

Statistiques appliquées (L3 d'économie) - Cours de Patrick Sevestre - TD 5 - Corrigé

Marc Sangnier - marc.sangnier@ens-cachan.fr

6 janvier 2008

Exercice 1

Question 1 - Test bilatéral

L'estimateur du loyer moyen calculé est sur le premier échantillon (de taille $n_1 = 6$) est $\hat{m}_1 = 493$. On veut tester au seuil de 5% l'hypothèse selon laquelle la vraie moyenne de cet échantillon est égale à 500. Soit m_1 cette vraie valeur. Soit $\alpha = 0,05$ le risque de première espèce.

Hypothèses du test

L'hypothèse nulle est la suivante : le vrai loyer moyen est égal à 500 euros :

$$H_0 : m_1 = 500$$

L'hypothèse alternative est la suivante : le vrai loyer moyen n'est pas égal à 500 euros :

$$H_1 : m_1 \neq 500 \iff m_1 < 500 \cup m_1 > 500$$

Risques du test

Le risque de première espèce est celui de refuser à tort l'hypothèse nulle : considérer que m_1 n'est pas égal à 500 alors qu'il l'est.

Le risque de seconde espèce est celui d'accepter à tort l'hypothèse nulle : considérer que m_1 est égal à 500 alors qu'il ne l'est pas.

Région critique

La région critique correspondant à ce test s'écrit :

$$W = \left\{ (x_1; \dots; x_{n_1}) / \left| \frac{\hat{m}_1 - m_1}{\sigma_{\hat{m}_1}} \right| \geq F^{-1}\left(1 - \frac{\alpha}{2}\right) \right\}$$

Comment "lire" cette expression ? La région critique définit l'ensemble des réalisations observées (à partir desquelles on calcule notre estimateur \hat{m}_1) pour lesquelles on va rejeter l'hypothèse nulle faite sur m_1 .

Dans l'expression précédente, on appelle $F^{-1}(\cdot)$ l'inverse de la fonction de répartition de la loi Normale centrée-réduite.

Application

$$\begin{aligned} W &= \left\{ (x_1; \dots; x_{n_1}) / \left| \frac{\hat{m}_1 - m_1}{\sigma_{\hat{m}_1}} \right| \geq 1,96 \right\} = \left\{ (x_1; \dots; x_{n_1}) / \left| \frac{\hat{m}_1 - 500}{\sigma_{\hat{m}_1}} \right| \geq 1,96 \right\} \\ &\iff W = \left\{ (x_1; \dots; x_{n_1}) / \left| \frac{493 - 500}{76,4} \right| \geq 1,96 \right\} = \left\{ (x_1; \dots; x_{n_1}) / \frac{7}{76,4} \geq 1,96 \right\} \\ &\iff W = \{(x_1; \dots; x_{n_1}) / 0,09 \geq 1,96\} \end{aligned}$$

Notre estimateur \hat{m}_1 ne vérifie donc pas cette inégalité, il ne se trouve donc pas dans la région critique, il ne nous permet pas de rejeter l'hypothèse nulle, nous l'acceptons.

Question 2 - Test bilatéral (2)

L'estimateur du loyer moyen calculé est sur le premier échantillon (de taille $n_2 = 300$) est $\hat{m}_2 = 550$. On veut tester au seuil de 5% l'hypothèse selon laquelle la vraie moyenne de cet échantillon est égale à 500. Soit m_2 cette vraie valeur. Soit $\alpha = 0,05$ le risque de première espèce.

Hypothèses du test

L'hypothèse nulle est la suivante : le vrai loyer moyen est égal à 500 euros :

$$H_0 : m_2 = 500$$

L'hypothèse alternative est la suivante : le vrai loyer moyen n'est pas égal à 500 euros :

$$H_1 : m_2 \neq 500$$

Risques du test

Le risque de premier espèce est celui de refuser à tort l'hypothèse nulle : considérer que m_2 n'est pas égal à 500 alors qu'il l'est.

