Université Claude Bernard Lyon 1

M1IGAPAS (Semestre 1) statistique
Session 1 28 Novembre 2011

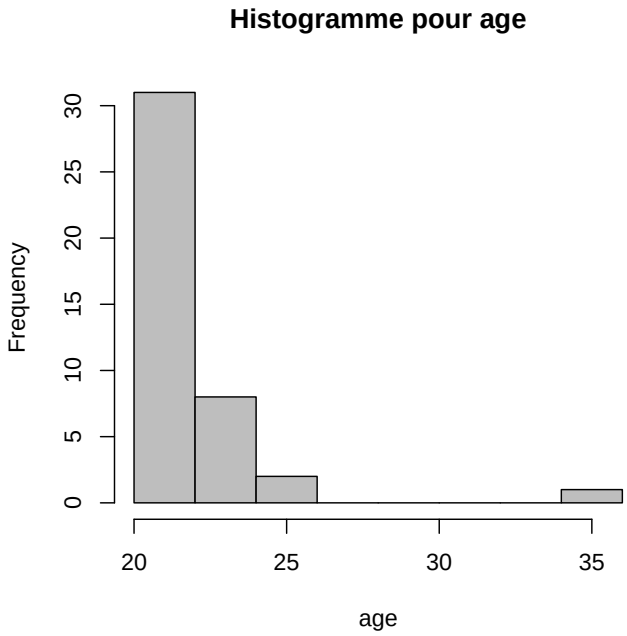
Corrigé de l'examen CCF2 de statistiques

Correction de l'exercice 1.

(1)
- On étudie la variable quantitative (ou numérique) '`age`'. Pour les manipulations avec R, on renvoie donc à la section 3.4 et aux sections récapitulatives 7.1.1 et 7.1.3 du document de cours.
- Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	22.29
écart-type	2.4
Q_1 (quartile à 25 %)	21
médiane	22
Q_3 (quartile à 75 %)	22.75
minimum	20
maximum	35
nombre	42

-



Voir les deux graphiques ci-dessus pour la variable 'age'.

(2) Vu que le nombre de données est "grand" (42 individus), les graphes les plus adaptés sont l'histogramme et la boîte à moustache.

(3) On les trace avec les commandes

```
M1IGAPASA11data<-read.table("M1IGAPASA11data,h=T)
hist(M1IGAPASA11data$age)
```

et

```
boxplot(M1IGAPASA11data$age)
```

(4) On calcule la moyenne, l'écart-type et les quartiles de cette variable en tapant

```
summary(M1IGAPASA11data$age)
sd(M1IGAPASA11data$age)
```

ce qui fournit

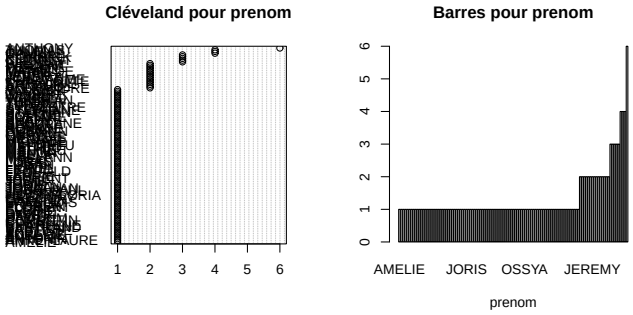
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   21.00   22.00   22.29   22.75   35.00
[1] 2.40209
```

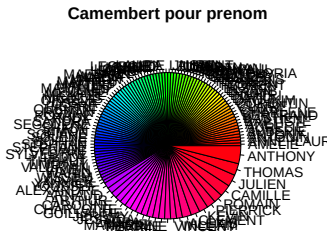
Correction de l'exercice 2.

(1) Les variables sont '`prenom`', '`groupe`' et '`note`'

(2) (a) • On étudie la variable qualitative (ou catégorielle) '`prenom`'. Pour les manipulations avec R, on renvoie donc à la section 3.3 et aux sections récapitulatives 7.1.1 et 7.1.2 du document de cours.

•





Voir les trois graphiques illisibles ci-dessus pour la variable '`prenom`'.

(b) Cette analyse n'est pas du tout pertinente, car seuls quelques prénoms sont présents plusieurs fois, la très grande majorité d'entre eux n'apparaissant qu'une seule fois.

(c) En tapant

```
sort(table(CCF1L2biomecaA11$prenom))
```

et en prenant les derniers, on peut donc voir ceux qui apparaissent 3 fois ou plus.

Mieux, on peut taper

