Université Claude Bernard Lyon 1

M1IGAPAS (Semestre 1) statistique/gestion
Session 1 30 Novembre 2009

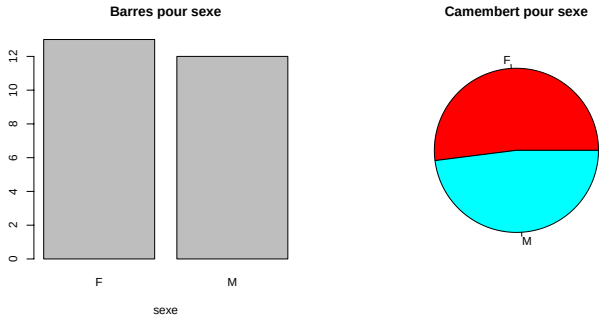
Corrigé de l'examen CCF2 de statistiques

Correction de l'exercice 1.

- On étudie la variable qualitative (ou catégorielle) '`sexe`'. Pour les manipulations avec R, on renvoie donc à la section 3.3 et à la section récapitulative 7.1.2 du document de cours.
- Les effectifs et les pourcentages déterminés par R sont donnés dans le tableau suivant

	effectifs	pourcentages
F	13	52.000
M	12	48.000

-



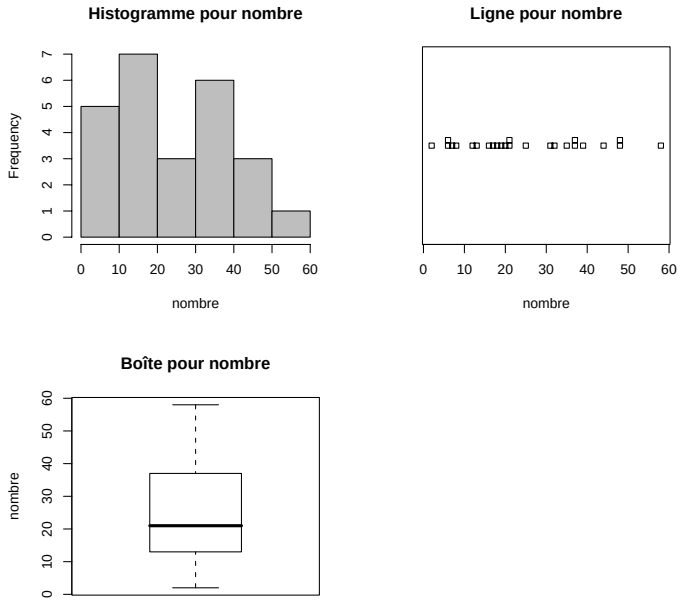
Voir les deux graphiques ci-dessus pour la variable '`sexe`' (un seul des deux suffisait). Il y a à peu près autant de femmes que d'hommes dans ce groupe d'étudiants.

Correction de l'exercice 2.

On étudie le fichier de données '`M1IGAPASA09data.txt`'.

(1) (a)
- On étudie la variable quantitative (ou numérique) '`nombre`'. 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	24.8
sd	15.32971
Q_1 (quartile à 25 %)	13
médiane	21
Q_3 (quartile à 75 %)	37
minimum	2
maximum	58
nombre	25

- 

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

(b) L'histogramme nous indique que les nombres entre 0 et 20 et 30 à 40 sont plus représentés que les autres.

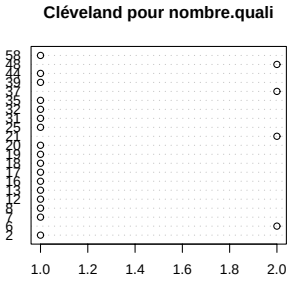
Remarque 1. Cette variable est qualitative. Pour en savoir un peu plus sur elle ici (ce qui n'était pas exigé !) et ses valeurs, en nombre fini, on pouvait la rendre quantitative grâce à la commande

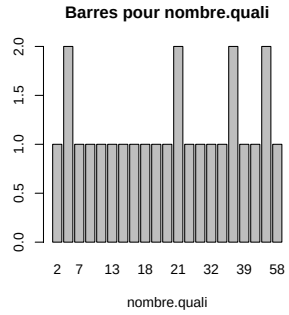
```
nombre.quali <- as.factor(M1IGAPASA09data$nombre)
```

