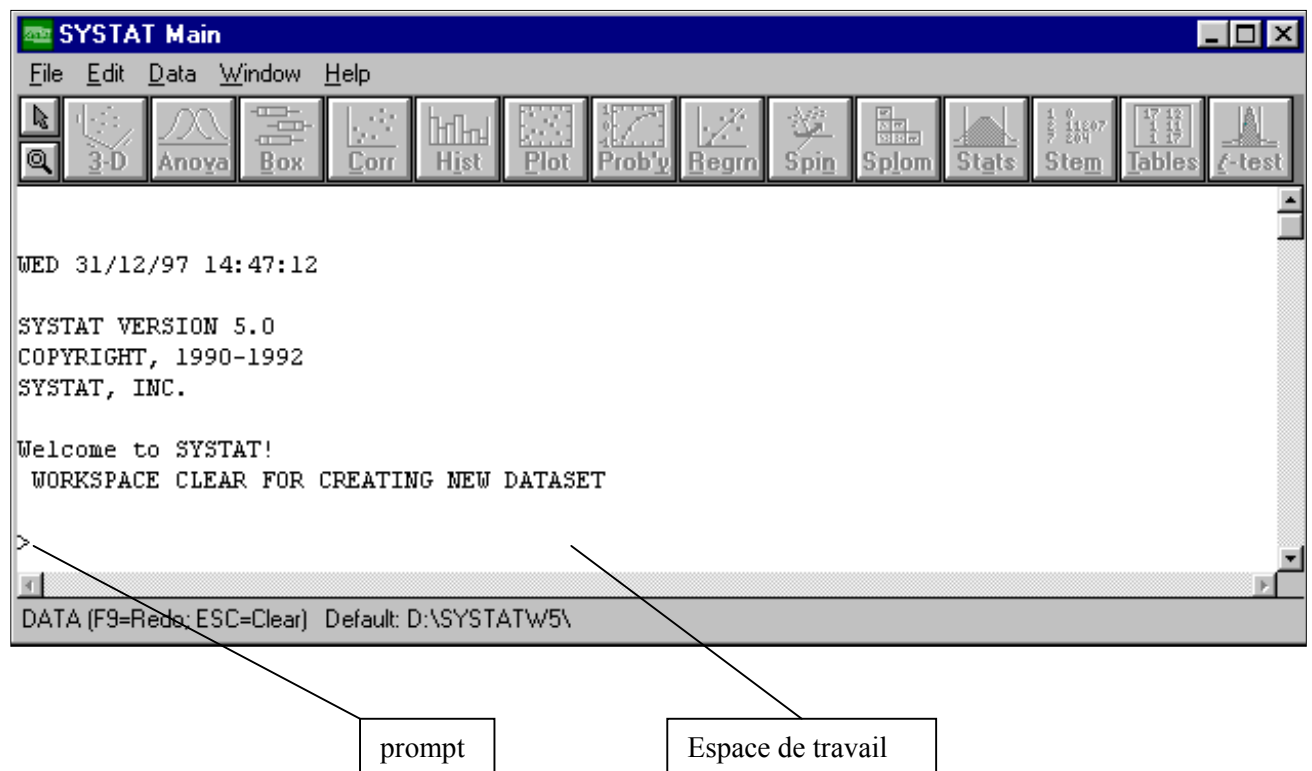
Deuxième Partie : Introduction à l'utilisation de SYSTAT

Systat est un logiciel statistique qui permet de préparer les graphiques et de faire des calculs nécessaires à l'analyse des données. Avec sa vaste panoplie de fonctionnalités, SYSTAT est un des produits d'analyse statistique les plus importants. Cette présentation de SYSTAT est loin d'être exhaustive mais permet de donner un aperçu sur les différents outils statistiques qu'offre SYSTAT.

Dans une première partie nous donnons une présentation rapide des différents menus de SYSTAT, ainsi que la gestion des différentes fenêtres générées par SYSTAT

Lancement de Systat

Pour lancer Systat, cliquer deux fois sur l'icône **SYSTAT**. Si vous avez bien lancé Systat, vous obtenez la fenêtre suivante :



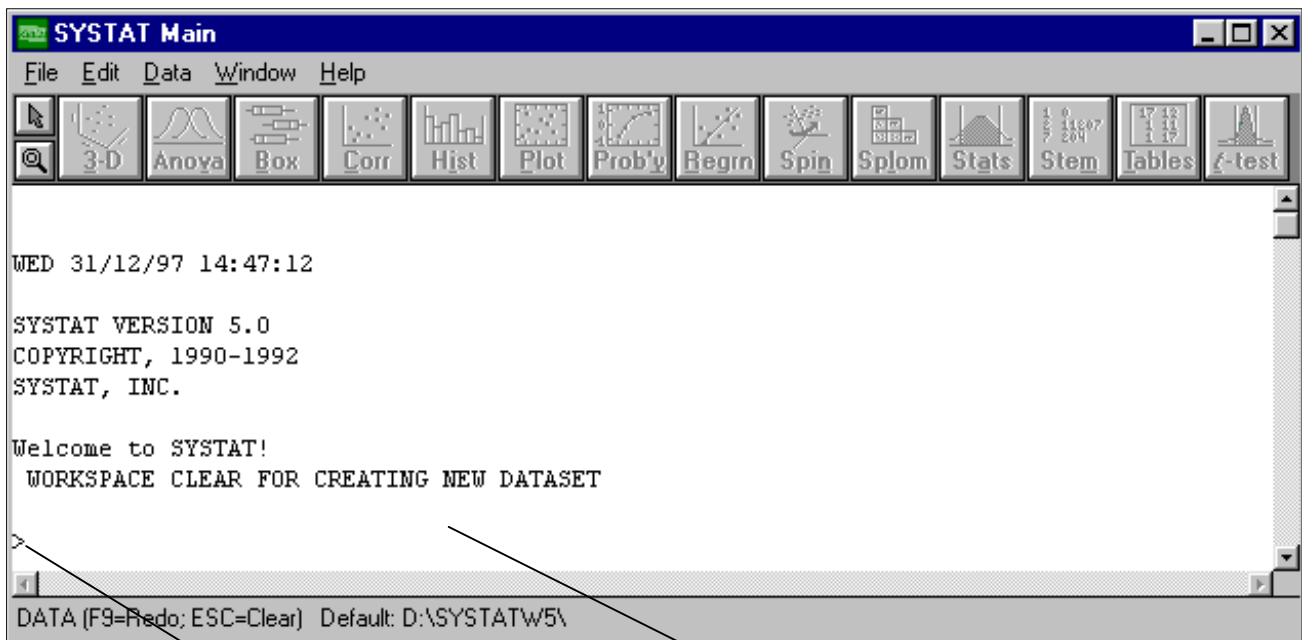
Une présentation rapide des menus et fenêtres de SYSTAT

Systat profite pleinement de l'interface Windows avec menus déroulants, zones de dialogues, fenêtres redimensionnables.....vous pouvez faire la plupart du travail à travers les menus et les boîtes de dialogue. Ce chapitre vous donne un aperçu des items du menu principal (fenêtre **Main**) et donne brièvement la liste des tâches possibles.

Des types différents de fenêtres sont utilisés pour diverses tâches. Les graphes sont tracés dans la fenêtre Graphe (**Graph**), les résultats d'analyse statistiques sont générés dans la fenêtre principale (**Main**), et ainsi de suite. Nous allons décrire les diverses fenêtres de Systat et les mécanismes nécessaires à leur utilisation.

