Université Claude Bernard Lyon 1

M1PPMR (Semestre 1) statistique/gestion
Session 2 04 Janvier 2010

Corrigé de l'examen CT de statistiques

Correction de l'exercice 1.

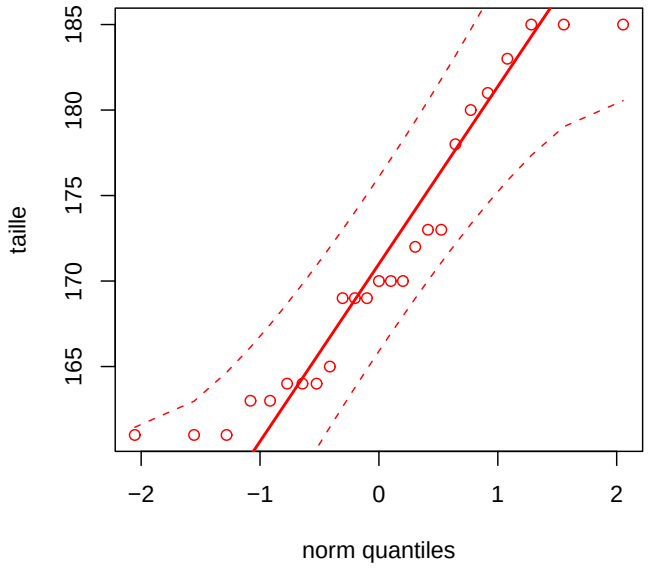
(1) (a)
- On étudie la variable quantitative (ou numérique) '`taille`'. Pour les manipulations avec R, on renvoie donc aux sections 3.2, 3.3 et 3.4 du document de cours.
- Les différents résultats déterminés par R sont donnés dans le tableau suivant

noms	valeurs
moyenne	171.12
sd	8.156388
Q_1 (quartile à 25 %)	164
médiane	170
Q_3 (quartile à 75 %)	178
minimum	161
maximum	185
nombre	25

-

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

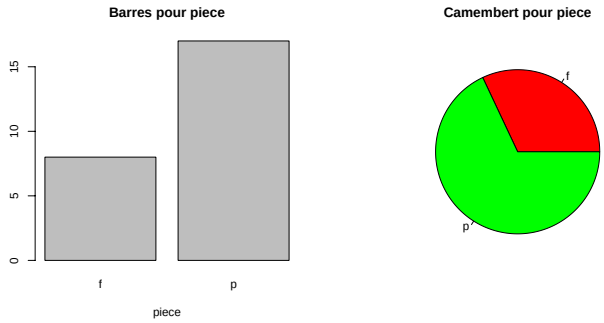
(b) On peut voir sur ce graphique que la distribution est n'est pas allure très normale, ce qui est confirmé par le graphe quantile-quantile en particulier (voir figure ci-dessus).



(2) (a)
- On étudie la variable qualitative (ou catégorielle) '`piece`'. 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
f	8	32.000
p	17	68.000

-



Voir les deux graphiques ci-dessus pour la variable '`piece`'. On constate que les '`face`' représentent la moitié environ des '`pile`'.

(b) Normalement, ces données ont été choisies au hasard par les étudiants et chacune des modalités '`pile`' ou '`face`' devrait représenter environ la moitié des effectifs, ce qui n'est pas le cas ! soit, les étudiants n'ont pas déterminé ces valeurs au hasard, soit, ce qui est plus vraisemblable, ils ne sont pas assez nombreux pour que cette répartition uniforme des valeurs '`pile`' ou '`face`' apparaisse clairement.

Remarque 1. La séquence suivante

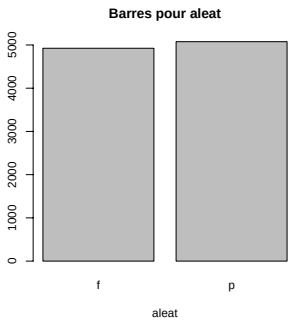
```
n <- 10000
aleat <- sample(as.factor(c("f", "p")), replace = T, size = n)
```

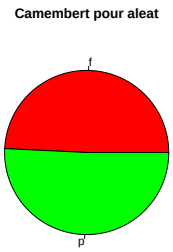
crée un tableau de type catégoriel, avec deux modalités '`f`' et '`p`' grâce à 10000 tirages aléatoires dans l'ensemble {'`f`','`p`'}.