Le risque de seconde espèce est celui d'accepter à tort l'hypothèse nulle : considérer que m_2 est égal à 500 alors qu'il ne l'est pas.

Région critique

La région critique correspondant à ce test s'écrit :

$$W = \left\{ (x_1; \dots; x_{n_2}) / \left| \frac{\hat{m}_2 - m_2}{\sigma_{\hat{m}_2}} \right| \geq F^{-1}\left(1 - \frac{\alpha}{2}\right) \right\}$$

Comment "lire" cette expression ? La région critique définit l'ensemble des réalisations observées (à partir desquelles on calcule notre estimateur \hat{m}_2) pour lesquelles on va rejeter l'hypothèse nulle faite sur m_2 .

Dans l'expression précédente, on appelle $F^{-1}(\cdot)$ l'inverse de la fonction de répartition de la loi Normale centrée-réduite.

Application

$$\begin{aligned} W &= \left\{ (x_1; \dots; x_{n_2}) / \left| \frac{\hat{m}_2 - m_2}{\sigma_{\hat{m}_2}} \right| \geq 1,96 \right\} = \left\{ (x_1; \dots; x_{n_2}) / \left| \frac{\hat{m}_2 - 500}{\sigma_{\hat{m}_2}} \right| \geq 1,96 \right\} \\ \iff W &= \left\{ (x_1; \dots; x_{n_2}) / \left| \frac{550 - 500}{17,3} \right| \geq 1,96 \right\} = \left\{ (x_1; \dots; x_{n_2}) / \frac{50}{17,3} \geq 1,96 \right\} \\ \iff W &= \{(x_1; \dots; x_{n_2}) / 2,89 \geq 1,96\} \end{aligned}$$

Notre estimateur \hat{m}_2 vérifie donc cette inégalité, il se trouve donc dans la région critique, il nous permet de rejeter l'hypothèse nulle. Nous retenons donc provisoirement (avant un autre test) l'hypothèse alternative.

Question 3 - Tests unilatéraux

On veut tester au seuil de 5% l'hypothèse selon laquelle la vraie moyenne est supérieure ou égale à 580 (test A); inférieure ou égale à 580 (test B). Soit m la vraie valeur moyenne des loyers.

Hypothèses du test A

L'hypothèse nulle est la suivante : le vrai loyer moyen est supérieur ou égal à 580 euros :

$$H_0 : m \geq 580$$

L'hypothèse alternative est la suivante : le vrai loyer moyen est strictement inférieur à 580 euros :

$$H_1 : m < 580$$

Risques du test A

Le risque de première espèce est celui de refuser à tort l'hypothèse nulle : considérer que m n'est pas supérieur ou égal à 580 alors qu'il l'est.

Le risque de seconde espèce est celui d'accepter à tort l'hypothèse nulle : considérer que m est supérieur ou égal à 580 alors qu'il ne l'est pas.

Région critique pour l'échantillon 1, test A

$$\begin{aligned} W &= \left\{ (x_1; \dots; x_{n_2}) / \hat{m}_1 \leq m - \sigma_{\hat{m}_1} * F^{-1}(1 - \alpha) \right\} = \left\{ (x_1; \dots; x_{n_2}) / \hat{m}_1 \leq 580 - 76,4 * 1,64 \right\} \\ \iff W &= \{(x_1; \dots; x_{n_2}) / 493 \leq 454,7\} \end{aligned}$$

Cette inégalité n'est pas vérifiée, l'hypothèse nulle n'est donc pas rejetée, elle est acceptée.

Région critique pour l'échantillon 2, test A

$$W = \{(x_1; \dots; x_{n_2})/\hat{m}_2 \leq m - \sigma_{\hat{m}_2} * F^{-1}(1 - \alpha)\} = \{(x_1; \dots; x_{n_2})/\hat{m}_2 \leq 580 - 17,3 * 1,64\}$$

$$\iff W = \{(x_1; \dots; x_{n_2})/550 \leq 551,6\}$$

Cette inégalité est vérifiée, l'hypothèse nulle est donc rejetée. On retient donc l'hypothèse alternative.