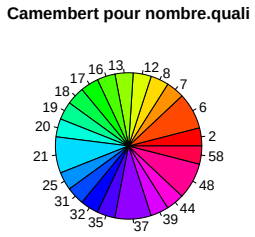
- On étudie la variable qualitative (ou catégorielle) '`nombre.quali`'. Pour les manipulations avec R, on renvoie donc à la section 3.3 et à la section récapitulative 7.1.2 du document de cours.
- Les effectifs et les pourcentages déterminés par R sont donnés dans le tableau suivant

-

	effectifs	pourcentages
2	1	4.000
6	2	8.000
7	1	4.000
8	1	4.000
12	1	4.000
13	1	4.000
16	1	4.000
17	1	4.000
18	1	4.000
19	1	4.000
20	1	4.000
21	2	8.000
25	1	4.000
31	1	4.000
32	1	4.000
35	1	4.000
37	2	8.000
39	1	4.000
44	1	4.000
48	2	8.000
58	1	4.000



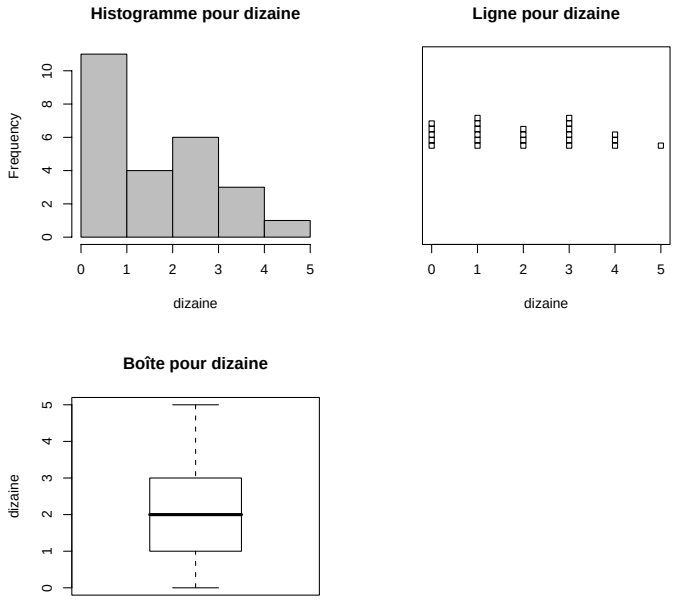




Voir les trois graphiques ci-dessus pour la variable '`nombre.quali`'. Sur le diagramme en barre ou le graphe de Cléveland, on peut constater que 4 valeurs sont représentés 2 fois et le reste, une quinzaine, est représenté une fois seulement.

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

noms	valeurs
moyenne	1.96
sd	1.485485
Q_1 (quartile à 25 %)	1
médiane	2
Q_3 (quartile à 75 %)	3
minimum	0
maximum	5
nombre	25

- 

  Voir les trois graphiques ci-dessus pour la variable '`dizaine`'. On constate sur l'histogramme que le chiffre zéro est plus représenté que les autres.

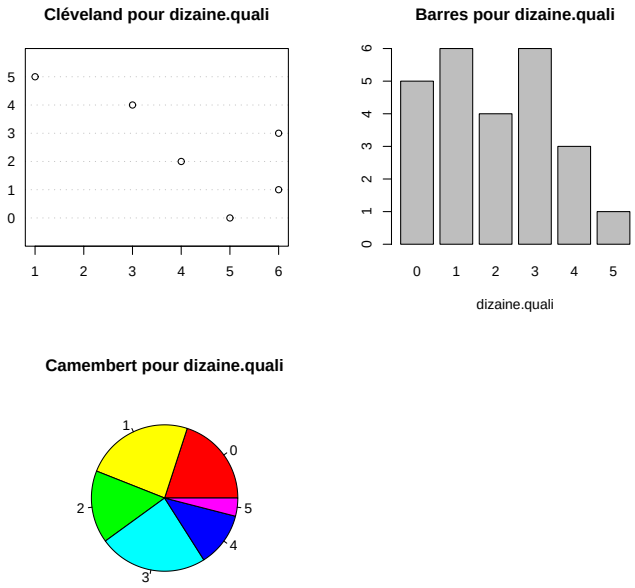
Remarque 2. Cette variable est quantitative. Comme dans la remarque 1 page 2, on peut aussi la rendre qualitative grâce à la commande

```
dizaine.quali <- as.factor(dizaine)
```

