TISD – FICHE 4

Quelques tests statistiques

Adrien Hardy, `adrien.hardy@math.univ-lille1.fr`

1 Test de Student (sous R)

On utilisera la commande `t.test` pour faire un test de Student.

1. Simuler un échantillon de 5 gaussiennes indépendantes $\mathcal{N}(12, 9)$ puis tester sur cet échantillon l'hypothèse $H_0 = \{\mu = 0\}$ contre $H_1 = \{\mu \neq 0\}$ à l'aide du test de Student. Le faire d'abord "manuellement", c'est-à-dire calculer avec R la statistique de test ξ et la region de confiance R_α avec $\alpha = 0.05$ et conclure, puis utiliser la commande `t.test`.

2. Refaire ce test avec l'hypothèse $H_0 = \{\mu = 12\}$ contre $H_1 = \{\mu \neq 12\}$.

3. A l'aide de l'option `alternative` de `t.test`, tester $H_0 = \{\mu \geq 0\}$ contre $H_1 = \{\mu < 0\}$, puis $H_0 = \{\mu \leq 0\}$ contre $H_1 = \{\mu > 0\}$. Commenter les résultats.

4. On s'intéresse au jeu de données `Iris` disponible sous R. Si μ représente la largeur moyenne des sépales, tester $H_0 = \{\mu \geq 3\}$ contre $H_1 = \{\mu < 3\}$. Commentez votre résultat.

2 Tests de gaussianité (Sous R)

Les commandes pour les tests de Kolmogorov-Smirnov, Lilliefors, et Shapiro-Wilks sont respectivement `ks.test`, `lillie.test`, et `shapiro.test`.

NB: Pour le test de Lilliefors, il faut appeler la librairie "nortest" en tapant : `library(nortest)`, qu'il faut installer au préalable si ce n'est pas déjà fait : `install.packages("nortest")`.

1. Soit $X_1, \ldots, X_{20}$ un échantillon de loi exponentielle $\mathcal{E}(1)$, on considère la variable aléatoire

$$Y = \frac{\sqrt{20}}{\sqrt{\mathrm{Var}(X_1)}}\big(\overline{X}_{20} - \mathbb{E}(X_1)\big).$$

A l'aide d'un échantillon $Y_1, \ldots, Y_{50}$ de Y, tester l'hypothèse : $H_0 = \{Y \sim \mathcal{N}(0, 1)\}$.

2. Est-ce que l'hypothèse de gaussianité de la variable "largeur des sépales" de `Iris` est raisonnable ? Et que peut-on dire de la variable "longueur des sépales" ?

3 Test d'équilibre d'un dé (sous R)

Dans le fichier `des.txt` sur Moodle vous trouverez 100 jets du même dé.

1. Rédiger un protocole scientifique pour vérifier si ce dé est truqué ou non.

2. Implémenter cette procédure sous R : Le faire d'abord manuellement, c'est-à -dire en calculant la statistique de test et les quantiles associés, puis en utilisant la commande `chisq.test` (consulter l'aide). Conclure.

3. Donner une estimation de la moyenne théorique du résultat du dé avec intervalle de confiance.

4 Pics d'O_3 (sous SAS)

On s'intéresse de nouveau à la série des pics d'Ozone qui se trouve dans le jeu de données Ozone du TP3 ; la variable maxO3 correspond au maximum d'Ozone pour chaque jour de la table.

1. Après avoir importé ce jeu de données, utiliser la `proc univariate` pour obtenir un histogramme (`histogram`) de `maxO3`.

2. Superposer à cet histogramme des densités de lois théoriques qui vous semblent pertinentes pour modéliser la distribution de `maxO3` (regarder dans l'aide), et interpréter les résultats des tests d'adéquation fournis par SAS.

