TP1 : Analyse univariée

Durée : 6h

L'objectif de ce TP est d'exposer les outils élémentaires permettant de présenter de façon synthétique et claire une série de données brute, et d'en résumer les principales caractéristiques. Ce TP est illustré grâce au langage de programmation et environnement statistique  .

Quelques mots sur R

***R** est un langage et un environnement logiciel open source pour les calculs statistiques et graphiques. R devient incontournable dans le traitement exploratoire et statistique des données.*

Site Internet : Statistiques avec R
http://zoonek2.free.fr/UNIX/48_R_2004/all.html

Il est recommandé de travailler dans l'environnement RStudio.

Les commandes nécessaires aux traitements des données seront précisées au fur-et-à-mesure des exercices. Voici quelques commandes de base :

- `# pour écrire un commentaire`
- `help(`*`fonction`*`)  # pour obtenir l'aide sur une fonction`
- `ls()  # permet d'afficher tous les objets de la session de travail`
- `rm()  # efface les objets de la session de travail`
- `q()  # permet de quitter R`

La session de travail avec tous les objets qui auront été définis dedans (matrices, dataframes, fonctions, …) peut être sauvegardée puis rechargée à chaque utilisation :

- `getwd()  # affiche le répertoire courant`
- `setwd("K:/ING1/GI/Statistiques Descriptives")  # permet de fixer le chemin d'accès au répertoire. attention d'utiliser / et non \`
- `save.image(file="nom.Rdata") # sauvegarde la session de travail dans le fichier nom.RData du répertoire courant`
- `load("nom.Rdata") # recharge la session de travail`

Exercice 1 — *Questions générales*

1) Pour chaque variable, déterminer sa nature et préciser son échelle de mesure. Il est possible de proposer plusieurs échelles de mesure afin de changer la nature de la variable : Age, Département, Note à un examen, Date de naissance.

- Les réponses ne sont pas uniques et dépendent de l'échelle de mesure. Il faut faire discuter les étudiants suivant différentes échelles. Ex. Age de 0 à 99 est un variable continue, Age = 1 2 ou 3 ans est une variable discrète, Age = sénior, adulte, junior est une variable ordinale.
- Le département même sous son codage numérique reste une variable nominale
- La date de naissance n'est pas exploitable. Il faut la coder suivant l'usage souhaité (un age, un mois, …)

2) Normalement, on ne fait des calculs de moyennes que sur les caractères quantitatifs. Si on code un caractère dichotomique avec les codes 0 pour l'absence de la caractéristique et 1 pour la présence de la caractéristique. Expliquer pourquoi dans ce cas précis, la moyenne des codes sur une série donnée a vraiment un sens.

Réponse : La moyenne des codes est égale à la proportion d'individus présentant la caractéristique.

3) Comment peut-on transformer un caractère quantitatif en caractère ordinal ? On différenciera les caractères discrets des caractères continus. Quelle peut être la conséquence néfaste de cette transformation.

Réponse : Pour un caractère quantitatif discret, il n'y a rien à changer. Chaque valeur de la série devient une modalité du caractère nominal. L'ensemble R étant ordonné, le caractère est donc ordinal. Pour un caractère quantitatif continu, il faut le transformer en définissant un nombre fini d'intervalles contigus et disjoints. Inconvénient : on perd de l'information.

4) Pourquoi est-il a priori incorrect de transformer un caractère ordinal en un caractère quantitatif discret ?

Réponse : avec un caractère quantitatif, certains indicateurs sont obtenus en faisant des différences entre les valeurs. Si on transforme les modalités d'un caractère ordinal par des nombres on fait l'hypothèse (forte) que les écarts entre les modalités successives sont égaux aux écarts des nombres correspondants.

5) Démontrer qu'une série centrée-réduite est de moyenne nulle et de variance égale à 1.