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

	effectifs	pourcentages
f	4924	49.240
p	5076	50.760

-





Voir les deux graphiques ci-dessus pour la variable '`aleat`'. Ici, les proportions observées pour chacune des modalités '`f`' et '`p`' sont proches de la moitié. Les effecitfs sont proches de 5000, ce qui correspond à 10000/2. En théorie des probabilités, chaque modalité '`f`' ou '`p`' "a autant de chance de sortir", ce qui justifie la valeur de 1/2, probabilité d'apparition de chacune des modalités.

(3) (a) On procéde par exemple comme indiqué dans la remarque 6.26 du cours.

On détermine dans un premier temps, grâce à Rcmdr,

- `mu` : moyenne mesurée
- `sd` : écart-type (déviation standart) mesuré ;
- `n` : la taille de l'échantillon.

Pour le niveau de confiance 0.95, on obtient alors

```
[1] 167.7532 174.4868
```

et donc un intervalle de confiance donné par

$$[167.7532085, 174.4867915]$$

Plus rapidement, on pouvait aussi procéder comme indiqué dans la remarque 6.25 du cours.

De même, pour le niveau de confiance 0.99, on obtient un intervalle de confiance donné par

$$[166.5574152, 175.6825848]$$

(b)
- Reprenons le raisonnement de la section 6.5.1 du cours : la norme $\mu_0 = 1.8$ n'appartient à aucun des deux intervalles de confiance ; cela signifie, sans autre calcul supplémentaire, que la théorie des tests, que la moyenne μ est différente de la norme !
- Confirmons cela par les deux calculs suivants :
  - On procède au *test de Student pour un échantillon* (à la moyenne).
    On fait l'hypothèse nulle H_0 : $\mu = \mu_0$. avec μ_0 =1.8. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les données de l'échantillon étudié, est différente de μ_0.
    On fait donc l'hypothèse alternative suivante : H_1 : $\mu \neq \mu_0$.
    Grâce à R, on trouve la valeur suivante de la statistique

$$t = \frac{m - \mu_0}{sd/\sqrt{n}} = 103.795944$$

    La probabilité critique $P(|T| \geq |t|) = 2P(T \geq |t|)$ (pour la loi de Student à $ddl = 24$ degrés de libertés) est égale à

$$p_c = 2.34621e - 33$$

    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 est différente de* $\mu_0 = 1.8$, au risque 0.05.
  - On procède au *test de Student pour un échantillon* (à la moyenne).
    On fait l'hypothèse nulle H_0 : $\mu = \mu_0$. avec μ_0 =1.8. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les données de l'échantillon étudié, est différente de μ_0.
    On fait donc l'hypothèse alternative suivante : H_1 : $\mu \neq \mu_0$.
    Grâce à R, on trouve la valeur suivante de la statistique

$$t = \frac{m - \mu_0}{sd/\sqrt{n}} = 103.795944$$

    La probabilité critique $P(|T| \geq |t|) = 2P(T \geq |t|)$ (pour la loi de Student à $ddl = 24$ degrés de libertés) est égale à

$$p_c = 2.34621e - 33$$

    Puisque p_c est inférieure au égal au niveau de signification 0.01, on rejette l'hypothèse nulle H_0. Ainsi, H_1 est vraie et *la moyenne est différente de* $\mu_0 = 1.8$, au risque 0.01.
- Pour décider si la moyenne est supérieure à $\mu_0 = 1.8$, il faut faire le test complet ici !
  - On procède au *test de Student pour un échantillon* (à la moyenne).
    On fait l'hypothèse nulle H_0 : $\mu = \mu_0$. avec μ_0 =1.8. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les données de l'échantillon étudié, est plus grande que μ_0. On fait donc l'hypothèse alternative suivante : H_1 : $\mu > \mu_0$.
    Grâce à R, on trouve la valeur suivante de la statistique

$$t = \frac{m - \mu_0}{sd/\sqrt{n}} = 103.795944$$

    La probabilité critique $P(T \geq t)$ (pour la loi de Student à $ddl = 24$ degrés de libertés) est égale à

$$p_c = 1.17311e - 33$$

    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 est plus grande que* $\mu_0 = 1.8$, au risque 0.05.

