Statistique descriptive – 1re année [1]

Feuille de T.P. 5

Objectifs. Etude de la liaison entre une variable qualitative et une variable quantitative. Traitement des données manquantes.

Instructions. Ce T.P. et votre compte-rendu sont à finir d'ici **dimanche 3 décembre** minuit. Le compte-rendu devra être déposé sur *moodle*. Il devra être nommé :

`TP5_Prenom_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.

Les deux sections de ce T.P. sont totalement indépendantes. Pour ce T.P., vous devez récupérer sur Moodle les documents suivants :
- DonneesTP1 (jeu de données *Recensement.sta*), si vous ne l'avez plus dans vos documents,
- TP5FichierExcel TP5VarianceInterIntra_Recensement.xls,
- DonneesTP5 (jeu de données*EnfantsAvecDonneesManquantes.sta*).

1 Lien variables quantitative-qualitative

Nous reprenons dans cette section l'étude du jeu de données *Recensement.sta* commencée lors des T.P. 1 et 2. On commencera par créer un nouveau classeur *Statistica*, et par enregistrer à nouveau les données dans ce classeur, comme pour chaque T.P.

Nous nous intéressons dans ce T.P. au lien entre *variable quantitative* et *variable qualitative*.

1. Nous étudions tout d'abord le lien entre le *salaire* et la *catégorie professionnelle*.
   - Décrire précisément les deux variables.
   - Calculer les moyennes et variances conditionnelles et marginales du salaire.
   - Remplir le fichier Excel TP5VarianceInterIntra_Recensement.xls.
   - Commenter le rapport de corrélation entre ces deux variables.
   - Quelle est la signification d'un rapport de corrélation égal à 0 ? Dans quel cas le rapport de corrélation aurait-il pu être égal à 1 ?
   - On va maintenant calculer l'indicateur de Fisher avec *Statistica*. Utiliser le menu *Statistiques > ANOVA*. Choisir *ANOVA à un facteur* dans le menu *Type d'analyse*. Après avoir sélectionner les variables, choisir *R modèle complet* dans l'onglet *Synthèse*.
   - Proposer une représentation graphique de la loi marginale du salaire et des lois conditionnelles du salaire sachant la catégorie professionnelle.

1. Enseignant responsable des TP : G.Chagny (gaelle.chagny@parisdescartes.fr) et C.Laclau (charlotte.laclau@parisdescartes.fr)

2. Etudier à l'aide des mêmes indicateurs d'autres couples de variables quantitative–qualitative.

3. Conclure sur l'étude de ces différents couples de variables.

2 Jeu de données avec données manquantes

Dans cette deuxième partie, nous nous intéressons au problème des données manquantes. Le jeu de données étudié contient des mesures relevées sur des enfants de 4 à 7 ans (genre, poids, taille). Mais certaines observations sont manquantes pour les variables « Taille » et « Poids ». Nous allons voir les conséquences possibles d'une telle situation.

On commencera par créer un nouveau classeur, et enregistrer la feuille de données *EnfantsAvecDonneesManquantes.sta* dans ce nouvel objet.

Dans le fichier *EnfantsAvecDonneesManquantes.sta*, les données manquantes ont été codées par la valeur 0. On cherche à comprendre l'intérêt de coder de préférence les données manquantes par une valeur que l'on pourra déclarer à *STATISTICA* comme caractéristique des données manquantes.

1. Essayer de représenter graphiquement la distribution d'une des deux variables « Taille » ou « Poids ».
2. Créer deux nouvelles variables « TailleBis » et « PoidsBis », contenant les mêmes valeurs que « Taille » et « Poids » respectivement, mais où la valeur 0 a été remplacée par le symbole reconnu par défaut comme caractéristique des données manquantes sous *STATISTICA* : -999999998 pour le codage, case vide dans une feuille de données. On pourra créer deux variables, copier-coller les valeurs des deux anciennes variables, et utiliser la boîte de dialogue *Données>Recodifier*, pour remplacer le 0 par le *Code des VM*, en n'oubliant pas de laisser les autres données inchangées.
3. Représenter à nouveau la distribution de la variable choisie à la première question. Que se passe-t-il alors ?
4. Quelle est la proportion de données manquantes pour chacune des variables ?
5. Quel est le nombre de données manquantes pour chaque individu ? On pourra créer une nouvelle variable, prenant les valeurs 0 (si aucune valeur n'est manquante), 1 (si la valeur du poids ou de la taille est manquante) ou 2 (si les deux valeurs sont manquantes).
6. Quelle est la proportion d'individus présentant des données manquantes ?