Réponse :

$$\bar{\tilde{x}} = \frac{1}{n}\sum_{k=1}^{n}\tilde{x}_k = \frac{1}{n}\sum_{k=1}^{n}\frac{x_k-\bar{x}}{s_x} = \frac{1}{s_x}\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x}) = \frac{1}{s_x}\left(\frac{1}{n}\sum_{k=1}^{n}x_k-\bar{x}\right) = \frac{1}{s_x}(\bar{x}-\bar{x})$$

$$s_{\tilde{x}}^2 = \frac{1}{n}\sum_{k=1}^{n}(\tilde{x}_k-\bar{\tilde{x}})^2 = \frac{1}{n}\sum_{k=1}^{n}\tilde{x}^2 = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{x_k-\bar{x}}{s_x}\right)^2 = \frac{1}{s_x^2}\left(\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2\right) = \frac{1}{s_x^2}s_x^2 = 1$$

6) Soit une série d'observations $x_1,\dots,x_n$ d'une variable aléatoire quantitative. Notons $\bar{x}$ sa moyenne et s^2 sa variance.
   a) Montrer que la variance minimise l'écart quadratique moyen,

$$f(a)=\frac{1}{n}\sum_{i=1}^{n}(x_i-a)^2$$

Réponse : La fonction est dérivable avec $f'(a)=\frac{1}{n}\sum_{i=1}^{n}-2(x_i-a)$. D'où $f'(a)>0 \Leftrightarrow a>\frac{1}{n}\sum_{i=1}^{n}x_i$.
Donc $a=\bar{x}$ est le minimum de la fonction.

   b) A l'aide d'un exemple, regardez ce qu'il se passe pour l'écart absolu moyen,

$$f(a)=\frac{1}{n}\sum_{i=1}^{n}|x_i-a|$$

Réponse : La fonction n'est pas dérivable. On ne peut pas trouver son minimum comme précédemment. Afin de ne pas compliquer les choses on va juste regarder ce qu'il se passe sur un exemple.

Soit S la série statistique dont le tableau des effectifs est le suivant.

Modalité	1	2	3	4
Effectif	4	3	2	1

Soit f la fonction donnant la somme des distances d'un nombre à chacun des termes de S.
f est définie pour tout $x \in [1; 4]$ par

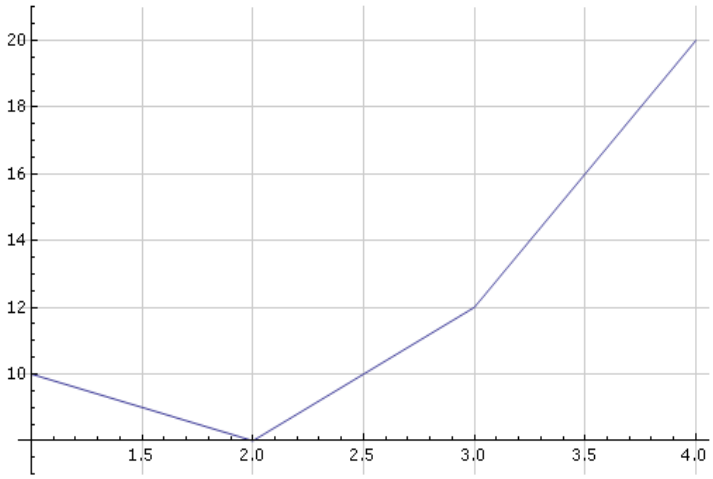
$$f(x) = |x-4| + 2|x-3| + 3|x-2| + 4|x-1|$$

et, pour tout $x \in [1; 4]$,

$$\begin{cases} f(x) = -2(x-6) & \text{si } x \in [1; 2] \\ f(x) = 4x & \text{si } x \in [2; 3] \\ f(x) = 8x-12 & \text{si } x \in [3; 4] \end{cases}$$

Le minimum de f est atteint en 2 qui représente la médiane de S.

Une représentation graphique de f sur [1; 4] est donnée ci-contre.



   c) Notons n_a le nombre d'observations x_i telles que $|x_i-\bar{x}|>a$, c'est-à-dire le nombre d'observations dont l'écart à la moyenne est supérieur à a. Montrer l'inégalité suivante,

$$\frac{n_a}{n} \le \frac{s^2}{a^2}.$$

Qu'est-ce que cela signifie si on prend $a=2s$ ou $a=3s$?

