Master 1 GSI - Mentions ACCIE et RIM - ULCO, La Citadelle, 2013/2014

Mesures et Analyses Statistiques de Données - Probabilités

CORRECTION TP 1.3 - Statistiques descriptives avec le logiciel R

Exercice 1 Sauvegarde de la fenêtre graphique sous format pdf (Boîte à moustaches)

```
> europe<-read.table("europe.csv",dec=".",sep=";",quote="¨",header=TRUE)
> pdf(file="boxplot.pdf",width=6,height=6,onefile=TRUE,family="Helvetica",
+ title="Europe boxplot",paper="special")
> boxplot(europe$Durée.heures.,ylab="Durée (heures)")
> points(1,mean(europe$Durée.heures.),pch=2)
> dev.off()
```

Exercice 2 Travail personnel

Exercice 3

1. • Données :

   ```
   > Femmes<-scan()
   > Hommes<-scan()
   ```

   • Regroupement en classes :

   ```
   > classes<-c(104,114,...,184)
   > regroupFemmes<-cut(Femmes,classes)
   > regroupHommes<-cut(Hommes,classes)
   ```

   • Effectifs :

   ```
   > effFemmes<-table(regroupFemmes)
   > effHommes<-table(regroupHommes)
   ```

   • Fréquences :

   ```
   > freqFemmes<-effFemmes*100/length(Femmes)
   > freqHommes<-effHommes*100/length(Hommes)
   ```

2. Histogrammes :

   ```
   > hist(Femmes,classes)
   > hist(Hommes,classes)
   ```

3. Moyennes distributions initiales :

   ```
   > mean(Femmes)
   ```
   132.9333
   ```
   > mean(Hommes)
   ```
   158.8667
   ```
   > mean(c(Femmes,Hommes))
   ```
   145.9

4. Moyennes après regroupement :

   • On regroupe les femmes et les hommes :

   ```
   > Ens<-c(Femmes,Hommes)
   > regroupEns<-cut(Ens,classes)
   > effEns<-table(regroupEns)
   ```

- On crée un data.frame contenant toutes les informations :

```
> tab<-data.frame(effFemmes,effHommes,effEns)
> tab
```

- On calcule les moyennes des trois groupes :

```
> moyF=0
> for(i in 1 :length(tab$Freq)){
>     moyF=moyF+tab$Freq[i]*(classes[i]+classes[i+1])/2}
> moyF=moyF/length(Femmes) 132.6667
```

```
> moyH=0
> for(i in 1 :length(tab$Freq.1)){
>     moyH=moyH+tab$Freq.1[i]*(classes[i]+classes[i+1])/2}
> moyH=moyH/length(Hommes) 158
```

```
> moyEns=0
> for(i in 1 :length(tab$Freq.2)){
>     moyH=moyH+tab$Freq.2[i]*(classes[i]+classes[i+1])/2}
> moyEns=moyEns/length(Ens) 145.3333
```

5. Quartiles :

```
> summary(sort(Femmes)) Q1=121.2 Q2=133.5 Q3=144.2
> summary(sort(Hommes)) Q1=151.5 Q2=158.0 Q3=165.8
> summary(sort(Ensemble)) Q1=133.8 Q2=149.5 Q3=158.5
```

6. Calcul de l'intervalle interquartile à l'aide du calcul explicite des quantiles :

```
> iiqFemmes<-quantile(Femmes,0.75)-quantile(Femmes,0.25) 23
> iiqHommes<-quantile(Hommes,0.75)-quantile(Hommes,0.25) 14.25
> iiqEnsemble<-quantile(Ensemble,0.75)-quantile(Ensemble,0.25) 24.75
```

7. Calcul des variances et écart-types des trois distributions initiales :

```
> var(Femmes) 208.2023
> sd(Femmes) 14.42922
> var(Hommes) 96.3954
> sd(Hommes) 9.818116
> var(Ens) 320.7017
> sd(Ens) 17.90815
```

8. Calcul des variances et écart-types des trois distributions après regroupement :

```
> varF=0
> for(i in 1 :length(tab$Freq)){
>     varF=varF+tab$Freq[i]*((classes[i]+classes[i+1])/2)^2}
> varF=varF/length(Femmes)-moyF^2 196.5556
```

```
> varH=0
> for(i in 1 :length(tab$Freq.1)){
>     varH=varH+tab$Freq.1[i]*((classes[i]+classes[i+1])/2)^2}
> varH=varH/length(Hommes)-moyH^2 109
```

```
> varEns=0
> for(i in 1 :length(tab$Freq.2)){
```

```
>     varEns=varEns+tab$Freq.2[i]*((classes[i]+classes[i+1])/2)^2}
> varEns=varEns/length(Ens)-moyEns^2
```
313.2222

9. (Question supplémentaire) Calcul des moments et des moments centrés $E((X-E(X))^k)$ pour $k \in \{1,\dots,4\}$.

• Calcul des moments d'ordre 3 (= coefficient d'asymétrie, *skewness* en anglais) des femmes (sans regroupement) soit

$$E(X^3) \simeq \frac{1}{n}\sum_{i=1}^{n} x_i^3 \text{ et } E((X-E(X))^3) \simeq \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3 :$$

```
> mom3F=0 ; mom3centF=0
> for(i in 1 :length(Femmes)){
>     mom3F=mom3F+((classes[i]+classes[i+1])/2)^3
>     mom3centF=mom3centF+((classes[i]+classes[i+1])/2-moyF)^3}
> mom3F=mom3F/length(Femmes)
```
3212784
```
> mom3centF=mom3centF/length(Femmes)
```
19305.70

• Calcul des moments d'ordre 4 (= coefficient d'aplatissement, *kurtosis* en anglais) des femmes (sans regroupement) soit

$$E(X^4) \simeq \frac{1}{n}\sum_{i=1}^{n} x_i^4 \text{ et } E((X-E(X))^4) \simeq \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4 :$$

```
> mom4F=0 ; mom4centF=0
> for(i in 1 :length(Femmes)){
>     mom4F=mom4F+((classes[i]+classes[i+1])/2)^4}
>     mom4centF=mom4centF+((classes[i]+classes[i+1])/2-moyF)^4}
> mom4F=mom4F/length(Femmes)
```
495785721
```
> mom4centF=mom4centF/length(Femmes)
```
906723

• Test de Fisher exact : il permet de tester si les fréquences entières observées sur 2 échantillons sont identiques ou non :

```
> mat<-matrix(c(tab$Freq,tab$Freq.1),ncol=2)
> Fisher <- fisher.test(mat)
```

Exercice 4

11. Le nuage de points a une allure longiligne. Il semble donc y avoir un lien linéaire entre les deux variables.
12. On peut faire la même étude avec le "sépale" :
    ```
    > plot(iris$Sepal.Length,iris$Sepal.Width,xlab="Longueur du sépale", ylab="Largeur du sépale",
    + main="Nuage de points, pch=20)
    ```
    Il n'y a par contre pas de lien linéaire entre ces deux variables.
13. Chaque espèce a sa dispersion propre. L'espèce "virginica" est l'espèce pour laquelle la dispersion est le plus importante.
14. `> boxplot(iris$Petal.Width iris$Species,col=grey(0.6))`
15.
16.