```
u<-table(CCF1L2biomecaA11$prenom)
indd<-as.numeric(u)>=3
u[indd]
```

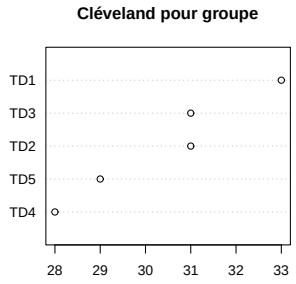
et on obtient

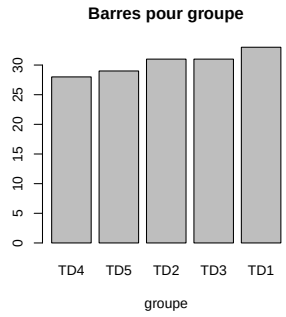
```
 ALEXIS  ANTHONY  CAMILLE  CLEMENT   JULIEN    KEVIN PIERRICK   ROMAIN   THOMAS
      3        6        4        3        4        3        3        3        4
```

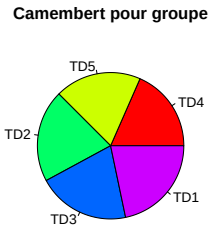
(3) (a)
- On étudie la variable qualitative (ou catégorielle) '`groupe`'.
- Les effectifs et les pourcentages déterminés par R sont donnés dans le tableau suivant

	effectifs	pourcentages
TD1	33	21.711
TD2	31	20.395
TD3	31	20.395
TD4	28	18.421
TD5	29	19.079

-







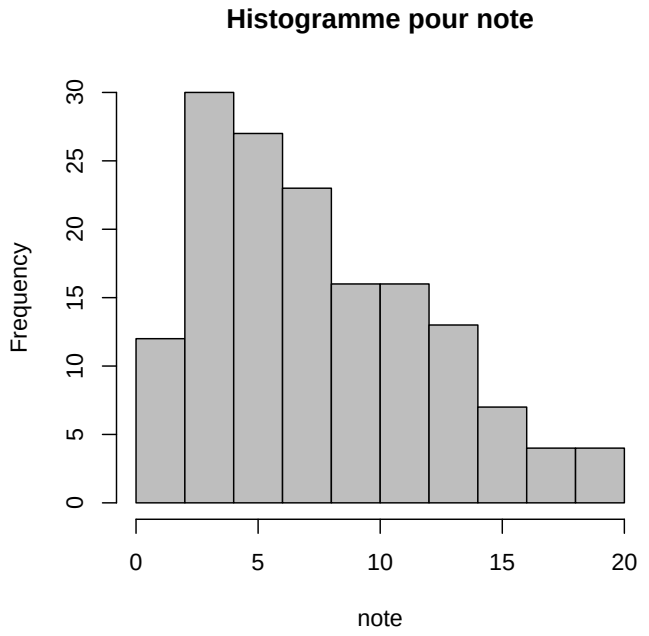
Voir les trois graphiques ci-dessus pour la variable '`groupe`'.

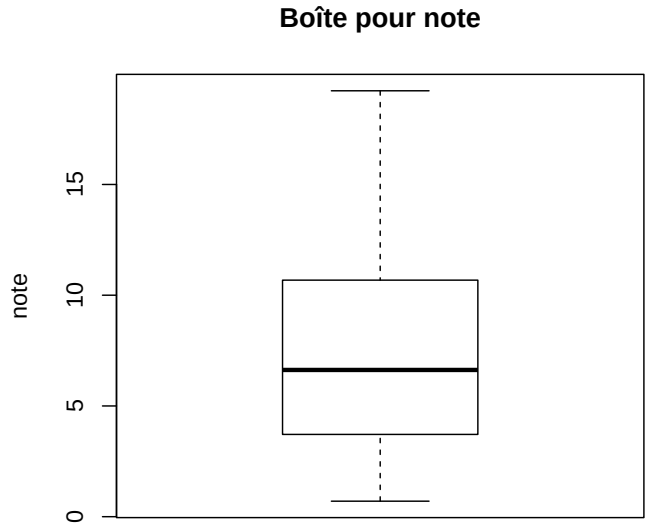
(b) Les 5 groupes de TD contiennent à peu près le même nombre d'étudiants : autour de 30, ce qui est raisonnable !

(4)
- On étudie la variable quantitative (ou numérique) '`note`'.
- Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	7.51
écart-type	4.48
Q_1 (quartile à 25 %)	3.75
médiane	6.62
Q_3 (quartile à 75 %)	10.66
minimum	0.7
maximum	19.23
nombre	152

-

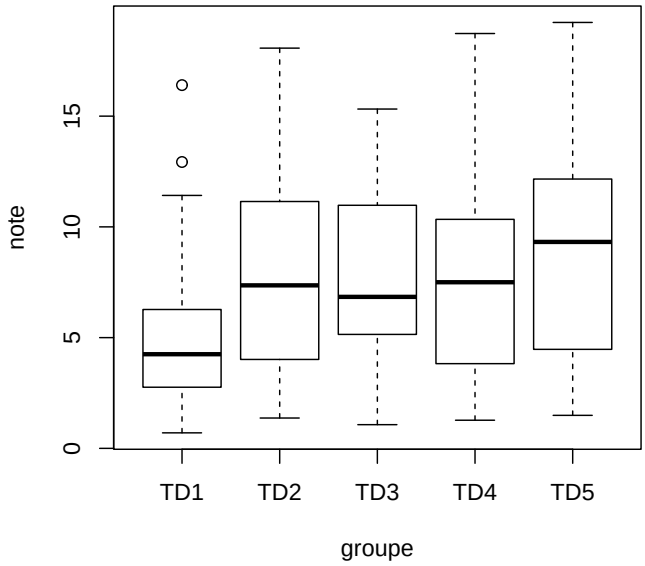


