Université Claude Bernard Lyon 1

M1PPMR (Semestre 1) statistique/gestion
Session 1 14 Décembre 2009

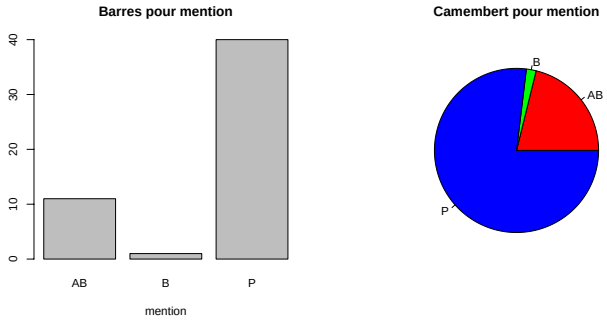
Corrigé de l'examen CCF2 de statistiques

Correction de l'exercice 1.

(1)
- On étudie la variable qualitative (ou catégorielle) '`mention`'. Pour les manipulations avec R, on renvoie donc aux sections 2.3 et 2.4 du document de cours.
- Les effectifs et les pourcentages déterminés par R sont donnés dans le tableau suivant

	effectifs	pourcentages
B	1	1.923
AB	11	21.154
P	40	76.923

-



Voir les deux (un seul d'entre eux suffit) trois graphiques ci-dessus pour la variable '`mention`'.

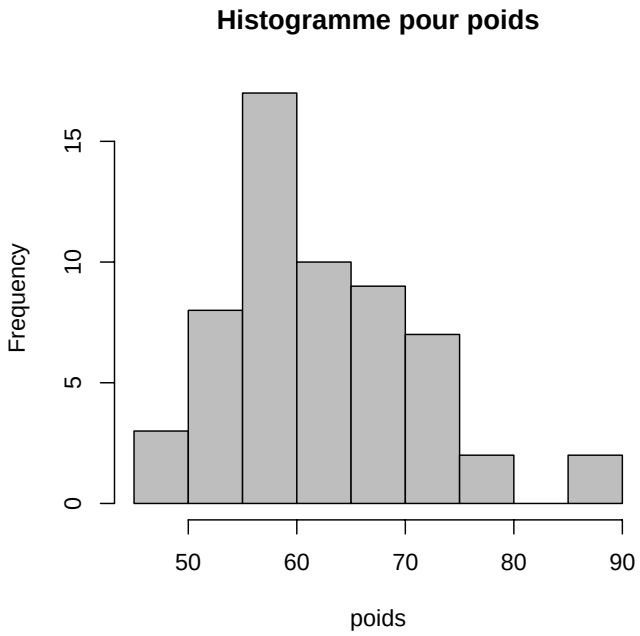
(2) Les graphiques et les pourcentages nous montrent que
- la mention la plus représentée est la mention '`P`' (76.923% des effectifs),
- puis vient la mention '`AB`' (21.154% des effectifs),
- et enfin, marginale, vient la mention '`B`' (1.923% des effectifs).

Correction de l'exercice 2.

(1)
- On étudie la variable quantitative (ou numérique) '`poids`'. Pour les manipulations avec R, on renvoie donc aux sections 3.2, 3.3 et 3.4 du document de cours.
- Les différents résultats déterminés par R sont donnés dans le tableau suivant

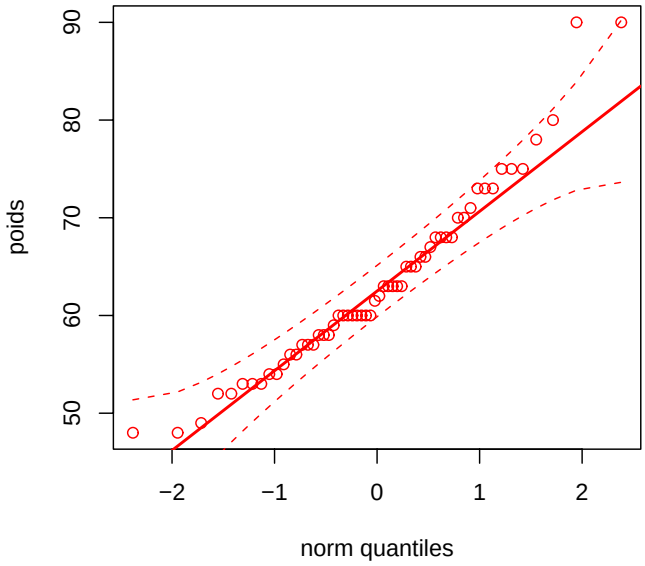
noms	valeurs
moyenne	63.181034
sd	9.173139
Q_1 (quartile à 25 %)	57
médiane	61.75
Q_3 (quartile à 75 %)	68
minimum	48
maximum	90
nombre	58

