Statistique descriptive – 1re année [1]

Feuille de T.P. 1

Objectifs. Première utilisation du logiciel *STATISTICA*. Etude des différents graphiques (statistique univariée).

Instructions. Nous travaillerons sur les données que vous trouverez sur *moodle*. Copiez les fichiers du répertoire TP1 sur votre compte. Nous poursuivrons parfois l'étude des mêmes données durant plusieurs T.P.

La page internet `www.statsoft.fr` contient des informations intéressantes sur le logiciel *STATISTICA*. En particulier, on pourra trouver un document de prise en main du logiciel à l'adresse `www.statsoft.fr/pdf/StatSoftQuickRefFR.pdf`.

Vous répondrez aux questions de ce T.P. dans un fichier à part. Ce T.P. est à terminer pour le **lundi 15 octobre**. Le compte-rendu devra être déposé sur *moodle* (avant minuit) et devra être nommé :

`TP1_Prénom_Nom.pdf`

Si le compte-rendu a été fait en binôme, les deux étudiants doivent le déposer sur *moodle* et spécifier **explicitement** dans le texte les deux noms.

Vous devez expliquer votre démarche et commenter les résultats obtenus. Pensez à illustrer votre compte-rendu de graphiques lorsque vous le jugez utile et pertinent. Les graphiques devront comporter un titre et une légende. Reportez-vous au document *Conseils de rédaction pour les comptes rendus de T.P.* pour plus de détails.

1 Description du jeu de données

Depuis plus de 50 ans, aux États-Unis, le Bureau du Recensement et le Bureau des Statistiques du Travail réalisent chaque mois une étude appelée Current Population Survey en recueillant des informations sur les membres de plus de 15 ans d'environ 50 000 foyers. Cette étude est la principale source de renseignements sur les caractéristiques de la population active du pays. Les données du fichier *Recensement85.sta* ont été recueillies au cours de l'enquête de mai 1985 et portent sur des personnes de plus de 16 ans en activité au moment de l'étude.

Objectif : décrire la population dans son ensemble à l'aide de graphiques uniquement.

2 Prise en main de *STATISTICA*

2.1 Ouverture d'un jeu de données, Création d'un classeur

Ouvrir le logiciel *STATISTICA*. Deux modes d'affichages des menus et onglets sont proposés :

- Affichage "*Menus Classiques*"

1. Enseignants responsables des TP : G. Chagny (`gaelle.chagny@parisdescartes.fr`) et C. Laclau (`charlotte.laclau@parisdescartes.fr`)

- Affichage "*Ruban*"

Il est possible de passer d'un mode à l'autre à tout moment (voir le menu *Affichage*). Le chemin d'accès aux commandes et boîtes de dialogues peut différer légèrement d'un mode à l'autre. Dans la suite de l'énoncé de ce TP (et pour les suivants), les instructions sont indiquées pour le mode d'affichage *Ruban*.

Ouvrir le fichier de données *Recensement85.sta*. Le fichier qui s'affiche a la forme d'un tableau à double entrée. Cette grille contient les données :

- Chaque ligne correspond à un individu.
- Chaque colonne correspond à une variable.
- Chaque case est la valeur prise par une variable pour l'individu correspondant.

Dans la suite, nous allons créer de nombreux objets (graphiques, tableaux). Il est possible de tout enregistrer dans un document *STATISTICA* se présentant dans le logiciel comme un dossier avec arborescence. C'est un document appelé *Classeur* dont l'extension est "*.stw*".

En utilisant les onglets *Accueil>Nouveau*, créer un *classeur*. Ajouter la feuille de données *Recensement85.sta* dans le classeur. Vérifier que la feuille de données dite *active* est bien l'exemplaire de *Recensement85.sta* qui est dans le classeur : on pourra par exemple utiliser un clic droit sur le nom de cette feuille dans l'arborescence du classeur, et cliquer sur *Feuille de données active*.

2.2 Description des données

Décrire précisément la population étudiée, sa taille et l'unité statistique. Décrire avec soin les variables en indiquant leur type, l'ensemble de leurs valeurs et l'unité de mesure le cas échéant.

On pourra à cet effet utiliser les différentes possibilités de l'onglet *Affichage>Nom des Variables*, et s'aider de la boîte de dialogue *Données>Specifications* pour afficher les propriétés des variables.

Dans le logiciel, le *type* d'une variable désigne la manière dont la variable est codée : *Double* (format par défaut des valeurs numériques), *Entier*, *Octet* (entiers compris entre 0 et 255). Modifier le 'type' de chacune des variables du jeu de données, lorsque vous le jugez nécessaire.

3 Analyse des variables qualitatives

L'objectif de cette section est de décrire la population en se basant uniquement sur les variables qualitatives. Pour cela, vous pourrez utiliser les procédures adéquates du menu *Statistiques>Statistiques Elémentaires*, les boîtes de dialogue du menu *Graphiques>Graph. en 2D*.