3. Faire un `qqplot` avec la loi qui vous semble la plus adaptée.

4. Faire la même analyse pour la variable `T12` (température à midi).

5 Syndicats aux U.S. (sous R)

Les données de la table `cps85.txt` (cf. Moodle) proviennent d'une enquête auprès des ménages effectuée par le U.S. Census Bureau en mai 1985. Le fichier contient 534 individus. Nous nous intéresserons aux variables suivantes :

- NONWH qui vaut 1 si l'individu interrogé n'est ni Blanc ni Hispanique
- HISP qui vaut 1 si l'individu interrogé est Hispanique
- FE qui vaut 1 si l'individu interrogé est une femme
- UNION qui vaut 1 si l'individu est syndiqué

1. Importer le fichier avec `read.table`, en rajoutant l'option `header=TRUE`.

2. Pour chacune des variables, donner les fréquences empiriques à l'aide de la commande `table`), puis tracer le diagramme en secteurs associé (commande `pie`).

3. On s'intéresse maintenant aux variables NONWH et UNION.

(a) Donner la table de contingence de ces variables, avec l'aide de la commande `table`.

(b) Tester numériquement la liaison entre les deux variables avec la commande `chisq.test` : Expliquez votre démarche et conclure.

Remarque : En France, la réalisation de traitements de données à caractère personnel faisant apparaître directement ou indirectement les origines raciales ou ethniques des personnes est interdite depuis 2007 par décision du Conseil constitutionnel.

6 Passagers du Titanic (sous R)

Les données sur les passagers du Titanic, fournies par R, peuvent-être chargées par la commande `data(Titanic)`.

1. Demander la description de cette table via `help("Titanic")`, explorer la table (en particulier sa structure, à l'aide de `dim(Titanic)`), et décrire les différentes variables. Combien d'observations en tout ?

2. Donner les fréquences marginales de chaque variable : on utilisera la commande `apply` en choisissant "`FUN = sum`". Tracer ensuite leurs diagrammes en secteurs.

3. Obtenir la table de contingence des variables `Survived` et `Class` puis faire un test d'indépendance.

4. Faire de même avec `Survived` et `Sex`, puis `Survived` et `Age`. Avec quelle variable `Survived` semble-elle le plus liée ?

5. Ecrire un résumé de votre analyse.

7 Travail des femmes (sous SAS)

Importer `travf.sas7bdat` dans **SAS** depuis Moodle (c'est un fichier construit avec **SAS** : il suffit de le déposer dans le dossier physique de la librairie de votre choix pour l'importer). Ces données sont issues de l'Enquête Emploi en Continu 2005 de l'INSEE. Nous disposons de 99 959 individus et des variables suivantes (voir le descriptif, lui aussi sur Moodle) :

- identifiant (IDENT),
- activité (ACT6),
- âge (AGE)
- niveau d'étude (CITE97),
- type de ménage dans laquelle vit la personne (TYPMEN5), indicatrice de vie en couple (COHAB)
- de la CSP de la personne (CSTOT), du chef de ménage ou du conjoint (CSTOTCJ), et de la personne de référence dans le ménage (CSTOTPR),

- du nombre d'enfants (NBENF18)
- de la tranche de salaire (SALREDTR),
- du temps de travail (TPPRED) et du nombre d'heures travaillées pendant une semaine de référence (EMPNBH).

1. Tracer un diagramme en tuyaux d'orgue pour la variable ACT6 avec la `proc gchart` et l'option `hbar` ou `vbar`.

2. Toujours en utilisant la `proc gchart` et avec l'option `pie`, dessiner un diagramme à secteurs pour la variable CITE97.

3. A l'aide de la `proc freq` et l'option `tables`, obtenir la liste des modalités et les fréquences pour les variables ACT6, CITE97.

4. Toujours à l'aide de `proc freq`, croiser les variables ACT6 et CITE97 et donner la valeur des indicateurs de liaison suivants : chi-deux, Φ^2 et coefficient de Cramer. Faire de même avec ACT6 et COHAB.