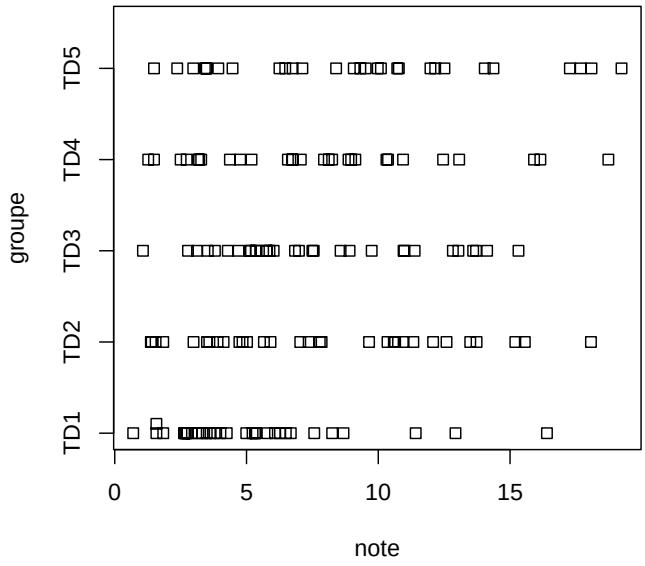


Voir les deux graphiques ci-dessus pour la variable '`note`'.

(5) (a) • On étudie le croisement de la variable qualitative (ou catégorielle) '`groupe`' et de la variable quantitative (ou numérique) '`note`'. Pour les manipulations avec ®, on renvoie donc aux sections 6.2 et 6.3 et la section récapitulative 7.2.3 du document de cours.

•





Voir la figure ci-dessous.

- Avec ®, on obtient les statistiques par groupes données dans le tableau suivant ;

On rappelle que, dans ce tableau :

- le nombre noté `0%` est le quartile à 0 % (c'est le minimum) ;
- le nombre noté `25%` est le quartile à 25 % (c'est Q_1) ;
- le nombre noté `50%` est le quartile à 50 % (c'est la médiane) ;

	moyenne	écart-type	0%	25%	50%	75%	100%	n
TD1	5.17	3.40	0.70	2.76	4.25	6.27	16.40	33
TD2	7.89	4.68	1.37	4.01	7.36	11.14	18.07	31
TD3	7.74	3.86	1.07	5.14	6.84	10.98	15.32	31
TD4	7.80	4.57	1.27	4.10	7.50	10.32	18.73	28
TD5	9.22	5.08	1.49	4.47	9.32	12.16	19.23	29

- le nombre noté `75%` est le quartile à 75 % (c'est Q_3) ;
- le nombre noté `100%` est le quartile à 100 % (c'est le maximum).

Les graphiques et les statistiques par groupes montrent une certaine hétérogénéité entre les notes des différents TD.
