Université de Strasbourg
UFR de Mathématiques et d'Informatique
Han-Ping LI

Année 2018/2019
Statistique
Etude de cas L3

Fiche 4 Tests d'hypothèses

Exercice 1 : Choix de l'hypothèse nulle

Que préférez vous choisir comme l'hypothèse nulle

1) s'il s'agit du diagnostic d'une maladie grave :
a) $\mathbf{H}_0$: "le patient en question est malade" contre $\mathbf{H}_1$: "le patient en question est sain"

b) $\mathbf{H}_0$: "le patient en question est sain" contre $\mathbf{H}_1$: "le patient en question est malade"

2) 1) s'il s'agit de la culpabilité d'un défendeur devant le Tribunal correctionnel :
a) $\mathbf{H}_0$: "le défendeur est innocent " contre $\mathbf{H}_1$: "le défendeur est coupable "
b) $\mathbf{H}_0$: "le défendeur est coupable" contre $\mathbf{H}_1$: "le défendeur est innocent "

Exercice 2 :

L'étiquette d'une bouteille de 75 cl de jus d'orange (d'une certaine marque) indique que le jus d'orange contient en moyenne, au plus un gramme de matière grasse. On prélève $n = 30$ bouteilles de la même marque, en trouve

0.99	1.19	1.03	1.10	0.97	0.79	0.87	1.46	1.02	0.95
1.09	0.85	1.18	0.81	0.96	1.22	0.72	1.13	1.23	1.05
1.36	1.32	1.21	1.02	1.36	0.97	1.21	1.06	1.31	1.01

Après une étude statistique (test de la normalité), on peut raisonnablement supposer qu'il s'agit $\mathcal{N}(\mu, \sigma^2)$. On souhaite tester

$\mathbf{H}_0 : \mu \leq 1$ contre $\mathbf{H}_1 \; : \mu > 1$ au seuil de $\alpha = 0.05$,.

Quelle sera votre conclusion ?

Exercice 3 :

Une université a reçu un envoi en masse de $n = 400$ mails de xx@xxxx.fr. On souhaite savoir si xx@xxxx.fr est un spammeur (c'àd que le score > 2500). Si, se basé sur les scores de ces 400 mails, on a $\overline{x} = 2505$ et $s_c^2 = 3293$, Que doit on conclure au seuil de $\alpha = 0.05$?

a) par un test unilatéral à gauche

b) par un test unilatéral à droit

c) commenter les avantages et inconvénients de a) et b).

Exercice 4 :

On s'intéresse à la taille du lobe frontal des crabes. On prend la variable la taille du lobe frontal notée "FL" du fichier "crabs" de la librairie "MASS" avec 200 comme étant le cardinal de la population . On note μ l'espérance de la variable "FL". On souhaite, basé sur un échantillon de taille $n = 18$, tester l'hypothèse nulle

$\mathbf{H_0}$ $\mu \leq 14,5$ contre $\mathbf{H_1}$: $\mu > 14,5$ au seuil de $\alpha = 5\%$.

a)`install.packages("MASS"); library(MASS)`

Sauvegarder ces 200 valeurs de "FL" du fichier "crabes" dans un vecteur nommé **Population**.

b) Générer **une** réalisation de l'échantillon de taille 30 $(X_1, \ldots, X_{30})$, puis tester l'hypothèse de la normalité :
$\mathbf{H}_0$: "il s'agit d'une loi normale" contre $\mathbf{H}_1$: "il s'agit d'une loi non normale" au seuil de 10%.

On utilise pour cela shapiro.test() .

Donner la p-valeur du test et conclure.

($n = 30$ *uniquement dans cette question.*)

On suppose dans la suite que la population suit une loi normale $N(\mu, \sigma^2)$ *avec les deux paramètres inconnus. On fixe la taille de l'échantillon à* $n = 18$ *dans toute la suite.*

c) Générer $M = 1000$ réalisations de l'échantillon $(X_1, \ldots, X_n)$. Calculer $M = 1000$ réalisations de la statistique T du test en précisant la région de rejet.

d) Tracer l'histogramme de ces $M = 1000$ réalisations de la statistique T.

e) Calculer la vraie valeur de la moyenne théorique μ en utilisant la totalité des 200 valeurs de la population. Déterminer la proportion des réalisations du test (parmi $M = 1000$) qui conduisent à conserver $\mathbf{H_0}$.

f) Interpréter les résultats de e) en utilisant les termes spécifiques du test statistique.

