Université Claude Bernard Lyon 1

M1IGAPAS (Semestre 1) statistique/gestion
Session 2 04 Janvier 2010

Corrigé de l'examen CT de statistiques

Correction de l'exercice 1.
Cette exercice provient de [DR07].

(1) (a) On prendra garde au fait que la taille du fichier de donnée est en cm. On utilisera donc la formule suivante de l'IMC :

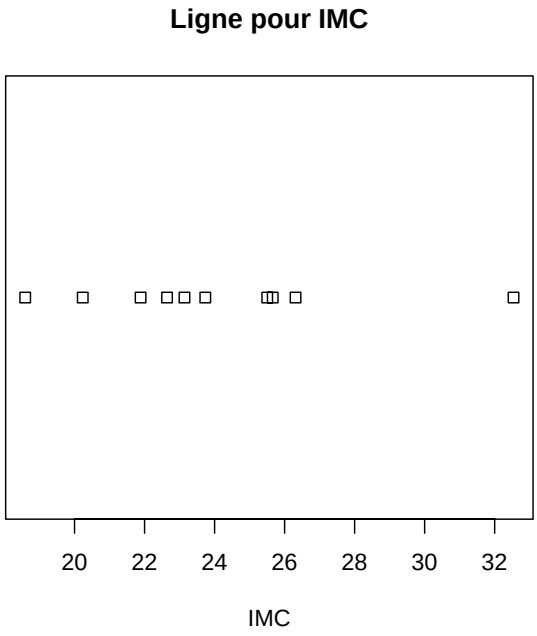
$$\text{IMC} = \frac{\text{poids}}{(\text{taille}/100)^2}, \tag{1}$$

Pour déterminer la colonne '`IMC`' du fichier '`OrJOsydney2000.txt`', on a introduit une nouvelle variable, nommée '`IMC`', définie par la formule précédente.

(b)
- On étudie la variable quantitative (ou numérique) '`IMC`'. Pour les manipulations avec R, on renvoie donc à la section 3.4 et à la section récapitulative 7.1.3 du document de cours.
- Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	24.0245
sd	3.856438
Q_1 (quartile à 25 %)	22.07575
médiane	23.4375
Q_3 (quartile à 75 %)	25.62175
minimum	18.59
maximum	32.539
nombre	10

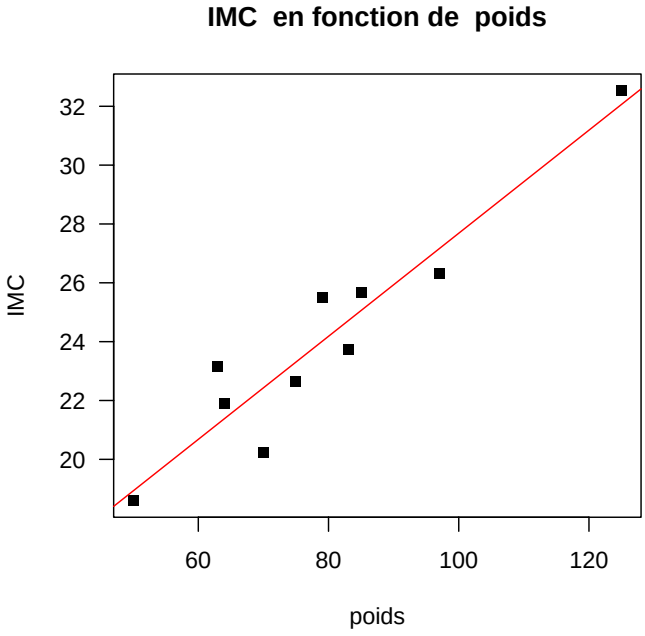
-



Voir le graphique ci-dessus pour la variable '`IMC`'.

(c) *Non commenté !*

(2) (a)
- On étudie le croisement de la variable quantitative (ou numérique) '`poids`' et de la variable quantitative (ou numérique) '`IMC`'. Pour les manipulations avec R, on renvoie donc à la section 4.5 et la section récapitulative 7.2.1 du document de cours.
- Voir la figure ci-dessous.



Sur cette figure, les points semblent bien alignés.