-



Voir l'histogramme ci-dessus pour la variable '`poids`'.

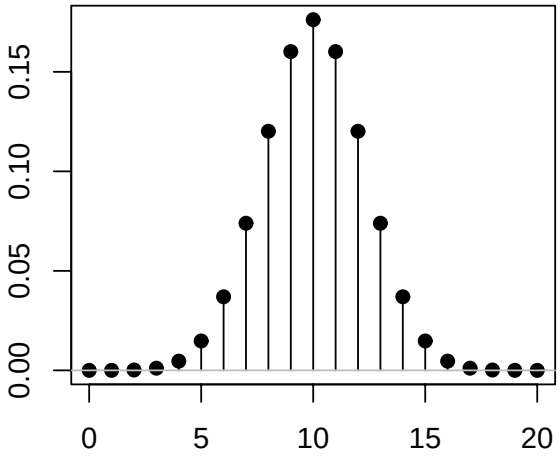
(2) On peut voir sur ce graphique que la distribution est clairement symétrique et d'allure normale.



Ceci peut être confirmé par d'autres graphiques (graphe quantile-quantile en particulier, voir figure ci-dessus). De plus, il semble que le poids typique soit de l'ordre de 60 ce qui est confirmé par la valeur de la médiane : $Q_2 = 61.75$. Les poids s'étendent de 48 à 90.

Correction de l'exercice 3.
Ce sujet avait été donné en Janvier 2008 (session CT) par Stéphane Champely aux M1PPMR.

(1)
- Le nombre de réponses justes données par le sourcier est de 12 ce qui est proche de la moitié du nombre de containers (10).
  Ce résultat est à comparer avec celui qu'il aurait obtenu "par hasard". Si le sourcier n'a pas de compétences, on peut faire l'hypothèse que la probabilité de succès sur chaque container est de $\pi = 0.5$. Un nombre $n = 20$ d'essais sont tentés et on fait donc l'hypothèse que le nombre de réponses justes suit une loi binomiale.



On peut tracer la loi de probabilité comme sur la figure ci-dessus. On renvoie aux manipulations avec Rcmdr 5.33 page 31 du cours.
La question que l'on se pose est "est-ce que le sorcier fait mieux que le hasard". Déterminons la probabilité que le score du sorcier soit inférieur celui du hasard, c'est-à-dire $P(X \geq 12)$ où X suit une loi binomiale de paramètres $\pi = 0.5$ et $n = 20$. On renvoie aux manipulations avec Rcmdr 5.46 page 34 du cours. On écrit $P(X \geq 12) = P(X > 11)$, pour passer par l'aire à droite avec Rcmdr, ce qui fournit 0.25172. On obtient donc

$$p = 0.25172. \tag{1}$$

Cette probabilité est supérieure au seuil usuel de $\alpha = 0.05$. Si elle avait été inférieure, on aurait pu la considérer comme suffisamment proche de zéro pour considérer le score du sourcier comme très peu probable, et donc non dû au hasard ! Donc, ici, on rejette au contraire le fait que le score du sourcier est non dû au hasard et on conclue donc que *le sourcier n'a pas de compétences !*.

