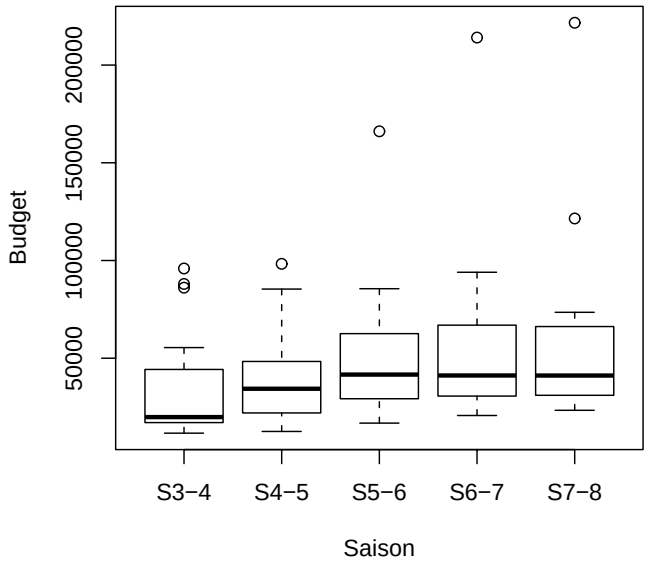
Université Claude Bernard Lyon 1

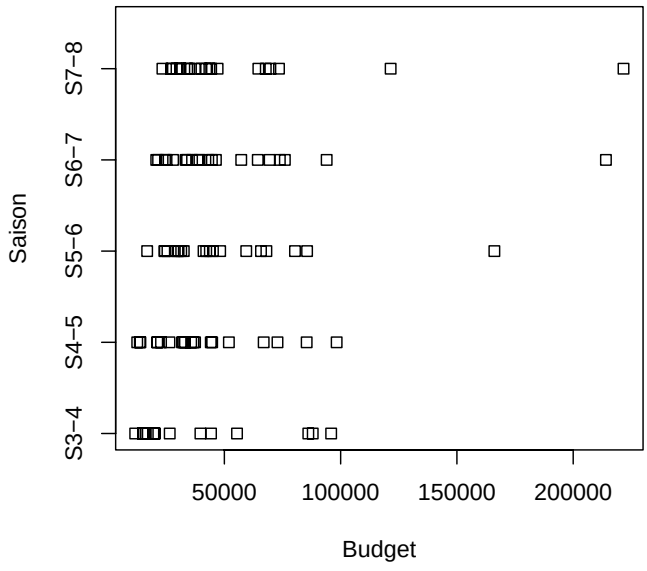
M2IGAPAS (Semestre 1) statistique
Session 1 01 décembre 2014

Corrigé de l'examen CCF2 de statistiques

Correction de l'exercice 1.
Ce sujet provient d'un vieil examen de statistique donné par Stéphane Champely.
On étudie le fichier de données '`VictBudgL1.txt`'.

(1) (a)
- On étudie le croisement de la variable quantitative (ou numérique) '`Budget`' et de la variable qualitative (ou catégorielle) '`Saison`'. Pour les manipulations avec R, on renvoie donc aux sections H.2 et H.3 du document de cours.
-





Voir la figure ci-dessous.

- Avec R, 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
S3-4	35888.94	28376.96	11654.00	17072.00	19885.00	44258.00	95923.00	17
S4-5	40144.40	24032.63	12530.00	22383.00	34385.00	46552.75	98287.00	20
S5-6	49469.70	33598.86	16804.00	29561.25	41627.00	60993.50	166110.00	20
S6-7	54197.65	42847.31	20702.00	31985.50	41174.50	65641.00	214077.00	20
S7-8	55579.00	45422.57	23362.00	31081.25	41145.50	65414.25	221642.00	20

- 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 homogénéité entre les différents budgets en fonction des saisons.
Confirmons cela grâce à R.
Les autres résultats donnés par R sont les suivants :