Hypothèses du test B

L'hypothèse nulle est la suivante : le vrai loyer moyen est inférieur ou égal à 580 euros :

$$H_0 : m \leq 580$$

L'hypothèse alternative est la suivante : le vrai loyer moyen est strictement supérieur à 580 euros :

$$H_1 : m > 580$$

Risques du test B

Le risque de premier espèce est celui de refuser à tort l'hypothèse nulle : considérer que m n'est pas inférieur ou égal à 580 alors qu'il l'est.

Le risque de seconde espèce est celui d'accepter à tort l'hypothèse nulle : considérer que m est inférieur ou égal à 580 alors qu'il ne l'est pas.

Région critique pour l'échantillon 1, test B

$$W = \{(x_1; \dots; x_{n_2})/\hat{m}_1 \geq m + \sigma_{\hat{m}_1} * F^{-1}(1 - \alpha)\} = \{(x_1; \dots; x_{n_2})/\hat{m}_1 \geq 580 + 76,4 * 1,64\}$$

$$\iff W = \{(x_1; \dots; x_{n_2})/493 \geq 705,3\}$$

Cette inégalité n'est pas vérifiée, l'hypothèse nulle n'est donc pas rejetée, elle est acceptée.

Région critique pour l'échantillon 2, test B

$$W = \{(x_1; \dots; x_{n_2})/\hat{m}_2 \geq m + \sigma_{\hat{m}_2} * F^{-1}(1 - \alpha)\} = \{(x_1; \dots; x_{n_2})/\hat{m}_2 \geq 580 + 17,3 * 1,64\}$$

$$\iff W = \{(x_1; \dots; x_{n_2})/550 \leq 608,4\}$$

Cette inégalité n'est pas vérifiée, l'hypothèse nulle n'est donc pas rejetée, elle est acceptée.

Question 4 - Test d'égalité des moyennes

On veut tester au seuil de 5% l'hypothèse d'égalité des moyennes calculées sur les deux échantillons.

Hypothèses

L'hypothèse nulle est la suivante : les loyers moyens des deux échantillons sont égaux :

$$H_0 : m_1 = m_2$$

L'hypothèse alternative est la suivante : les loyers moyens des deux échantillons ne sont pas égaux :

$$H_1 : m_1 \neq m_2$$

Risques du test

Le risque de premier espèce est celui de refuser à tort l'hypothèse nulle : considérer que les loyers moyens sont différents alors que ce n'est pas le cas.

Le risque de seconde espèce est celui d'accepter à tort l'hypothèse nulle : considérer que les loyers moyens sont égaux alors que ce n'est pas le cas.

Région critique

$$W = \left\{ (x_1; \dots; x_{n_2}) / \left| \frac{\hat{m}_1 - \hat{m}_2}{\sigma_{\hat{m}_1 - \hat{m}_2}} \right| \geq F^{-1}\left(1 - \frac{\alpha}{2}\right) \right\}$$

Rappelons que (voir les corrigés des séances précédentes pour le calcul de la variance d'une somme) :

$$\sigma_{\hat{m}_1 - \hat{m}_2} = \sqrt{V(\hat{m}_1 - \hat{m}_2)}$$

Il vient alors :

$$W = \left\{ (x_1; \dots; x_{n_2}) / \left| \frac{493 - 550}{78,3} \right| \geq 1,96 \right\} = \{(x_1; \dots; x_{n_2}) / 0,73 \geq 1,96\}$$

Cette inégalité n'est donc pas vérifiée, on accepte donc l'hypothèse nulle : les deux populations envisagées ont la même moyenne.

Question 5 - Test d'égalité des variances

On veut tester au seuil de 5% l'hypothèse d'égalité des variances calculées sur les deux échantillons.

Hypothèses

L'hypothèse nulle est la suivante : les variances des deux échantillons sont égales :

$$H_0 : \sigma_1 = \sigma_2$$

L'hypothèse alternative est la suivante : les variances des deux échantillons ne sont pas égales :

$$H_1 : \sigma_1 \neq \sigma_2$$

Risques du test

Le risque de premier espèce est celui de refuser à tort l'hypothèse nulle : considérer que les loyers des deux échantillons n'ont pas la même variance alors que c'est le cas.