Confirmons cela grâce à R.
Les autres résultats donnés par R sont les suivants :

Noms des indicateurs	Valeurs
Rapport de corrélation RC	0.090498
probabilité critique p_c	0.00717507

On compare le rapport de corrélation RC=0.090498 aux seuils de Cohen (0.01,0.05,0.15) (voir [Coh92]) et la probabilité critique p_c=0.00717507 à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison :

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

- On peut donc affirmer qu'il existe une relation entre les variables '`groupe`' et '`note`'.

(b) On constate aussi que les moyennes des trois groupes 2, 3 et 4 sont à peu près égales (autour de 7.8), tandis que le premier groupe est en dessous des autres (moyenne de 5.17) et le dernier est au-dessus des autres (moyenne de 9.22). Cela explique donc le travail beaucoup plus important du premier groupe et beaucoup plus faible du dernier. Les autres groupes ont l'air d'être plus homogènes. Cela est confirmé par la dépendance entre la note et le groupe.

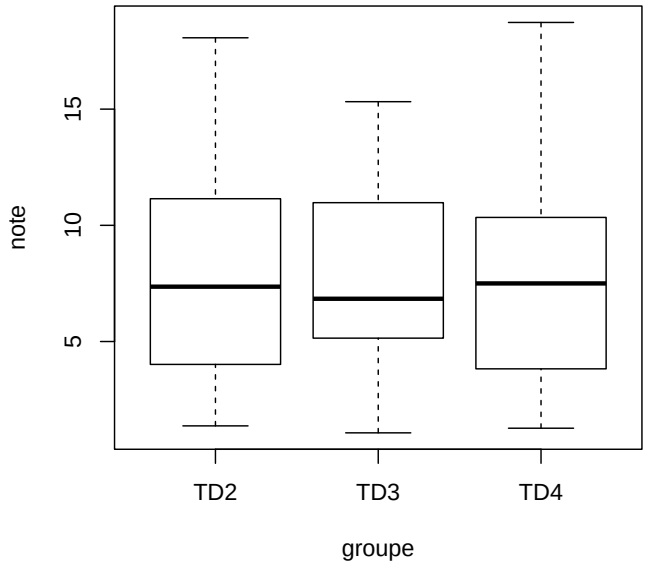
(c) (i) Voir question 5b

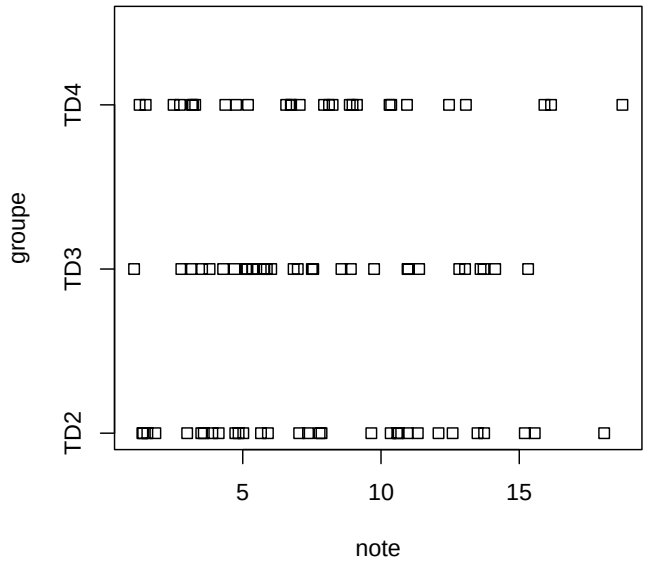
(ii) Les commandes suivantes :