Exercice 5

On s'intéresse aux rendements journaliers de bourse de l'indice Standard and Poors 500 de 1990 à 1999. On prend 2780 valeurs du fichier "SP500" de la librairie "MASS" comme population. On note μ et σ^2 la moyenne théorique et la variance théorique du rendement. On souhaite, basé sur un échantillon de taille $n = 50$, tester l'hypothèse nulle

$\mathbf{H_0}$ $\mu \geq 0.1$ contre $\mathbf{H_1}$: $\mu < 0.1$ au seuil de $\alpha = 5\%$.

a)install.packages("MASS"); library(MASS)
Sauvegarder ces $N = 2780$ valeurs du rendement du fichier "SP500" dans une variable nommée **Population**.

b) Générer **une** réalisation de l'échantillon de taille 30 $(X_1, \dots, X_{30})$, puis tester l'hypothèse de la normalité au seuil de 10% (shapiro.test). Donner la p-valeur du test et conclure. ($n = 30$ uniquement dans cette question)

On suppose dans la suite que la population suit une loi normale $N(\mu, \sigma^2)$ avec les deux paramètres inconnus. On fixe la taille de l'échantillon à $n = 28$ dans toute la suite.

c) Générer $M = 1000$ réalisations de l'échantillon $(X_1, \dots, X_n)$. Calculer $M = 1000$ réalisations de la statistique T du test en précisant la région de rejet.

d) Tracer l'histogramme de ces $M = 1000$ réalisations de la statistique T.

e) Calculer la vraie valeur de la moyenne théorique μ en utilisant la totalité des 200 valeurs de la population. Déterminer la proportion des réalisations du test (parmi $M = 1000$) qui conduisent à conserver $\mathbf{H_0}$.

f) Interpréter les résultats de e) en utilisant les termes spécifiques du test statistique.

Exercice 6 :

Soit $(X_1, \dots, X_{16})$ un échantillon de taille $n = 16$ d'une loi normale $\mathcal{N}(\mu, \sigma^2)$. On considère le test unilatéral droite suivant : $\mathbf{H_0}$: $\mu \leq 10$ contre $\mathbf{H_1}$: $\mu > 10$ au seuil $\alpha = 0,05$.
On souhaite étudier le risque de première espèce et celui de deuxième espèce.

a) Déterminer la valeur critique ainsi que la zone de rejet du test.

b) Que vaut la valeur de référence μ_0 dans l'expression $T_s = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S_c}$?

c) On simule M=10000 réalisations d'un échantillon avec $\mu = 9.7$ et $\sigma = 5$. On n'est pas censé de connaitre les deux paramètres $\mu = 9.7$ et σ=5. On sauvegarde les 10000 réalisations des statistique $\overline{X}$, S_c ainsi que celles de $T_s = \dfrac{\sqrt{n}(\overline{X}-\mu_0)}{S_c}$..

Se basé suer ces 10000 réalisations de T_s, évaluer le risque. De quel risque il s'agit ? Comparer avec round(pt(qt(0.95, df=n-1), ncp = -0.24, df = n-1),5).

d) Tracer l'histogramme de T_s, puis superposer avec la graphe de la densité dt(x, ncp = -0.24, df = n-1).

e) On simule $M = 10000$ réalisations d'un échantillon avec $\mu = 12$ et $\sigma = 5$. On n'est pas censé de connaitre les deux paramètres $\mu = 12$ et $\sigma = 5$. On sauvegarde les 10000 réalisations des statistique $\overline{X}$, S_c ainsi que celles de $T_s = \dfrac{\sqrt{n}(\overline{X}-\mu_0)}{S_c}$..

Se basé suer ces 10000 réalisations de T_s, évaluer le risque. Comparer De quel risque il s'agit ?

Comparer avec round(pt(qt(0.95, df=n-1), ncp = 1.6, df = n-1), 5);

f) Tracer l'histogramme de la statistique T_s basé sur $M = 10000$ réalisations obtenues en e). Superposer avec la densité dt(x, ncp = 1.6, df=n-1).

g) Superposer les trois densités suivantes
$h_0(x)$= dt(x, ncp = -0.24, df=n-1),
$f_0(x)$= dt(x, ncp = 0, df=n-1),
$h_1(x)$= dt(x, ncp = 1.6, df=n-1)
avec la droite verticale $x = \mathrm{qt}(0.95, \mathrm{df} = \mathrm{n}-1)$. Interpréter les graphes.

Exercice 7 :

Soit $(X_1, \ldots, X_{20})$ un échantillon de taille $n = 20$ d'une loi normale $\mathcal{N}(\mu, \sigma^2)$. On considère le test bilatéral suivant : $\mathbf{H_0}$: $\sigma^2 = 4$ contre $\mathbf{H_1}$: $\sigma^2 \neq 4$ au seuil $\alpha = 0,05$.