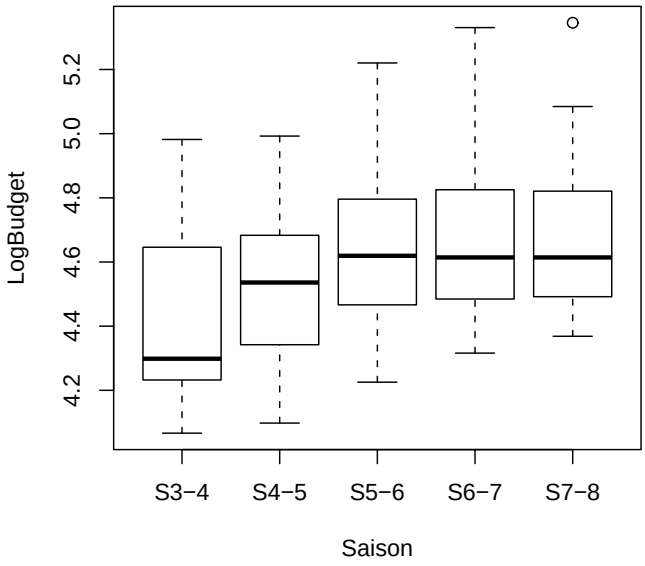
Noms des indicateurs	Valeurs
Rapport de corrélation RC	0.04521
probabilité critique p_c	0.366659

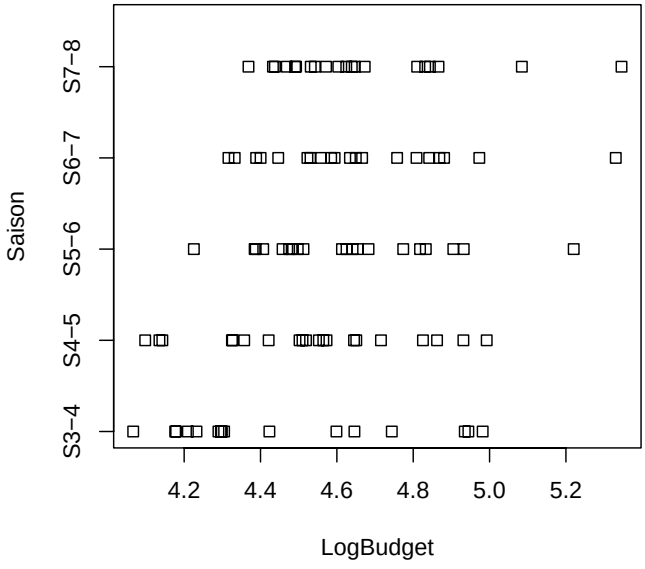
On compare le rapport de corrélation RC=0.04521 aux seuils de Cohen (0.01,0.05,0.15) (voir [Coh92]) et la probabilité critique p_c=0.366659 à la valeur seuil de la probabilité critique 0.05 et on déduit les résultats suivants sur la significativité de la liaison :

significativité pratique	**moyenne**
significativité statistique	**non**

- On peut donc affirmer qu'il existe une relation moyenne entre les variables '`Budget`' et '`Saison`', mais attention au fait que c'est peut-être dû au hasard !

(b)
- On étudie le croisement de la variable quantitative (ou numérique) '`LogBudget`' et de la variable qualitative (ou catégorielle) '`Saison`'.
-





Voir la figure ci-dessous.

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

Les graphiques et les statistiques par groupes montrent une certaine hétérogénéité entre les types.
Confirmons cela grâce à R.

	moyenne	écart-type	0%	25%	50%	75%	100%	n
S3-4	4.4484	0.2987	4.0665	4.2323	4.2985	4.6460	4.9819	17
S4-5	4.5326	0.2572	4.0980	4.3497	4.5360	4.6670	4.9925	20
S5-6	4.6264	0.2357	4.2254	4.4707	4.6193	4.7849	5.2204	20
S6-7	4.6543	0.2485	4.3160	4.5037	4.6142	4.8169	5.3306	20
S7-8	4.6653	0.2407	4.3685	4.4925	4.6142	4.8156	5.3457	20

Les autres résultats donnés par R sont les suivants :

Noms des indicateurs	Valeurs
Rapport de corrélation RC	0.094687
probabilité critique p_c	0.0551509

On compare le rapport de corrélation RC=0.094687 aux seuils de Cohen (0.01,0.05,0.15) (voir [Coh92]) et la probabilité critique p_c=0.0551509 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	**non**