Le risque de seconde espèce est celui d'accepter à tort l'hypothèse nulle : considérer que les loyers des deux échantillons ont la même variance alors que ce n'est pas le cas.

Région critique

$$W = \left\{ (x_1; \dots; x_{n_2}) / \left| \frac{\max(\hat{\sigma}_1^2; \hat{\sigma}_2^2)}{\min(\hat{\sigma}_1^2; \hat{\sigma}_2^2)} \right| \geq F_{(dlmax; dlmin)}^{-1}(1 - \alpha) \right\}$$

Avec $F_{(dlmax; dlmin)}^{-1}(\cdot)$ l'inverse de la fonction de répartition de la loi de Fisher à $dlmax$ et $dlmin$ degrés de liberté où $dlmax$ est le plus grand degré de liberté et $dlmin$ le plus petit degré de liberté parmi les deux échantillons ($n_1 - 1$ et $n_2 - 1$).

Ici on a $dlmax = n_2 - 1 = 299$ et $dlmin = n_1 - 1 = 5$

Application

$$W = \left\{ (x_1; \dots; x_{n_2}) / \left| \frac{90000}{35066} \right| \geq F_{(299; 5)}^{-1}(0, 95) \right\} = \{(x_1; \dots; x_{n_2}) / 2, 57 \geq 4, 37\}$$

Cette inégalité n'est donc pas vérifiée, on accepte donc l'hypothèse nulle : les deux populations envisagées ont la même variance.

Question 6 - Modification du seuil critique

Plus le seuil critique est faible, plus on réduit la taille de la région de rejet, plus on évite de se tromper en refusant à tort l'hypothèse nulle.

Exercice 2

Question 1 - La distribution des loyers suit-elle une loi normale ?

Dans un premier temps, on calcule les fréquences empiriques : $f_i = \frac{n_i}{N}$ où n_i est le nombre d'observation de la classe i et $N = 306$.

Centre de classe x_i	250	350	450	550	800
Fréquence empirique	0,075	0,203	0,464	0,203	0,056

Au passage on remarque qu'on a bien : $\sum f_i = 1$

Calculons la moyenne empirique \bar{X} de la distribution :

$$\bar{X} = \sum_{i=1}^5 f_i * x_i = 454,41$$

On peut retrouver ce résultat en faisant :

$$\bar{X} = \frac{1}{N} \sum_{i=1}^5 x_i * n_i = 454,41$$

Calculons maintenant la variance empirique $\bar{\sigma}^2$ de la distribution :

$$\bar{\sigma}^2 = \sum_{i=1}^5 (x_i - \bar{X})^2 * f_i = 13844,91$$

On en déduit l'écart-type $\bar{\sigma}$:

$$\bar{\sigma} = \sqrt{\bar{\sigma}^2} = 117,66$$

Calculons maintenant la distribution théorique des loyers en tenant compte de la moyenne et de l'écart type :

Soit X la variable aléatoire représentant le loyer.

Pour tout t , $P(\frac{X-\bar{X}}{\bar{\sigma}} \leq t)$ est lu dans la table de la loi Normale centrée réduite.

$X \leq \check{x}_i$	$X \leq 200$	$X \leq 300$	$X \leq 400$	$X \leq 500$	$X \leq 600$	$600 \leq X$
$\frac{\check{x}_i - \bar{X}}{\bar{\sigma}} = t_i$	-2,16	-1,31	-0,46	0,38	1,23	
$P(\frac{\check{x}_i - \bar{X}}{\bar{\sigma}} \leq t_i)$	0,0154	0,0951	0,323	0,652	0,892	0,108

