T.P. de Statistique Descriptive

1 Description du jeu de données et préliminaires

Depuis 1948, aux Etats-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 150 000 foyers. Cette étude est la principale source de renseignements sur les caractéristiques de la population active du pays. Les données du fichier *Recensement.txt* sont extraites de l'enquête de juillet 2012.

Ouvrir le fichier *Recensement.txt* : `read.table('Recensement.txt',header=T)` L'option `header` permet d'indiquer au logiciel R que le fichier contient des noms pour les colonnes. Il s'agit des noms des différentes variables : chaque ligne correspond à un individu, chaque colonne à une variable.

Les variables du jeu de données sont :

- ◇ AGE : Âge en années.
- ◇ SEXE :
  - — F : Féminin,
  - — M : Masculin.
- ◇ REGION :
  - — NE : Nord-Est.
  - — MW : Mid-Ouest.
  - — S : Sud.
  - — W : Ouest.
- ◇ STAT_MARI : statut marital.
  - — C : Célibataire.
  - — M : Marié(e).
  - — V : Veuf(ve).
  - — D : Divorcé(e).
  - — S : Séparé(e).
- ◇ SAL_HOR : Salaire horaire en dollars.
- ◇ SYNDICAT : Appartenance à un syndicat.
- ◇ CATEGORIE : Catégorie professionnelle.
  - — 1 : Activités de gestion, commerciales et financières.
  - — 2 : Profession libérale.
  - — 3 : Activités de services.
  - — 4 : Vente.
  - — 5 : Employés de bureau, administration.
  - — 6 : Agriculture, pêche, forêts.
  - — 7 : Activités de construction et d'extraction.
  - — 8 : Activités d'installation, de maintenance et de réparation.
  - — 9 : Activités de production.
  - — 10 : Activités de transport de marchandises et de matériaux.
- ◇ NIV_ETUDES : niveau d'études.
  - — 32 : Au plus 4 années de primaire.
  - — 33 : Entre 5 et 6 années.
  - — 34 : Entre 7 et 8 années.
  - — 35 : 9 années.
  - — 36 : 10 années.
  - — 37 : 11 années.
  - — 38 : 12 années, sortie du lycée sans diplôme.
  - — 39 : 12 années, diplômé à la sortie du lycée.
  - — 40 : Etudes à l'université sans diplôme.
  - — 41 : Associate degree, parcours professionnel (équivalent à DUT, BTS).
  - — 42 : Associate degree, parcours académique (équivalent à L1–L2).
  - — 43 : Bachelor (équivalent à licence ou maîtrise).

  - 44 : Master.
  - 45 : Diplôme d'école spécifique.
  - 46 : Doctorat.
- ◇ NB_PERS : Nombre de personnes au sein du foyer.
- ◇ NB_ENF : Nombre d'enfants au sein du foyer.
- ◇ REV_FOYER : Classes de revenu annuel du foyer en dollars.
  - 1 : < 5000.
  - 2 : $[5000; 7500[$.
  - 3 : $[7500; 10000[$.
  - 4 : $[10000; 12500[$.
  - 5 : $[12500; 14500[$.
  - 6 : $[15000; 20000[$.
  - 7 : $[20000; 25000[$.
  - 8 : $[25000; 30000[$.
  - 9 : $[30000; 35000[$.
  - 10 : $[35000; 40000[$.
  - 11 : $[40000; 50000[$.
  - 12 : $[50000; 60000[$.
  - 13 : $[60000; 75000[$.
  - 14 : $[75000; 100000[$.
  - 15 : $[100000; 150000[$.
  - 16 : ≥ 150000.

Déterminer la taille de l'échantillon et donner l'unité statistique. Quelles sont les variables qualitatives du jeu de données ? Et les variables quantitatives ?

2 Analyse des variables qualitatives

2.1 Analyse de l'ensemble de l'échantillon

L'objectif de cette section est de décrire l'échantillon en se basant uniquement sur les variables qualitatives.