La fenêtre principale de Systat

La figure suivante montre la fenêtre Main de systat, qui a des caractéristiques communes à toutes les fenêtres.



prompt

Espace de travail

Dans le coin supérieur droit de la fenêtre, il y a trois boutons. En cliquant sur le premier, vous pouvez mettre Systat en veille.

En cliquant sur le deuxième bouton, vous pouvez réduire ou agrandir la fenêtre de Systat. Le troisième vous permet de fermer la fenêtre de Systat.

La barre de menu SYSTAT

la ligne en dessous de la barre de titre de l'écran principal de SYSTAT est une barre de menu qui contient les opérations possibles.

Chaque menu de cette barre contient des commandes qui accomplissent certaines tâches. Pour utiliser une commande de menu, placer le curseur sur le nom du menu, cliquer avec la souris pour ouvrir le menu, et cliquer sur la commande de menu que vous voulez utiliser.

Pour un aperçu rapide, cliquer avec la flèche sur chaque menu comme expliqué, puis regarder ses commandes.

La plupart des commandes de menu sont en gris clair, vous ne pouvez pas les sélectionner pour l'instant parce que vous n'avez pas encore ouvert un fichier de données.

Les boutons de commandes raccourcis

En dessous de la barre de menu de SYSTAT, il y a une ligne de boutons à utilisation immédiate. Parce que vous n'avez pas ouvert un fichier de données, la barre de boutons à utilisation immédiate sur gris clair. Les boutons rapides sont utilisés pour effectuer des calculs statistiques et des graphes simples. Des options plus sophistiquées sont disponibles en utilisant toutes les possibilités des autres menus de SYSTAT. Nous verrons l'utilisation détaillée de ces boutons un peu plus loin.



Caractéristiques des menus

Quelques commandes du menu finissent par des points d suspension. Par exemple, sous le menu **File**, la commande **Save As.....** a des points de suspension contrairement à la commande **New**.

Chaque commande de menu suivie de points de suspension édite une boîte de dialogue vous demandant plus d'informations.

A droite de quelques commandes menu, vous pouvez trouver une abréviation de combinaison de touches comme, par exemple, à droite de la commande **Print...**, il y a «Ctrl+P». D'autres commandes menu, ont un menu supplémentaire, appelé menu emboîté. A droite de ces commandes se trouve un triangle. Par exemple la commande **New**, est suivie de trois autres sous-commandes Data, Main et Command. Vous pouvez créer un nouveau fichier de données, soit en utilisant la commande **Data**, soit, comme on le verra plus loin la commande la commande **Data** du menu **Window**.

Le Menu File

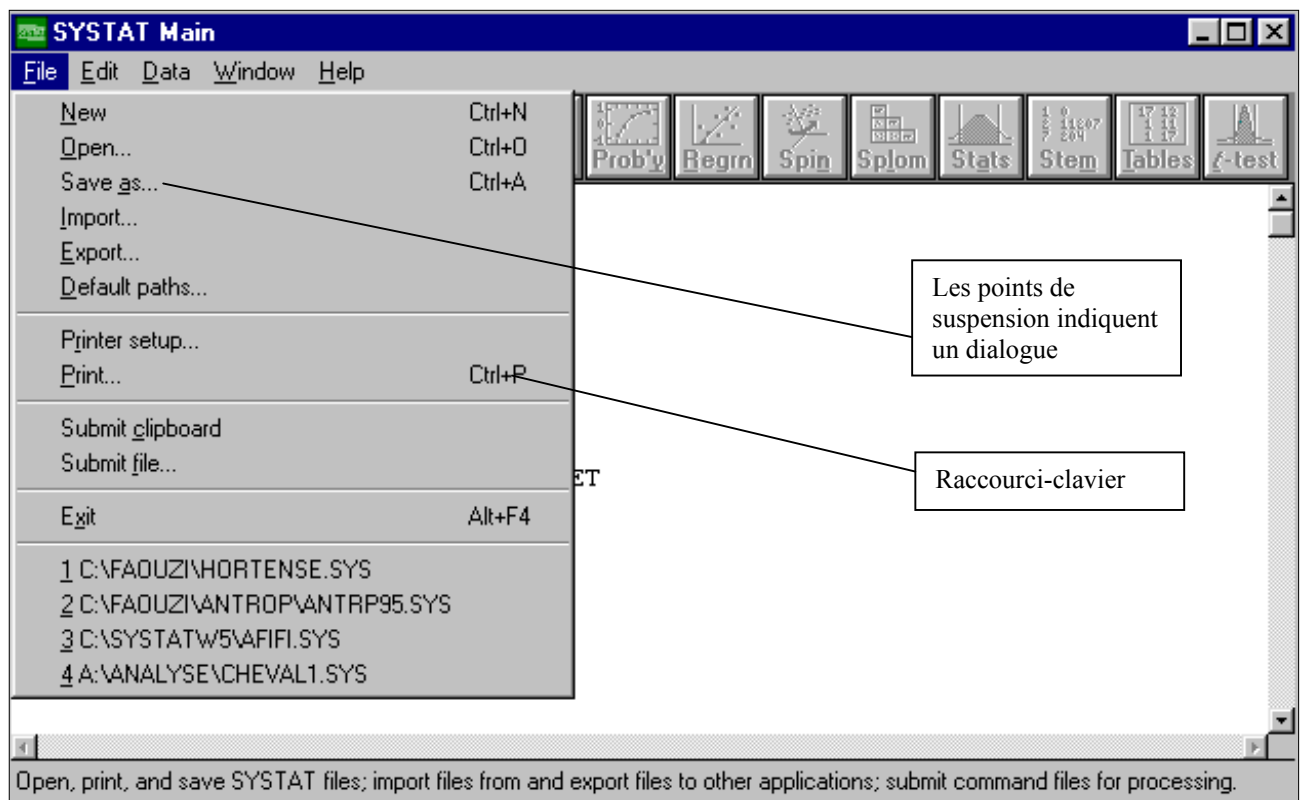
Pour voir le menu File :

Cliquer et maintenez appuyé le bouton gauche de la souris quand le curseur est sur **File** dans la barre du menu Main de SYSTAT.

Quand le menu File se déroule, la ligne d'état (ligne en bas de la fenêtre) change pour décrire ce qui se passera si vous cliquez sur le choix en surbrillance dans le menu. Quand la commande **Data** de l'option **New** est en surbrillance, la barre du bas inscrit « create a new Systat data file ». Si vous voulez connaître ce que les autres choix du menu font, pointez la commande **File**, puis maintenez appuyé le bouton de la souris. Vous pouvez maintenant pointer chaque option (tant que vous gardez le bouton de la souris appuyé) et lire la barre du bas pour voir ce qu'elle fait.

