Université Claude Bernard Lyon 1

M1APA (Semestre 1) statistique/gestion
Session 1 26 novembre 2007

Corrigé de l'examen CCF2 de statistique

IMPORTANT : Seules sont rédigées les corrections des exercices 1 à 4, proposés par J. Bastien

L'ensemble des script R est disponible sur `http://utbmjb.chez-alice.fr/UFRSTAPS/` **sous la forme d'un fichier zip appelé** `corexamA07.zip`

Correction de l'exercice 1.

(1) Les instructions permettant de donner les formules de base suivantes

$$\overline{n} = \frac{1}{p}\sum_{i=1}^{p} n_i,$$

puis l'écart type

$$\sigma = \sqrt{\frac{1}{p}\sum_{i=1}^{p} (n_i - \overline{n})^2},$$

et l'écart type estimé (déviation standart)

$$s = \sqrt{\frac{1}{p-1}\sum_{i=1}^{p} (n_i - \overline{n})^2}$$

sont données dans le fichier `exo1_corexamA07.R`. On rappelle les rappelles ici (plusieurs versions sont proposées) : si la variable `note` contient les notes n_1 à n_p :

```
sum(note)/length(note)
mean(note)

sqrt(sum((note-mean(note))^2)/length(note))
sd(note)*sqrt((length(note)-1)/(length(note)))

sqrt(sum((note-mean(note))^2)/(length(note)-1))
sd(note)
```

(2)

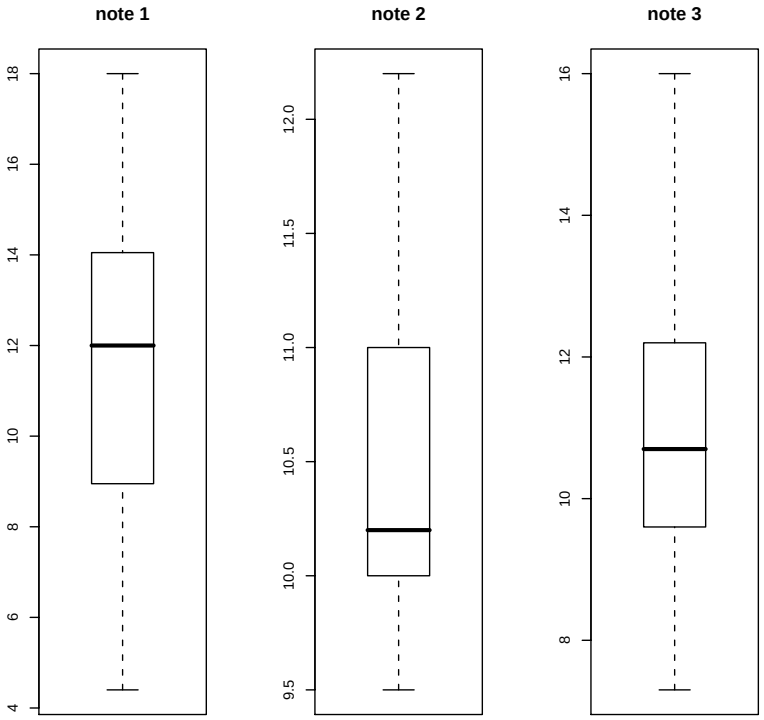

FIG. 1. trois boîtes de dispersion

Les commandes permettant de réaliser le graphique 1 (appelé boîte de dispersion ou à moustache) ou le graphique 2 page suivante (histogramme en densité) sont données dans le même fichier `exo1_corexamA07.R`. On les rappelle ici (plusieurs versions sont proposées) :

```
# déclaration des 3 vecteurs de données.
note1<-c(14.9,12.0,9.5,7.3,8.4,9.8,11.0,13.8,14.3,
    5.0,4.4,14.3,13.7,18.0,12.4)
note2<-c(10.9,10.1,10.0,12.2,10.0,11.1,10.3,9.5,
    9.6,10.0,10.9,11.2)
note3<-c(13.0,12.1,8.7,10.9,12.7,9.5,10.5,12.2,
    16.0,10.3,9.6,10.9,7.3,9.8)

notes<-c(note1, note2, note3)
groupe<- rep(c(1, 2, 3), c(length(note1), length(note2), length(note3)))

par(mfrow=c(1,3))
for (i in 1:3) boxplot(notes[groupe==i],main=paste("note",i))
pause()

# la même chose plus simplement (mais moins beau informatiquement)
```