Pour chaque variable qualitative du fichier :
- Etablir la distribution en effectifs et en fréquences à l'aide des fonctions `table` et `prop.table`. La fonction `round` avec l'option `digit=k` permet d'arrondir les valeurs des fréquences à k décimales.
- Proposer un graphique pour représenter la distribution en fréquences de la variable. On pourra utiliser les fonctions `barplot` et `pie`.

Lorsque cela semble pertinent, plusieurs modalités pourront être regroupées. Les fonctions `colSums` et `rowSums` permettant de sommer les élements d'une matrice par colonne ou par ligne peuvent être utiles dans ce cadre, ainsi qu'une instruction du type `v[v==i]=k` (modalité i de v remplacée par k) associée à l'opérateur logique *ou* `|`.

2.2 Analyse selon la catégorie professionnelle

Le but de cette section est de comparer des groupes d'individus de catégories professionnelles différentes. On pourra par exemple comparer les cadres (1) aux individus exerçant une profession libérale (2) et aux employés de bureau (5).

Noter qu'une instruction de la forme `tab1=tab[tab[,j]==i,]` permet de sélectionner les individus de `tab` ayant la modalité i pour la j^e variable. Les opérateurs logiques peuvent être utiles pour sélectionner des sous-ensembles d'individus.

Reprendre les questions du paragraphe précédent pour les 3 modalités de la catégorie professionnelle. Représenter la distribution de chaque variable pour les 3 catégories sur un même graphique si le type de graphique s'y prête, et sur 3 graphiques juxtaposés sinon.

3 Analyse de variables quantitatives

3.1 Analyse de l'ensemble des salariés

Analyser les variables quantitatives du jeu de données en proposant et commentant plusieurs représentations graphiques de la distribution de ces variables. Un diagramme en bâtons se trace à l'aide de la fonction `plot(table())` . Une fonction de répartition empirique est obtenue avec `plot(ecdf())`. La fonction `cumsum` peut être employée pour construire un tableau de fréquences cumulées, à partir duquel on trace ensuite la courbe des fréquences cumulées. L'option `breaks` de la fonction `hist`, qui sert à construire un histogramme, permet de préciser le nombre de bornes ou les bornes elles-mêmes. Mettre en évidence l'influence du choix des classes pour la variable SAL_HOR.

3.2 Analyse selon la catégorie professionnelle

Nous nous intéressons ici aux 3 groupes étudiés dans la section précédente. Etudier les variables quantitatives pour chaque catégorie professionnelle et commenter.

4 Indicateurs statistiques du jeu de données Recensement

Pour étudier les différences de tendance pour les variables considérées en fonction de la catégorie professionnelle des individus, nous utiliserons divers indicateurs statistiques.

4.1 Analyse de l'ensemble des salariés

L'objectif de cette section est de décrire l'échantillon dans son ensemble.

— Pour chaque variable du fichier, qualitative ou quantitative, calculer les indicateurs statistiques pertinents. On pourra utiliser les fonctions `mean`, `median`, `sd` et `var`, `max`, `min`, `quantile` ou encore la fonction `summary`.

— Trouver le mode des variables quantitatives discrètes à l'aide des fonctions `sort` (option decreasing=TRUE) et `table`.

— Quelle est la formule utilisée par R pour le calcul de la variance ? Calculer la variance comme définie dans le cours à l'aide des fonctions `var` et `length`.

Lorsque cela a un sens, tracer les boîtes à moustaches à l'aide de la fonction `boxplot`.

4.2 Analyse par catégorie professionnelle

Dans cette section, on cherche à comparer les cadres (1) aux professions libérales (2) et aux employés de bureau (5).

Calculer les indicateurs statistiques pour ces 3 modalités de la catégorie professionnelle.

Tracer sur un même graphique les boîtes à moustaches pour les différentes catégories professionnelles.