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

	effectifs	pourcentages
0	5	20.000
1	6	24.000
2	4	16.000
3	6	24.000
4	3	12.000
5	1	4.000

•



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

(b) On constate par rapport à la question 1, que les distributions des chiffres des dizaines (mis à part le 4 et le 5) semblent être réparties de façon plus uniforme (autant pour chacun d'entre eux).

(c) En fait, la distribution de nombres tirés au hasard doit être "à peu près "uniforme ; pour les nombres simples, ils sont trop nombreux (seulement 25 tirages, c'est-à-dire de nombres d'étudiants pout 60 possibilités) pour que cet aspect uniforme apparaisse. Pour les chiffres des dizaines (seulement 25 tirages, c'est-à-dire de nombres d'étudiants pour 6 possibilités), cet aspect est déjà plus observable. La séquance suivante

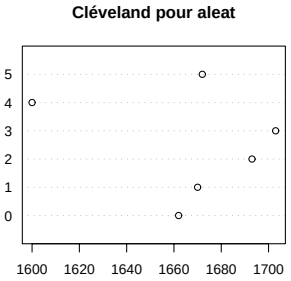
```
n <- 10000
aleat <- as.factor(sample(0:5, replace = T, size = n))
```

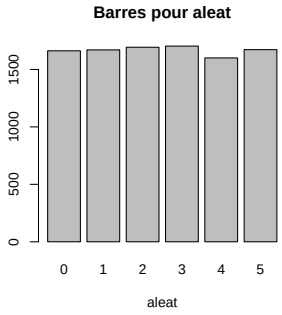
crée un tableau de type catégoriel grâce à 10000 tirages aléatoires valeurs dans l'ensemble $\{0, \ldots, 5\}$.

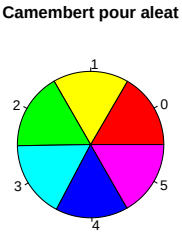
- On étudie la variable qualitative (ou catégorielle) '`aleat`'.
- Les effectifs et les pourcentages déterminés par sont donnés dans le tableau suivant

	effectifs	pourcentages
0	1662	16.620
1	1670	16.700
2	1693	16.930
3	1703	17.030
4	1600	16.000
5	1672	16.720