- On peut aussi, en utilisant les manipulations avec Rcmdr 5.46 page 34 du cours, calculer les probabilités cumulées $P(X \geq k)$ pour chaque valeur de $k \in \{0, \ldots, 20\}$:
  On obtient le tableau 1 page suivante.
  Dans ce tableau, on voit deux sous-ensembles :
  • l'ensemble $\{0, ..., 14\}$, où la probabilité cumulée $P(X \geq k)$ est strictement supérieure à 0.05.
  • l'ensemble $\{15, \ldots, 20\}$, où la probabilité cumulée $P(X \geq k)$ est inférieure ou égale à 0.05.
  Comme on vient de faire, dans la première région, appelée R_c, on accepte le fait que le résultat du sourcier est dû au hasard et dans la seconde, on accepte qu'il ait des compétences.
  Bref, *au seuil* 0.05*, on accepte les compétences du sourcier pour un score supérieur ou égal à à 15.*
  Graphiquement, on peut tracer le graphe des probabilités cumulées (avec l'aire à droite) comme dans la figure 1 page ci-contre à gauche (voir manipulations avec Rcdmr 5.42 page 33 du cours). Sur ce graphe, on a rajouté en rouge la droite d'ordonnée 0.05 et en bleu pointillé la droite d'abscisse 15.

k	probabilités cumulées
0	1.00000000
1	0.99999905
2	0.99997997
3	0.99979877
4	0.99871159
5	0.99409103
6	0.97930527
7	0.94234085
8	0.86841202
9	0.74827766
10	0.58809853
11	0.41190147
12	0.25172234
13	0.13158798
14	0.05765915
15	0.02069473
16	0.00590897
17	0.00128841
18	0.00020123
19	0.00002003
20	0.00000095

TAB. 1. Les différentes probabilités cumulées

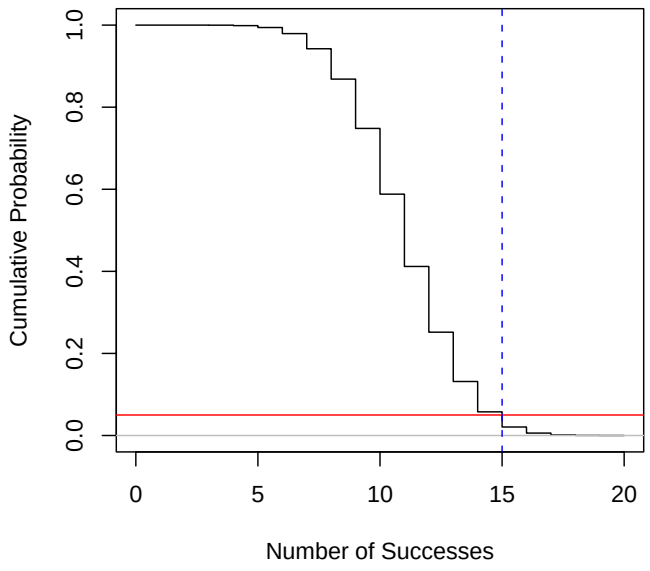


FIG. 1. Le graphes des probabilités cumulées $P(X \geq k)$ pour $n = 20$ et $\pi = 0.5$.

Voir aussi la figure 2 page suivante où on a représenté le graphe des probabilités simples avec la région du "vrai don du sourcier" en rouge. Sur ce graphe, on a rajouté en rouge la droite la droite d'abscisse 15.

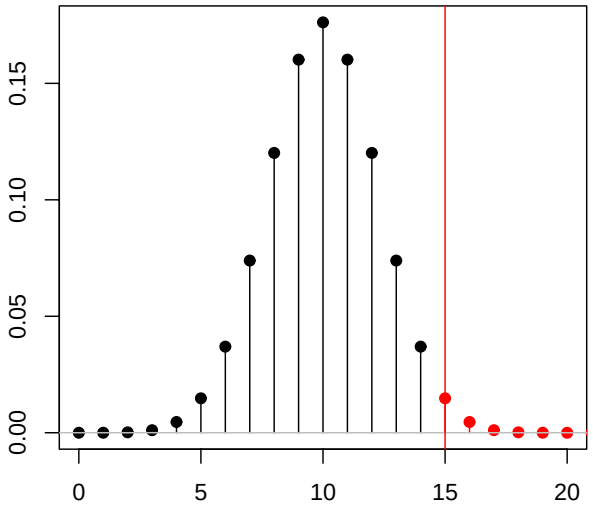


FIG. 2. Le graphes des probabilités simples avec la région du "vrai don du sourcier" en rouge.