- Confirmons cela grâce à R.
  Les résultats donnés par R sont les suivants :

Noms des indicateurs	Valeurs
pente a	0.175147
ordonnée à l'origine b	10.170411
corrélation linéaire r	0.946743
probabilité critique p_c	3.29949e-05

  On compare la valeur absolue de la corrélation linéaire r =0.946743 aux seuils de Cohen (0.1,0.3,0.5) (voir [Coh92]) et la probabilité critique p_c=3.29949e-05 à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison linéaire :

significativité pratique	**très forte**
significativité statistique	**oui**

- On peut donc affirmer il existe une relation entre les variables '`poids`' et '`IMC`'.

(b) Les sportifs sont tous issus de sports assez différents (voir la variable '`specialite`'); leur morphologie sont assez différentes, et donc *a priori*, cette droite de regression ne devrait pas être pertinente. Curieusement, la relation est en fait très forte !

(3) Les points les moins bien représentés sont ceux qui s'écartent le plus de la droite de régression. Graphiquement, à partir du nuage de points, il semblerait que le point soit d'abcisse 70 (difficile de trouver avec précision le plus éloigné ...), soit encore (grâce au poids) l'aviron.

On peut procéder autrement pour classer par le calcul les points selon leur éloignement à la droite : on tape (copie-colle)

```
x <- OrJOsydney2000$poids
y <- OrJOsydney2000$IMC
lmd <- lm(y ~ x)
IMCtheo <- predict(lmd)
ecart <- abs(IMCtheo - y)
ecartsort <- sort(ecart, index.return = T, decreasing = T)
ordre <- ecartsort$ix
ecart[ordre]
OrJOsydney2000$poids[ordre]
OrJOsydney2000$specialite[ordre]
OrJOsydney2000$nom[ordre]
```

ce qui donne les différentes valeurs absolues des écarts à la droite, classés par ordre décroissant, les poids, les spécialités correspondantes et les noms des champions :

```
        3         2         8         7         1         6        10
2.1966667 1.9353589 1.4970147 0.9725714 0.8466226 0.6643993 0.6031356
        5         4         9
0.5072124 0.4752750 0.3377365
 [1]  70  63  79  83  97  75  85  64 125  50
 [1] Aviron            Boxe              Cyclisme_sur_Piste
 [4] Escrime           Aviron            Canoe_Kayak
```

```
 [7] Cyclisme_sur_Piste Tir                Judo
[10] VTT
Levels: Aviron Boxe Canoe_Kayak Cyclisme_sur_Piste Escrime Judo Tir VTT
 [1] Bette     Asloum    Gané      Ferrari   Andrieux  Estanguet Rousseau
 [8] Dumoulin  Douillet  Martinez
10 Levels: Andrieux Asloum Bette Douillet Dumoulin Estanguet Ferrari ... Rousseau
```

Bref,

- le point le plus éloigné de la droite correspond à
  - écart : 2.1966667
  - poids : 70
  - sport : '`Aviron`'
  - nom :'`Bette`'
- puis vient le point correspond à
  - écart : 1.9353589
  - poids : 63
  - sport : '`Boxe`'
  - nom :'`Asloum`'
- pour finir par le plus proche de la droite
  - écart : 0.3377365
  - poids : 50
  - sport : '`VTT`'
  - nom :'`Martinez`'

Correction de l'exercice 2.

(1)
- On étudie le croisement de la variable qualitative (ou catégorielle) '`sexe`' et de la variable qualitative (ou catégorielle) '`mecriture`'. Pour les manipulations avec R, on renvoie donc à la section 5.5 et la section récapitulative 7.2.2 du document de cours.
- La table de contingence déterminée par R est donnée dans le tableau suivant

	droite	gauche
féminin	28	3
masculin	23	4

Les autres résultats donnés par R sont les suivants :

Noms des indicateurs	Valeurs
χ^2	0.358898
coefficient de Cramer V	0.078663
taille d'effet w	0.078663
probabilité critique p_c	0.549119

On compare la taille d'effet w=0.078663 aux seuils de Cohen (0.1,0.3,0.5) (voir [Coh92]) et la probabilité critique p_c=0.549119 à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison :

significativité pratique	**faible**
significativité statistique	**non**

- On peut donc affirmer qu'il n'existe pas de relation entre les variables 'sexe' et 'mecriture'.

(2) *non commenté*

Références

[Coh92] J Cohen. A power primer. *Psychological bulletin*, 112(1) :155–159, 1992.

[DR07] AB Dufour et M Royer. Fiche de td 206 : Croisement de deux variables quantitatives. Disponible sur `http://pbil.univ-lyon1.fr/R/enseignement`, 2007.