Tests Statistiques 1

L. Noiret, A. Rogozan

1 Rappels

L'objectif d'un test statistique est de déterminer si une hypothèse H0 doit être rejetée ou non. Un test comporte deux risques :

- un risque de première espèce α qui consiste à déclarer que H0 est fausse (c'est à dire conclure à une différence) alors qu'en réalité H0 est vraie
- un risque de deuxième espèce β qui consiste à ne pas déclarer que H0 est fausse (conclure qu'il n'y a pas de différence) alors qu'en réalité H0 est fausse.

Dans la pratique :

1. Formulez l'hypothèse que vous voulez tester H0 et l'hypothèse alternative H1. La formulation de unilatérale ou bilatérale de H1 va déterminer les bornes de la région de rejet.
2. Choisissez le type de test à implémenter. Il existe souvent plusieurs tests pour répondre à la même question. Le choix du test est notamment guidé par ses conditions d'application. Exemple : si vous souhaitez tester que deux échantillons indépendants sont issus de la même population, vous pouvez utiliser un test de Student (T test) ou un test non paramétrique tel que Mann Whitney. Comment choisir? Vous choisissez un test de Student si vos données sont gaussiennes ou si la taille de vos 2 échantillons est assez grande ($n_1 > 30$ et $n_2 > 30$), afin que le théorème central limite s'applique. Sinon vous utilisez le test de Mann Whitney.
3. Interprétez les sorties du logiciel : on rejette H0, si la p-value est inférieur au risque de 1ère espèce (en général $\alpha = 5\%$).

Les tests non paramétriques ne font aucune hypothèses sur la distribution sous-jacente, mais sont moins puissants que les tests paramétriques, c'est-à-dire qu'ils détectent moins souvent une différence lorsqu'il y en une. La puissance (1-β) est la probabilité de déclarer que H0 est fausse à tort (conclure qu'il n'y a pas de différence lorsqu'il y en a une).

Ce TP peut se faire sous Matlab ou R (au choix). Des indications sont fournies pour l'utilisation de R.

On souhaite comparer des données statistiques de la délinquance en Ile de France (IDF) avec celles de la région Provence-Alpes-Côte d'Azur (PACA). Les données, publiées par le journal l'Express, fournissent le nombre de vols de voitures ($VolVoiture$) et de cambriolages ($Cambriolage$) dans différentes circonscriptions. Suivant les indicateurs, les délits sont exprimés pour 1000 ou 10000 habitants.

1.1 Etapes préliminaires

1. Chargez les données

```
setwd("~/coursINSA/ASI/stat/TP5") # changer le repertoire courant
mydata <- read.table("data_securite.csv",header=TRUE,sep="|")
```

2. Faites une analyse descriptive rapide de la variable $Cambriolage$.

```
# le symbole # indique un commentaire
# charger la bibliotheque "psych" permet d'utiliser la fonction "describe"
library(psych)
# la fonction "describe" fourni les indicateurs les plus courants
stat_des<-describe(mydata$Cambriolage) # utiliser $ pour acceder a une colonne
stat_des$mean # recuperer la moyenne
```

3. Créez 2 nouveaux tableaux : l'un contenant les données pour les circonscriptions d'Ile de France (IDF, départements : 75, 77, 78, 91, 92, 93, 94, 95) et l'autre pour celles de la région Provine Alpe Côte d'Azure (PACA, départements : 04, 05, 06, 13, 83 , 84)

```
# exemple selection de la 1ere ligne : mydata[1,]
# exemple selection de la colonne Dpt : mydata[,"Dpt"]
# exemple selection avec une condition : mydata[mydata$Habitants >10000,]

list_dpt <-c(75,77,78,91,92,93,94,95) # c() permet de creer un vecteur
IDF <-mydata [mydata$Departement %in% list_dpt, ]
```

1.2 Comparaison d'une moyenne à une valeur théorique

1. Le nombre moyen de cambriolages en IDF est-il différent de la la moyenne nationale?

   - Donnez H0 et H1. Quelle est la loi de la statistique de test sous H0 ?
   - Effectuez le test et interprêtez la p-value au risque 5%.