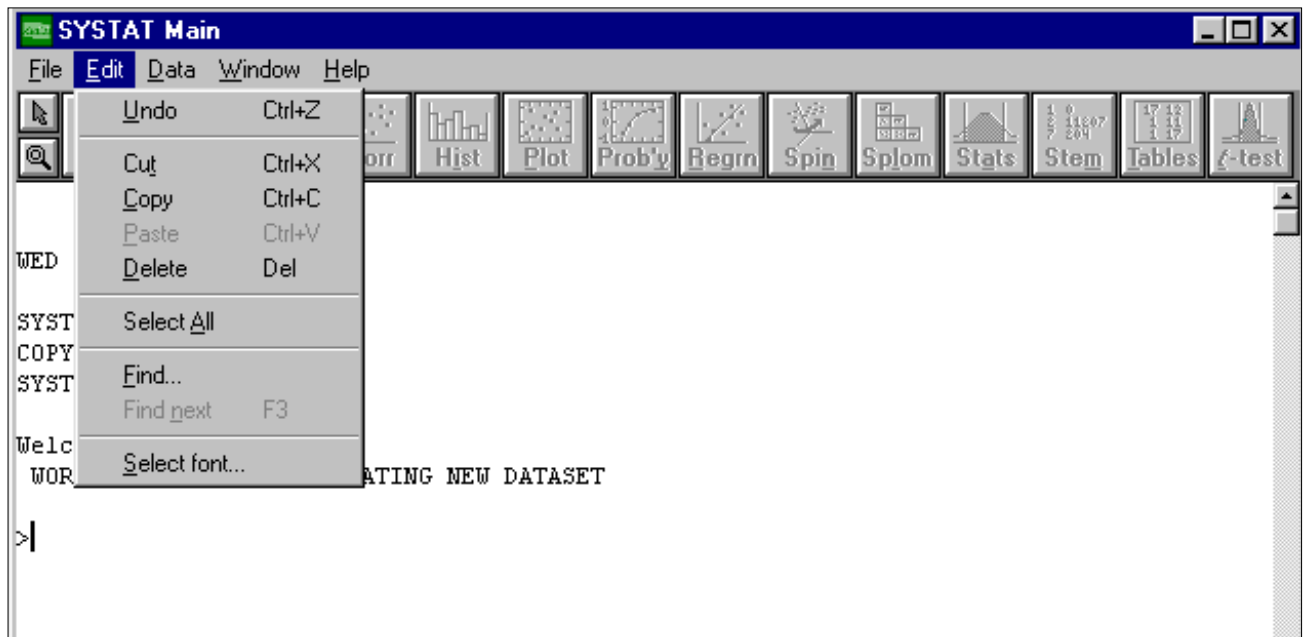
Vous utilisez le menu File pour :

- Créer de nouveaux fichiers de données SYSTAT, à l'aide de la commande **New**
- Ouvrir un fichier SYSTAT, **Open**
- importer ou exporter des fichiers Lotus, dBase, Excel..etc., **Open**
- Sauvegarder une sélection de texte ou un fichier de données, **Save As** (pour un nouveau fichier) ou **Save** (pour un fichier existant)
- Imprimer, **Print**



Le menu Edit

Pour voir le menu Edit, cliquer sur le menu Edit comme le montre la figure suivante :



Le menu Edit vous permet de :

- Annuler la plus récente commande d'édition de texte
- Couper (**cut**), copier (**copy**), coller (**paste**), des graphiques, des données et des commandes
- Effacer (**delete**) des observations ou des variables dans la fenêtre **Data**
- Effacer des sorties de la fenêtre Main
- Sélectionner le contenu d'une fenêtre
- Trouver un mot ou une phrase dans une fenêtre

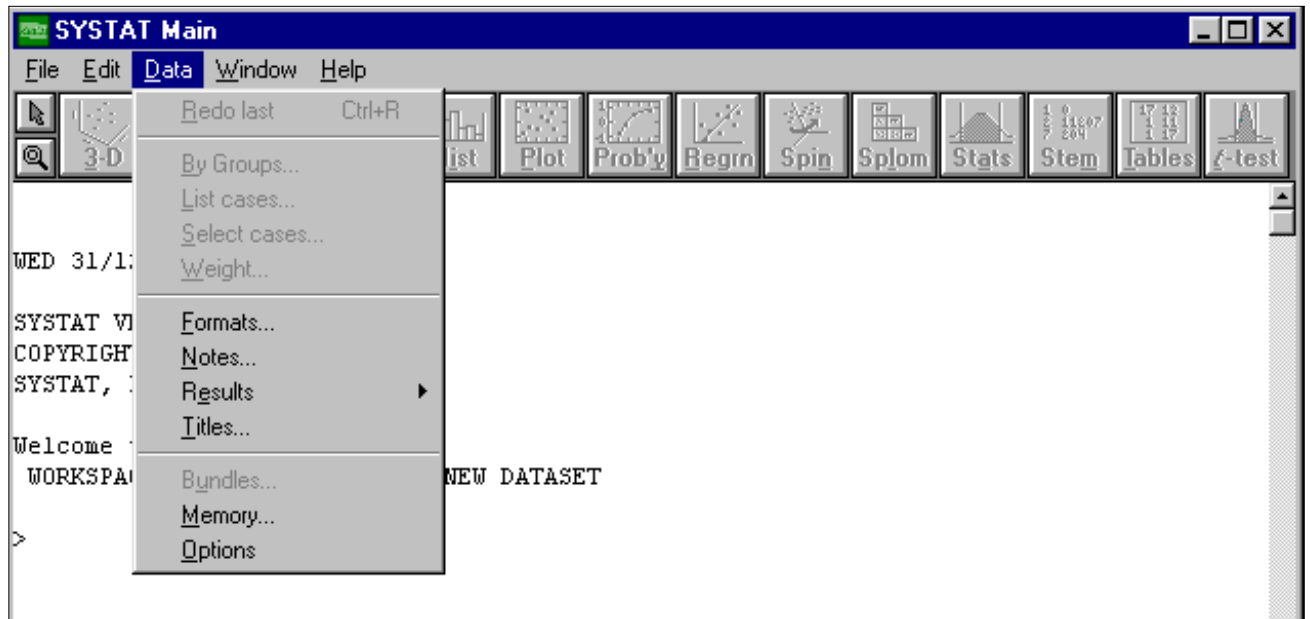
Le menu Data

Pour voir le menu **Data** :

Cliquer sur le menu Data qui apparaît comme dans la figure suivante :

Utiliser le menu Data pour :

- Transformer ou créer des données à l'aide de la commande **Let** et **If then Let** de l'option **transform**
- Trier les individus de vos fichiers selon les valeurs d'une ou plusieurs variables à l'aide de la commande **Select cases**
- Spécifier un ou plusieurs groupes de variables ; une analyse séparée est faite pour chaque sous-groupe à l'aide de la commande **By Groups**
- Trier les données dans l'ordre croissant des valeurs d'une variable donnée à l'aide de la commande **Sort**
- Spécifier une variable pondération ; Systat reproduit chaque donnée autant de fois que spécifié par la variable de pondération à l'aide de la commande **Weight**
- Annuler une sélection, tout le fichier de données est alors utilisé
- En cliquant sur la commande **Results**, vous pouvez diriger les sorties vers une imprimante, un fichier ou à l'écran (l'option par défaut)
- En cliquant sur la commande **Formats...**, vous définir le format de vos sorties : nombre de décimales, sorties longues, moyennes ou courtes.