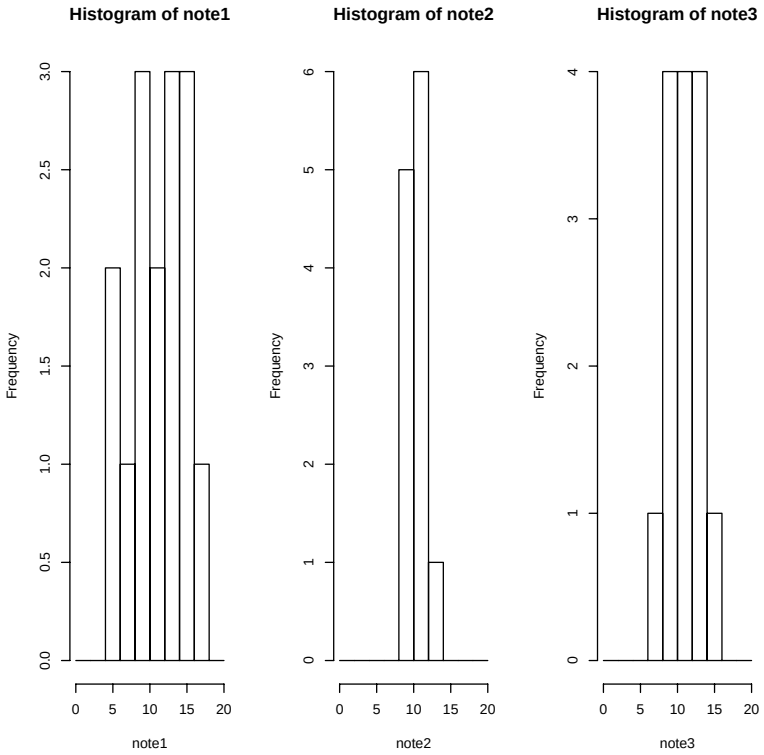

FIG. 2. trois histogrammes

```
par(mfrow=c(1,3))
boxplot(note1,main="note1")
boxplot(note2,main="note2")
boxplot(note3,main="note3")
pause()

cla<-seq(from=0, to=20, by=2)
par(mfrow=c(1,3))
hist(note1,breaks=cla)
hist(note2,breaks=cla)
hist(note3,breaks=cla)
```

Correction de l'exercice 2.
On consultera le fichier `exo2_corexamA07.R`.

(1) La variable `sexe` est qualitative (et n'a que deux modalité, `masculin` ou `feminin`). La variable `taille` est en revanche quantitative (discrète, puisque la taille a été arrondie au centimètre près).

(2)

On obtient la taille de contingence donnée dans le tableau 1 page suivante grâce à l'instruction

taille	féminin	masculin
$[155, 160[$	4	0
$[160, 165[$	11	0
$[165, 170[$	9	0
$[170, 175[$	5	9
$[175, 180[$	2	6
$[180, 185[$	0	9
$[185, 190[$	0	1
$[190, 195[$	0	2

TAB. 1. table de contingence taille-sexe.

```
table(taillecat,sexe)
```

où la variable `(taillecat` contient la distribution de taille qualitative du tableau 2.

taille	$[155, 160[$	$[160, 165[$	$[165, 170[$	$[170, 175[$	$[175, 180[$	$[180, 185[$	$[185, 190[$	$[190, 195[$
nombre	4	11	9	14	8	9	1	2

TAB. 2. étude de la taille.

Elle a pu être déterminée grâce (question non posée !) à

```
taillecat<-cut(taille,breaks=seq(from=155, to=195, by=5),right=FALSE)
```

(3) On constate dans le tableau 1 que, pour les femmes, les effectifs non nuls sont groupés autour des classe de taille $[155, 160[$ à $[175, 180[$ avec un pic pour la classe $[160, 165[$; au contraire pour les hommes, les effectifs non nuls sont groupés autour des classe de taille $[170, 175[$ à $[190, 195[$ avec un pic pour la classe $[170, 175[$. Cela confirme l'opinion répandue *a priori* que les femmes sont plus petites que les hommes et donc que la taille dépend du sexe.

(4)

Confirmons cela grâce à R !