```
CCF1 <- read.table("CCF1L2biomecaA11.txt", h = T)
CCF1 <- CCF1[CCF1[, 2] != "TD1" & CCF1[, 2] != "TD5", ]
```

chargent le fichier '`CCF1L2biomecaA11.txt`' dans la variable '`CCF1`' et ne gardent que les TD différents de 1 et de 5.

(iii) •





Voir la figure ci-dessous.

- Avec R, on obtient les statistiques par groupes données dans le tableau suivant ;

les trois groupes de TD ont l'air d'être homogènes.
Confirmons cela grâce à R.
Les autres résultats donnés par R sont les suivants :

	moyenne	écart-type	0%	25%	50%	75%	100%	n
TD2	7.89	4.68	1.37	4.01	7.36	11.14	18.07	31
TD3	7.74	3.86	1.07	5.14	6.84	10.98	15.32	31
TD4	7.80	4.57	1.27	4.10	7.50	10.32	18.73	28

Noms des indicateurs	Valeurs
Rapport de corrélation RC	0.000237
probabilité critique p_c	0.98974

On compare le rapport de corrélation RC=0.000237 aux seuils de Cohen (0.01,0.05,0.15) (voir [Coh92]) et la probabilité critique p_c=0.98974 à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison :

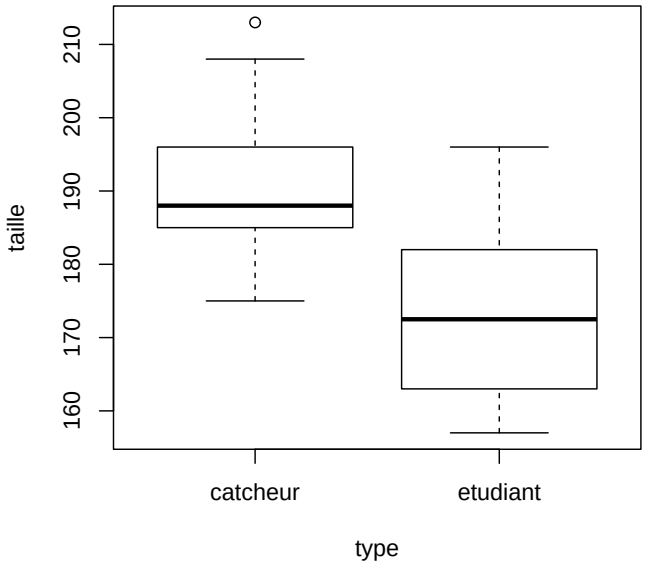
significativité pratique	**faible**
significativité statistique	**non**

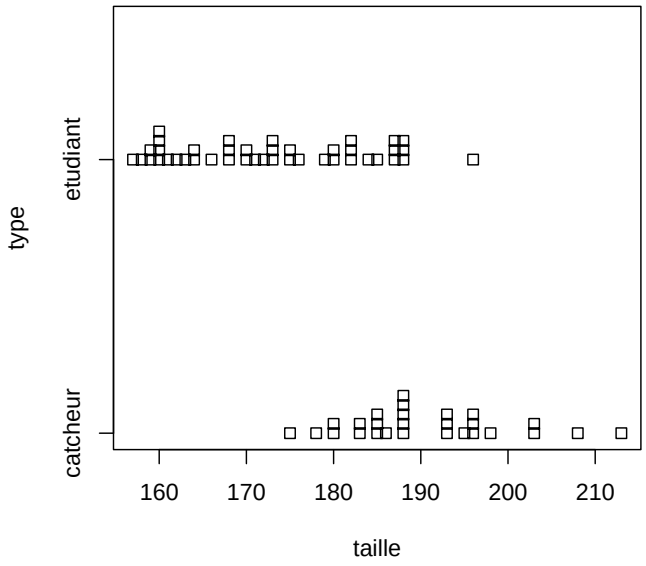
- On peut donc affirmer qu'il n'existe plus de relation entre les variables '`groupe`' et '`note`'. C'est normal, on a supprimé les deux groupes de TD différents des autres !

Correction de l'exercice 3.

(1) Tout la première question revient en fait, directement à faire le croisement de la variable '`taille`' et de la variable '`type`'.

- On étudie le croisement de la variable quantitative (ou numérique) '`taille`' et de la variable qualitative (ou catégorielle) '`type`'.
-





Voir la figure ci-dessous.

- Avec R, on obtient les statistiques par groupes données dans le tableau suivant ;

	moyenne	écart-type	0%	25%	50%	75%	100%	n
catcheur	190.63	9.16	175.00	185.00	188.00	196.00	213.00	27
etudiant	172.86	10.68	157.00	163.25	172.50	182.00	196.00	42