•



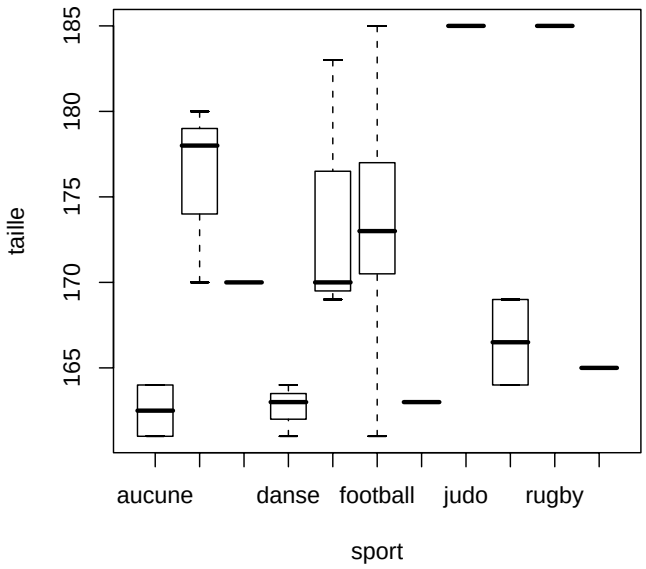


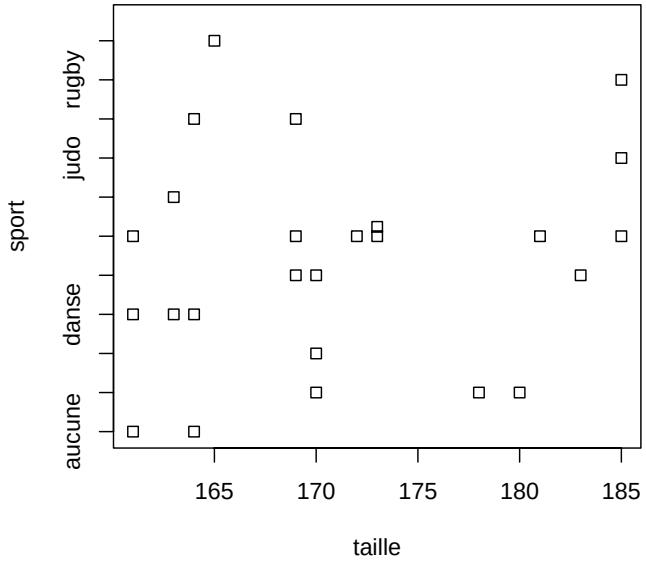


Voir les trois graphiques ci-dessus pour la variable '`aleat`'. Ici, les proportions observées pour chacun des nombres est proche de 16, ce qui correspond à 1/6. Les effecitfs sont proches de 1600, ce qui correspond à 10000/10000. En théorie des probabilités, chaque nombre "a autant de chance de sortir", ce qui justifie la valeur de 1/6, probabilité d'apparition de chacun des nombres.

Correction de l'exercice 3.

(1)
- On étudie le croisement de la variable qualitative (ou catégorielle) '`sport`' et de la variable quantitative (ou numérique) '`taille`'. Pour les manipulations avec R, 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 R, on obtient les statistiques par groupes données dans le tableau suivant ;

	moyenne	écart-type (sd)	0%	25%	50%	75%	100%	n
aucune	162.500	2.121	161.000	161.750	162.500	163.250	164.000	2
basket	176.000	5.292	170.000	174.000	178.000	179.000	180.000	3
basket_ball	170.000		170.000	170.000	170.000	170.000	170.000	1
danse	162.667	1.528	161.000	162.000	163.000	163.500	164.000	3
escalade	174.000	7.810	169.000	169.500	170.000	176.500	183.000	3
football	173.429	7.829	161.000	170.500	173.000	177.000	185.000	7
gymnastique	163.000		163.000	163.000	163.000	163.000	163.000	1
judo	185.000		185.000	185.000	185.000	185.000	185.000	1
natation	166.500	3.536	164.000	165.250	166.500	167.750	169.000	2
rugby	185.000		185.000	185.000	185.000	185.000	185.000	1
tir_sportif	165.000		165.000	165.000	165.000	165.000	165.000	1

On rappelle que :
- le quartile à 0 % correspond au minimum ;
- le quartile à 25 % correspond à Q_1 ;
- le quartile à 50 % correspond à la médiane ;
- le quartile à 75 % correspond à Q_3 ;
- le quartile à 100 % correspond au maximum.

On constate dans les statistiques par groupes que certains des écart-types ne sont pas définis. De plus, les effectifs par groupe sont très faibles (entre 1 et 7, 2 ou 3 en moyenne). En fait, les sports où il n'y a qu'un seul individus ne peuvent donner lieu à un écart-type, puisque l'on divise par $n - 1 = 0$. Les différentes boîtes de dispersions sont très variables.
Confirmons cela grâce à R.
Les autres résultats donnés par R sont les suivants :

Noms des indicateurs	Valeurs
Rapport de corrélation RC	0.644641
probabilité critique p_c	0.054326

On compare le rapport de corrélation RC=0.644641 aux seuils de Cohen (0.01,0.05,0.15) (voir [Coh92]) et la probabilité critique p_c=0.054326 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	**non**

(2) *A priori*, d'après les calculs précédents, la relation entre les variables '`sport`' et '`taille`' est très forte mais non statistiquement significative. Cela provient des très faibles effectifs par groupe ; d'autre part, d'après notre remarque précédente, on ne peut calculer certains écart-type. Ce calcul est donc non pertinent, compte tenu des faiblesses d'effectifs. On ne peut donc affirmer l'absence ou la présence de relation entre les variables '`sport`' et '`taille`'.