Indication : Dans l'expression de la variance, on pourra considérer séparément les termes tels que $|x_i - \bar{x}| \leq a$ et $|x_i - \bar{x}| > a$.

Réponse : $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i:|x_i-\bar{x}|\leq a}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i:|x_i-\bar{x}|>a}(x_i - \bar{x})^2$

$\geq \frac{1}{n}\sum_{i:|x_i-\bar{x}|>a}(x_i - \bar{x})^2$ car la première somme est positive

Par ailleurs, pour chaque terme de la somme restante, on a

$$|x_i - \bar{x}| > a \Rightarrow (x_i - \bar{x})^2 > a^2$$

Donc en remplaçant dans la somme,

$$s^2 \geq \frac{1}{n}\sum_{i:|x_i-\bar{x}|>a} a^2 = \frac{a^2}{n}\sum_{i:|x_i-\bar{x}|>a} 1 = \frac{a^2}{n}\text{card}\{i : |x_i - \bar{x}| > a\} = \frac{a^2}{n} n_a$$

$$\Leftrightarrow \frac{n_a}{n} \leq \frac{s^2}{a^2}$$

Si on pose a=2s alors on obtient $\frac{n_{2s}}{n} \leq \frac{1}{4}$. Cela signifie que la proportion d'observations x_i qui s'éloignent de plus de 2s de la moyenne est inférieure à ¼, autrement dit, plus de 75% des observations sont dans l'intervalle $[\bar{x} - 2s ; \bar{x} + 2s]$. De la même façon, on peut dire que plus de 89% des observations sont dans l'intervalle $[\bar{x} - 3s ; \bar{x} + 3s]$.

Exercice 2 — *Série synthétique*

Pour la série de valeurs observées ci-dessous,

4	0	2	2	8	5

1) Calculer la moyenne, la médiane, le mode, les quartiles, l'étendue inter-quartiles, l'écart-type et le coefficient de variation. Les représenter graphiquement.

```
> x <- c(4,0,2,2,8,5)        # création d'un vecteur
> sort(x)                    # tri par valeurs croissantes
> median(x)                  # médiane
> mean(x)                    # moyenne
> quantile(x)                # quantiles
> Q1 <- quantile(x)[2]
> Q3 <- quantile(x)[4]
> etendue <- Q3-Q1           # étendue inter-quartiles
> sd(x)                      # écart-type(standard deviation)
> var(x)                     # variance
> boxplot(x)                 # trace la boite de Tuckey
> points(mean(x))            # ajoute la moyenne

# il est possible de modifier les paramètres graphiques dans
toutes les fonctions graphiques : col="red",lwd=5 …
```

2) Refaire la même chose en ajoutant 10 à la série. Conclusion.
3) Refaire la même chose en multipliant la série par 2. Conclusion.
4) Refaire la même chose en ajoutant la valeur supplémentaire 25 à la série. Conclusion.

```
> y <- x+10     # 10 s'ajoute à chaque composante de x
> z <- x*2      # chaque composante de x est multipliée par 2
> w <- c(x,25)  # concaténation
```

	Série	Série +10	Série *2	Série avec 25
moyenne	3,5	13,5	7	6,57
médiane	3	13	6	4
mode	2	12	4	2
Q1	2	12	4	2
Q3	4,75	14,75	9,5	6,5
variance	6,58	6,58	26,33	62,24
ecart-type	2,57	2,57	5,13	7,89
ecart Q1-Q3	2,75	2,75	5,5	4,5

- Ajouter 10 augmente d'autant les caractéristiques de position mais ne change rien à ceux de dispersion
- Multiplier par 2 la série multiplie toutes les caractéristiques par 2 sauf la variance qui est multipliée par 2^2.
- Ajouter la valeur extrême 25 ne change quasiment rien aux quantiles mais augmente nettement la moyenne et la variance.

Exercice 3 *Elèves*

Le fichier ElevesData.csv contient les résultats d'un sondage d'une classe de 58 élèves concernant des caractères physiques.