- On procède au *test de Student pour un échantillon* (à la moyenne).
  On fait l'hypothèse nulle H_0 : $\mu = \mu_0$. avec μ_0 =1.8. On cherche à montrer que la moyenne de la loi normale, dont proviendraient les données de l'échantillon étudié, est plus grande que μ_0. On fait donc l'hypothèse alternative suivante : H_1 : $\mu > \mu_0$.
  Grâce à R, on trouve la valeur suivante de la statistique

$$t = \frac{m - \mu_0}{sd/\sqrt{n}} = 103.795944$$

  La probabilité critique $P(T \geq t)$ (pour la loi de Student à $ddl = 24$ degrés de libertés) est égale à

$$p_c = 1.17311e - 33$$

  Puisque p_c est inférieure au égal au niveau de signification 0.01, on rejette l'hypothèse nulle H_0. Ainsi, H_1 est vraie et *la moyenne est plus grande que* $\mu_0 = 1.8$, au risque 0.01.

Correction de l'exercice 2.
Cet exercice a été donné par Stéphane Champely lors du CCF2 des M2 FIAPS en Décembre 2009. Il est très proche de la question 3 de l'exercice 3 de votre CCF2 de décembre 2009 !
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 : ici, le succés est d'avoir un bébé séropositif.

(1) "à la main" : La proportion observée de succés est $pr = 13/164 = 0.079268$.

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 $\pi_0 = 40/160 = 0.25$ et $n = 164$. 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. La différence de proportion est statistiquement significative si $\pi \neq \pi_0 = 0.25$. l'hypothèse alternative choisie est donc $H_1 : \pi \neq \pi_0$. Comme dans la définition 6.33 dans le document de cours (ou définition 12.5 page 204 de [Cha04]), on calcule $2P(Z \geq |t|)$ où Z suit une loi normale centrée réduite.

Sous R, on obtient successivement

$$z = \frac{pr - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.079268 - 0.25}{\sqrt{\frac{0.25\times(1-0.25)}{164}}} = -5.04935,$$

puis

$$p_c = 4.4332e - 07. \tag{1}$$

Enfin, comme dans la définition 6.33 dans le document de cours, on constate que $p_c \leq \alpha = 0.05$; ainsi, on refusera H_0 et on accepte H_1 donc la proportion observée est, au risque 0.05, différente de 0.25.

Le traitement à l'AZT semble donc efficace !

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

```
prop.test(x=13,n=164,p=0.25,alternative="two.sided",correct=F,conf.level=0.95)
```

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

```
        1-sample proportions test without continuity correction

data:  13 out of 164, null probability 40/160
X-squared = 25.4959, df = 1, p-value = 4.433e-07
alternative hypothesis: true p is not equal to 0.25
95 percent confidence interval:
 0.04690774 0.13088777
sample estimates:
        p
0.0792683
```

soit une probabilité critique

$$p_c = 4.433e-07. \tag{2}$$

à comparer à $p_c = 4.4331597637922e-07.$ donné par (1) !

Correction de l'exercice 3. On veut montrer que la hausse est significative, c'est-à-dire que la moyenne du second groupe est inférieure à celle du premier groupe. Ici, on n'a pas accès aux données ; il faut donc faire les calculs "à la main".
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. On fait donc l'hypothèse alternative suivante : $H_{a(2)}$: $\mu_x - \mu_y < 0$
Grâce à ®, on trouve la valeur suivante de la statistique

$$t = \frac{m_d}{sd_d/\sqrt{n}} = -140.84566$$

La probabilité critique $P(T \leq t)$ (pour la loi de Student à $ddl = 149$ degrés de libertés) est égale à

$$p_c = 1.02963e-160$$

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.
La hausse est donc significative au seuil de 0.05.

Remarque 2. Cela pouvait se deviner vu la petitesse de l'écart-type (la probabilité critique est alors elle-même toute petite !).
En réalité, il y avait une coquille (ce qui ne changeait en fait rien à la démarche ...) et il aura fallu lire $sd = 200$:
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. On fait donc l'hypothèse alternative suivante : $H_{a(2)}$: $\mu_x - \mu_y < 0$
Grâce à ®, on trouve la valeur suivante de la statistique

$$t = \frac{m_d}{sd_d/\sqrt{n}} = -1.408457$$

La probabilité critique $P(T \leq t)$ (pour la loi de Student à $ddl = 149$ degrés de libertés) est égale à

$$p_c = 0.0805396$$

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

La hausse n'est donc pas significative au seuil de 0.05.

Correction de l'exercice 4. On renverra à l'exercice 11.3 du cours, très proche !
En utilisant, par exemple, la fonction `int.conf.prop.R`, on trouve l'intervalle de confiance

$$[0.46618, 0.55382]$$

Ici, ce n'est guère exploitable, car cet intervalle contient la valeur 50 % !

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.