Feuille de Travaux Dirigés n$^{\text{o}}$ 10
Problèmes à deux échantillons

Exercice X.1. Les mètres ou les pieds ?
Le professeur T.Lewis a réalisé une expérience en Australie dans les années 70 peu après l'introduction officielle du système métrique. Tous les étudiants d'un groupe de 44 ont dû estimer la largeur, en mètres, de l'amphithéâtre dans lequel ils étaient assis. On a demandé à un autre groupe de 69 étudiants de procéder à l'estimation de la largeur du même amphithéâtre mais cette fois-ci en pieds. L'objectif de cette expérience était de déterminer si le fait d'estimer des mesures en pieds et en mètres donnait des résultats différents.

1. Télécharger le fichier « roomwidth » du package « HSAUR »
2. Afficher-le à l'écran.
3. Combien de lignes ? Combien de colonnes ? Quel est le type de chaque colonne ?
4. Une des colonnes est un « factor ». Quel(s) type(s) de renseignements devez-vous avoir lorsque vous traitez un facteur en statistique ?
5. Quel est le nombre de niveaux de ce facteur ? Comment obtenez-vous les différents noms des différents niveaux de ce facteur ?
6. Nous souhaiterions tester l'hypothèse que la moyenne de la population des largeurs estimées en mètres, puis converties en pieds est égale à la moyenne de la population des largeurs estimées en pieds. Quel type de test proposez-vous ?
7. Il y a un problème à identifier. Quel est ce problème ?
8. Exécuter la ligne de commande suivante :

   $$> convert <- ifelse(roomwidth\$unit == "feet", 1, 3.28)$$

   Que fait-elle ?
9. Exécuter les lignes de commande suivantes :

   $$> tapply(roomwidth\$width * convert, roomwidth\$unit, summary)$$

   et

   $$> tapply(roomwidth\$width * convert, roomwidth\$unit, sd)$$

   Que font-elles ? Réessayer en remplaçant *tapply* par *sapply*. Qu'observez-vous ?
10. À cette étape, il est suggéré de faire des représentations graphiques. Quel(s) type(s) de représentation envisagez-vous ?

11. Que font les lignes de commande suivantes ? (Recommandation : taper les unes à la suite des autres puis comparer le résultat obtenu avec le Graphique 1 page 3.)

$> \ layout(matrix(c(1, 2, 1, 3), nrow = 2, ncol = 2, byrow = F))$
$> \ boxplot(I(width * convert) \sim unit, data = roomwidth,$
$+ \ ylab = "Estimatedwidth(feet)", var.width = T,$
$+ \ names = c($"Estimates in feet", "Estimates in metres (converted to feet)"$))$
$> \ feet <- roomwidth\$unit == "feet"$
$> \ qqnorm(roomwidth\$width[feet], ylab = "Estimatedwidth(feet)")$
$> \ qqline(roomwidth\$width[feet])$
$> \ qqnorm(roomwidth\$width[!feet], ylab = "Estimatedwidth(metres)")$
$> \ qqline(roomwidth\$width[!feet])$

Remarque : Dans la seconde ligne, nous avons écrit $I(width * convert) \sim unit$, peut-être aviez-vous eu l'idée de n'écrire que $width * convert \sim unit$. Nous verrons que dans la définition d'une formule associée à un modèle statistique, le symbole $*$ a un sens particulier. Lorsque l'on souhaite se servir du symbole $*$ pour indiquer une multiplication entre deux termes d'un modèle, il est conseillé, voire nécessaire dans certains cas, d'utiliser la fonction $I()$ qui indique à `R` de conserver au symbole $*$ son sens de multiplication.

12. À cette étape, est-ce que vous envisagez encore un test de Student ? Si oui lequel ?
13. Réaliser un test de Shapiro-Wilk sur les deux groupes d'estimations de largeur.

$$> shapiro.test(roomwidth\$width[feet])$$

et

$$> shapiro.test(roomwidth\$width[!feet])$$

Pourquoi dans cette situation suggérera-t-on un test non-paramètrique ?

14. Quel est l'équivalent non paramètrique du test de Student ?
15. Que fait la ligne de commande ?

$$> wilcox.test(I(width * convert) \sim unit, data = roomwidth, conf.int = T)$$

16. Lorsque l'on exécute un test de Wilcoxon, quelle est la différence principale avec un test de Student qu'il faut noter ?