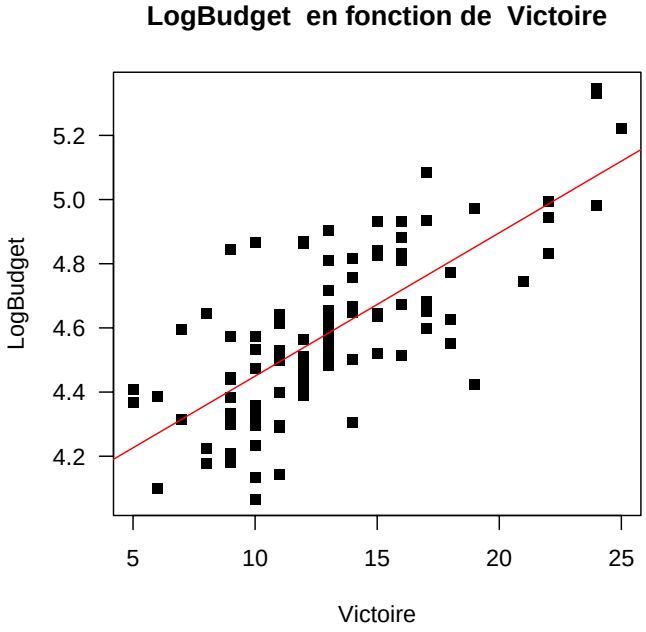
- On peut donc affirmer qu'il existe une relation forte entre les variables '`Budget`' et '`Saison`', mais attention au fait que c'est peut-être dû au hasard ! La probabilité critique (0.0552) étant très proche de 0.05, c'est donc moins sensible que dans le cas précédent.

Si l'on compare les deux valeurs des RC (respectivement 0.0452 et 0.0947), on se rend compte que la liaison est meilleure dans le second cas.

En fait ce qui justifie ce changement de variable, c'est le fait que les données sont normales dans le second cas ; le calcul de la probabilité critique, issu de la théorie des tests et non justifié, a besoin de cette hypothèse fondamentale et donne donc plus de rigueur au second cas ♢

Dans tous les cas, concluons que le bugdet, peut-être à cause du niveau de l'équipe, est variable selon les années.

(2)
- On étudie le croisement de la variable quantitative (ou numérique) '`Victoire`' et de la variable quantitative (ou numérique) '`LogBudget`'. Pour les manipulations avec R, on renvoie donc à la section F.5 du document de cours.
- Voir la figure ci-dessous.



Sur cette figure, les points semblent alignés.

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

Noms des indicateurs	Valeurs
pente a	0.044679
ordonnée à l'origine b	4.002806
corrélation linéaire r	0.732409
probabilité critique p_c	1.52826e-17

On compare la valeur absolue de la corrélation linéaire r =0.732409 aux seuils de Cohen (0.1,0.3,0.5) (voir [Coh92]) et la probabilité critique p_c=1.52826e-17 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 très forte entre les variables '`Victoire`' et '`LogBudget`'; nous sommes rassurés par la significativité statistique.

Correction de l'exercice 2.

(1) (a) $P(X \leq 1) = 0.000399$
(b) $P(X > 2) = 0.994115$
(c) $P(3 \leq X \leq 4) = 0.217631$
(d) $P(X \leq 5) = 0.62285$

(2) (a) $P(X \leq 1) = 0.5$

(b) $P(X > 2) = 0.02275$

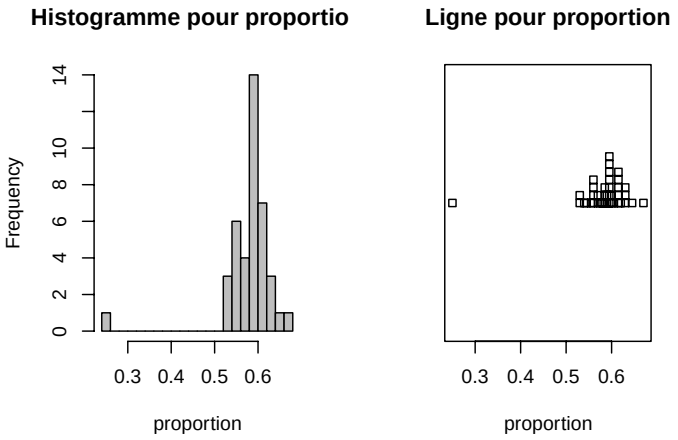
(c) $P(1 \leq X \leq 2) = P(X \leq 2) - P(X \leq 1)) = 0.97725 - 0.841345 = 0.47725$

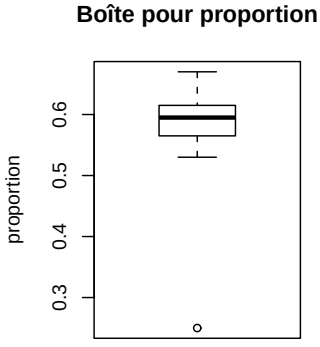
Correction de l'exercice 3.