On souhaite étudier le risque de deuxième espèce.

a) Déterminer la valeur critique ainsi que la zone de rejet du test.

b) On simule M=10000 réalisations d'un échantillon avec $\mu = 10$ et $\sigma = 3$. On n'est pas censé de connaitre les deux paramètres $\mu = 10$ et $\sigma^=3$. On sauvegarde les 10000 réalisations des statistique $\overline{X}$, S_c ainsi que celles de$K^2 = \frac{(n-1)S_c^2}{4}$..

Se basé suer ces 10000 réalisations de T_s, évaluer le risque. De quel risque il s'agit ?

c) Comparer avec la valeur
(β_1=round(pchisq(4*32.85233/3 , df = 19) - pchisq(4*8.906516/3 , df = 19), 3))

d) On simule M=10000 réalisations d'un échantillon avec $\mu = 10$ et $\sigma = 6$. On n'est pas censé de connaitre les deux paramètres $\mu = 10$ et $\sigma^=6$. On sauvegarde les 10000 réalisations des statistique $\overline{X}$, S_c ainsi que celles de$K^2 = \frac{(n-1)S_c^2}{4}$.. Se basé suer ces 10000 réalisations de T_s, évaluer le risque. De quel risque il s'agit ? Commenter vos résultats.

e) Comparer avec la valeur
(β_2=round(pchisq(4*32.85233/6 , df = 19) - pchisq(4*8.906516/6 , df = 19), 3))

Exercice 8 :

Nous souhaitons comparer la taille du lobe frontal des "crabes bleus" et celle des "crabes oranges". Pour simplifier l'étude, notre population des crabes bleus est constituée des 100 valeurs (en mm) de la variable "la taille du lobe frontal", notée "FL" du fichier "**crabs**" de la librairie "MASS" dont l'espèce (la couleur), notée "sp", est **B** et notre population des crabes orange est constituée des 100 valeurs (en mm) de la même variable "FL" des crabes dont l'espèce (la couleur), notée "sp", est **O** du même fichier.

Nous voulons savoir si la taille de crabes " bleu" est significativement plus petite que celle des crabes "orange" ou elles sont compatibles.

Pour ce faire, nous prélevons un échantillon de taille $n_1 = 28$ de la première population, et un échantillon de taille $n_2 = 29$ de la seconde population. Nous notons μ_1 et μ_2 les deux espérances de la variable "FL" selon la couleur des crabes.

1) Sauvegarder les 100 valeurs de "FL" des crabes bleus ("**B**") dans un vecteur nommé **Population1** et les 100 valeurs de de "FL" des crabes orange ("**O**") dans un vecteur nommé **Population2**. (rien à reporter sur la copie).

2) Si les résultats de deux commandes $\mathbf{sample}(1:100,\, n_i, \mathbf{replace} = \mathbf{TRUE})$ pour i=1, 2 sont les suivants (voir le fichier "Indices2" sur le site Moodle.) :

$\text{indices}_1 \quad \big(= \mathbf{sample}(1:100,\, 28, \mathbf{replace} = \mathbf{TRUE})\big)$
$= (7, 13, 12, 78, 57, 34, 15, 73, 56, 9, 23, 74, 29, 74, 26, 70, 82, 73, 24, 68, 77, 4, 87, 13, 82, 68, 15, 64)$
$\text{indices}_2 \quad \big(= \mathbf{sample}(1:100,\, 29, \mathbf{replace} = \mathbf{TRUE})\big)$
$= (50, 13, 43, 68, 90, 99, 76, 62, 46, 65, 23, 14, 30, 35, 100, 87, 85, 36, 10, 46, 84, 85, 69, 73, 84, 10, 76, 58, 95)$,

sauvegarder alors le vecteur $x = (x_1, \ldots, x_{28}))$ qui est une réalisation d'un échantillon de la première population ("crabes bleus") et le vecteur $y = (y_1, \ldots, y_{29})$ qui est une réalisation d'un échantillon de la deuxième population ("crabes oranges"). (rien à reporter sur la copie).

3) Effectuer trois tests préliminaires au seuil de $\tilde{\alpha} = 0,10$. Quelles seront vos conclusions ?

On suppose dans la suite que les deux populations admettent pour lois deux lois normales $N(\mu_1, \sigma_1^2)$ et $N(\mu_2, \sigma_2^2)$ respectivement avec les quatre paramètres inconnus.

4) Formuler l'hypothèse nulle $\mathbf{H}_0$ ainsi que l'hypothèse opposée $\mathbf{H}_1$.

5) Effectuer le test d'hypothèse principal, sans utiliser la commande "t.test()" en R, avec le seuil $\alpha = 0,05$.

- Préciser la forme de la statistique utilisée.

- Déterminer ses degrés de liberté.

- Déterminer la (ou les) valeur(s) critique(s).

- Conclure.