1) Importer le fichier ElevesData.csv dans votre session de travail R.

*L'importation des données des données avec R se fait sous forme d'un objet appelé **data.frame**. Il s'agit d'un tableau avec en ligne les observations (individus) et en colonne les variables observées. A la différence d'une matrice, un data.frame est un objet dont on peut extraire des renseignements (nom et type des variables, résumé,…).*

```
> tab <- read.table("ElevesData.csv",header=TRUE,sep=";",dec=".")
                                                  # importe le fichier
> names(tab)     # affiche le nom des variables
> tab$TAILLE     # affiche uniquement la variable TAILLE
> dim(tab)       # donne la dimension du tableau (nrow(tab) et ncol(tab)
                   pour le nb de lignes ou de colonnes)
> tab[,1:3]      # affiche les 3eres colonnes
> tab[10:12,]    # affiche les lignes 10, 11, 12
```

2) Déterminer la nature de chaque caractère et faire la représentation graphique des effectifs appropriés.

```
> ######
> str(tab)       # affiche la nature des variables
> summary(tab)   # donne un résumé numérique des variables

> #### Variables qualitatives nominales
> tab$SEXE <- as.factor(tab$SEXE)    # transforme la variable SEXE en
                                       variable qualitative
> effSEXE <- table(tab$SEXE)         # calcul des effectifs
> pie(effSEXE,labels=c("H","F"))     # trace le diagramme circulaire
> title("Répartition H/F")           # titre du graphique
> effYEUX <- table(tab$YEUX)
> barplot(as.matrix(effYEUX),col=heat.colors(3),legend.text=TRUE)
                                                    # diagramme en barre
> #### Variables ordinales ou discrètes
> effPOINTURE <- table(tab$POINTURE)
> barplot(effPOINTURE, main="Distribution Pointures") # diagramme en
                                                                bâtons
> #### Variables continues
> hist(tab$POIDS,breaks=5,freq=FALSE,xlab="POIDS",main="Histogramme")
```

3) Transformer le caractère taille en caractère qualitatif en considérant le découpage suivant : <150cm = petit, >180 = grand, sinon moyen. Faire sa représentation graphique et comparer avec la précédente.

```
> vect <- hist(tab$TAILLE,breaks=c(131,149,180,198))$counts
> barplot(vect,names.arg=c("petits","moyens","grands"))
```

Réponses : on constate une perte d'information …

4) Créer un tableau croisé des effectifs des couleurs des yeux chez les hommes et chez les femmes. Peut-on comparer ?

```
> tableau <- table(tab$SEXE,tab$YEUX)
> addmargins(tableau)     # ajoute les totaux en ligne et colonne
> prop.table(tableau)     # calcule les pourcentages
```

Réponses : Le tableau croisé des effectifs (ou %) ne permet pas de faire des comparaisons entre les deux sexes car il n'y a pas le même nombre d'hommes et de femmes.

Il faut faire un tableau des proportions par sexe. On peut alors faire des conclusions du type "chez les hommes, 34,78% ont les yeux bleus contre 17,14% chez les femmes".

Exercice 4 — *Dépenses éducation USA*

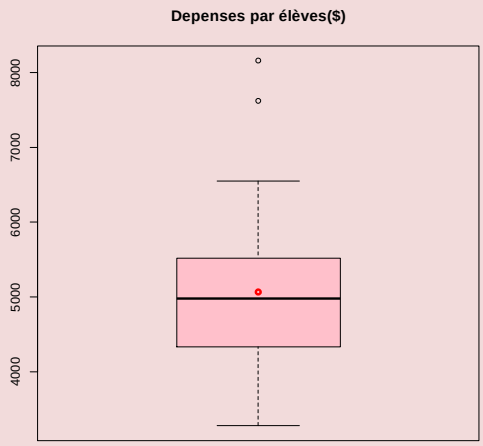
Le fichier EducationUSAData.csv présente les dépenses moyennes (en $) par élève et les résultats d'évaluation nationale pour chacun des états unis d'Amérique.