On peut en déduire les fréquences \tilde{f}_i et effectifs \tilde{n}_i théoriques de la façon suivante :

$$\tilde{f}_i = P(\frac{\check{x}_{i+1} - \bar{X}}{\bar{\sigma}} \leq t_{i+1}) - P(\frac{\check{x}_i - \bar{X}}{\bar{\sigma}} \leq t_i)$$

$$\tilde{n}_i = \tilde{f}_i * N$$

$X \leq \check{x}_i$	$X \leq 200$	$200 \leq X \leq 300$	$300 \leq X \leq 400$	$400 \leq X \leq 500$	$500 \leq X \leq 600$	$600 \leq X$
\tilde{f}_i	0,0154	0,0797	0,2279	0,329	0,24	0,108
\tilde{n}_i	4,71	24,38	69,73	100,67	73,44	33,04

Calculons maintenant la statistique du χ^2 qui nous intéresse :

$X \leq \check{x}_i$	$X \leq 200$	$200 \leq X \leq 300$	$300 \leq X \leq 400$	$400 \leq X \leq 500$	$500 \leq X \leq 600$	$600 \leq X$
Effectif observé n_i	0	23	62	142	62	17
$(n_i - \tilde{n}_i)^2$	22,20	1,92	59,86	1707,83	130,87	257,53
$\frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i}$	4,71	0,079	0,858	16,96	1,782	7,792

$$\chi^2 = \sum \frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i} = 32,18$$

On compare cette statistique à la loi du *Khi - deux* à $m - k - 1$ degrés de liberté où m est le nombre de classes de l'échantillon et k le nombre de paramètres qui ont été estimés.

Ici $m = 6$ et $k = 2$, donc $m - k - 1 = 3$.

La valeur limite lue dans la table du *Khi - deux* est 7,82.

L'hypothèse nulle est : la distribution suit une loi Normale.

La région critique est $W = \{\chi^2 \geq 7,82\}$

On se trouve donc dans la région critique. On peut donc rejeter l'hypothèse de la loi Normale au seuil de 5%.

Question 2 - Test d'indépendance

Soit n_{ij} le nombre d'observations possédant le caractère X avec la modalité i et le caractère Y avec la modalité j . Soit n_i le nombre d'observations possédant le caractère X avec la modalité i . Soit n_j le nombre d'observations possédant le caractère Y avec la modalité j . N est le nombre total d'observations. Ici le caractère X est le loyer et le caractère Y la surface.

On calcule l'effectif théorique \tilde{n}_{ij} de chaque groupe (modalités i et j) de la façon suivante :

$$\tilde{n}_{ij} = \frac{n_i * n_j}{N}$$

Sous l'hypothèse d'indépendance, la statistique D suit une loi du *Khi - deux* à d degrés de liberté.

$$D = \sum_{i \& j} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}$$

$$d = (r - 1)(s - 1)$$

Avec r et s les nombres de modalités des caractères X et Y .

Ici, $D = 37,46$ et $d = 12$.

La valeur limite lue dans la table du *Khi - deux* est 21,03. La région critique est $W = \{D \geq 21,03\}$.

On se trouve donc dans la région critique, on peut donc refuser l'hypothèse d'indépendance entre la surface et le loyer.

Exercice 3

Question 1 - Tests bilatéraux

On a estimé $\hat{p} = 0,78$ lors des séances précédentes. On souhaite tester au seuil de 5% l'hypothèse selon laquelle la probabilité de réussite est de 0,75.

Hypothèses du test

Hypothèse nulle : la probabilité p de obtenir un CDI moins d'un an après l'obtention d'un master est de 0,75 :

$$H_0 : p = 0,75$$

Hypothèse alternative : la probabilité p de obtenir un CDI moins d'un an après l'obtention d'un master n'est pas égale à 0,75 :

$$H_1 : p \neq 0,75$$

Région critique et application

$$W = \left\{ (x_1; \dots; x_{n_2}) / \left| \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \right| \geq F^{-1}\left(1 - \frac{\alpha}{2}\right) \right\}$$

Avec $F^{-1}(\cdot)$ l'inverse de la fonction de répartition de la loi Normale centrée-réduite.