Remarque 1. On pourra consulter la partie statistique de l'observatoire de zététique `http://www.zetetique.fr/stats/` plus particulièrement faire tourner le cas correspondant à celui que l'on vient de traiter grâce à `http://www.zetetique.fr/stats/stats.php?cas=11`.

- Une autre façon de procéder est la suivante :
  On fait un test d'hypothèse en proportion, non vu en toute rigueur en cours. Cela n'était pas traité explicitement dans le cours, mais en appliquant les idées des chapitres 5 et 6 du document de cours avec un peu d'astuce, on pouvait s'en sortir ! On renvoie aussi aux pages 203 et 204 (définition 12.5) de l'ouvrage de S. Champely [Cha04].
  Deux façon de procéder :

  (a) "à la main" : La proportion observée de succés est $pr = 12/20 = 0.6$.

  On choisit un risque $\alpha = 0.05$. Nous allons adapter la définition 6.33 dans le document de cours à une variable aléatoire qui n'est pas normale mais qui suit une loi binomiale de paramètres $\pi_0 = 0.5$ et $n = 20$. Comme dans la section 5.7 du document de cours, on approchera la loi suivie par la proportion par une loi normale (puisque n est "grand").

  On fait donc l'hypothèse H_0 que la proportion de succés (observée) pr suit une loi normale de moyenne $\mu = \pi_0$ et d'écart-type $\sqrt{\pi_0(1-\pi_0)/n}$. Ainsi la statistique

  $$z = \frac{pr - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

  suit une loi normale centrée réduite. Le sourcier a de réelles compétence si $\pi > \pi_0 = 0.5$. l'hypothèse alternative choisie est donc $H_1 : \pi > \pi_0$. Comme dans la définition 6.33 dans le document de cours (ou définition 12.5 page 204 de [Cha04]), on calcule la probabilité que $P(Z > z)$, où Z suit une loi normale centrée réduite.

Sous R, on obtient successivement

$$z = \frac{pr - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.6 - 0.5}{\sqrt{\frac{0.5\times(1-0.5)}{20}}} = 0.894427,$$

puis

$$p_c = 0.18555. \tag{2}$$

Cette valeur est légèrement différente de la valeur donnée par (1). Enfin, comme dans la définition 6.33 dans le document de cours, on constate que $p_c \geq \alpha = 0.05$; ansi, on acceptera H_0 donc la proportion observée est, au risque 0.05, égale à 0.5. Le sourcier n'a donc pas de réelles compétences ! !

(b) "avec R" On pourra utiliser directement la commande donnée en examen :

```
prop.test(x=12,n=20,p=0.5,alternative="greater",correct=F,conf.level=0.95)
```

adaptée à l'hypothèse alternative $H_1 : \pi > \pi_0$, qui donne

```
[1] 0.1855467
```

à comparer à $p_c = 0.185546684761349.$ donné par (2) !

Remarque 2. Si on tape

```
prop.test(x=12,n=20,p=0.5,alternative="greater",correct=T,conf.level=0.95)
```

ou directement

```
prop.test(x=12,n=20,p=0.75,alternative="greater",conf.level=0.95)
```

on obtient une valeur légèrement différente de la probabilité critique

$$p_c = 0.251167.$$

Cette valeur est presque identique à la valeur donnée par (1). Pour ce calcul, puisque l'on a choisit une valeur de '`correct`' égale à `T`, le calcul est exact, l'approximation par la loi normale (valable pour les grandes valeurs de n) non utilisée.

(2) (a) Pour les manipulations avec R, on renvoie à la manipulation avec Rcmdr 5.32 dans document de cours. Les paramètres de la loi binomiale sont $n = 20$ et $\pi = 0.5$. La probabilité pour le sourcier d'avoir exactement 10 réponses justes est égale à 0.1762.

(b) Pour calculer la probabilité d'avoir plus de 13 réponses justes, on écrit que cette probabilité vaut $P(X \geq 13) = P(X > 12)$, pour passer par l'aire à droite avec Rcmdr, ce qui fournit 0.13159.