Remarque : Par souci de complétude, nous indiquons les lignes de commande pour exécuter les tests de Student sur cet échantillon bien qu'il ne semble pas approprié de les utiliser :

$$> t.test(I(width * convert) \sim unit, data = roomwidth, var.equal = T)$$

ou encore

$$> t.test(I(width * convert) \sim unit, data = roomwidth, var.equal = F)$$

Qu'observez-vous ? Qu'appelle-t-on un « Two Sample » ? Que faites-vous comme différence entre un « Two Sample t-test » et un « Welch Two Sample t-test » ?

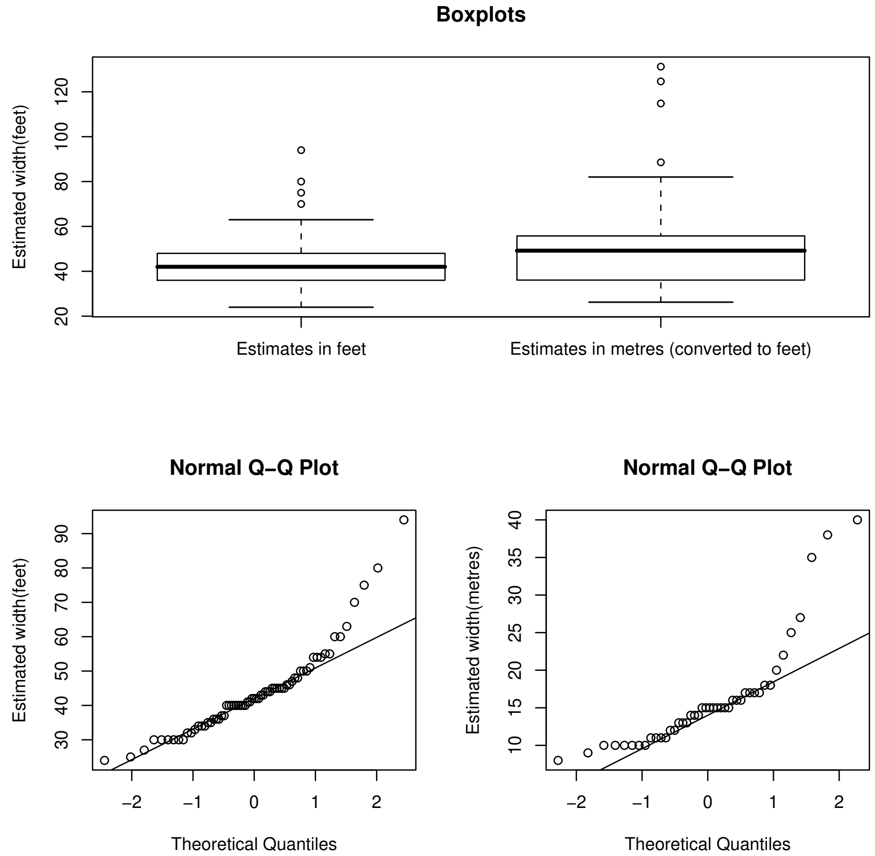

FIG. 1. Graphique 1

Exercice X.2. Comparaison de deux méthodes d'amarrage.
Dans une expérience contrôlée réalisée pour la construction d'une usine marémotrice, l'influence de deux méthodes d'amarrage sur la tension résultant dans les cables d'amarrage a été étudiée. Une vaste gamme d'états différents de l'océan a été reproduite à l'identique avec ces deux méthodes d'amarrage et la tension mesurée reproduites dans le jeu de données « waves ».

1. Télécharger le fichier « waves » du package « HSAUR ».
2. Afficher-le à l'écran.
3. Combien de lignes ? Combien de colonnes ? Quel est le type de chaque colonne ?
4. Exécuter la ligne de commande suivante

$$> mooringdiff <- waves\$method1 - waves\$method2$$

5. Sur la même image,
   (i) dessiner une boxplot de la variable « mooringdiff »,
   (ii) mettre une légende « Difference (Newton metres) » sur l'axe des ordonnées,

(iii) mettre un titre « Boxplot »,

(iv) tracer une ligne horizontale d'équation $y = 0$ dans cette représentation « Boxplot », à l'aide de la commande suivante :

$$> abline(h = 0, lty = 2)$$

(v) tracer le QQ-plot de la variable « mooringdiff » à côté de la « Boxplot »,

(vi) mettre une légende « Difference (Newton metres) » sur l'axe des ordonnées,

(vii) tracer la QQ-line dans cette même représentation,

(viii) comparer votre graphique avec le graphique 2 page 5.

6. Quel test envisagez-vous ? Pourquoi ?

$$> shapiro.test(mooringdiff)$$

7. Exécuter la ligne de commande :

$$> t.test(mooringdiff).$$

Que fait-elle ? Quelle est la différence avec le test de Student de l'exercice 1 ?