Il vient, sous l'hypothèse nulle : $\sqrt{p(1-p)/n} = \sqrt{0,75(0,25)/300} = 0,025$

$$W = \left\{ (x_1; \dots; x_{n_2}) / \left| \frac{0,78 - 0,75}{0,025} \right| \geq 1,96 \right\} = \{(x_1; \dots; x_{n_2}) / 1,2 \geq 1,96\}$$

Cette inégalité n'est pas vérifiée, on ne rejette donc pas l'hypothèse nulle, on l'accepte.

Autres tests

– On souhaite tester l'hypothèse selon laquelle la probabilité de réussite est de 0,80.

Il vient alors : $\sqrt{p(1-p)/n} = \sqrt{0,80(0,20)/300} = 0,023$

$$W = \left\{ (x_1; \dots; x_{n_2}) / \left| \frac{0,78 - 0,80}{0,023} \right| \geq 1,96 \right\} = \{(x_1; \dots; x_{n_2}) / 0,87 \geq 1,96\}$$

Cette inégalité n'est pas vérifiée, on ne rejette donc pas l'hypothèse nulle, on l'accepte.

– On souhaite tester l'hypothèse selon laquelle la probabilité de réussite est de 0,85.

Il vient alors : $\sqrt{p(1-p)/n} = \sqrt{0,85(0,15)/300} = 0,021$

$$W = \left\{ (x_1; \dots; x_{n_2}) / \left| \frac{0,78 - 0,85}{0,021} \right| \geq 1,96 \right\} = \{(x_1; \dots; x_{n_2}) / 3,33 \geq 1,96\}$$

Cette inégalité est vérifiée, on peut donc rejeter l'hypothèse selon laquelle la probabilité de trouver un CDI moins d'un an après l'obtention du master est égale à 0,85.

Question 2 - Lien avec l'intervalle de confiance

Lors des séances précédentes, on a construit pour p l'intervalle de confiance à 95% suivant :

$$[0, 736; 0, 830]$$

On remarque qu'on a accepté l'hypothèse nulle des tests pour lesquels la valeur se trouvait à l'intérieur de cet intervalle et qu'on a rejeté l'hypothèse nulle sinon. Attention cependant à ne pas trop généraliser ce résultat, l'intervalle ayant été construit en utilisant la valeur de \hat{p} pour estimer l'écart-type alors que dans les tests on utilise l'hypothèse nulle à la place.

Question 3 - Tests unilatéraux

- On veut tester au seuil de 5% l'hypothèse selon laquelle la probabilité de réussite est inférieure ou égale à 0,80.

$$H_0 : p \leq 0,80$$

$$H_1 : p > 0,80$$

$$W = \left\{ (x_1; \dots; x_{n_2}) / \hat{p} \geq p + \sqrt{p(1-p)/n} * F^{-1}(1-\alpha) \right\} = \{(x_1; \dots; x_{n_2}) / 0,78 \geq 0,80 + 0,023 * 1,64\}$$

$$\iff W == \{(x_1; \dots; x_{n_2}) / 0,78 \geq 0,837\}$$

Cette inégalité n'est pas vérifiée, on accepte donc l'hypothèse nulle.

- On veut tester au seuil de 5% l'hypothèse selon laquelle la probabilité de réussite est supérieure ou égale à 0,80.

$$H_0 : p \geq 0,80$$

$$H_1 : p < 0,80$$

$$W = \left\{ (x_1; \dots; x_{n_2}) / \hat{p} \leq p - \sqrt{p(1-p)/n} * F^{-1}(1-\alpha) \right\} = \{(x_1; \dots; x_{n_2}) / 0,78 \leq 0,80 - 0,023 * 1,64\}$$

$$\iff W == \{(x_1; \dots; x_{n_2}) / 0,78 \leq 0,762\}$$

Cette inégalité n'est pas vérifiée, on accepte donc l'hypothèse nulle.

On ne peut donc pas trancher alors que les deux hypothèses faites ici sont opposées. Ceci se produit lorsque les hypothèses sont trop proches de l'estimation et/ou que l'estimation n'est pas assez précise.