On contruit la table de contingence (tableau 3 page ci-contre) sous l'hypothèse que les deux variables `taille` et `sexe` sont indépendantes en tapant sous R :

```
tabletailsexe<-table(taillecat,sexe)
res<-chisq.test(tabletailsexe)
expect<-res$expected
print(expect)
```

(5) On obtient les totaux lignes par lignes et colonnes par colonnes en tapant

taille	féminin	masculin
[155, 160[	2.1379310	1.8620690
[160, 165[	5.8793103	5.1206897
[165, 170[	4.8103448	4.1896552
[170, 175[	7.4827586	6.5172414
[175, 180[	4.2758621	3.7241379
[180, 185[	4.8103448	4.1896552
[185, 190[	0.5344828	0.4655172
[190, 195[	1.0689655	0.9310345

TAB. 3. table de contingence taille-sexe, supposée indépendantes.

```
margin.table(tabletailsexe,1))
margin.table(tabletailsexe,2))
```

```
margin.table(expect,1))
margin.table(expect,2))
```

On constate que l'on obtient les même totaux lignes par lignes et colonnes par colonnes ! Cela est normal puisque la construction de la table de contingence sous l'hypothèse que les deux variables `taille` et `sexe` sont indépendantes se fait grâce à ces totaux !

(6) Pour calculer le coefficient V de Cramer, on utilise la fonction fournie `cramer` en tapant

```
source("cramer.R")
cramer(taillecat,sexe))
```

(7) On obtient un V fort, ce qui corrobore donc (*a posteriori*) que les variables sont fortement liées.

(8) Cette méthode est criticable dans la mesure on a dû transformer la variable `taille` *a priori* quantitative en variable qualitative. On peut se demander si les résultats obtenus dépendent de la largeur des intervalles considéres lors de la transformation. Il serait peut-être plus judicieux de croiser directement la variable quantitative `taille` avec la variable qualitative `sexe`, en utilisant par exemple les théories de la feuille de TDR208 (non vue ce semestre !) Voir `http://pbil.univ-lyon1.fr/R/fichestd/tdr208.pdf`.

Correction de l'exercice 3.
On consultera le fichier `exo3_corexamA07.R`.

(1) (a) La fonction `read.table` permet de lire un fichier (dont on indique le nom entre double guillement ou dont indique l'URL absolue au même format) au format texte et de le stocker dans une variable. La fonction `head` appliquée alors à cette variable donne le haut du tableau.

sport	effectif
athlétisme	20
basket	2
foot	33
hand	44
judo	13
natation	24
volley	19

TAB. 4. Nombre de pratiquants par sport.

(b)

La variable `sport` est qualitative et contient 7 modalités comme l'indique le tableau 4. On peut l'obtenir en tapant

```
levels(sport)
summary(sport)
```

(c)

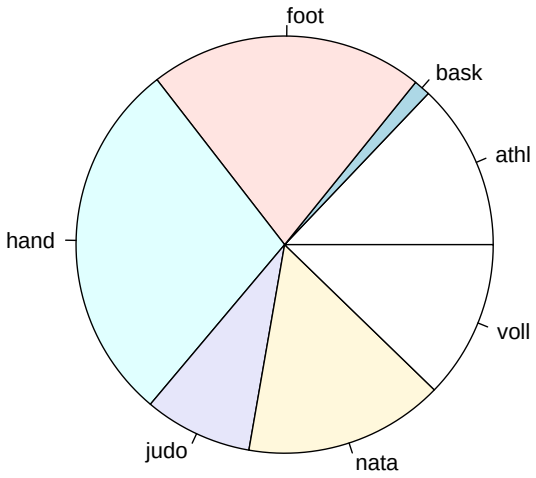

FIG. 3. le camembert des sports pratiqués

On obtient le camembert de la figure 3 en tapant

```
pie(summary(sport))
```

(d)

L'histogramme des IMC (voir figure 4 page suivante) est obtenu en tapant

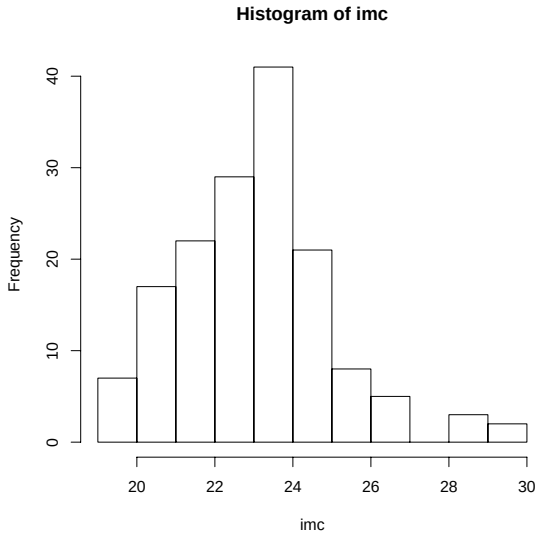

FIG. 4. Histogramme des IMC

```
imc<-poids/taille^2
hist(imc)
```

(2) (a) On calcule les IMC des joueurs de foot en tapant :

```
IMCfoot<-imc[sport=='foot']
mean(IMCfoot)
sd(IMCfoot)
```