Correction de l'exercice 4.
Cet exemple pédagogique a été mis au point par Anscombe [Ans64] et provient de l'ouvrage [Cha04].

(1) On étudie le croisement de la variable quantitative (ou numérique) 'X' et de la variable quantitative (ou numérique) 'Y1'. Pour les manipulations avec R, on renvoie donc à la section 4.5 et la section récapitulative 7.2.1 du document de cours.

On a indiqué en figure 1 page 11 et 2 page 12, les quatres nuages de points et les droites de régression linéaire.

- Le premier présente un nuages de points qui semblent être à peu près alignés, pour lequel la régression linéaire a l'air pertinente.
- Le deuxième graphique nous indique un nuage de point en forme de parabole tournée vers le bas ; la régression linéaire n'est donc pas pertinente.
- Sur le troisième graphique, on peut constater, qu'hormis le dernier point, les points ont l'air d'être alignés. Cependant, ce dernier point, mesure extrême, a tendance à attirer la droite et la modifie par rapport au nuage de point sans cette donnée extrême ; la régression linéaire n'est donc pas pertinente.
- Enfin, sur le quatrième graphique, on constate que tous les points sauf un, on la même abscisse. Il n'existe donc pas de droite de régression pour les premiers points. Le dernier point modifie sensiblement la droite de régression ; la régression linéaire n'est donc pas pertinente.

(2) Étudions le croisement des variables 'X' et 'Y1'

Les résultats donnés par R sont les suivants :

Noms des indicateurs	Valeurs
pente a	0.812452
ordonnée à l'origine b	0.451378
corrélation linéaire r	0.786901
probabilité critique p_c	0.000298198

On compare la valeur absolue de la corrélation linéaire r =0.786901 aux seuils de Cohen (0.1,0.3,0.5) (voir [Coh92]) et la probabilité critique p_c=0.000298198 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**

Croisement	pente a	ordonnée à l'origine b	corrélation linéaire r	probabilités critique p_c
('X','Y1')	0.81245168	0.45137803	0.78690107	0.0002982
('X','Y2')	0.80852812	0.52367369	0.78539584	0.00031191
('X','Y3')	0.80619049	0.55752951	0.78427149	0.00032249
('Xp','Yp')	0.78255247	1.30087118	0.78384874	0.00032654

Dans le tableau ci-dessous, on a indiqué les quatre pentes et ordonnées à l'origine, ainsi que les quatre coefficients de corrélation linéaire et les quatre probabilités critiques obtenues. Les trois premières pentes et ordonnées l'origine sont à peu près égales ! Ainsi, les quatre nuages de points donnent mêmes

corrélations linéaires et mêmes probabilités critiques ! Cependant, d'après nos observations graphiques précédentes, seule la première régression linéaires est pertinente.

(3) La morale de l'histoire, c'est qu'il convient donc toujours de commencer par une visualisation des données avant de continuer les calculs de corrélation linéaire et de probabilité critique !

Remarque 3. Les données étudiées ici, créées de façon pédagogiques par Anscombe, sont en fait déjà présentes dans ℝ ! Il suffit de taper dans ℝ :

```
data(anscombe)
anscombe
```

Les variables du data frame '`anscombe`' sont : '`x1`', '`x2`', '`x3`', '`x4`', '`y1`', '`y2`', '`y3`' et '`y4`'.
On obtient des nuages de points un peu différents que ceux crées par le fichier de données, mais leurs propriétés sont les mêmes !