Les graphiques et les statistiques par groupes montrent une légère différence entre les types.
Confirmons cela grâce à R.
Les autres résultats donnés par R sont les suivants :

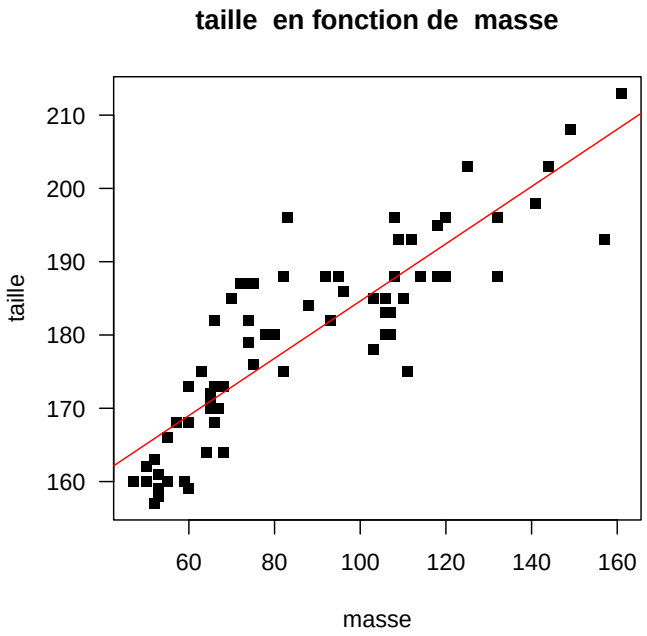
Noms des indicateurs	Valeurs
Rapport de corrélation RC	0.430921
probabilité critique p_c	9.12864e-10

On compare le rapport de corrélation RC=0.430921 aux seuils de Cohen (0.01,0.05,0.15) (voir [Coh92]) et la probabilité critique p_c=9.12864e-10 à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison :

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

- On peut donc affirmer qu'il existe une faible relation entre les variables '`taille`' et '`type`'.

(2) (a)
- On étudie le croisement de la variable quantitative (ou numérique) '`masse`' et de la variable quantitative (ou numérique) '`taille`'. 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	0.390598
ordonnée à l'origine b	145.569184
corrélation linéaire r	0.864839
probabilité critique p_c	9.95936e-22

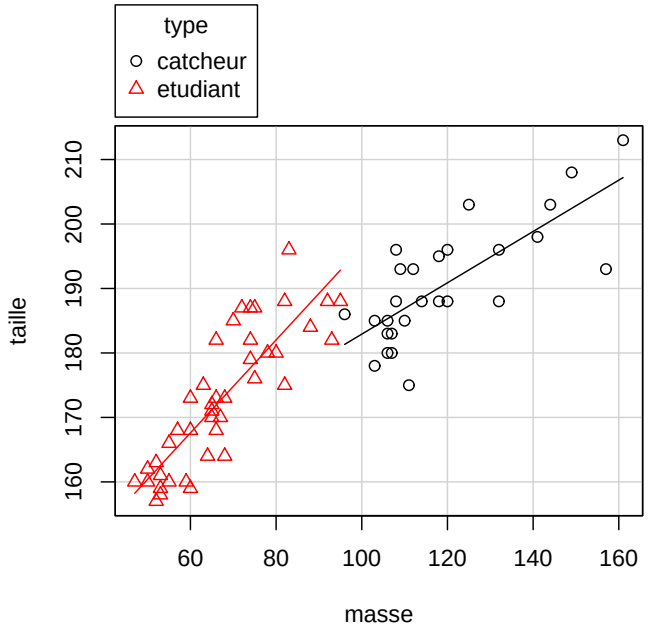
On compare la valeur absolue de la corrélation linéaire r =0.864839 aux seuils de Cohen (0.1,0.3,0.5) (voir [Coh92]) et la probabilité critique p_c=9.95936e-22 à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison linéaire :

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

- On peut donc affirmer il existe une relation faible entre les variables '`masse`' et '`taille`'.

En fait, cette analyse n'est pas pertinente, car les tailles des catcheurs et des étudiants sont très différentes (comme cela a été vu en question 1) et les masses sont très différentes. Ainsi, dans le nuage de point, se cachent en fait deux droite de régression linéaire, une par type de population, comme va le montrer la question 2b.

(b)



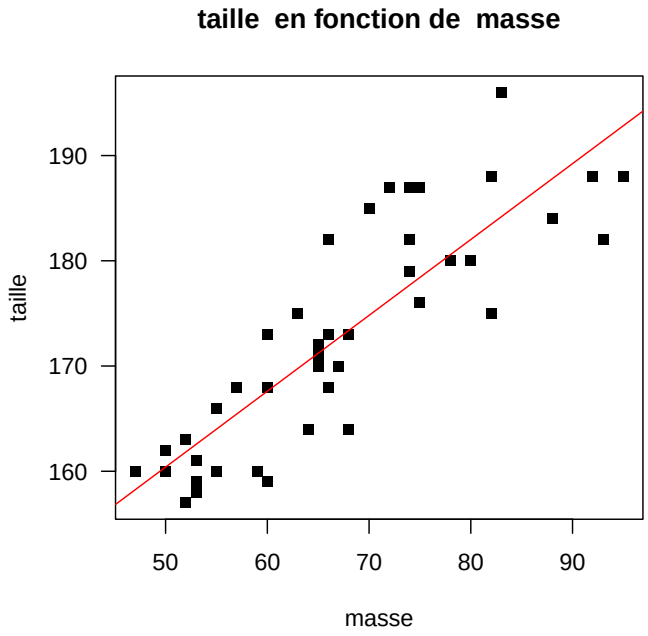
On obtiendrait le graphique ci-dessus.