(3) (a)
- Les paramètres de la loi binomiale sont $n = 20$ et $\pi = 0.75$. La probabilité d'avoir plus de 13 réponses justes est maintenant égale à 0.89819 (supérieure à 0.13159).
- Si la probabilité de succés est effectivement de $\pi = 0.75$, la probabilité d'avoir plus de 13 réponses justes est égale à $P(X \geq 13) = P(X > 12)$, soit 0.89819. Ainsi, si $\pi = 0.75$, dans 89.8 % des cas, le sourcier obtiendra plus de 13 réponses justes ! Cependant, le nombre 13 ne prouve pas *a posteriori* que $\pi = 0.75$.
- On fait un test d'hypothèse en proportion, non vu en toute rigueur en cours. Comme précédemment, cela n'était pas traité explicitement dans le cours, mais en appliquant les idées des chapitres 5 et 6 du document de cours avec un peu d'astuce, on pouvait s'en sortir ! On renvoie aussi aux pages 203 et 204 (définition 12.5) de l'ouvrage de S. Champely [Cha04].
  Deux façon de procéder :

(i) "à la main" : La proportion observée de succés est $pr = 13/20 = 0.65$.
On choisit un risque $\alpha = 0.05$. Nous allons adapter la définition 6.33 dans le document de cours à une variable aléatoire qui n'est pas normale mais qui suit une loi binomiale de paramètres $\pi_0 = 0.75$ et $n = 20$. Comme dans la section 5.7 du document de cours, on approchera la loi suivie par la proportion par une loi normale (puisque n est "grand").
On fait donc l'hypothèse H_0 que la proportion de succés (observée) pr suit une loi normale de moyenne $\mu = \pi_0$ et d'écart-type $\sqrt{\pi_0(1-\pi_0)/n}$. Ainsi la statistique

$$z = \frac{pr - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

suit une loi normale centrée réduite. On n'a, *a priori*, pas confiance dans le sourcier et on va montrer qu'il ment et donc que la valeur π du paramètre est strictement inférieure à $\pi_0 = 0.75$; l'hypothèse alternative choisie est donc $H_1 : \pi < \pi_0$. Comme dans la définition 6.33 dans le document de cours (ou définition 12.5 page 204 de [Cha04]), on calcule la probabilité que $P(Z < z)$, où Z suit une loi normale centrée réduite.
Sous R, on obtient successivement

$$z = \frac{pr - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.65 - 0.75}{\sqrt{\frac{0.75\times(1-0.75)}{20}}} = -1.032796,$$

puis

$$p_c = 0.15085. \tag{3}$$

Enfin, comme dans la définition 6.33 dans le document de cours, on constate que $p_c \geq \alpha = 0.05$; ansi, on acceptera H_0 donc la proportion observée est, au risque 0.05, égale à 0.75. Le sourcier a donc de réelles compétences ! !

(ii) "avec R" On pourra utiliser directement la commande donnée en examen :

```
prop.test(x=13,n=20,p=0.75,alternative="less",correct=F,conf.level=0.95)
```

adaptée à l'hypothèse alternative $H_1 : \pi < \pi_0$, qui donne

```
[1] 0.1508498
```

à comparer à $p_c = 0.150849791239174$. donné par (3) !

Remarque 3. Si on tape

```
prop.test(x=13,n=20,p=0.75,alternative="greater",correct=T,conf.level=0.95)
```

ou directement

```
prop.test(x=13,n=20,p=0.75,alternative="greater",conf.level=0.95)
```

on obtient une valeur légèrement différente de la probabilité critique

$$p_c = 0.219289.$$

Pour ce calcul, puisque l'on a choisit une valeur de '`correct`' égale à `T`, le calcul est exact, l'approximation par la loi normale (valable pour les grandes valeurs de n) non utilisée.

Remarque 4. On vient de montrer avec la théorie des tests d'hypthèse en proportion que la valeur de $n = 20$ permettait d'accepter que $\pi = 0.75$. Mais la question était plus précise : "Prendre comme critère de décision que le sourcier ait des compétences si on enregistre plus de 13 succès vous semble-t-il raisonnable". Autrement dit, on cherche la valeur minimale de n qui permette d'accepter que $\pi = 0.75$!
Ici, on raisonne en fait en terme de région critique (voir section 6.5.2.2).
Voir le tableau 2 page suivante. Dans ce tableau, on voit deux sous-ensembles :

k	probabilités critiques
0	0.00000000
1	0.00000000
2	0.00000000
3	0.00000000
4	0.00000001
5	0.00000012
6	0.00000168
7	0.00001805
8	0.00015030
9	0.00097289
10	0.00491164
11	0.01943355
12	0.06066763
13	0.15084979
14	0.30278831
15	0.50000000
16	0.69721169
17	0.84915021
18	0.93933237
19	0.98056645
20	0.99508836