Croisement	pente a	ordonnée à l'origine b	corrélation linéaire r	probabilités critique p_c
('x1','y1')	0.50009091	3.00009091	0.81642052	0.00216963
('x2','y2')	0.5	3.00090909	0.81623651	0.00217882
('x3','y3')	0.49972727	3.00245455	0.81628674	0.00217631
('x4','y4')	0.49990909	3.00172727	0.81652144	0.0021646

Dans le tableau ci-dessous, on a indiqué les quatre pentes et ordonnées à l'origine, ainsi que les quatre coefficients de corrélation linéaire et les quatre probabilités critiques obtenues pour les données '`anscombe`'.
On a indiqué en figure 3 page 13 et 4 page 14, les quatres nuages de points et les droites de régression linéaire.
On pourra aussi consulter la rubrique de Wikipédia sur le quartet d'anscombe : `http://fr.wikipedia.org/wiki/Quartet_d'Anscombe`

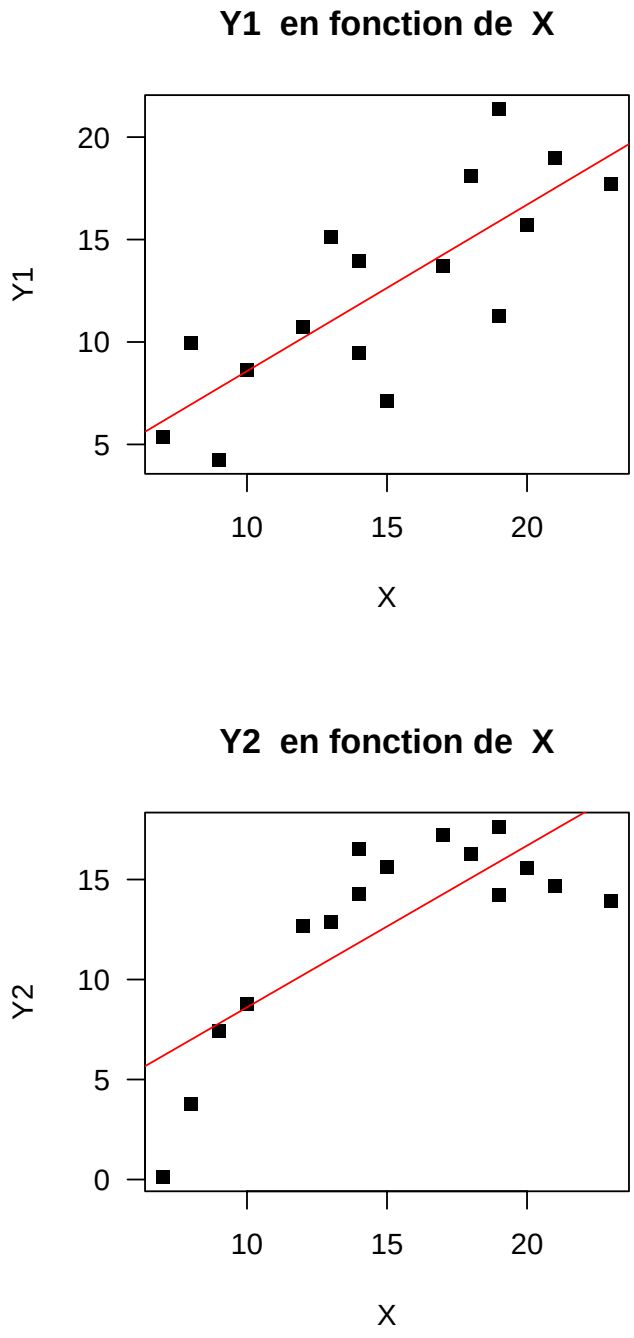


FIG. 1. Les deux premiers nuages de points et les droites de régression linéaire.

Références

[Ans64] F.J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27 :17–21, 1964.