(1) (a)
- On étudie la variable quantitative (ou numérique) '`proportion`'. Pour les manipulations avec R, on renvoie donc aux sections E.2 et E.3 et aux sections récapitulatives I.1.1 et I.1.4 du document de cours.
- Les différents résultats déterminés par R sont donnés dans le tableau suivant

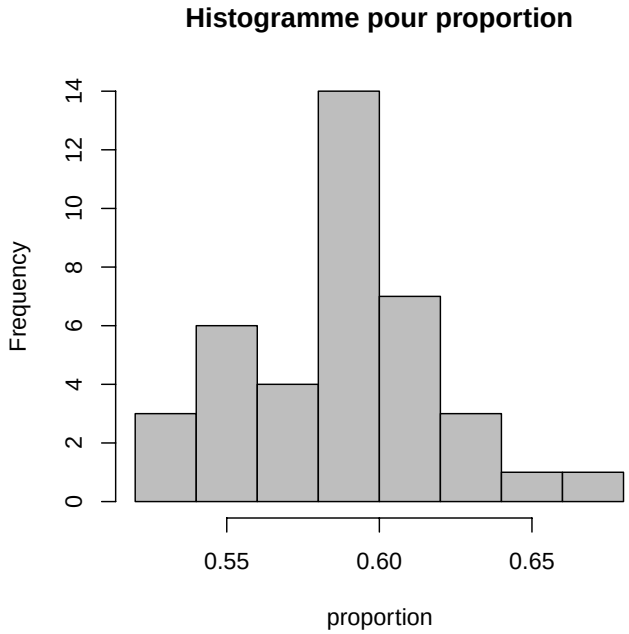
noms	valeurs
moyenne	0.58225
écart-type	0.061975
Q_1 (quartile à 25 %)	0.5675
médiane	0.595
Q_3 (quartile à 75 %)	0.615
minimum	0.25
maximum	0.67
nombre	40

-





Voir les trois graphique ci-dessus pour la variable '`proportion`' (l'histogramme est réalisé avec 20 classes).

- 

Voir l'histogramme si on enlève la valeur minimale (0.25).

(b) Mise à part la valeur minimale, ces statistiques et ces graphiques nous montrent les résultats déjà observés à la section 4.3.1 page 38 du document de cours : excepté la valeur minimale (0.25), qui est une valeur exceptionnellement basse !

- l'histogramme a l'allure d'une densité normale (c'est-à-dire, "en cloche") ;
- la distribution de proportions observées est centrée autour de la véritable valeur du paramètre $p = 0.6$.

On constate aussi en outre que la moyenne est proche de $p = 0.6$; en revanche, l'écart-type estimé ici (0.0619755) n'est pas proche de $\sqrt{p(1-p)/n} = 0.034641$! En fait, si on enlève la valeur minimale, on trouve un écart-type égal à 0.031025, plus proche $\sqrt{p(1-p)/n}$.

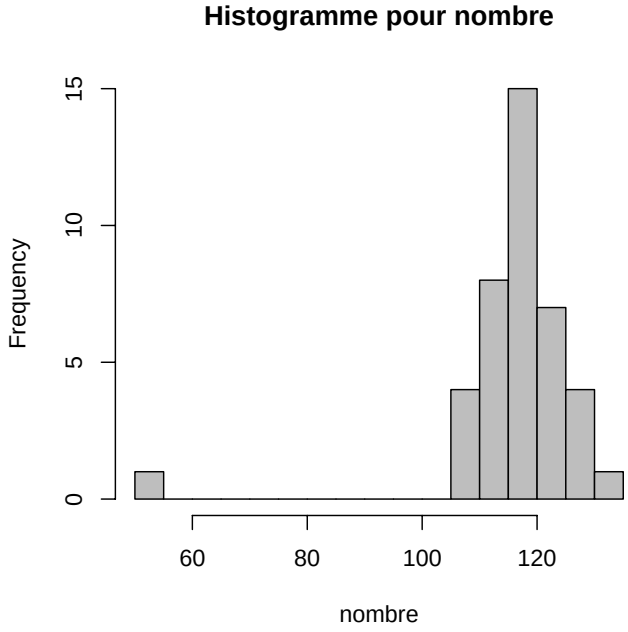
Ainsi, on a constaté expérimentalement que

$$p_r \rightsquigarrow \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Bref, hormis la valeur minimale qui correspond à un jour "sans chance", les résultats observés sont conformes à la théorie !

