ANALYSE DE DONNÉES
TRAVAUX DIRIGÉS ET PRATIQUES

Fiche n^{o}1 : Rappels de Statistiques Descriptives

Exercice 1

1) Pour chacune des variables suivantes, préciser son type.

- Revenu annuel.
- Sexe.
- État matrimonial.
- Lieu de résidence.
- Pointure des chaussures.
- Citoyenneté.
- Couleur des yeux.
- Nombre de langues parlées.
- Âge.
- Tour de taille.

2) Quel est le principal défaut de la moyenne, en tant que caractéristique de la tendance centrale ?

3) Quel est le principal défaut de la variance, en tant que caractéristique de dispersion ?

4) Dans une distribution symétrique, la moyenne, la médiane et le mode sont-ils confondus ?

5) Quelle est la différence entre valeurs manquantes, aberrantes et extrêmes ?

Exercice 2 Nettoyage et prétraitement des données

Cet exercice sera effectué sous le logiciel R.

1) Télécharger les données *Recensement_12.csv* sur Moodle ou sur le répertoire COMMUN. Ces données sont un échantillon de 599 foyers du recensement effectué en 2012 aux Etats Unis, décrits par 11 variables.

2) Ouvrir RStudio et se placer dans le répertoire où les données ont été enregistrées grâce à la commande `setwd("`*chemin du répertoire*`")`.

3) Charger les données dans R en cliquant sur `Import Dataset`, ou en tapant dans la console les commandes

```
> Recensement_12=read.csv2("Recensement_12.csv")
> View(Recensement_12)
```

4) Décrire les variables suivant leurs types. Eliminer toute variable n'apportant aucune information statistique.

5) Utiliser la fonction `summary` pour avoir une première description rapide des données. Existe-t'il des valeurs manquantes ou aberrantes ?

6) Calculer le pourcentage de valeurs manquantes dans la table grâce aux fonctions `is.na` et `sum`.

7) Calculer le pourcentage de valeurs manquantes par variable grâce à la fonction `colSums`. Éliminer toutes les variables ayant plus de 70% de valeurs manquantes (utiliser la fonction `which` afin de récupérer les indices de colonnes contenant plus de 70% de valeurs manquantes).

8) Effectuer la même opération sur les individus (fonction `rowSums`) en éliminant les lignes ayant plus de 60% de valeurs manquantes.

9) Séparer la table de données obtenue en 2 tables
— une table `quali` contenant toutes les variables qualitatives,
— une table `quanti` contenant toutes les variables quantitatives

10) Imputer les valeurs manquantes dans la table `quanti` et enregistrer la table complétée sous un autre nom. Vérifier que cette nouvelle table ne contient plus de valeurs manquantes.

Exercice 3 Statistiques descriptives univariées

Cet exercice sera effectué sous le logiciel R.

1) Variables qualitatives

a- Pour chaque variable qualitative, indiquer le(s) type(s) de graphique à utiliser pour représenter leurs distributions statistiques.

b- Suivant le type de graphique choisi, utiliser les fonctions suivantes :

Diagramme circulaire	Diagramme en barres
`pie`	`barplot`

Ces fonctions prennent en entrée les fréquences de chaque modalité. Utiliser la commande `prop.table(table(x))` pour calculer ces fréquences, où `x` est la variable considérée.

c- Interpréter les graphiques obtenus.

2) Variables quantitatives

a- Interpréter les différents indicateurs centraux obtenus en sortie de la commande
`> summary(dat)`,
où `dat` est la table construite après imputation des valeurs manquantes.

b- Taper les commandes
`> var(dat[,1])`
`> sum((dat[,1]-mean(dat[,1]))^2)/nrow(dat)`
Que remarquez-vous ? Utiliser l'aide de la fonction `var` afin de trouver une explication.

c- Pour chaque variable quantitative, indiquer le(s) type(s) de graphique à utiliser pour représenter leurs distributions statistiques.

d- Suivant le type de graphique choisi, utiliser les fonctions suivantes :

Diagramme en barres	Histogramme	Boîte à moustaches
`barplot`	`hist`	`boxplot`

e- Interpréter les graphiques obtenus.

f- Utiliser la fonction `density` pour estimer la densité de probabilité des variables continues. Comparer avec les représentations graphiques des distributions statistiques.

Exercice 4 Statistiques descriptives bivariées

Cet exercice sera effectué sous le logiciel R.

1) Indiquer les différents types d'outils statistiques utilisés pour interpréter les liens éventuels entre 2 variables.

2) Parmi les variables du jeu de données *Recensement_12*, sélectionner plusieurs couples de variables et utiliser les outils statistiques adéquats afin d'interpréter leurs relations. Les commandes suivantes pourraient être utiles :

Table de contingence	Corrélation	Boîte à moustaches	Nuages de points
`table(x,y)`	`cov(dat), cor(dat)`	`boxplot(x ~ y)`	`plot(x,y), pairs(dat)`

où `x` et `y` sont les deux variables croisées, `dat` est la table de données considérée.