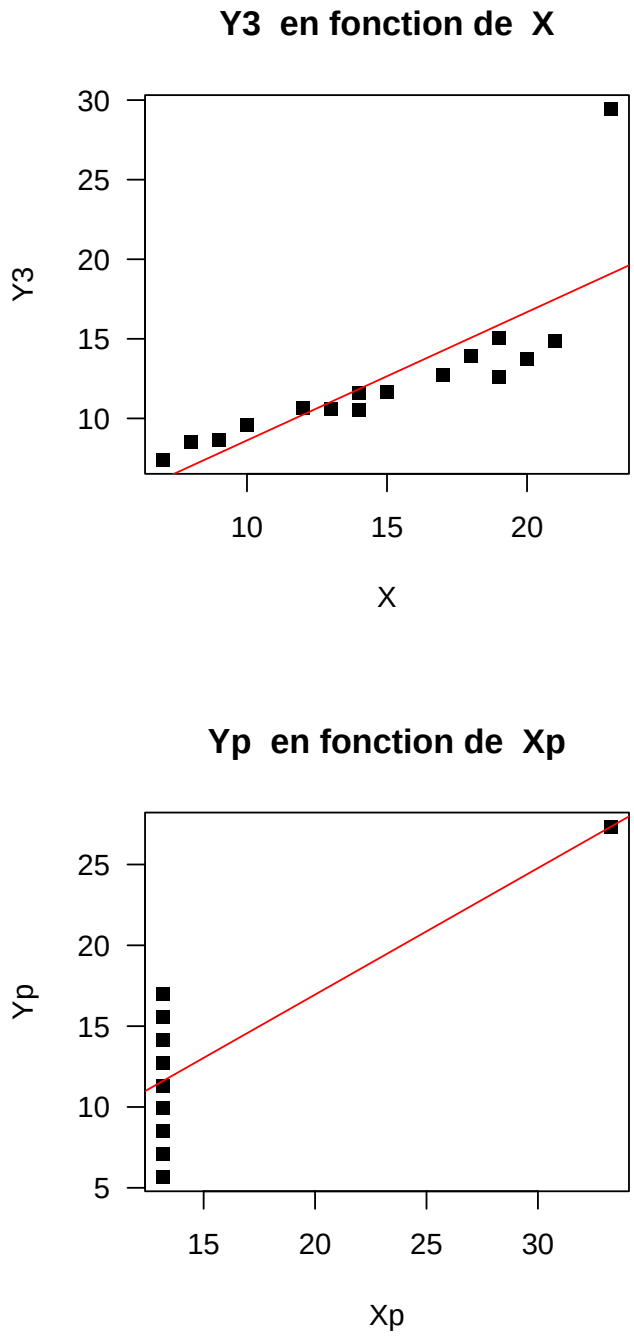


FIG. 2. Les deux derniers nuages de points et les droites de régression linéaire.

[Cha04] Stéphane Champely. *Statistique vraiment appliquée au sport*. de Boeck, 2004. disponible à la BU de Lyon I sous la cote 519.5 CHA.