TAB. 2. Les différentes probabilités critiques en fonction du nombre de réponses justes

- l'ensemble $\{0, ..., 11\}$, où la probabilité critique est inférieure ou égale à 0.05.
- l'ensemble $\{12, \ldots, 20\}$, où la probabilité critique est strictement supérieure à 0.05.

Dans la première région, appelée R_c, on rejette H_0 et donc $\pi < \pi_0$. Dans la seconde, on acceptera H_0 et donc $\pi = \pi_0$.
On constate que 13 appartient à la seconde région et de plus, que jusqu'à 11, on peut considérer le sourcier comme capable ! Autrement dit, au seuil 0.05, la bonne décision est "Prendre comme critère de décision que le sourcier ait des compétences si on enregistre plus de 11 succès " et non "si on enregistre plus de 13 succès" !

Remarque 5. *Attention*, on vient de montrer dans cette question, que, au delà de 11 succès, le sourcier a de réelles compétences, au sens où, au seuil 0.05, la proportion annoncée $\pi = 0.75$ est conforme au nombre de succès. En revanche dans la question 1, on a montré que le sourcier obtient des résultat non dûs au hasard pour un score supérieur à 15, ce qui n'est pas identique !

(b) Si π=0.75, on peut s'attendre, en moyenne à $n\pi = 0.75 \times 20 = 15$ succès.

Correction de l'exercice 4.

(1) On sait qu'à chaque âge, la taille suit une loi normale. D'après le cours (voir chapitre 6), la loi normale est une lois symétrique : la moyenne est au milieu des tailles qui se trouve à $\pm$ écart-type.

Ainsi, la taille moyenne à 0 mois est par exemple donné par

$$\frac{48 + 52}{2} = 50.$$

De même, la taille moyenne à 2 mois est par exemple donné par

$$\frac{54 + 59}{2} = 56.5.$$

De même, la taille moyenne à 3 mois est par exemple donné par

$$\frac{56.5+62}{2}=59.25.$$

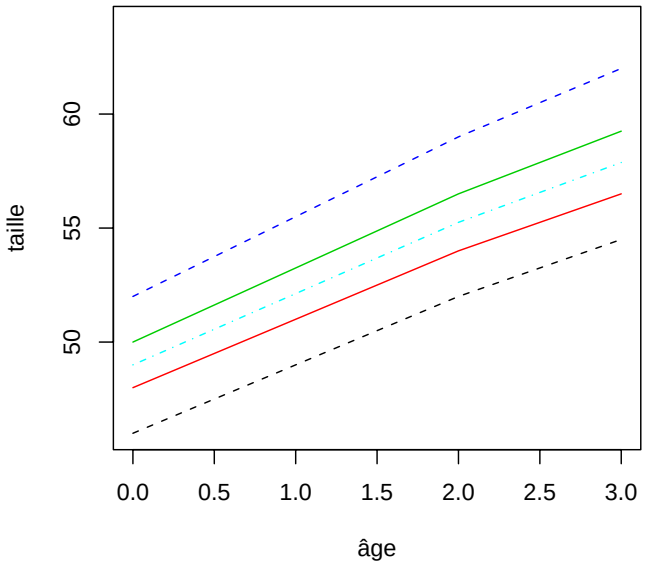
On peut aussi faire le calcul en prenant les courbes à ± 2 écart-types ; par exemple, la taille moyenne à 3 mois est par exemple donné par

$$\frac{54.5+64}{2}=59.25.$$

Voir le tableau suivant :

âge	-2 sigma	-1 sigma	moyenne	+1 sigma	+2 sigma
0	46.0	48.0	50.0	52	54
2	52.0	54.0	56.5	59	61
3	54.5	56.5	59.2	62	64

On peut faire figurer cette nouvelle courbe sur le graphique précédent (en pointillés-tirets) :



(2) On s'intéresse aux deux courbes en pointillés, celles qui se trouvent à ± 2 écart-types de la moyennes.

