UGA – L2/STA401

TP2 : Statistiques descriptives élémentaires

Objectifs : *Savoir faire une première description numérique et graphique de chaque variable d'un jeu de données présenté dans un tableau de données brutes stockées dans un* `data.frame`.

Préparer un rendu de TP en R markdown

R markdown permet de préparer un rapport dynamique, c'est à dire qui se génère automatiquement, comportant: du code R, les résultats du code R et des commentaires.

- Ouvrez un nouveau fichier R markdown en allant dans **File** > **New file** > **Rmarkdown**. Mettez comme titre "TP2" et votre nom comme auteur. Laissez le format de sortie par défaut en HTML. Enregistrez ce fichier sous le nom TP2. L'extention .Rmd est automatiquement ajoutée.
- Lisez le document généré. Après une entête et la mise en place des options par défaut des chunks, vous observez la présence de :
  - chunks contenant du code R
  - parties de texte en dehors des chunks au format markdown
- Lancez la génération du document html: cliquez sur la pelotte de laine **Knit**. Un fichier html TP2.html est alors créé.
- Remplacez le texte et le code donnés en exemples par les réponses à ce TP. Notez que vous pouvez exécuter les chunks au fur et à mesure à l'aide de la flèche verte à droite du chunk.

Les données analysées ici sont celles proposées dans le package de base de R appelées `mtcars` qui présentent pour 32 véhicules des mesures décrivant les particularités de leur moteur et de leur désign. On peut afficher la description de ce jeu de données avec la commande `help("mtcars")`. Il s'agit d'un échantillon de 32 véhicules pour lesquels ont été observées les 11 variables suivantes :

`mpg`: consommation (Miles/(US) gallon)
`cyl`: nombre cylindres
`disp`: volume air déplacé (cu.in.)
`hp`: puissance
`drat`: couple (Rear axle ratio)
`wt`: poids (en tonnes)
`qsec`: 1/4 mile time
`vs`: dispo.cyl. (0 pour V-engine et 1 pour straight engine)
`am`: transmission (automatique (0) ou manuelle (1))
`gear`: nombre vitesses
`carb`: nombre carburateurs

1 Un premier aperçu des données

Exercice 1 : résumés numériques

1. Afficher le data.frame en exécutant :
   `mtcars`
2. Pour obtenir les noms des variables on peut utiliser la fonction générique `names()`:
   `names(mtcars)`
3. Pour obtenir les premières lignes du tableau on utilise la fonction `head()`:
   `head(mtcars)`.
   Pour extraire l'échantillon de la variable contenue dans une colonne du data.frame on utilise : `madataframe$nomdecolonne`. Par exemple `mtcars$cyl` retourne l'échantillon de la variable nombre de cylindres nommée `cyl` dans le data.frame `mtcars`.

   Quelles sont les variables qualitatives, quantitatives discrètes et quantitatives continues (Dans le doute entre quantitatif discret ou quantitatif continu on regardera le tableau en effectifs des échantillons avec `table()`) ? Préciser les modalités ou ensemble des valeurs prises par `cyl`, `am`, `mpg`, `disp`.
4. Que fait la fonction `summary()` ?
   Que vaut le consommation moyenne de l'échantillon ? Quelle est la valeur maximale observée dans `disp` ? Quelle est la part de véhicules automatiques dans l'échantillon observé ?

Exercice 2 : boites de distribution

Un premier graphique permettant de décrire très rapidement un échantillon d'une variable quantitative est la boite de distribution appelée boxplot en anglais. La fonction de R qui trace un boxplot s'appelle `boxplot()`.

1. Extraire la variable `mpg` du data.frame et l'affecter à `mpg` avec :
   `mpg <- mtcars$mpg`
   Construire de manière analogue les variables `disp`,`cyl` et `am`.
2. Faire le boxplot de la variable `mpg` en ajoutant un titre et une légende sur chaque axe (on pourra consulter l'aide en ligne pour savoir comment utiliser les options de `boxplot()`).
3. Construire l'échantillon des consommations des véhicules automatiques que l'on nommera `mpga` et celui des véhicules manuels que l'on nommera `mpgm`. Pour créer `mpga` on utilise la syntaxe :
   `mpga <- mtcars[am==0,"mpg"]`
4. Représenter simultanément et côte à côte, les boxplots des deux sous-échantillons précédemment extraits :
   `boxplot(mpga,mpgm,names=c("automatique","manuelle"),main="mpg selon la transmission")`
   Lesquelles sont-elles les plus économiques ?