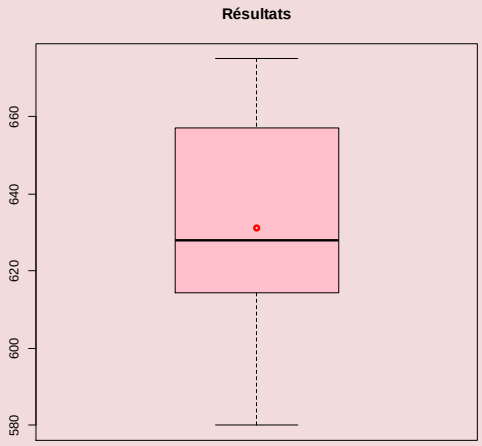
```
tab=read.table("EducationUSAData.csv",header=T,sep=";")
> str(tab)
'data.frame':   35 obs. of  2 variables:
 $ Depenses: int  3777 4041 4060 4917 4772 7629 6208 4934 4663 5532 ...
 $ Resultats: int  604 618 615 580 644 657 615 611 611 580 ...
```

1) Calculer la médiane, la moyenne et l'écart-type de chacune des deux séries (dépenses par élève et résultats d'évaluation). Faire une boite de Tukey pour chacune des séries.

```
> summary(tab)
   Depenses      Resultats
 Min.   :3280   Min.   :580.0
 1st Qu.:4339   1st Qu.:614.5
 Median :4985   Median :628.0
 Mean   :5069   Mean   :631.2
 3rd Qu.:5524   3rd Qu.:657.0
 Max.   :8162   Max.   :675.0
> boxplot(tab$Depenses,col="pink",main="Depenses par élèves($)")
> points(mean(tab$Depenses),col="red",lwd=3)
```

```
> boxplot(tab$Resultats,col="pink",main="Résultats")
> points(mean(tab$Resultats),col="red",lwd=3)
```





2) Peut-on comparer les degrés de dispersion de chacune des séries à partir de leur écart-type.

```
> sd(tab$Depenses)
[1] 1085.691
> sd(tab$Resultats)
[1] 27.57502
```

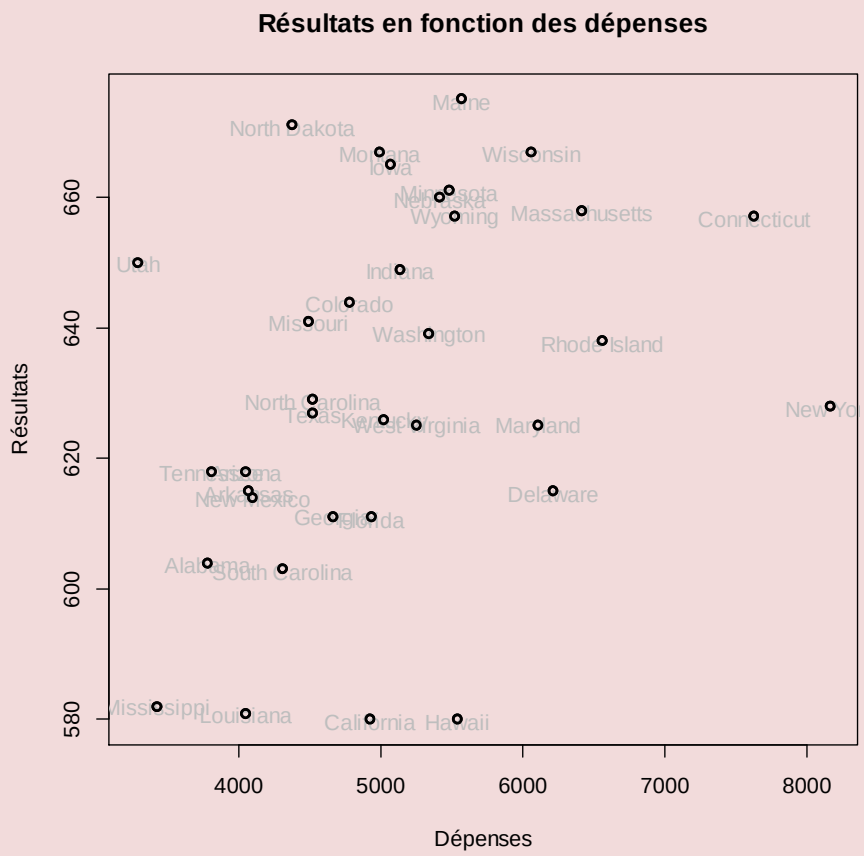
Les valeurs des écarts-types ne permettent de comparer les dispersions car les séries n'ont pas la même échelle.