(a) On cherche à déterminer la probabilité p tel que

$$P\left(-2\leq\frac{X-\mu}{\sigma}\leq 2\right)=p$$

où X suit une loi normale de moyenne μ et d'écart-type σ. Il suffit de reprendre le raisonnement de la page 70 : On cherche donc à "résoudre"

$$P(-z\leq Y\leq z)=p \tag{4}$$

où Y suit une loi de probabilité normale centrée réduite et ici $z = 2$ est connu Comme dans le raisonnement de la note en petit caractère 5.7 page 43, on utilise l'équation du cours (5.27) page 43 : ainsi, (4) est équivalent à

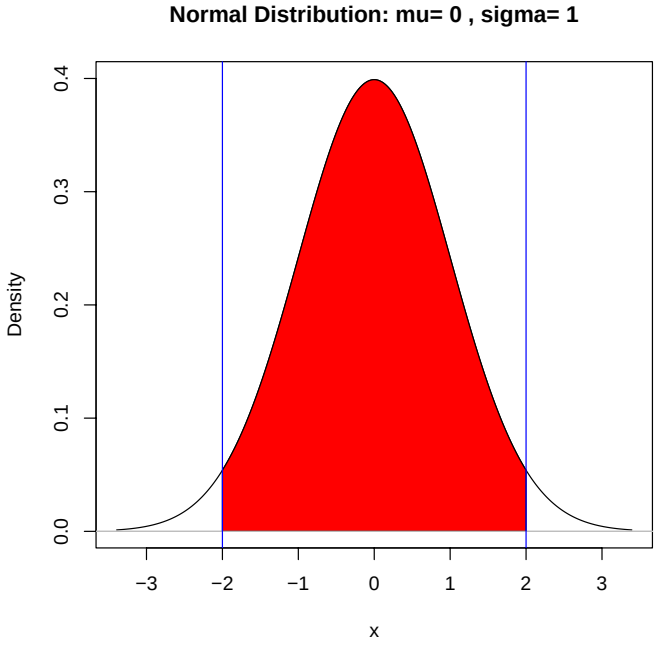
$$P(Y \leq z) = \frac{1 + NC}{2}$$

On connaît $z = 2$ et on trouve alors grâce à Rcmdr

$$\frac{1 + NC}{2} = 0.97725$$

et donc

$$p = 2 \times 0.97725 - 1 = 0.9545 \approx 0.95.$$

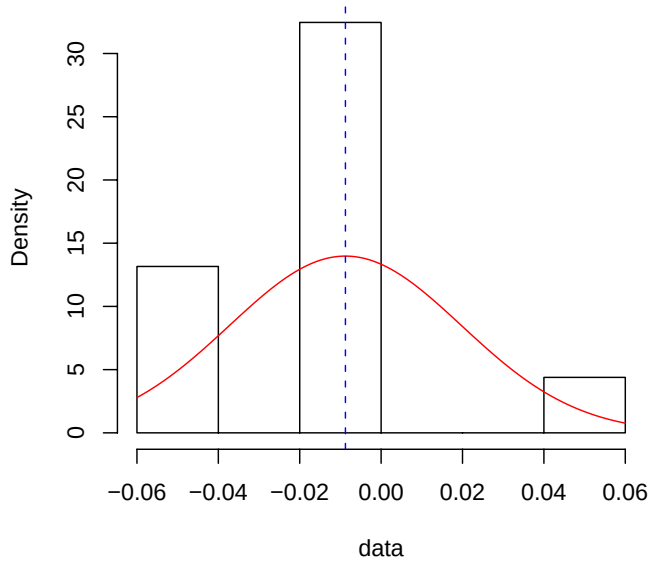


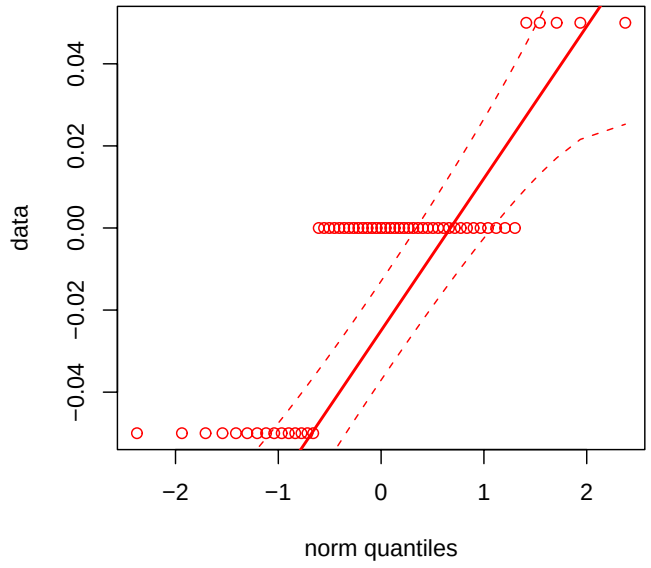
On retrouve donc le fait déjà vu que les tailles sont à moins de deux écart-types dans 95% des cas (voir la figure ci-dessus, où l'aire rouge représente 95% du total, entre les deux lignes bleues verticales aux abscisses ± 2, c'est-à-dire à ± 2 écart-types de l'origine.) !

(b) On a donc $1 - p$ proportion de bébé à l'extérieur des deux courbes (symétriques par rapport à la moyenne), en particulier $(1 - p)/2 = 0.02275$ de proportion p de bébés sous la première courbe et $(1 - p)/2 = 0.02275$ de proportion p de bébés au -dessus de la deuxième courbe ; soit encore 2.275 % de bébés sous la première courbe et 2.275 % de bébés au-dessus de la deuxième courbe.

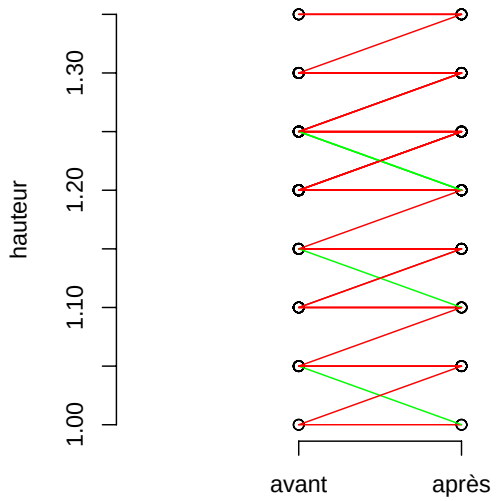
Correction de l'exercice 5.
On veut montrer que le stage est bénéfique, c'est-à-dire que la moyenne du second groupe est supérieure à celle du premier groupe.