Remarque 1. Le tableau `sport=='foot'` est en effet un tableau de booléens (vrais ou faux) qui sont vrais pour les individus pratiquant du football et faux sinon. Ainsi l'instruction

```
IMCfoot<-imc[sport=='foot']
```

stocke dans la variable `IMCfoot` l'ensemble des IMC des joueurs de foot (ce dernier point n'était pas exigible !)

(b)

La boîte de dispersion des IMC des joueurs de foot (voir figure 5 page suivante) est obtenue en tapant

```
boxplot(IMCfoot)
```

(c) Elle fait apparaître la médiane (trait horizontal plein) , le premier et le troisième quartile (Q_1 et Q_3 : bords supérieur et inférieur de la boîte), les quantités $Q_1 - 1.5(Q_3 - Q_1)$ et $Q_3 + 1.5(Q_3 - Q_1)$ (traits inférieur et supérieur) et des points extrêmes (ronds) qui sont en dessous de $Q_1 - 1.5(Q_3 - Q_1)$ et au dessus de $Q_3 + 1.5(Q_3 - Q_1)$ On constate donc que la médiane vaut à peu près 22.4, Q_1 vaut à peu près 21.8, Q_3 vaut à peu près 23.3, informations que l'on peut aussi obtenir en tapant

```
summary(IMCfoot)
```

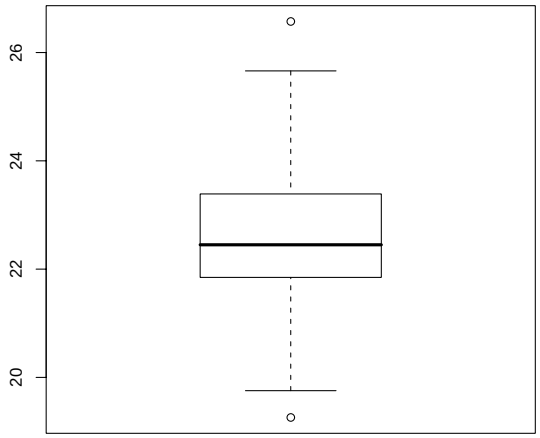

FIG. 5. Boîte de dispersion des IMC des joueurs de foot

Sur ce graphique, on constate donc que 50% des IMC sont compris entre 21.8 et 23.3.

Correction de l'exercice 4. On consultera le fichier `exo4_corexamA07.R`.

(1) Les instructions

```
cor(tde,poids)
plot(tde,poids,pch=20)
abline(lm(poids~tde),col="red")
```

affiche le coefficient de corrélation linéaire (qui vaut 0.6253807, ce qui fait un nuage de point moyennement alignés), trace le nuage de point (poids en fonction des taille debout) en rond noirs et rajoute sur ce graphique la droite de régression linéaire, qui passe «le plus près possible» de ce nuage de point (voir figure 6).

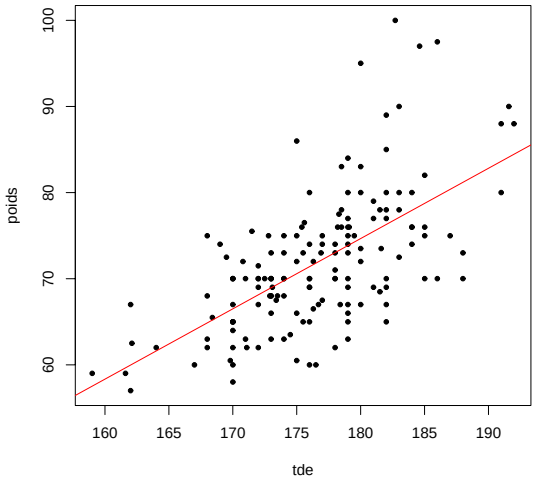

FIG. 6. le nuage de point (taille,poids) et la droite de régression linéaire pour l'ensemble des sportifs.