8. Quel est son équivalent en non paramétrique ?
9. Exécuter la ligne de commande suivante :

$$> wilcox.test(mooringdiff)$$

Que fait-elle ? Qu'observez-vous ? Il y a quelque chose de nouveau par rapport aux autres tests...

Remarque : Si nous avions voulu uniquement procédé aux tests de Student ou de Wilcoxon sur ces deux échantillons appariés, il aurait suffi de rajouter l'option « paired=T » dans la ligne de commande concernée :

$$> t.test(waves\$method1, waves\$method2, paired = T).$$

$$> wilcox.test(waves\$method1, waves\$method2, paired = T)$$

Exercice X.3. Engrais.

On traite 3 échantillons de plantes avec 3 engrais différents. On obtient les résultats suivants :

	Engrais A	Engrais B	Engrais C
Ont fleuri	40	75	63
N'ont pas fleuri	15	12	12

1. Introduire les données dans `R`.

$> \ flor <- matrix(c(40, 75, 63, 15, 12, 12), nrow = 2, byrow = T)$
$> \ rownames(flor) < -c("Feuri", "Pasfleuri")$
$> \ colnames(flor) < -c("EngraisA", "EngraisB", "EngraisC")$
$> \ flor <- as.table(flor)$

2. Afficher-les à l'écran.
3. Combien de lignes ? Combien de colonnes ?

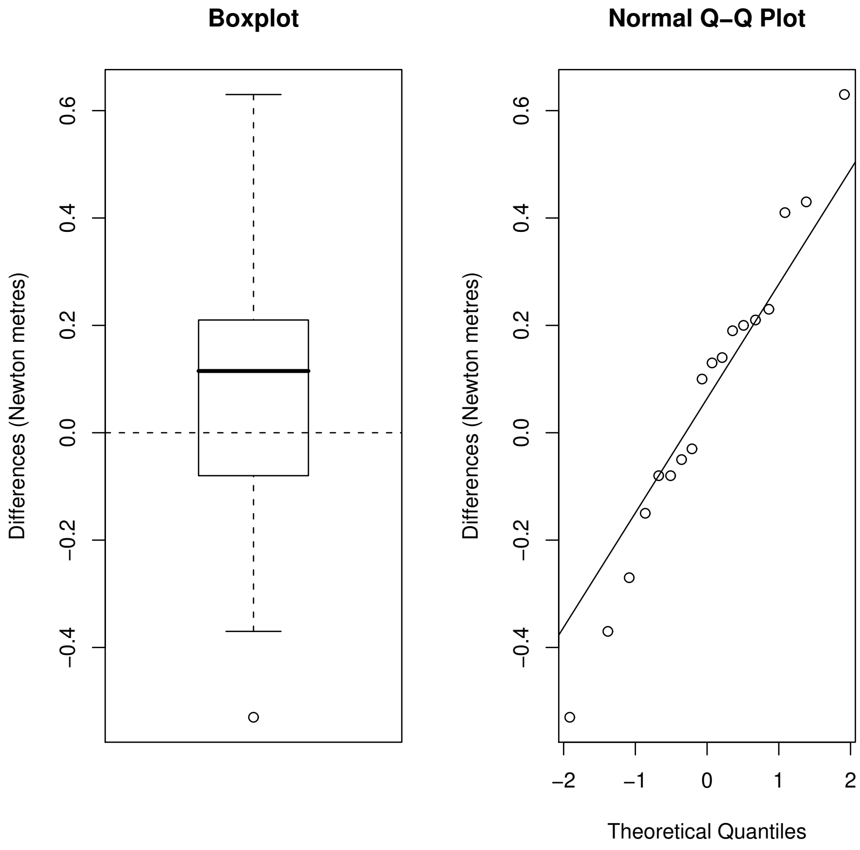

FIG. 2. Graphique 2

4. On souhaite savoir s'il y a une dépendance entre le type d'engrais et la présence ou l'absence de floraison. Quel test envisagez-vous ? Pourquoi ?
5. Exécuter la ligne de commande suivante :

$$> chisq.test(flor)$$

Que fait-elle ? Les conditions d'utilisation du test sont-elles vérifiées ?

$$> chisq.test(flor)\$expected$$

Il est également possible de calculer une $p-$valeur exacte pour le test à l'aide de techniques de Monte-Carlo. Celle-ci est particulièrement utile lorsque les conditions d'utilisation de l'approximation asymptotique de la $p-$valeur du test ne sont pas remplies.

