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

Identifier l'erreur de première espèce du test selon le choix. Quel est le choix selon lequel l'erreur dont la conséquence la plus grave correspond à l'erreur de première espèce ?

1) Test pour diagnostiquer une maladie grave :

a) $\mathbf{H}_0$ : "la maladie est présente" contre $\mathbf{H}_1$ : "la maladie est absente".

b) $\mathbf{H}_0$ : "la maladie est absente" contre $\mathbf{H}_1$ : "la maladie est présente".

2) Test pour juger 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 |

1. 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 le code suivant :

   `shapiro.test()` .

2. En supposant 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 $\alpha = 0.05$,.

Le test est-il significatif ? Si vous vous trompez avec votre conclusion, quelle espèce d'erreur commettrez-vous ? Que pouvez-vous dire sur le risque correspond ?

## 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'est-à-dire 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 $\alpha = 0.05$ si on utilise :

a) un test unilatéral à gauche ; b) un test unilatéral à droit.

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" (sans la lettre "e") de la librairie "MASS" avec 200 comme étant le cardinal de la population . On note $\mu$ 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 $\alpha = 5\%$.

1. library(MASS)

   Sauvegarder ces 200 valeurs de "FL" du fichier "crabs" dans un vecteur nommé **Population** :

   Population=crabs$FL

2. Générer **une** réalisation de l'échantillon $(X_1, \ldots, X_{30})$ de taille $n = 30$, 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 le
   shapiro.test() .
   Donner la p-valeur du test et conclure.

   *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.*

3. Générer $M = 1000$ réalisations de l'échantillon $(X_1, \ldots, X_n)$ avec $n = 18$. Calculer $M = 1000$ réalisations de la statistique $T$ du test en précisant la valeur critique $C$ ainsi que la région de rejet.

4. Tracer l'histogramme de ces $M = 1000$ réalisations de la statistique $T$. Superposer avec la droite verticale $x = C$ où $C$ est la valeur critique pour déterminer la zone de rejet :

   abline( v= C, add=T)

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

6. Interpréter les résultats de 5. 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 $\mu$ et $\sigma^2$ 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 $\alpha = 5\%$.

1. `library(MASS)`
   Sauvegarder ces $N = 2780$ valeurs du rendement du fichier "SP500" dans une variable nommée **Population**.

2. Générer **une** réalisation de l'échantillon de taille 30 $(X_1, \ldots, X_{30})$, 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.*

3. 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 valeur critique $C$ ainsi que la région de rejet.

4. Tracer l'histogramme de ces $M = 1000$ réalisations de la statistique $T$. Superposer avec la droite verticale $x = C$ où $C$ est la valeur critique pour déterminer la zone de rejet :

   ```
   abline( v= C, add=T)
   ```

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

6. Interpréter les résultats de 5) en utilisant les termes spécifiques du test statistique.

**Exercice 6 :**

Soit $(X_1, \ldots, X_{16})$ 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 dans cet exercice étudier **le risque de première espèce et celui de deuxième espèce**.

- Déterminer la valeur critique $C$ ainsi que la zone de rejet du test.
- Que vaut la valeur de référence $\mu_0$ dans l'expression $T_s = \dfrac{\sqrt{n}(\overline{X} - \mu_0)}{S_c}$?

I. On se propose d'abord d'étudier **le risque de première espèce**. Les données doivent être choisies de sorte que $\mathbf{H_0}$ soit vraie. On simule M = 10000 réalisations d'un échantillon avec $\mu = 9.7$ et $\sigma = 5$. On n'est pas censé de connaître les deux paramètres $\mu = 9.7$ et $\sigma = 5$. On sauvegarde les 10000 réalisations de la statistique $\overline{X}$, $S_c$ ainsi que celles de $T_s = \dfrac{\sqrt{n}(\overline{X} - \mu_0)}{S_c}$.

1. Se basé suer ces 10000 réalisations de $T_s$, évaluer le risque. De quel risque il s'agit ?
2. Comparer avec la valeur théorique donnée par round( pt( qt(0.95, df=n-1), ncp = -0.24, df = n-1),5).
3. Tracer l'histogramme de $T_s$, puis superposer avec la graphe de la densité dt(x, ncp = -0.24, df = n-1).

II. On se propose maintenant d'étudier **le risque de de deuxième espèce**. Les données doivent être choisies de sorte que que $\mathbf{H_0}$ soit fausse. 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 $T_s = \dfrac{\sqrt{n}(\overline{X} - \mu_0)}{S_c}$..

1. Se basé suer ces 10000 réalisations de $T_s$, évaluer le risque. De quel risque il s'agit ?
2. Comparer avec la valeur théorique donnée par round( pt( qt(0.95, df=n-1), ncp = 1.6, df = n-1), 5).
3. 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).
4. 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 $x = \mathtt{qt}(0.95, \mathtt{df} = \mathtt{n} - 1)$. Interpréter les graphes.