3) Calculer les rapports écart-type sur moyenne et donner une première impression sur la corrélation entre dépenses et résultats. Illustrer votre conclusion à l'aide d'un graphique.

```
> sd(tab$Depenses)/mean(tab$Depenses)
[1] 0.2141898
> sd(tab$Resultats)/mean(tab$Resultats)
[1] 0.04368863
```

Les rapports écart-type/moyenne indiquent que la disparité des valeurs de dépenses ne se justifie pas par les résultats assez resserrés.

```
> plot(tab$Depenses,tab$Resultats,main="Résultats en fonction des dépenses",xlab="Dépenses",ylab="Résultats",lwd=2)
> text(tab$Depenses,tab$Resultats,row.names(tab),col="grey")
```



Rien n'indique sur le nuage de points un lien entre les deux variables.

Exercice 5 — *Salariés d'une entreprise*

Le fichier SalariesData.csv présente les salariés d'une entreprise ayant 3 sites (A, B et C). On y indique leur sexe, leur salaire annuel (en milliers d'euros), leur catégorie (cadre supérieur, moyen ou ouvrier employé), leur âge et leur site.

```
tab <- read.table("SalariesData.csv",header=T,sep=";")
```

1) Calculer le salaire moyen et le salaire médian. Que signifie la différence entre ces deux valeurs?

```
> mean(tab$Salaire)
[1] 32.03846
> median(tab$Salaire)
[1] 23
> hist(tab$Salaire)
```

La différence entre salaire moyen et médian est due aux très gros salaires des cadres supérieurs.

2) Calculer le 1er et le 3eme quartile, ainsi que le 1er et le 9eme déciles de la série des salaires.

```
> quantile(tab$Salaire,prob=c(0.1,0.25,0.75,0.90))
```

```
 10%  25%  75%  90%
18.0 21.0 26.0 55.1
```

3) Lesquels faudrait-il mettre en avant lors d'une négociation salariale, et pourquoi?

Les dirigeants mettraient en avant les quartiles qui masquent les écarts de salaires, alors que les salariés parleraient des déciles.

4) Etablir un tableau avec effectifs ou fréquences pour la série des catégories et celle des établissements. Faire une représentation graphique adaptée.

```
> pEtab=table(tab$Etablissement)/130*100
> pEtab
    A       B       C
46.15385 30.76923 23.07692
> barplot(as.matrix(pEtab),legend=T,col=heat.colors(3))
```

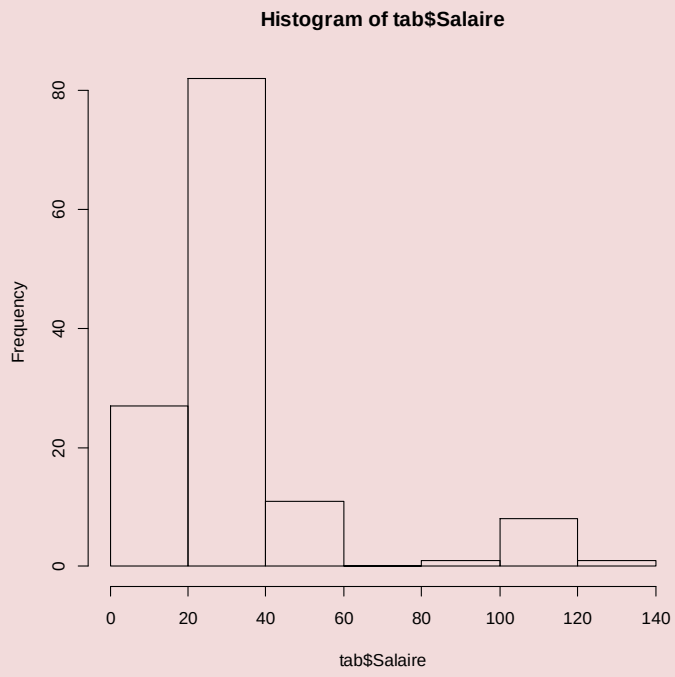
5) Représenter les variables discrètes salaire et âge? Cette représentation vous semble t'elle adaptée?

```
> barplot(tab$Salaire)
> barplot(tab$Age)
```