Les Menus Graph et Stats

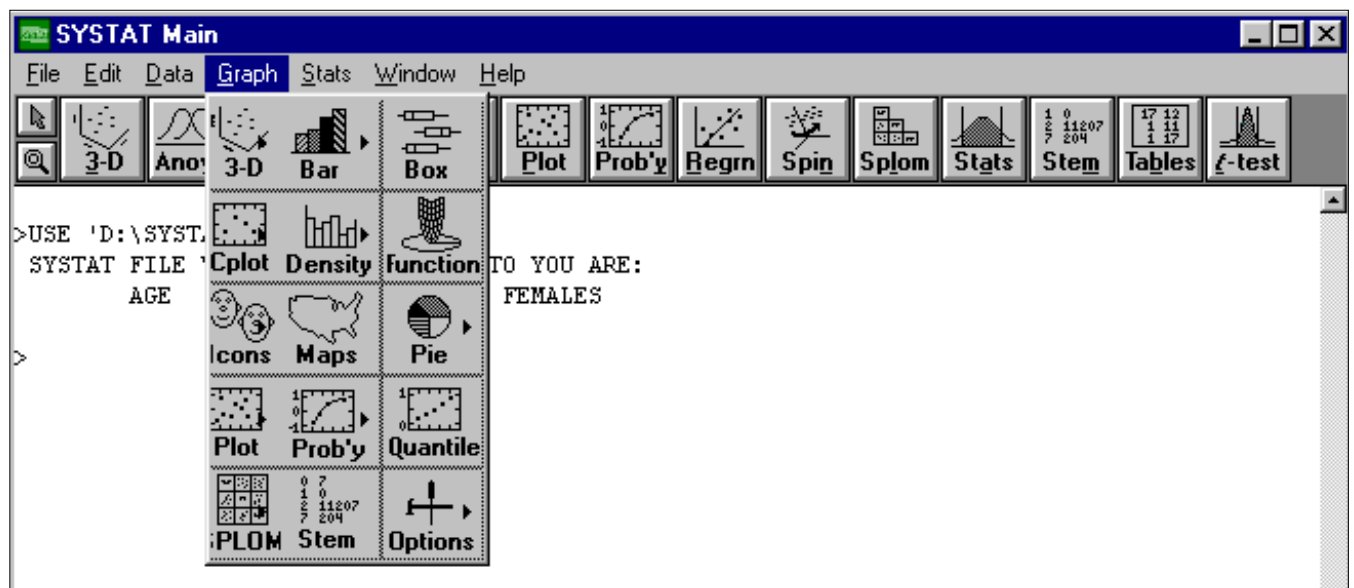
Il y a deux commandes supplémentaires dans la barre de menu qui apparaissent uniquement après que vous ayez chargé un fichier de données pour utiliser Systat. Ces commandes, qui sont Graph et Stats, ne sont utilisables que si les données sont disponibles. Vous ne pouvez voir ou accéder à ces menus maintenant, mais vous les utiliserez considérablement dans les chapitres suivants.

Le menu Graph

Le menu Graph, que vous ne verrez pas sur votre écran si vous n'avez pas ouvert un fichier de données, contient des boutons qui produisent des graphiques : boîtes à moustaches (**boxplot**), histogrammes, nuage de points (**scatterplot**).....etc. Le menu Graph permet de créer des graphiques plus sophistiqués que ceux réalisés par l'intermédiaire des boutons rapides. Le menu Graph permet de choisir les options graphiques (échelle, couleur, légende.....etc.).

Utiliser le menu Graph pour :

- Créer un graphe
- Sélectionner les options graphiques

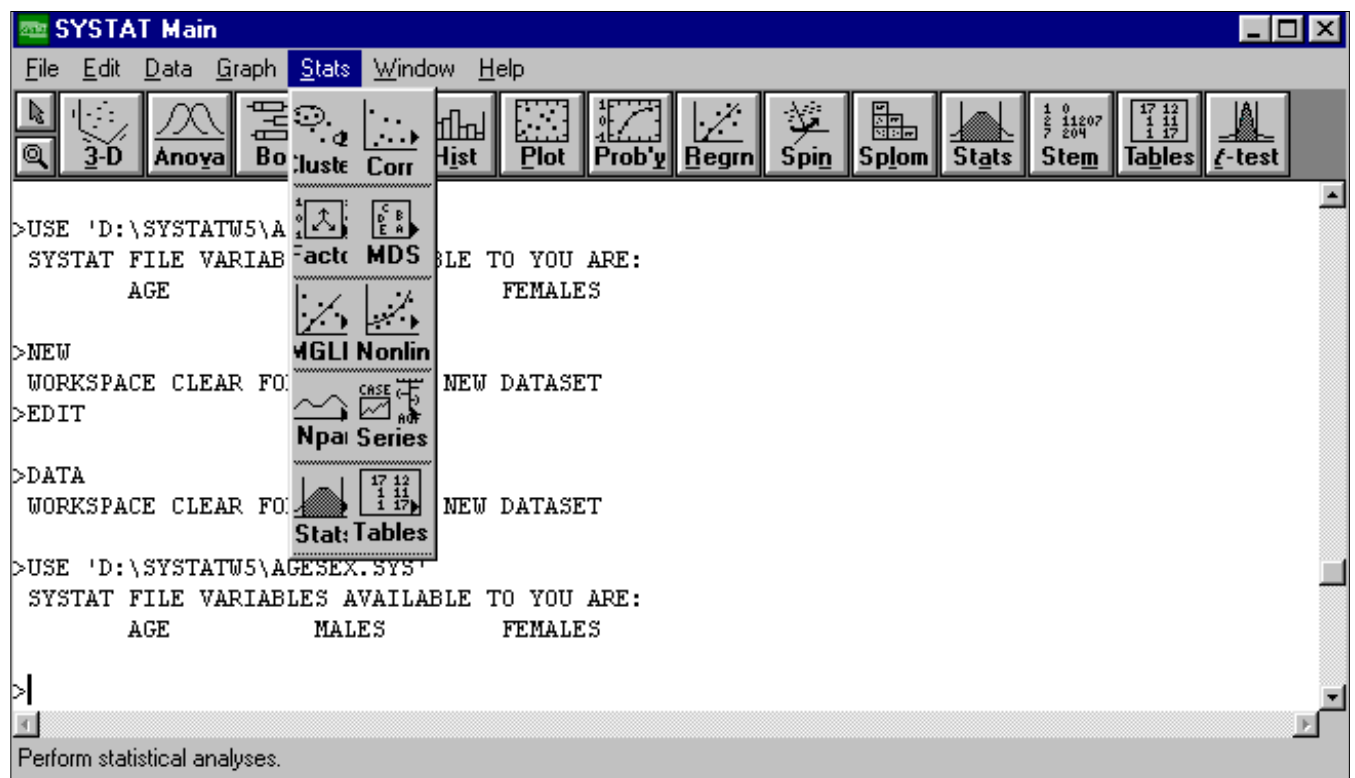


Le menu Stats

Le menu Stats, que vous ne verrez pas tant que vous n'ouvrez pas un fichier de données, contient des commandes d'analyses statistiques. La plupart de ces commandes font appel à des menus emboîtés (rappelez-vous que les commandes qui possèdent un menu emboîté ont un triangle à leur droite).