- Pour les manipulations sous ®, on renvoie à la remarque 7.4 du document de cours.
-





On peut tracer l'histogramme, en densité, des différences et le graphe quantile-quantile . Sur l'histogramme en densité, on peut ajouter en rouge la loi normale. On obtient les deux graphes de la figure ci-dessus.



On peut aussi tracer le graphe obtenu avec la fonction `parallelplot` sur la figure ci-dessus.

Le graphe quantile-quantile nous montre un mauvais accord avec la loi normale, qui provient probablement du fait que les hauteurs de saut sont déterminées avec un pas de 5 cm. Cependant, on applique abusivement la technique de test !

- On procède au *test de Student apparié* (à la différence des moyennes).

  On fait l'hypothèse nulle H_0 : $\mu_x - \mu_y = 0$. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les différences entre le premier et le second échantillon (qui sont appariés) est strictement négative. 0n fait donc l'hypothèse alternative suivante : $H_{a(2)}$: $\mu_x - \mu_y < 0$

  Grâce à R, on trouve la valeur suivante de la statistique

$$t = \frac{m_d}{sd_d/\sqrt{n}} = -2.320477$$

  La probabilité critique $P(T \leq t)$ (pour la loi de Student à $ddl = 56$ degrés de libertés) est égale à

$$p_c = 0.0119914$$

  Puisque p_c est inférieure au égal au niveau de signification 0.05, on rejette l'hypothèse nulle H_0. Ainsi, H_1 est vraie et *la moyenne μ_x du premier échantillon est strictement inférieure à celle du second échantillon μ_y*, au risque 0.05.

  Les effets du stage sont donc statistiquement significatifs au seuil de usuel de 0.05.

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.