Les représentations par un diagramme en bâtons ne sont pas adaptées car il y a trop de valeurs dans chaque série. Il faut faire un regroupement par classe.

6) Effectuer un regroupement par classe des salaires puis des âges (pas plus de 5 classes). Recalculer la moyenne et comparer avec la série d'origine.

```
> h <- hist(tab$Salaire, freq=TRUE,breaks=5)
```



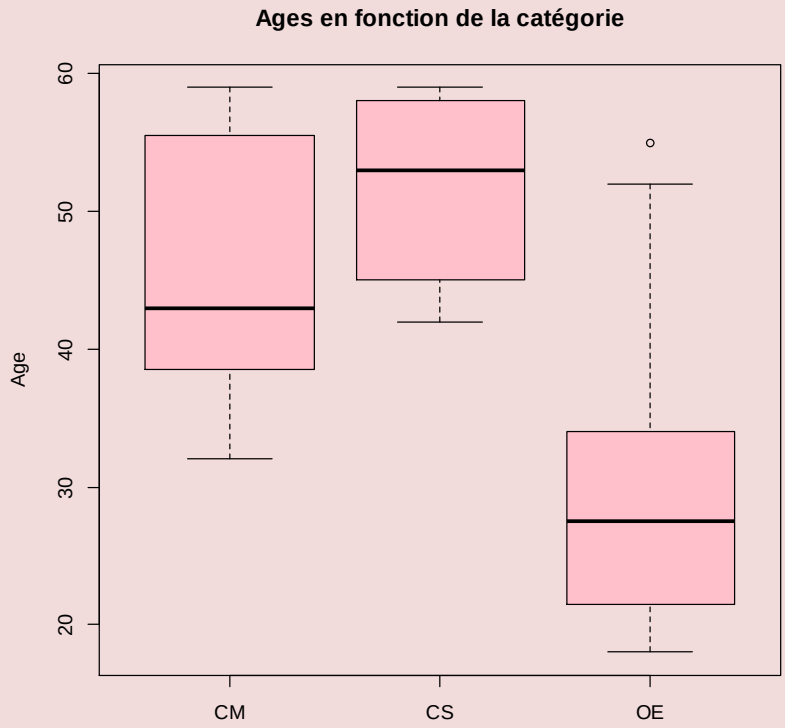

```
> seg <-h$breaks
> seg
```

```
[1]   0  20  40  60  80 100 120 140
> eff <-h$counts
> eff
[1] 27 82 11  0  1  8  1
> nbclasses <- length(eff)
> nbclasses
[1] 7
> lclasses=max(tab$Salaire)/nbclasses
> lclasses
[1] 20
> midclasses <- seg+lclasses/2
> midclasses <- midclasses[1:nbclasses]
> midclasses
[1]  10  30  50  70  90 110 130
> sum(midclasses*eff)/sum(eff)
[1] 33.69231
> mean(tab$Salaire)
[1] 32.03846
```

La différence entre les moyennes est due à l'hypothèse de répartition uniforme dans chaque classe.

7) Faire une étude comparative des âges de chaque catégorie de salariés (CS, CM, OE), en termes de tendances centrales et mesures de dispersion. Faire une représentation graphique de cette comparaison.

```
boxplot(tab$Age~tab$Categorie,main="Ages en fonction de la catégorie",ylab="Age",col="pink")
```



Les cadres supérieurs sont plus agés et les ouvriers plus jeunes. Les ages des cadres moyens sont plus dispersés.