Ce menu de SYSTAT est le plus important. Il contient toutes les analyses statistiques que l'on peut faire avec SYSTAT, et parmi celles qui sont les plus utilisées dans la pratique :

- Statistique descriptive, à l'aide de la commande **Stats**
- Table de contingence, **tables**
- Test de Student, **t-test**
- Analyse de la variance (ANOVA), **Analysis of variance**
- Tests non paramétriques, **Npar**
- Régression linéaire, **Linear regression**
- Séries chronologiques, **Series**
- Modèle linéaire généralisé, **MGLH**
- Modèles non linéaires, **Nonlin**
- corrélation, **Corr**



Utiliser le menu Stats pour :

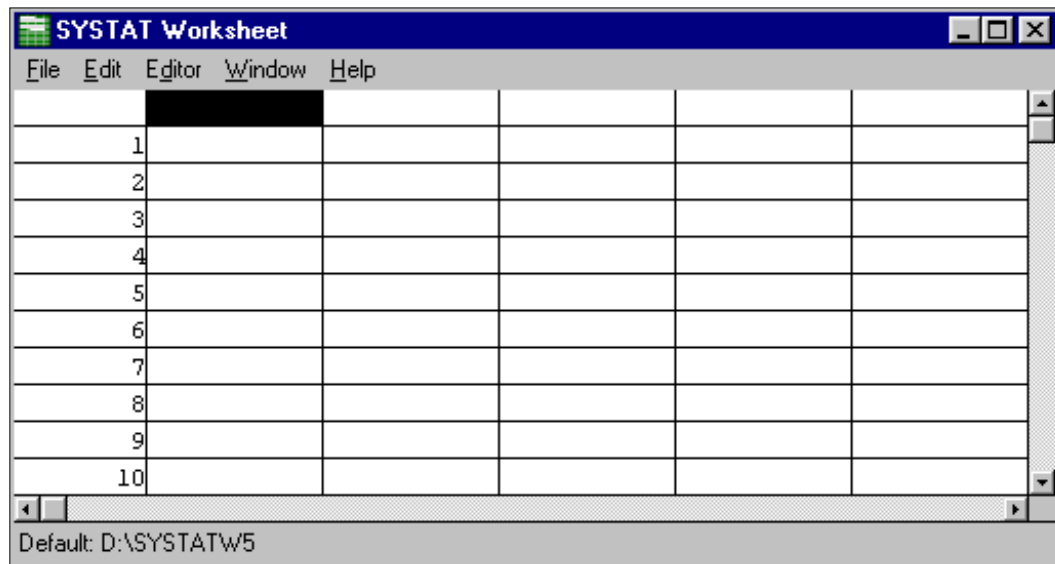
- Calculer des paramètres de statistiques descriptives
- Effectuer des analyses statistiques sophistiquées

Les Fenêtres SYSTAT

Tout logiciel Windows travaille avec un système de fenêtres. Dans SYSTAT, les données sont entrées, manipulées et visualisées dans la fenêtre **Worksheet**. Les résultats d'analyses statistiques sont exposés dans la fenêtre **Main**. Les graphes sont effectués dans la fenêtre **Systat Graph**. Les représentations graphiques en trois dimension (3-D) sont générés dans la même fenêtre.

La Fenêtre WORKSHEET

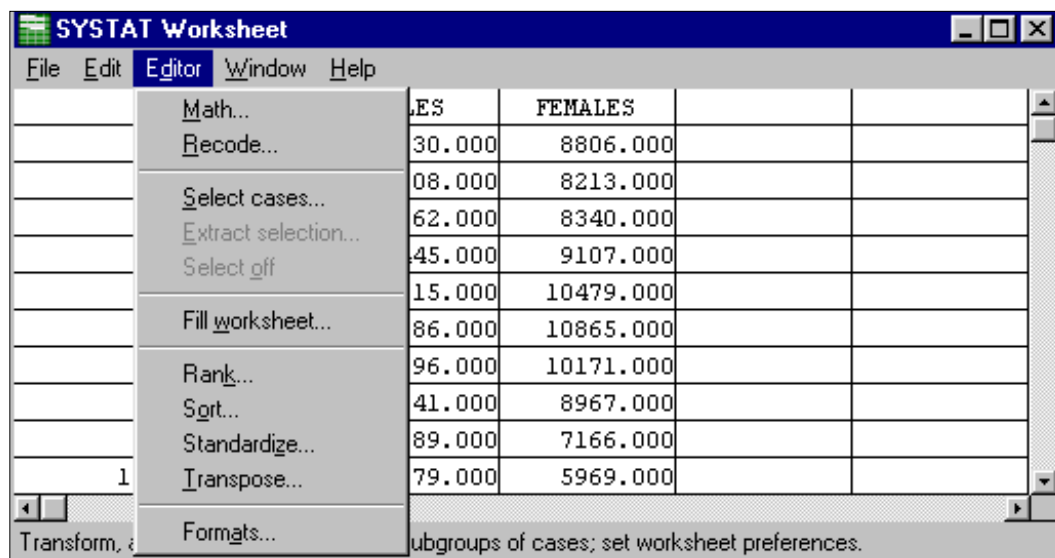
La fenêtre Worksheet présente vos données dans un format ligne-colonne. Chaque ligne correspond à un individu ou une observation, chaque colonne à une variable. L'intersection d'une ligne et d'une colonne est une cellule. Utiliser la barre de défilement en bas pour voir les autres variables et celle du côté pour voir les autres des individus.



Vous pouvez entrer de nouvelles données dans une fenêtre Worksheet vide, ou vous pouvez éditer ou transformer des données existantes.

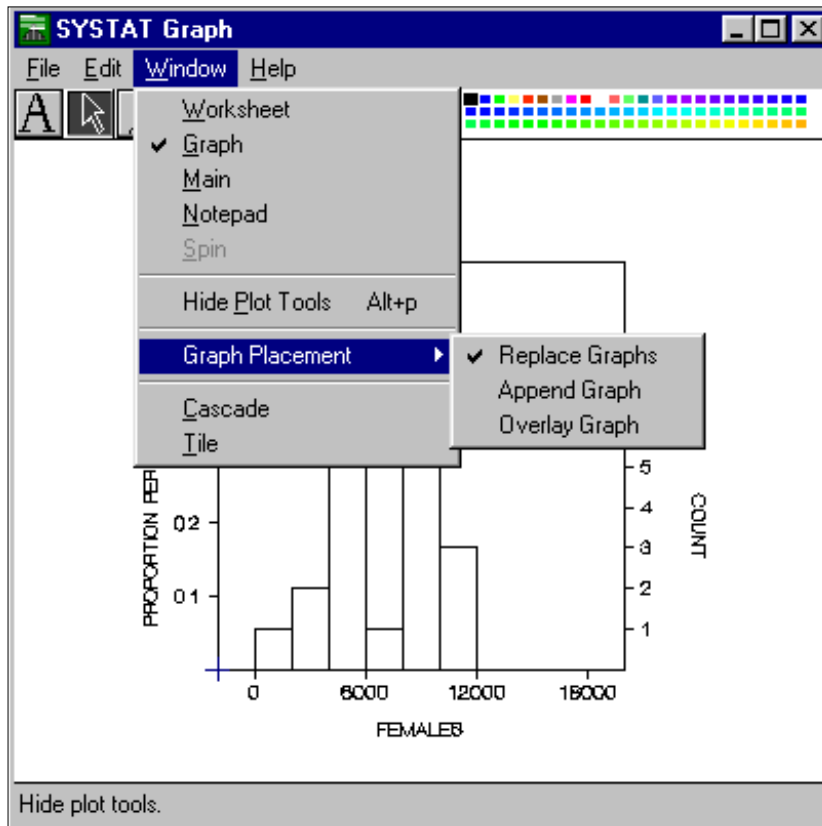
Dans la fenêtre Worksheet, vous pouvez utiliser le menu **Edit** pour couper, copier, coller et effacer des données en ligne, en colonnes. Vous pouvez aussi le menu **Editor** pour transformer des données, recoder des variables, chercher des individus satisfaisant certaines conditions, sélectionner des sous-ensembles de données selon des spécifications algébriques, et changer les attributs de la fenêtre. Vous pouvez aussi ordonner vos données suivant un critère, ou les transposer (les individus deviennent variables ...).