On travaille ensuite sur le poids en fonction de la taille en se restreignant aux athlètes en tapant :

```
cor(tde[sport=='athl'],poids[sport=='athl'])
plot(tde[sport=='athl'],poids[sport=='athl'],pch=20)
abline(lm(poids[sport=='athl']~tde[sport=='athl']),col="red")
coefficients(lm(poids[sport=='athl']~tde[sport=='athl']))
```

ce qui affiche le nuage de point et la droite de régression des athlètes seuls (voir figure 7).

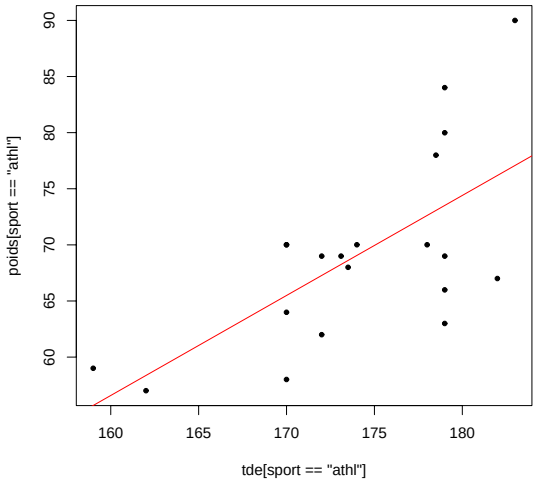

FIG. 7. le nuage de point (taille,poids) et la droite de régression linéaire pour les athlètes seuls.

Cela donne aussi la corrélation (0.6608253) et respectivement l'ordonnée à l'origine (-86.0094989) de la droite et sa pente (0.8911835). On constate donc que les athètes présentent un nuage de point un peu plus alignés que l'ensemble des joueurs (cela provient aussi du fait qu'ils sont moins nombreux !, voir tableau 4 page 6).

(2) On travaille ensuite sur le logarithme du poids en fonction du logarithme de la taille en se restreignant aux athlètes en tapant : (voir figure 8 page suivante)

```
cor(log(tde[sport=='athl']),log(poids[sport=='athl']))
plot(log(tde[sport=='athl']),log(poids[sport=='athl']),pch=20)
abline(lm(log(poids[sport=='athl'])~log(tde[sport=='athl'])),col="red")
coefficients(lm(log(poids[sport=='athl'])~log(tde[sport=='athl'])))
```

(3) On constate que la corrélation 0.675851 est légèrement plus élevée que sans le logarithme.

(4) Dans ce dernier cas, si la corrélation est proche de 1 (ou de -1), on peut écrire que

$$\log(\text{poids}) \approx p \times \log(\text{taille}) + C,$$

où p est la pente de la droite et C l'ordonnée à l'origine.

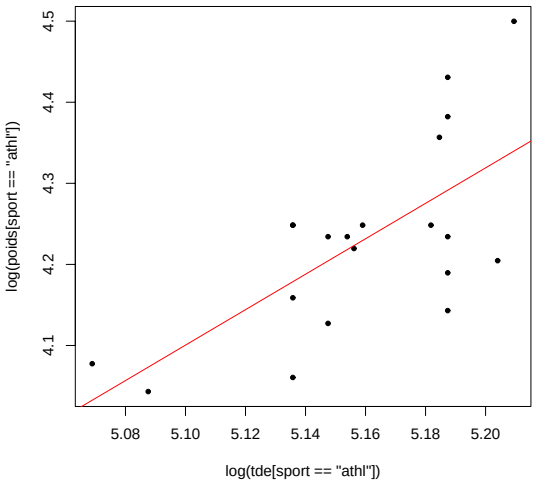

FIG. 8. le nuage de point (logarithme de la taille, logarithme du poids) et la droite de régression linéaire pour les athlètes seuls.

(5) (a) On peut réécrire l'équation précédente sous la forme

$$\log(\text{poids}) \approx p \times \log(\text{taille}) + \log(D),$$

où $C = \log(D)$ (soit $D = e^C$). Soit encore

$$\log(\text{poids}) \approx \log(\text{taille}^p \times D),$$

ce qui donne finalement

$$\text{poids} \approx D \times \text{taille}^p.$$

(b) Si cette relation est vraie, alors par définition de l'IMC, on a avec $p = 3$

$$\text{IMC} = \frac{\text{poids}}{\text{taille}^2} \approx \frac{D \times \text{taille}^3}{\text{taille}^2}$$

et donc

$$\text{IMC} \approx D \times \text{taille}.$$