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

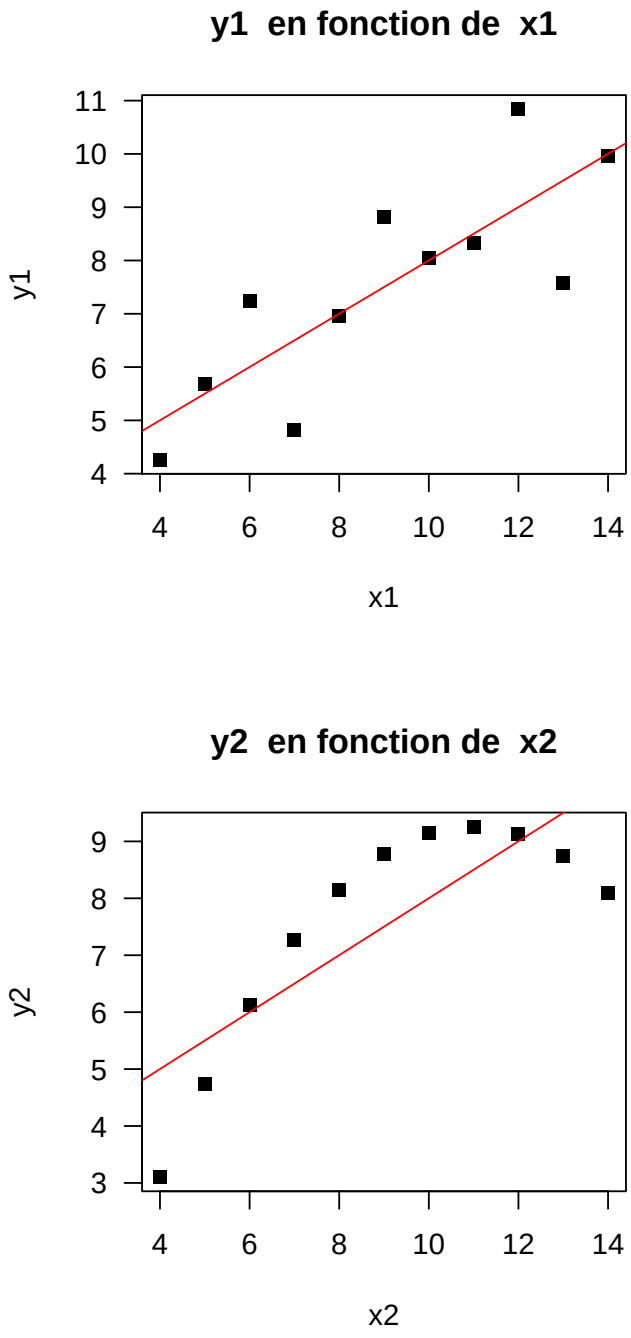


FIG. 3. Les deux premiers nuages de points et les droites de régression linéaire pour les données '`anscombe`'.

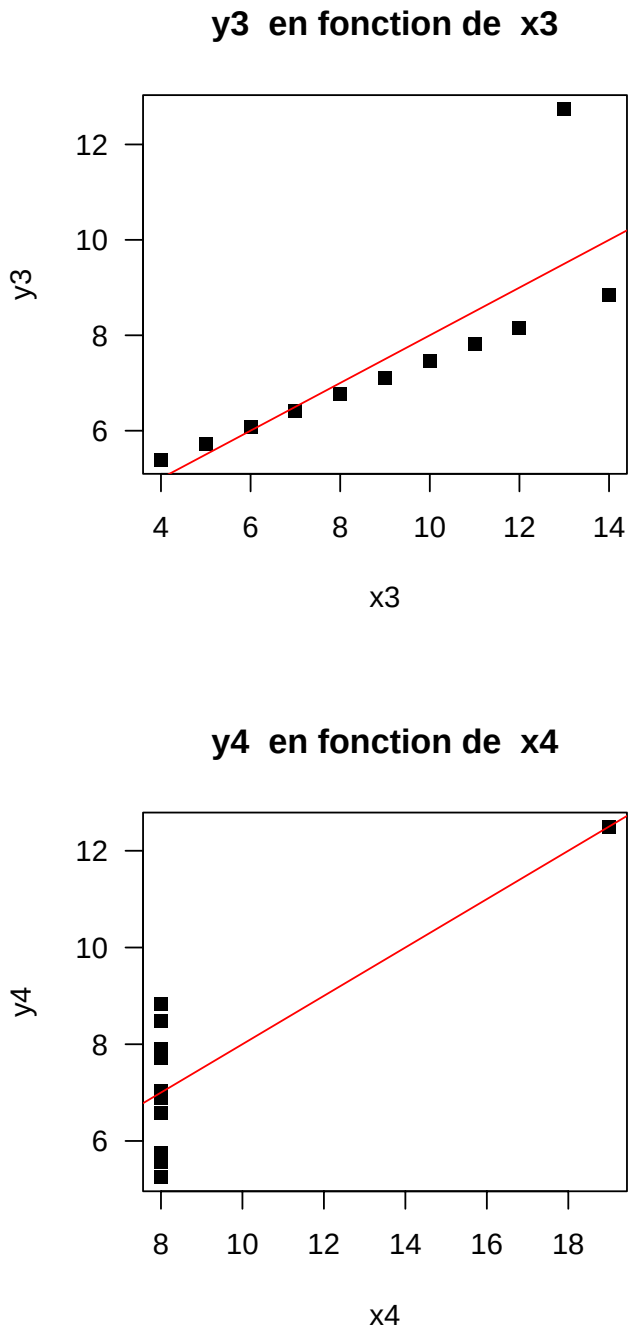


FIG. 4. Les deux derniers nuages de points et les droites de régression linéaire pour les données '`anscombe`'.