Il met en évidence qu'il existe des relations entre la taille et la masse, mais ces relations dépendent de la population, ce qui confirme l'aspect inadapté de la question 2a.

On pourrait faire une analyse bivariée pour chacune des deux population et montrer que pour chacune d'elle, il y a bien une relation entre masse et taille :

(i) Pour les étudiants :

- Voir la figure ci-dessous.



- Les résultats donnés par R sont les suivants :

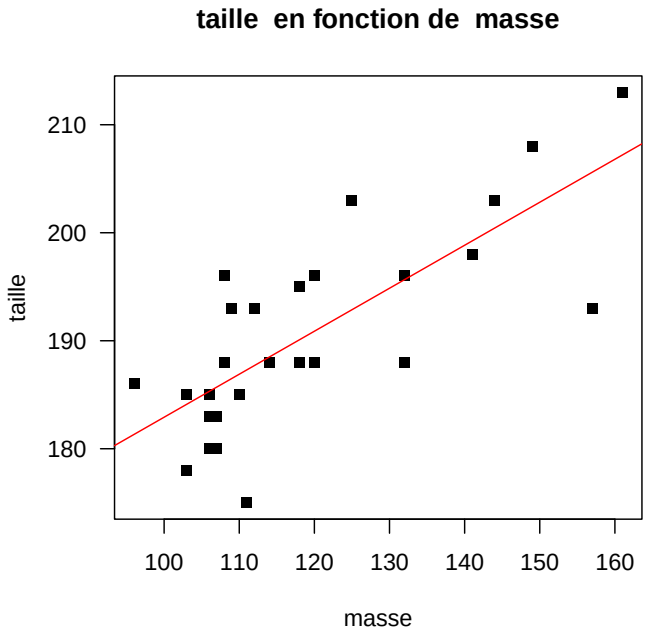
Noms des indicateurs	Valeurs
pente a	0.720898
ordonnée à l'origine b	124.351009
corrélation linéaire r	0.8492
probabilité critique p_c	1.183e-12

On compare la valeur absolue de la corrélation linéaire r =0.8492 aux seuils de Cohen (0.1,0.3,0.5) (voir [Coh92]) et la probabilité critique p_c=1.183e-12 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**

(ii) Pour les catcheurs :

- Voir la figure ci-dessous.



- Les résultats donnés par R sont les suivants :

Noms des indicateurs	Valeurs
pente a	0.398045
ordonnée à l'origine b	143.114816
corrélation linéaire r	0.761558
probabilité critique p_c	3.95301e-06

On compare la valeur absolue de la corrélation linéaire r =0.761558 aux seuils de Cohen (0.1,0.3,0.5) (voir [Coh92]) et la probabilité critique p_c=3.95301e-06 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**

Correction de l'exercice 4.

(1) On étudie une loi binomiale de paramètres $n = 15$ et $p = 0.8$.

(a) $P(X \leq 4) = 1.2e - 05$

(b) $P(X > 1) = 1$

(c) $P(2 \leq X \leq 10) = 0.164234$

(d) $P(X \leq -2) = 0$

(2) On étudie une loi normale centrée réduite (c'est-à-dire, de moyenne nulle et d'écart-type égal à 1).

(a) $P(X \leq -2.5) = 0.00621$

(b) $P(X > 0) = 0.5$

(c) $P(0.2 \leq X \leq 3.2) = P(X \leq 3.2) - P(X \leq 0.2) = 0.999313 - 0.57926 = 0.420053$

Correction de l'exercice 5.

(1) La moyenne de la taille est égale à 190.6296 et son écart-type est égal à 9.1616