La commande **Fill worksheet**, vous permet de décider du nombre d'observations (lignes) d'une variable que vous voulez créer ou importer d'une autre application.



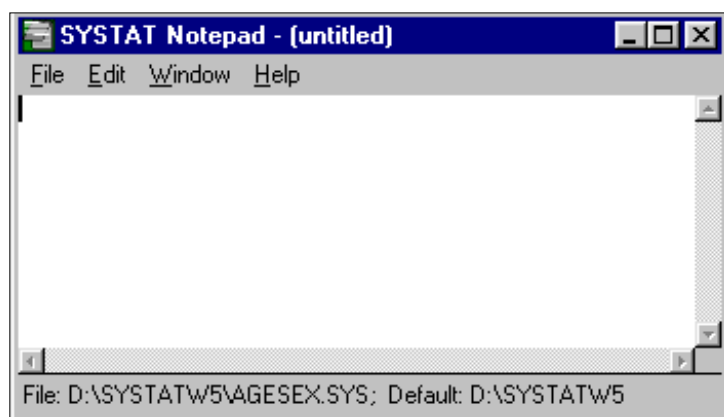
La fenêtre Systat Graph

Tous les graphiques sont effectués dans la fenêtre **Systat Graph**. Vous pouvez choisir de remplacer, ajouter, ou superposer des graphes avec la commande **Graph Placement** du menu Window de la fenêtre. **Graph placement** fournit un menu déroulant, montré dans la figure suivante, qui vous permet de choisir comment votre prochain graphique sera placé dans la fenêtre **Systat graph**. La commande **Append Graph** déroule la fenêtre **Systat Graph** pour masquer l'ancien graphe avant que le nouveau se trace. Vous pouvez alors dérouler la fenêtre pour voir les graphes précédents à l'aide de la petite flèche en bas à droite de la fenêtre **Systat graph**.



La fenêtre Notepad

Il y a aussi une fenêtre Editeur de commandes que vous pouvez utiliser pour créer, éditer et sauver des fichiers de commandes à exécuter. Cette fenêtre vous permet aussi d'ouvrir des fichiers texte.

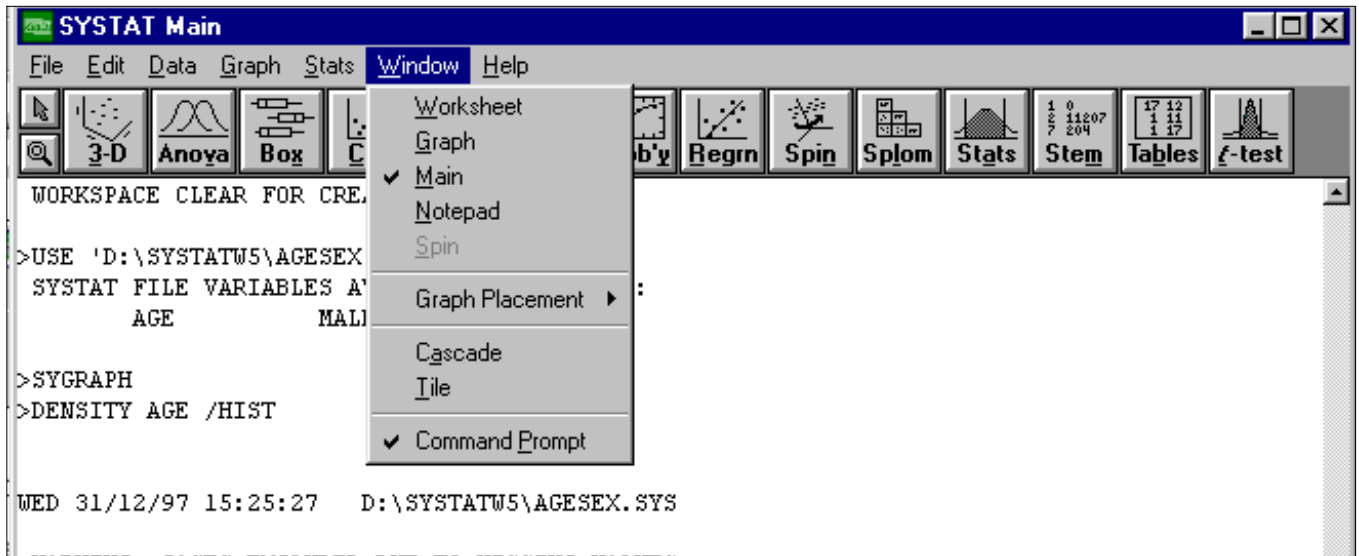


Une seule fenêtre peut être active à un instant donné. Les fenêtres inactives sont empilées derrière la fenêtre active. Une façon de rendre active est de déplacer la flèche sur une partie visible de la fenêtre que vous voulez activer et de cliquer. Finalement, vous pouvez utiliser le menu **Window** pour basculer d'une fenêtre à l'autre.

Le Menu Window

Le menu **Window**, montré ci-dessous, est utilisé pour :

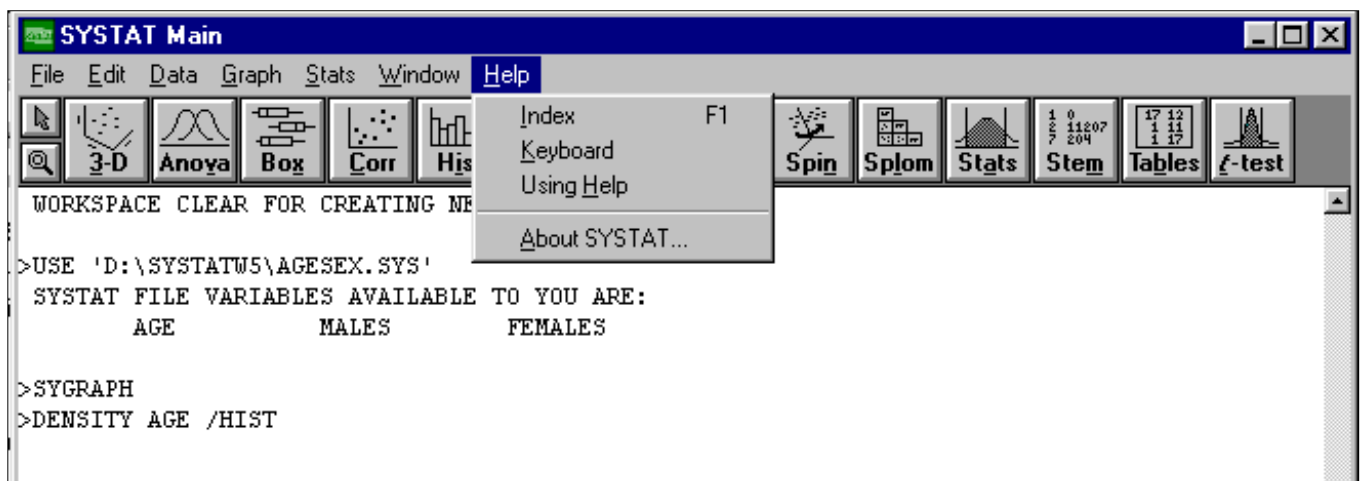
- Rendre actives ou non les Fenêtres **Worksheet**, **Main**, **Graph** et **Notepad**
- Passer en mode commande pour taper ou voir des commandes dans la fenêtre Main (**Command Prompt**).



