TP2 : Analyses univariées

Dans ce TP on va procéder à des analyses descriptives univariées, c'est-à-dire pour une seule variable à la fois, à partir du jeu de données *employes.txt*.

Ce jeu de données concerne 473 employés d'une entreprise américaine. Il contient les variables suivantes :

- *sexe* : sexe de l'employé (1 pour *féminin*, 2 pour *masculin*)
- *educ* : nombre d'années d'études
- *stat_pro* : statut professionnel (1 si *employé de bureau*, 2 si *agent de sécurité*, 3 si *manager*)
- *salembau* : salaire annuel à l'embauche dans l'entreprise (en dollars)
- *salaire* : salaire annuel courant (en dollars)
- *ancienne* : ancienneté dans l'entreprise (en mois)
- *exppasse* : expérience passée dans le type de poste (en mois)
- *national* : nationalité (0 pour *américaine*, 1 sinon)
- *age* : âge (en années)

Rappeler la population, l'échantillon et le type de chaque variable.

1 Préparation des données

Dans **RStudio**, créer un nouveau script *TP2.R*, dans lequel seront copiées les différentes commandes utilisées lors de ce TP.

Vérifier que le fichier *employes.RData* créé lors du TP1 se situe bien dans le répertoire *K :/Documents/TpStat* et le charger grâce à la commande `load("employes.RData")`.

Sinon, importer les données grâce à la commande `employes=read.table("employes.txt",header=TRUE)`.

On peut avoir des informations sur le jeu de données directement dans la fenêtre *Environment*, notamment en cliquant sur le bouton bleu ce qui permet d'obtenir des informations sur les variables du `data.frame`. Donner la taille de l'échantillon et le nombre de variables.

Pour afficher le fichier de données dans un onglet à côté du script, on peut cliquer sur le nom du `data.frame` dans la fenêtre *Environment*. Penser à fermer cet onglet avant de continuer.

Dans **R**, les variables qualitatives doivent être de type *factor* et les variables quantitatives de type *numeric* (numérique) ou *integer* (entier).

Dans le `data.frame` *employes*, quel est le type de chaque variable ? Est-ce que toutes les variables ont été importées au bon format ?

Pour transformer les variables qualitatives en *factor* et modifier leurs modalités (afin qu'elles soient

plus compréhensibles, faire attention à lister les labels dans l'ordre des valeurs), on utilise les commandes suivantes :

```
employes$sexe=factor(employes$sexe,labels=c("F","M"))
employes$stat_pro=factor(employes$stat_pro,labels=c("employé","agent de sécurité",
 "manager"))
employes$national=factor(employes$national,labels=c("américaine","autre"))
```

Afficher les dix premières et les dix dernières lignes du jeu de données grâce aux commandes : `head(employes,10)` et `tail(employes,10)`. Par défaut, ces commandes affichent six lignes.

Pour obtenir un résumé de l'ensemble du jeu de données, on utilise la commande `summary(employes)`. On obtient le tableau de distribution en effectifs pour les variables qualitatives (de type *factor*) et quelques résumés numériques pour les variables quantitatives (ici de type *integer*). Cette commande donne le nombre de données manquantes pour chaque variable en présentant.

Tester les commandes suivantes :

```
employes[1,]
employes[,1]
employes$sexe
employes$sexe[1:5]
employes[1:5,]
employes[employes$age>=65,]
which(employes$age>=65)
length(which(employes$age>=65))
employes$age[employes$age>=65]
```

On constate que pour accéder à la variable *sexe* on peut utiliser soit sa position avec `employes[,1]` soit son nom avec `employes$sexe`. Dans les deux cas, il faut préciser le nom du `data.frame` contenant la variable. Pour pouvoir accéder aux variables du `data.frame` *employes* directement par leur nom, on utilise la commande `attach(employes)`. Ainsi pour accéder à la variable *sexe* du `data.frame` *employes*, on pourra taper directement `sexe` plutôt que `employes$sexe`. Attention, il faut réexécuter la commande `attach` dès qu'on fait des modifications dans le fichier de données.

2 Description univariée d'une série de données

2.1 Variables qualitatives

Avec **R**, le tableau de distribution en effectifs s'obtient avec la fonction `table` et le tableau de distribution en fréquences avec la fonction `prop.table` appliquée à la sortie de `table`. Les commandes suivantes donnent ces deux tableaux pour la variable *sexe* :

```
table(sexe)
prop.table(table(sexe))
```

Pour arrondir les valeurs des fréquences à 3 décimales, on ajoute la fonction `round()` avec l'argument `digits` qui précise le nombre de décimales :

```
round(prop.table(table(sexe)),digits=3)
```

Le *diagramme en colonnes* s'obtient avec la fonction `barplot` appliquée au tableau de distribution en effectifs ou en fréquences. Le *diagramme circulaire* s'obtient par la fonction `pie`, appliquée au tableau de distribution en effectifs.