Les questions suivantes servent uniquement à guider votre analyse. N'hésitez pas à essayer différents choix de graphiques et à les comparer. Pensez également que votre compte-rendu ne doit pas prendre la forme d'une suite de réponses à ces questions (voir le document *Conseils de rédaction pour les compte-rendus de T.P.*).

1. Considérons la variable « Catégorie Professionnelle »
   a) Etablir la distribution en effectifs et en fréquences. On pourra utiliser la rubrique *Tables de fréquences* du menu *Statistiques>Statistiques Elémentaires*. Vous allez

créer ainsi une nouvelle feuille de données. Calculer les effectifs ou fréquences cumulés a-t-il un sens ? Rechercher comment construire la table des fréquences sans les afficher.
Pensez dans la suite, à chaque fois que vous souhaitez lancer une procédure, à vérifier quelle est la feuille de données *active*.

b) Construire un diagramme en barres pour représenter la distribution de la variable. Noter que sous *STATISTICA*, ce type de graphique est considéré (à tort !) comme un histogramme, avec l'option *Rupture entre les colonnes...*

2. On s'intéresse maintenant aux autres variables qualitatives. Pour chacune d'entre elles, établir la distribution en effectifs et en fréquences et proposer un graphique pour représenter la distribution en fréquences.

4 Analyse de variables quantitatives

4.1 Analyse de la variable « Wage »

Nous nous intéressons à la variable « Wage » (salaire). On souhaite définir huit catégories d'individus :

- Catégorie 0 : moins de 0 dollars de l'heure
- Catégorie 4 : entre 0 et 4 dollars de l'heure (0 exclu)
- Catégorie 8 : entre 4 et 8 dollars de l'heure (4 exclu)
- Catégorie 12 : entre 8 et 12 dollars de l'heure (8 exclu)
- ...
- Catégorie 48 : entre 44 et 48 dollars de l'heure (44 exclu)

1. Créer une variable « CatWage » de catégorie de salaire : on pourra d'abord créer la colonne correspondante puis utiliser la boîte de dialogue *Données>Recodifier*, et définir les bornes supérieures et inférieures des catégories ainsi que les étiquettes correspondantes.
2. Maintenant, créer un fichier ne contenant que les individus percevant moins de 12 dollars de l'heure. Pour cela, vous pourrez utiliser la boîte de dialogue *Outils>Filtres de Sélection>Edition*, et plus précisément l'onglet *Sous-Ensemble/Echantillonnage Aléatoire*. Ajouter cette nouvelle feuille de données dans votre classeur.
3. Construire un histogramme de la variable « Wage » : on pourra *spécifier les limites* des classes dans l'onglet *Avancé* de la boîte de dialogue de création d'un histogramme. Attention, dans *STATISTICA*, la seule possibilité est de construire un graphique dont les classes sont de même amplitude (quitte à avoir des classes vides...).
4. Représenter également la distribution de la variable « Wage » à l'aide de la courbe des fréquences cumulées. On fera deux tracés,
   - l'un utilisant les données brutes de la variable « Wage » : pour cela, utiliser le module *Statistiques>Distributions et Simulation*, en choisissant *Ajustement de Distributions* dans la boîte de dialogue. Sélectionner l'option *Fonction de Répartition empirique* après avoir choisi la variable à représenter. Sur le graphique qui apparait, il y a bien la fonction que l'on cherche, mais également trois autres courbes. En double-cliquant sur le graphique, chercher comment ne pas afficher leur tracé.
   - l'autre utilisant les données groupées (variable « CatWage ») : commencer par construire une feuille de données contenant les fréquences cumulées de cette variable, puis utiliser le type de graphique *Séquentiel/Empilé...* du menu *Graphique>Graph. en 2D*. Remarquer que le logiciel ne tient pas compte des classes vides. Que faut-il faire pour obtenir un graphique correct ?

4.2 Analyse de la variable « Education »

Nous nous intéressons à la variable « Education » (Nombre d'années d'études).

1. Etablir la distribution en effectifs et fréquences (cumulés si vous le jugez utile) de cette variable.
2. Construire un diagramme en bâtons pour illustrer la distribution de cette variable. Noter que sous *STATISTICA*, ce type de graphique est considéré (à tort !) comme un histogramme, dont les rectangles sont "séparés" et sur lequel chaque rectangle est remplacé par une ligne...
3. Représenter également la distribution de la variable « Education » à l'aide de la courbe des fréquences cumulées. On pourra utiliser la table des fréquences établie à la question 1., puis le type de graphique *Séquentiel/Empilé...* du menu *Graphique>Graph. en 2D*, avec l'option "*en escalier*" (onglet *Avancé* de la boîte de dialogue du graphique).

4.3 Analyse de la variable « Age »

On s'intéresse enfin à la variable « Age ». Proposer une étude cohérente de la distribution de cette variable.