Le Menu Help

Pour utiliser le menu **Help** :

Cliquer sur le menu **Help**, comme ci-dessous.



le menu **Help** est utilisé pour :

- Obtenir des informations sur les commandes Systat (**Index**)
- Découvrir comment utiliser **Help** (**Using Help**)

Troisième Partie : Analyse Statistique à l'aide de SYSTAT

Pour effectuer des calculs statistiques, les variables ont besoin d'être identifiées par un nom de telle sorte que l'on puisse les désigner facilement. Nous verrons que ceci dépend du fait que les variables sont quantitatives ou qualitatives. Leurs valeurs ont aussi besoin d'être enregistrées. Pour mettre en œuvre des modèles statistiques, on doit créer de nouvelles variables soit mathématiquement, soit à l'aide de variables déjà collectées. Finalement, les variables ont besoin d'être sauvegardées pour être utilisées dans des analyses futures. Nous allons voir à travers des exemples, comment Systat permet aux utilisateurs de faire tout cela.

Exercice 1 : La loi normale

Simulation d'un échantillon de nombres aléatoires
SYSTAT dispose de plusieurs fonctions qui génèrent des données qui proviennent des différentes lois classiques (loi normale, loi de student, etc.).

1. Pour obtenir un échantillon de 100 nombres aléatoires normaux :
 - Sélectionnez **File/New** dans la barre du menu Main pour obtenir une fenêtre **Worksheet** vide. Pour utiliser la fenêtre **Worksheet**, au moins une variable doit être nommée.
 - Cliquer sur la cellule du haut dans la première colonne de variable de la ligne 1 ; taper Z et appuyer sur [Entrée] pour appeler la première variable Z.
 - Sélectionner **Editor/Fill Worksheet...** dans la barre du menu Main. La boîte de dialogue de **Fill Worksheet** apparaît.
 - Dans cette boîte taper 100 pour le nombre de lignes (données).
 - Sélectionner **Editor/Math...** dans la barre de menu du **Worksheet** pour obtenir la boîte de dialogue de **Editor Math**.
 - Cliquer deux fois sur Z dans cette fenêtre en haut à gauche.
 - Taper **ZRN** dans la boîte variable or expression.
 - Sélectionner **Window/Worksheet** pour visualiser les 100 nombres aléatoires.
2. A l'aide des tables statistiques, calculer $P(-1 \leq U \leq 1)$ et $P(-2 \leq U \leq 2)$.
3. Ces probabilités sont-elles vérifiées par les données simulées ?
4. Calculer la moyenne et la variance de l'échantillon
5. Tracer l'histogramme et le Pplot de Z
6. Refaites la même chose pour plusieurs échantillons de taille 100. Que constatez-vous ?
7. Que se passe-t-il si la taille des échantillons augmente ?
8. Simuler un échantillon de 100 données provenant d'une loi normale de moyenne 50 et d'écart-type 12.
9. Tracer l'histogramme et le Pplot de cet échantillon

Exercice 2 : Théorème Central limit

La moyenne des données normales est normale !

Pour créer 1000 échantillons de 10 données normales :

- Sélectionner **File/New** dans la barre du menu **Worksheet** pour préparer le **worksheet** afin de saisir les données.
- Taper X1 (dans la cellule au dessus de la cellule #1) dans la fenêtre Worksheet et appuyer sur [Entrée] pour donner un nom à la première variable.
Maintenant, créer 10 variables ayant chacune 1000 composantes qui sont des observations normales, et calculer les milles moyennes de ces 10 composantes de ces variables que vous noterez MOY.
(cela va prendre un certain temps !).
- Tracer l'histogramme de MOY.
- Calculer la moyenne et la variance de MOY.
- Que constatez-vous ?

La moyenne des données qui ne sont pas normales peut-elle être normale ?

Pour le savoir, générer de la même façon que ci-dessus 1000 échantillons de 10 données uniformes (fonction **URN**).

1. Tracer l'histogramme et le Pplot de l'échantillon des moyennes
2. Calculer la moyenne et la variance de cet échantillon
3. Conclure

Exercice 3 : La loi de Student, Khi-deux et Fisher

On sait que la variable aléatoire $t = \frac{\bar{X} - m}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$ suit une loi de Student à $(n-1)$ degré de liberté.

1. Justifier cette affirmation

Dans SYSTAT cette distribution est désignée par **TDF**(t,df), où t est la valeur que prend la variable aléatoire et df est le degré de liberté.

2. Tracer plusieurs distributions de Student en faisant varier le degré de liberté (commande **Function** du menu **Graph**)
3. Comparez ces distributions à celle de la loi normale centrée réduite.
4. Donner un exemple de variable aléatoire qui suit un Chi-deux.

Dans SYSTAT, la loi du Khi-deux est désignée par **XDF**(x,df).

5. Tracer plusieurs distributions du Khi-deux en faisant varier le degré de liberté.
6. Donner un exemple de variable aléatoire qui suit une loi de Fisher.

Dans SYSTAT, cette distribution est désignée par **FDF**(f,df1,df2).

7. Tracer plusieurs distributions de Fisher en faisant varier les degrés de liberté.

Exercice 4 : Statistique descriptive

Ouvrez le fichier CK.XLS (vous l'avez déjà utilisé en informatique) qui se trouve dans votre répertoire. Si ce n'est pas le cas, copier le à partir du répertoire *Enonces* en utilisant *poste de travail*.

Copiez les valeurs de PICK des animaux en bonne santé (santé=1) de la classe d'âge 3 dans un fichier systat que vous nommerez CK.SYS

1°) Faites toutes les représentations graphiques que vous connaissez.

Choisissez le graphique qui à votre avis décrit le mieux ces données.

2°) Proposez un résumé numérique de ces données.

3°) Faites de même pour les variables du fichier EXE4.

Le fichier EXO4 contient les données que vous avez à résumer.

La variable VIE représente la durée de vie (en années) de 100 chiens.

La variable Ca contient la concentration en calcium dosé sur ces 100 chiens

La variable XX contient le résultat d'un dosage d'un autre paramètre biochimique.

Conseil :

faites tout d'abord des graphiques pour essayer de raccrocher la distribution que vous observez à des distributions que vous connaissez. Si vous arrivez à trouver une distribution qui colle à peu près, proposez les paramètres qui caractérisent cette distribution pour résumer les données. Si aucune distribution que vous connaissez ne "colle" aux données, il est possible qu'une transformation de la variable étudiée aura de "bonnes propriétés". Enfin si aucune transformation des données ne convient, donnez quelques percentiles de la distribution observée ainsi que le(s) mode(s) s'il(s) existe(nt). Dans ce dernier cas, il serait faux de résumer les données à MOYENNE \pm SD.