(2) En raisonnant comme dans le cours (cf page 29), on trouve que le nombre moyen de personnes que l'on peut espérer toucher est l'espérance de la loi binomiale de paramètres n et p, c'est-à-dire $\mathbb{E}(X) = np$, soit 120 personnes.

(3) Ce résultat est un nombre moyen de personnes contactée et ne garantit pas qu'il soit atteint dès la première vague ; cela est confirmé par l'étude de la seconde colonne du fichier de données, qui montre que le nombre de personnes oscille bien autour de 120, mais peut être plus faible !



Voir par exemple le graphique ci-dessus.

Voir aussi la question 2 page 17 de l'exercice 3.33 page 16 du cours.

(4) (a) On écrit que $P(X \geq N) = P(X > N-1)$ et on applique la manipulation avec ® 3.30 page 16 du cours.

(b) On a vu dans la question 2 page 17 de l'exercice 3.33 page 16 du cours. que la probabilité $P(X \geq n) = P(X \geq 120)$ valait 0.5307 et est donc plus faible que 0.95. Il faut donc remplacer n par une valeur plus grande n pour que cette probabilité soit plus grande. "par tatonnement", on trouve par exemple

$$\begin{aligned} &\text{pour } n = 200, \quad P(X \geq 120) = 0.53066, \\ &\text{pour } n = 210, \quad P(X \geq 120) = 0.82025, \\ &\text{pour } n = 230, \quad P(X \geq 120) = 0.99328. \end{aligned}$$

On prend la plus petite valeur possible de n qui garantisse une probabilité supérieure ou égale à NC et par tatonnement on trouve donc

$$n = 220. \tag{1}$$

De même, pour $NC = 0.999$, on trouve

$$n = 239. \tag{2}$$

(c) On suppose que

$$\frac{\frac{X}{n} - p}{\sigma} \rightsquigarrow \mathcal{N}(0,1), \text{ où } \sigma = \sqrt{\frac{p(1-p}{n}}. \tag{3}$$

On écrit

$$P(X \geq N) = NC \Longleftrightarrow P\left(\frac{\frac{X}{n} - p}{\sigma} \geq \frac{\frac{N}{n} - p}{\sigma}\right) = NC$$

où $z = (N/n - p)/\sigma$ suit une loi normale centrée réduite. En suivant ce qu'on a fait dans le cours (manipulation 4.1 page 33), on peut donc déterminer le quantile q tel que

$$P(z \geq q) = NC$$

où z suit une loi centrée réduite. On a donc

$$q = \frac{\frac{N}{n} - p}{\sigma}$$

En remplaçant σ par sa valeur, on constate que c'est successivement équivalent à

$$q = \frac{\frac{N}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \Longleftrightarrow q\sqrt{\frac{p(1-p)}{n}} = \frac{N}{n} - p,$$

$$\Longleftrightarrow q\frac{\sqrt{p(1-p)}}{M} = \frac{N}{M^2} - p,$$

où $M = \sqrt{n}$. C'est donc équivalent à

$$q\frac{\sqrt{p(1-p)}}{M} = \frac{N - pM^2}{M^2},$$

soit encore à

$$q\sqrt{p(1-p)} = \frac{N - pM^2}{M},$$

soit encore à

$$Mq\sqrt{p(1-p)} = N - pM^2$$

soit encore à l'équation du second degré

$$M^2p + q\sqrt{p(1-p)}M - N = 0$$

dont les deux racines sont

$$M = \frac{-q\sqrt{p(1-p)} \pm \sqrt{\Delta}}{2p}, \text{ où } \Delta = q^2p(1-p) + 4Np.$$

On en prendra la positive, puis on calculera $n = M^2$ et on prendra l'entier qui suit. On pourra donc, sous R, taper les commandes suivantes :

```
p<-0.6
N<-120
q<-qnorm(0.95,lower.tail=F)
```

puis

```
delta <- q^2 * (p * (1 - p)) + 4 * N * p
M <- (-q * sqrt(p * (1 - p)) + sqrt(delta))/(2 * p)
M^2
```

On a donc pour $NC = 0.95$, $n = 220$ et pour $NC = 0.999$, $n = 240$.