     ```
     # faire le test et enregistrer les resultats dans la variable res_moy
     res_1moy<-t.test(IDF$Cambriolage, mu = 5.4)
     # recuperer la p-value
     res_1moy$p.value
     ```

     La fonction *t.test* vous fournit : la valeur de la statistique (t), le nombre de degré de liberté de la loi de Student (df = degree of freedom), la p-value (probabilité d'observer une valeur au moins aussi extrême que la statistique si H0 est vrai), l'intervalle de confiance de la moyenne.

2. Le nombre moyen de cambriolages en IDF est-il supérieur à la moyenne nationale?

   - Forumulez H0 et H1.
   - Effectuez le test et interprêtez la p-value au risque 5%.

     ```
     # faire le test et enregistrer les resultats dans la variable res_moy
     res_1moy_unilateral<-t.test(IDF$Cambriolage, mu = 5.4,alternative = "greater")
     ```

   - Identifiez les différences entre les résultats du test bilatéral et unilatéral. Est-ce que la valeur de la statistique t et le nombre de degrés de libertés df sont différents? Quelle est la relation entre la p-value trouvée pour le test bilatéral et celle pour le test unilatéral? Comment l'expliquez-vous?

1.3 Comparaison de deux échantillons indépendants

On s'intéresse au nombre de cambriolages moyen par circonscription. Les moyennes dans les régions IDF et PACA sont-elles statistiquement différentes?

1. Formulez les hypothèses H0 et H1 permettant de répondre à cette question.

2. Quel test choisiriez-vous? Donner la statistique de test et sa loi sous H0.

3. Avant de procéder au test, vérifiez à l'aide d'un test de Fisher si les variances des cambriolages en IDF et PACA sont égales.

- Pourquoi a-t-on besoin de tester l'égalité des variances? Qu'est-ce que cela change-t-il?
- Rappelez la statistique de ce test, les hypothèses H0 et H1. Ainsi que la loi de la statistique sous H0.
- Récupérez la p-value et concluez au risque 5%

```
res_var<-var.test(IDF$Cambriolage, PACA$Cambriolage)
```

4. Réalisez le test de comparaison de moyennes (voir l'aide sur la fonction *t.test* pour comprendre les options). Concluez lorsque le risque α est de 5%, 1%.

```
res_moy<-t.test(IDF$Cambriolage, PACA$Cambriolage, var.equal = FALSE)
```

1.4 Comparaison de deux échantillons indépendants (formulation unilatérale)

Le nombre de vols de voiture moyen par circonscription en région PACA est-il supérieur au nombre de vols en région Ile de France ?

1. Comment devez-vous adapter H1 pour répondre à cette question?
2. Regarder les options disponibles de la fonction *t.test* (*help*("*t.test*")) pour répondre à cette question.

1.5 Comparaison de deux échantillons indépendants (version non paramétrique)

On souhaite maintenant comparer les cambriolages dans les Hauts-de Seine (Départemnt 92) avec ceux en Seine St Denis (Départemnt 93).

1. Créez 2 nouveaux tableaux : l'un contenant les données pour les circonscriptions des Hauts-de Seine $X92$ et l'autre celles de la Seine St Denis $X93$.
2. Quels tests choisiriez-vous pour comparer les deux départements? Pourquoi?
3. Réalisez ce test. Essayer de comprendre l'origine du message d'avertissement (Warning). Récupérez la p-value et concluez.

```
res_non_para=wilcox.test(X92$Cambriolage, X93$Cambriolage)
```

2 Sources

Méthode Statistiques Médecine - Biologie -Jean Bouyer, ESTEM éditions Inserm

L'Express - Sécurité 2013 - Classements des communes de France
https://lexpress.opendatasoft.com/explore/dataset/statistiques-securite-france-2013/?flg=fr
http://www.lexpress.fr/actualite/societe/insecurite-le-palmares-des-villes-de-france_1300974.html#m2ZPvIj0jcHs7c1g.99