Représenter le diagramme en colonnes en fréquences ainsi que le diagramme circulaire de la variable *stat_pro* grâce aux commandes suivantes :

```
barplot(prop.table(table(stat_pro)),main="Distribution du statut professionnel",
  ylab="Fréquence", xlab="Statut professionnel", ylim=c(0,1), col="blue")
pie(prop.table(table(stat_pro)),main="Distribution du statut professionnel")
```

Noter les arguments utilisés qui sont communs à la plupart des graphiques :
- `main` pour préciser le titre du graphique,
- `ylab` pour préciser le nom de l'axe des ordonnées,
- `xlab` pour préciser le nom de l'axe des abscisses,
- `ylim` pour choisir les bornes de l'axe des ordonnées (`xlim` pour l'axe des abscisses)
- `col` pour changer la couleur du graphique
- `\n` sert à revenir à la ligne lorsqu'un titre est trop long.

Remarque : la liste des couleurs reconnues par R est disponible sur `http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf`.

2.2 Variables quantitatives

A) Résumés numériques

On peut utiliser la commande `summary` pour une variable quantitative afin d'obtenir la plupart des résumés numériques (minimum, premier quartile Q1, médiane, moyenne, troisième quartile Q3, maximum) : `summary(age)`. On peut également les obtenir séparément avec des fonctions spécifiques (`min`, `max`, `mean`, `median`).

Les autres indicateurs s'obtiennent grâce aux fonctions suivantes :

indicateur	fonction **R**
variance	`var`
écart-type	`sd`
quantiles	`quantile` avec l'argument `probs`

Donner les principaux résumés numériques, la variance et l'écart-type de l'âge. Calculer le coefficient de variation de l'âge avec la commande `sd(age)/mean(age)` et qualifier la dispersion de cette variable.

Pour obtenir des quantiles précis, on utilise l'argument `probs` de la fonction `quantile` :

```
quantile(age,probs=seq(0.1,1,by=0.1))   # tous les déciles
quantile(age,probs=seq(0.01,1,by=0.01)) # tous les centiles
```

Exécuter ces commandes et observer les queues de distribution de l'âge.

B) Diagramme en bâtons pour une variable quantitative discrète

La représentation graphique spécifique d'une variable quantitative discrète est le *diagramme en bâtons*. Pour tracer le diagramme en bâtons en fréquences de la variable *educ*, on utilise la commande suivante :

```
plot(prop.table(table(educ)),type="h",col="red",ylab="Fréquence",
     xlab="Nombre d'années d'études après le 1st grade",
     main="Distribution du nombre d'années d'études")
```

C) Histogramme pour une variable quantitative continue

La représentation graphique spécifique d'une variable continue est *l'histogramme* pour lequel on représente en ordonnées la densité de proportion de chaque classe. Rappelons que la densité de proportion est égale à la fréquence relative divisée par l'amplitude.

Remarque : lorsque les classes sont d'amplitude égale (ce qui est le cas par défaut dans **R**), les densités de proportion et les fréquences relatives sont proportionnelles.

Pour tracer l'histogramme de la variable *age*, on utilise la commande suivante :

```
hist(age,freq=FALSE,xlab="Age",ylab="Densité de proportion",
     main="Histogramme de la variable age")
```

On utilise l'argument `freq=FALSE` pour représenter la densité de proportion en ordonnées.

On peut choisir les limites des classes grâce à l'argument `breaks` :

```
hist(age,freq=FALSE,breaks=c(30,33,36,38,40,45,50,55,60,65,75),xlab="Age",ylab="Densité
     de proportion",main="Histogramme de la variable age, classes personnalisées")
```

D) Boîte à moustaches

La *boîte à moustaches* (*boxplot* en anglais) permet de résumer la distribution d'une variable quantitative, discrète ou continue, et de repérer d'éventuelles valeurs extrêmes (ou atypiques ou outliers).

Pour tracer la boîte à moustaches de la variable *age*, on utilise la commande suivante :

```
boxplot(age, main="Boîte à moustaches de l'âge",ylab="Age (en années)")
```

Pour tracer la boîte à moustaches à l'horizontale, on utilise l'argument `horizontal=TRUE` :

```
boxplot(age, horizontal=TRUE, main="Boîte à moustaches de l'âge",
 xlab="Age (en années)")
```

2.3 Application

Réaliser l'étude univariée de chaque variable du fichier de données.