$$> chisq.test(flor, simulate.p.value = T, B = 100000)$$

6. Calculer les résidus associés à ce modèle. Pour cela, exécuter la ligne de commande suivante :

$$> chisq.test(flor)\$residuals$$

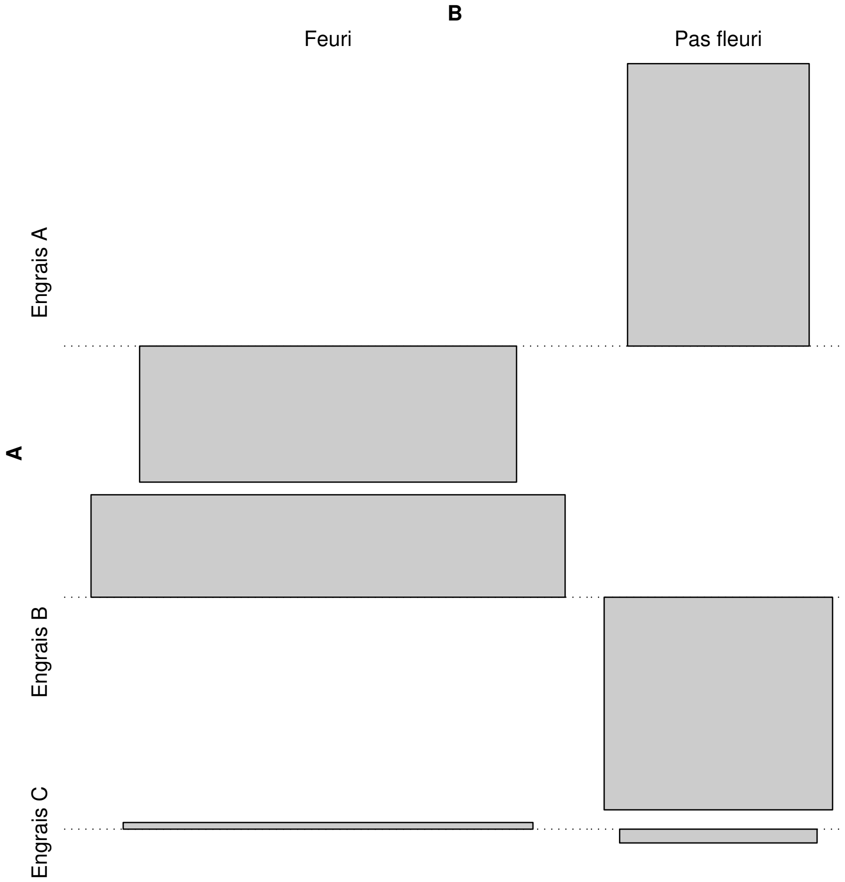

FIG. 3. Graphique 3

7. Une représentation graphique de ces résidus est appelée « association plot ». Pour cela, il faut charger en mémoire, et donc éventuellement télécharger, le package « vcd » (Meyer et al., 2005) pour accéder à la fonction « assoc ». Ensuite, exécuter la ligne de commande suivante :

$$> assoc(flor)$$

8. Le test du χ^2 suppose que les données récoltées sont indépendantes. Connaissez-vous son analogue pour un échantillon « paired » ?

Exercice X.4. Délinquance.
L'échantillon suivant est le résultat d'une expérience réalisée en Floride en 1987. Il s'agit de jeunes délinquants qui ont été associés en paire suivant des critères comme l'âge ou le nombre de leurs délits. L'un des deux membres de la paire était ensuite envoyé devant le tribunal des mineurs alors que le second était jugé par un tribunal pour adultes. On a reporté dans le tableau « rearrests » si le jeune délinquant a été à nouveau arrêté avant la fin de l'année 1988.

1. Télécharger le fichier « rearrests » du package « HSAUR ».
2. Afficher-le à l'écran.

3. Combien de lignes ? Combien de colonnes ?
4. On souhaite savoir s'il y a une dépendance entre le type de tribunal ayant jugé le mineur et sa récidive avant fin 1988. Quel test envisagez-vous ? Pourquoi ?
5. Exécuter la ligne de commande suivante :

   $$> mcnemar.test(rearrests, correct = F)$$

   Que fait-elle ? Que signifie « correct » ?
6. Il existe une version exacte de ce test. Savez-vous ce que signifie l'expression « version exacte de ce test » ? Exécuter la ligne de commande suivante :

   $$> binom.test(rearrests[2], n = sum(rearrests[c(2, 3)]))$$

   Que fait-elle ?

. . . .