On retrouve donc bien apeu près (puisqu'il y a approximation) les valeurs données par (1) et (2).

(d) Au vu de ces calculs, on peut affirmer que, en moyenne, si la vague d'appel st faite sur n personnes, le nombre de personnes réellement contacté sera supérieur à $N = 120$ dans $100NC = 95\%$ des cas pour $n = 220$ et sera supérieur à $N = 120$ dans $100NC = 99.9\%$ des cas pour $n = 239$! Bien entendu, il existera toujours des cas défavorable où le nombre de personnes réellement contacté sera inférieur à 120.

Dans le cas, où seules 200 personnes étaient contactées, 120 étaient réellement contactées seulement dans 53.066216% des cas.

Voilà qui a de quoi rassurer les finances de notre directeur !

Remarque 1.

(i) En fait, la réalité ne se passe pas ainsi : on arrête de contacter les personnes dès que le seuil de 120 personnes contactées est atteint. Mais ces calculs peuvent permette de prévoir les financements d'un central téléphonique dans la mesure où l'on connaît le nombre de personnes minimal à contacter.

(ii) Le nombre 120 est pris par défaut ; mais, il existe des lois qui permette de calculer le nombre minimal de personnes à sonder pour avoir une erreur inférieure à un seuil donné ; voir par exemple le tableau 3.3 page 32 de [Cha04].

♢

Correction de l'exercice 4.

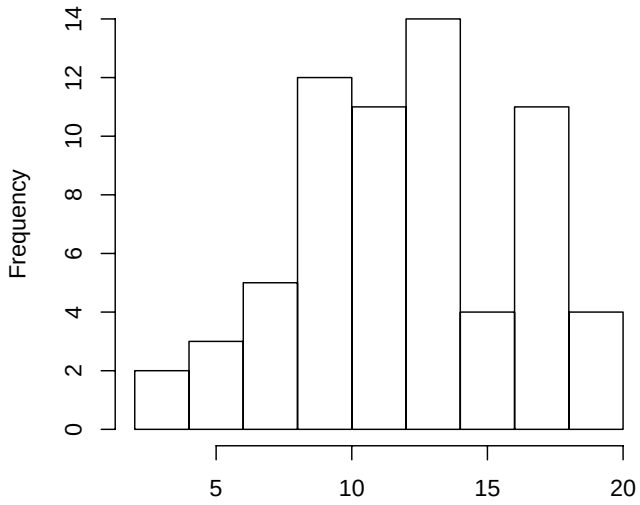
(1) (a) Le nombre total d'étudiants est $n = 66$. On trouve une moyenne égale à 12.45455 et un écart-type égal à 4.03897.

En procédant comme dans le cours (section 4.4.3 page 47), on trouve un intervalle de confiance égal à

$$[11.4616, 13.4474] \tag{4}$$

(b) La moyenne du groupe (12.45455) est supérieure à 10 !

(c)



Si on trace l'histogramme des notes, on obtient la figure ci-dessus, ce qui montre une répartition à peu près normale, excepté peut-être l'avant-dernière colonne.

Par ailleurs, d'après (4), dans 95 % des cas, la moyenne μ de la loi normale dont provient ce groupe (et qui est *inconnue*) est dans l'intervalle $[11.4616, 13.4474]$, intervalle dont toutes les valeurs sont supérieures à 10. Il n'est donc pas absurde de supposer que cette moyenne inconnue est supérieure à 10 (et ce dans 95 % des cas).

(d) La seconde façon de faire, à la base de la théorie des test d'hypothèse, est beaucoup plus précise que la première !

(2) (a) On trouve un intervalle de confiance égal à

$$[10.4211, 11.9789] \tag{5}$$

(b) La moyenne du groupe est supérieure à 10 !

(c) On ne peut tracer l'histogramme ici. On suppose tout de même la normalité des données.
Par ailleurs, d'après (4), dans 95 % des cas, la moyenne μ de la loi normale dont provient ce groupe (et qui est *inconnue*) est dans l'intervalle $[10.4211, 11.9789]$, intervalle dont toutes les valeurs sont supérieures à 10. Il n'est donc pas absurde de supposer que cette moyenne inconnue est supérieure à 10 (et ce dans 95 % des cas).

(d) La seconde façon de faire, à la base de la théorie des test d'hypothèse, est beaucoup plus précise que la première !

Références

[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.