2 variable qualitative

Pour décrire une telle variable on indique les modalités rencontrées et on décrit la distribution de la variable (appelée aussi facteur) avec les fréquences d'apparition (ou les effectifs) de chacune.

Exercice 3 : graphique de la distribution observée

1. Faire le tableau en effectifs de la variable `am` en utilisant la fonction `table`. En déduire les fréquences observées de chaque modalité. (soit à l'aide de la fonction `prop.table()` soit directement).
2. Représenter la distribution observée de `am` avec un diagramme en secteur (en utilisant `pie()`) et un diagramme en barre (avec `barplot()`). Commenter.

3. Représenter dans une même fenêtre graphique séparée en deux parties situées sur une même ligne, à gauche le diagramme en secteur et à droite le diagramme en barres. Pour partitionner la fenêtre graphique et la remplir en ligne on peut exécuter :
   `par(mfrow=c(1,2))`
   Comparer. Dans quelle situation préfèrera-t-on le diagramme en barres ?

4. On souhaite comparer la transmission (`am`) sur véhicule cylindres en avec la transmission sur véhicules cylindres alignés. Pour cela on construit le tableau des effectifs observés sur les quatre couples de modalités possibles pour le couple `(am,vs)` avec :
   `table(am,vs)`
   et les fréquences avec
   `prop.table(table(am,vs))`

5. Que donnent les commandes suivantes :
   ```
   prop.table(table(am,vs),1)
   barplot(prop.table(table(am,vs),1),beside=T)
   prop.table(table(am,vs),2)
   barplot(prop.table(table(am,vs),2),beside=T)
   ```
   Y a t-il un lien entre `vs` et `am` (justifier) ?

3 Variable quantitative

Dans cette partie sont étudiées les quantitatives discrètes (à valeurs entières et/ou n'ayant qu'un nombre faible de modalités rencontrées) et les quantitatives continues (à valeurs dans un intervalles et pour lesquelles on observe de très rares répétitions de valeurs dans l'échantillon). Dans les deux cas il s'agit de variable numériques pour lesquelles on dispose d'indicateurs de centrage et de dispersion de la répartition observée.

Exercice 4 : les principaux indicateurs numériques

1. Affecter à `cyl` la colonne `cyl` du data.frame `mtcars`.

2. Calculer la somme, la somme des carrés, la moyenne empirique et la variance empirique de l'échantillon `cyl` des nombres de cylindres.

3. Essayer les fonctions `mean()`, `var()`, `sd()` et `quantile()`. Commenter. En particulier préciser ce que R calcule avec la fonction `var()`.

4. Calculer moyenne et écart-type corrigé pour la variable continue `mpg`.

Pour les variables quantitatives discrètes ou continues on peut tracer les boites de distributions (avec `boxplot()`) en premier aperçu de la répartition des données. Ensuite pour les variables discrètes on utilise un diagramme en barres et pour les continues un histogramme, pour représenter la distribution observée (Exercice 5). On peut également tracer la fonction de répartition empirique d'une variable continue ou discrète qui représente les fréquences cumulées (Exercice 6).

Exercice 5 : représentations graphique des distributions

1. Représenter la répartition observée de `cyl` avec un diagramme en secteurs et avec un diagramme en barres (utiliser les fonctions `table()`, `pie()` et `barplot()` déjà rencontrées pour la représentation d'une variable qualitative). Que vaut la somme des hauteurs représentées sur le diagramme en barres ?

2. Représenter l'histogramme de l'échantillon de `disp` en veillant à ce que pour l'unité choisie sur l'axe des ordonnées la surface représentée vaille 1. On utilisera pour cela la fonction `hist()` et une option adaptée pour obtenir une surface sous la courbe valant 1 (consulter l'aide de R pour la fonction `hist()`).

3. Superposer à l'histogramme la densité de probabilité d'une loi normale convenablement choisie (en choisissant l'espérance et l'écart-type de façon adaptée aux données). Pour cela, on utilisera `curve()` avec la fonction `dnorm()` qui retourne la densité d'une loi de Gauss.

Exercice 6 : fonction de répartition empirique

1. Calculer les fréquences cumulées pour `cyl` (utiliser la fonction `cumsum`) et les affecter à un vecteur nommé `cumfreqcyl`.

2. Tracer la fonction de répartition empirique de `cyl` en utilisant l'option `type="s"` dans la fonction `plot()`.

3. Essayer à présent la fonction `ecdf()` sur `cyl` puis tracer le graphe avec `plot(ecdf(cyl))`