(2) On en déduit l'intervalle de confiance de la moyenne au niveau 0.95 :

$$[187.0054, 194.2538],$$

et au niveau 0.99. :

$$[185.7303, 195.5289],$$

(3) On suppose que la taille suit une loi normale dont les paramètres ont été estimés précédemment, c'est-à-dire

$$\mu = 190.6296, \quad \sigma = 9.1616.$$

(a)

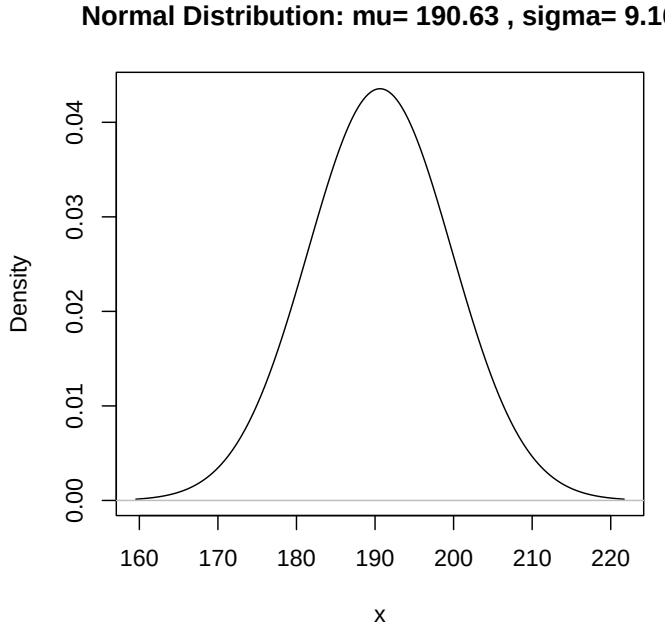


FIGURE 1. La loi normale.

Voir la figure 1.

(b) On calcule donc, pour cette loi normale, la probabilité :

$$P(X \geq 198) = 0.2106.$$

Correction de l'exercice 6.

(1) Pour les catcheurs, la moyenne de la taille est égale à 190.6296 et son écart-type est égal à 9.1616 et pour les étudiants, la moyenne est égale à 172.8571 et l'écart-type égal à 10.6761

(2) Pour les catcheurs, l'intervalle de confiance de la moyenne au niveau 0.95 est

$$[187.0054, 194.2538],$$

et au niveau 0.99. :

$$[185.7303, 195.5289],$$

Pour les étudiants, l'intervalle de confiance de la moyenne au niveau 0.95 est

$$[169.5302, 176.184],$$

et au niveau 0.99. :

$$[168.4073, 177.307],$$

(3)

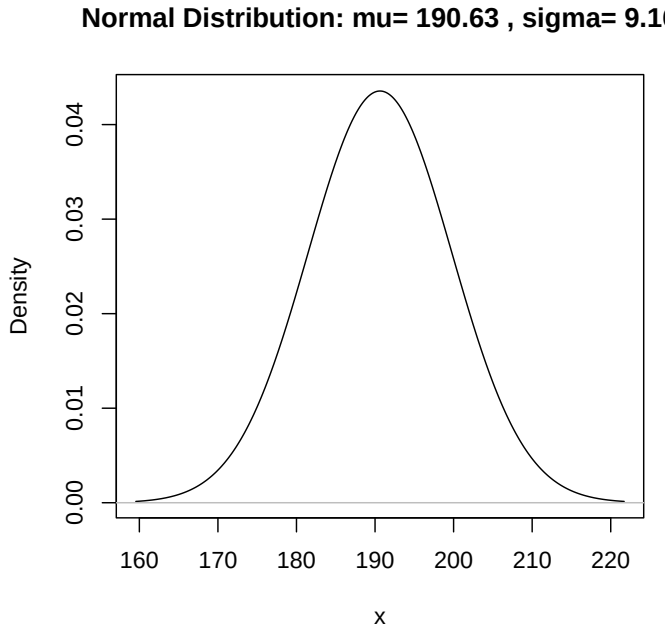


FIGURE 2. La loi normale pour les catcheurs.

Voir les deux figures 2 et 3. Cette deuxième courbe est moins faible à cause nombre de catcheurs (27) plus faible que le nombre d'étudiants (42).

(a) La probabilité pour un catcheur d'avoir une taille supérieure à 198. est égale à

$$P(X \geq 198) = 0.2106.$$

(b) La probabilité pour un étudiant d'avoir une taille supérieure à 198. est égale à

$$P(X \geq 198) = 0.0093.$$

(c) L'un est plus faible que l'autre, ce qui est normal car un individu extrait d'une population grande a plus de chance d'être grand qu'un individu extrait d'une population moins grande.

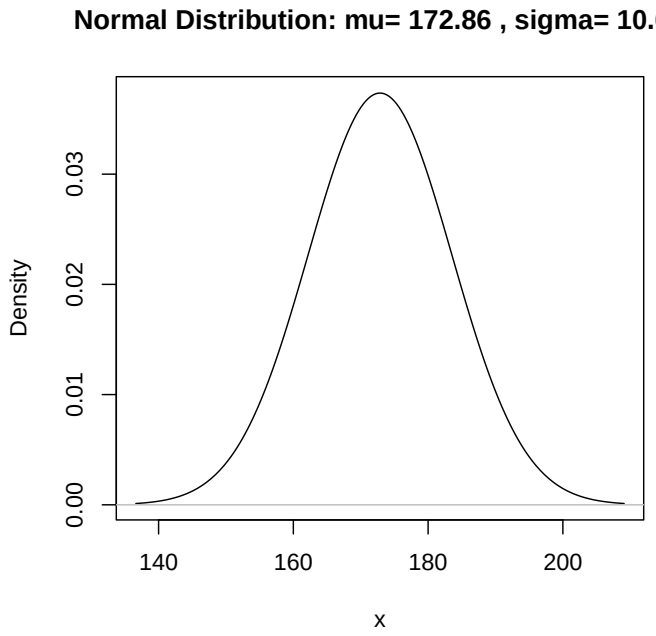


FIGURE 3. La loi normale pour les étudiants.

Références

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