Exercice 5 : Décision statistique

Il est courant en médecine d'avoir recours aux tests diagnostiques. C'est-à-dire à l'utilisation d'une variable qui mesure indirectement l'état du sujet. Le principe de base de l'utilisation du test repose sur le fait que la variable prend par exemple des valeurs plus élevées chez les sujets malades que sur les sujets sains.

En utilisant les données du fichier EXO5, trouver un seuil de CK qui permettra de discriminer « au mieux » les sujets sains (santé=1) des sujets atteints d'affections musculaires (santé = 2).

Remarques :

1°) Afin de trouver le « meilleur » seuil, vous devez en comparer plusieurs. Faites varier le seuil de 10 à 100 en incrémentant à chaque étape de 10, puis de 100 à 500 avec un incrément de 100, enfin essayer un seuil de 1000.

2°) Il est clair que vous devez justifier le seuil que vous proposez.

3°) Il est possible de créer dans l'éditeur de SYSTAT des variables qui satisfont une condition logique particulière. Par exemple vous disposez dans l'éditeur d'une variable X. Vous voulez créer une variable Y telles que

$Y=1$ si $X \geq 10$ et

$Y=2$ si $X < 10$

Il faut pour cela utiliser la fonction : (Condition logique)

la fonction (Condition logique) renvoie la valeur 1 si la condition logique entre parenthèses est vraie, et la valeur 0 sinon. Donc pour créer Y il suffit d'écrire : $(X < 10) + 1$

Exercice 6 : Intervalle de confiance

Le fichier EXO6 contient les concentrations en calcium (mg/dL) mesurées sur un échantillon de 51 chats

1°) Décrivez les données de cet échantillon.

2°) Donnez un intervalle dans lequel sont incluses les concentrations en calcium de 95 individus sur 100.

Donnez un intervalle dans lequel sont incluses les concentrations en calcium de 99 individus sur 100.

Supposons maintenant que la moyenne et l'écart-type que vous avez observés proviennent en fait d'un échantillon de 100 chats.

Donnez un intervalle dans lequel sont incluses les concentrations en calcium de 95 individus sur 100.

Donnez un intervalle dans lequel sont incluses les concentrations en calcium de 99 individus sur 100.

3°) On s'intéresse à la moyenne de population de ces concentrations en calcium.

Donnez un intervalle dans lequel on a 95 chances sur 100 de trouver cette moyenne.

Donnez un intervalle dans lequel on a 99 chances sur 100 de trouver cette moyenne.

Supposons maintenant que la moyenne et l'écart-type que vous avez observés proviennent en fait d'un échantillon de 500 chats.

Donnez un intervalle dans lequel on a 95 chances sur 100 de trouver cette moyenne.

Donnez un intervalle dans lequel on a 99 chances sur 100 de trouver cette moyenne.

Comment expliquez vous cette différence ?

4°) Est-il probable que la moyenne de population des concentrations en calcium soit égale à 10 mg/dL ?

Est-il probable que la moyenne de population des concentrations en calcium soit égale à 15 mg/dL ?

5°) Y a-t-il une valeur aberrante ?

Exercice 7 : Comparaison de pourcentages

L'objet de cet exercice est d'étudier les techniques de manipulation des pourcentages.

Un essai a été réalisé pour évaluer l'efficacité thérapeutique d'un nouveau produit. Ce nouveau produit est à comparer à un produit de référence. Les résultats suivants ont été obtenus :

	REFERENCE	TEST	Total
MALADES	10	17	27
GUERIS	40	33	73
Total	50	50	100

1°) Donnez à l'aide de 2 méthodes différentes un intervalle de confiance à approximativement 95% du pourcentage d'animaux malades avec le produit de référence.

2°) A votre avis, le pourcentage d'animaux malades avec le produit de référence est-il égal à 35%, à 22% ?

Donnez quand vous le pouvez le risque que vous prenez en tirant votre conclusion.

3°) En utilisant l'intervalle de confiance que vous avez construit avec la loi normale, calculez la taille de l'échantillon qu'il aurait fallu prendre pour avoir une "précision" de $\pm 2\%$.

Pour répondre à cette question, vous supposerez que le pourcentage d'animaux malades ne change pas avec la taille de l'échantillon.

4°) On s'intéresse à "l'équivalence éventuelle" de ces produits. Comparez (par deux méthodes différentes) les pourcentages d'animaux guéris avec les deux produits. Explicitez clairement les hypothèses que vous testez et précisez (ou calculez) le risque que vous prenez en faisant cette affirmation.

5°) Vous voulez savoir maintenant si le produit de référence est meilleur que le produit à tester.

Explicitez clairement les hypothèses que vous testez et précisez (ou calculez) le risque que vous prenez en faisant cette affirmation.

6°) Le nombre de sujet est-il suffisant pour répondre à la question 4). Si non, calculez le nombre de sujets qu'il aurait fallu prendre.

7°) Le nombre de sujets est-il suffisant pour répondre à la question 5°). Si non, calculez le nombre de sujets qu'il aurait fallu prendre.

Remarque :

Pour les questions 4,5,6,7, faites l'inventaire de toutes les informations dont vous avez besoin pour répondre à la question.

Exercice 8 : Comparaison de moyennes

Une expérience a été menée sur deux groupes de chiens pour mesurer l'effet de deux antipyrétiques sur la température centrale (un plan expérimental plus performant sera vu dans l'exercice suivant).

Le fichier EXO8 contient les variables suivantes :

- TRT : numéro du traitement utilisé
- TEMP : température observée sur chaque animal.

1°) Décrivez les données du fichier EXO8 avec la(es) méthode(s) appropriée(s).

2°) L'antipyrétique n°2 est-t-il plus efficace que le n°1 ?

Pour répondre à cette question, écrivez les 3 hypothèses fondamentales (en les classant par ordre d'importance décroissante) et vérifiez les quand vous le pouvez.

Donnez le risque que vous prenez en faisant cette affirmation.

3°) Les températures moyennes induites par ces deux antipyrétiques sont-elles égales ?

Donnez le risque que vous prenez en faisant cette affirmation.

Exercice 9 : Echantillons appariés

Un essai a été réalisé sur des chiens pour mesurer l'effet de deux anti-inflammatoires (fichier EX09)

Le critère de jugement adopté est le diamètre de l'articulation du grasset.

Une première mesure du critère de jugement est réalisée avant le traitement, une seconde mesure est réalisée 12 heures après le traitement.

1°) Représentez graphiquement (par traitement) les résultats après traitement en fonction des résultats avant traitement. Qu'en concluez-vous ?

2°) Les traitements ont-ils d'après vous des effets différents ?

Pour répondre à cette question vous devez

- justifier l'emploi du (des) critère(s) de jugement que vous adoptez
- justifier l'emploi des tests que vous utilisez.

3°) Y a-t-il un intérêt à travailler avec un essai organisé de cette manière (avant - après) ?

Justifiez votre réponse en comparant la puissance du test que vous utilisez à celle qui serait obtenue avec un essai organisé comme dans l'exercice précédent.

Exercice 10 : comparaisons multiples

Une expérience est menée pour comparer le rendement de 4 élevages. On a prélevé le même nombre de veaux dans chaque élevage. Le GMQ de chaque animal est mesuré. Les résultats de cette expérience sont consignés dans le fichier EX010.

1°) Y a-t-il un effet élevage ?

2°) Quel est le meilleur élevage ?

