Partiel TISD - Master Pro
Vendredi 7 novembre 2008

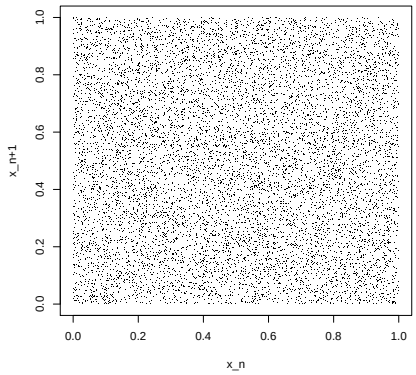
Durée 2 heures. Calculatrices autorisées. Seul un formulaire sur feuille double est autorisé.

Tran Viet Chi, `chi.tran@univ-lille1.fr`, bureau 316 (bâtiment M3).

Exercice 1 (Test du générateur uniforme de R)
Le but de cet exercice est d'étudier un test du générateur de nombres pseudo-aléatoires de **R**. On attend de ce générateur qu'il retourne des suites $X_1, \cdots X_N$ de variables aléatoires uniformes statistiquement indépendantes.

1. En vous appuyant sur les méthodes vues en TD, citer deux façons de vérifier graphiquement que les variables aléatoires générées par **R** suivent la loi uniforme.

2. Test des paires en série : Ce test s'assure qu'il n'y a pas de lien statistique entre deux nombres pseudo-aléatoires générés consécutivement. Lorsqu'on génère une suite de N nombres aléatoires $X_1, \cdots X_N$ et qu'on trace les $N-1$ points de coordonnées (X_i, X_{i+1}) pour $i \in [\![1, N-1]\!]$ on obtient le graphique suivant :



$X_i \backslash X_{i+1}$	A	B	C
A	622	629	1277
B	615	593	1273
C	1291	1259	2440

Que pensez-vous de la figure de gauche ?

3. Nous nous proposons de tester l'absence de liaison statistique entre X_i et X_{i+1}. Pour cela, on sépare l'intervalle $[0, 1]$ en trois : $A = [0, 0.25[$, $B = [0.25, 0.5[$, $C = [0.5, 1]$. On note par n_{AA}, n_{AB} etc. les nombres de points d'abscisse appartenant à A et d'ordonnée appartenant à A, d'abscisse appartenant à A et d'ordonnée appartenant à B etc. A partir des simulations qui ont servi à tracer la figure de la question **2**, on obtient le Tableau de la question **2**.

3.1. Enoncer l'hypothèse nulle H_0 que l'on cherche à tester.

3.2. Quel est le nombre total d'observations ? Calculer les fréquences marginales.

3.3. Calculer la distribution empirique de X_{i+1} conditionnellements à $X_i \in A$, et tracer l'histogramme de cette distribution. Comparer aux résultats de la question **3.2** et commenter.

3.4. Pour tester l'hypothèse H_0 donnée à la question **3.1**, on peut utiliser un test du χ^2. Comment peut-on procéder sous **R** ?

3.5. Nous réalisons maintenant ce test. Calculer la statistique de test donnée par :

$$\xi = \sum_{i=A}^{C}\sum_{j=A}^{C} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}, \quad \text{(avec les notations du cours).}$$

3.6. Quelle est la loi de ξ quand $n \to +\infty$ lorsque H_0 est vraie ? lorsque H_0 est fausse ?

3.7. On rappelle que la commande `pchisq(x,dl)` en **R** donne la valeur en x de la fonction de répartition d'une loi du χ^2 à dl degrés de liberté. Etant donné que `pchisq(xi,4)=0.69`, `pchisq(xi,6)=0.42`, `pchisq(xi,9)=0.14`, quelle est la conclusion du test si l'on veut que sous H_0 la probabilité d'erreur soit inférieure à $\alpha = 5\%$?

4. A partir des fréquences marginales pour X_i obtenues à la question **3.2**, calculer une approximation de l'espérance et de la variance des X_i.

Exercice 2 (Tremblements de terre)
Nous disposons d'une base de données relative à 1000 séismes de magnitude supérieure à 4 (échelle de Richter) s'étant produit depuis 1964 dans le voisinage de Fiji. Pour chacun des séismes, la variable `mag` nous donne la magnitude du séisme.

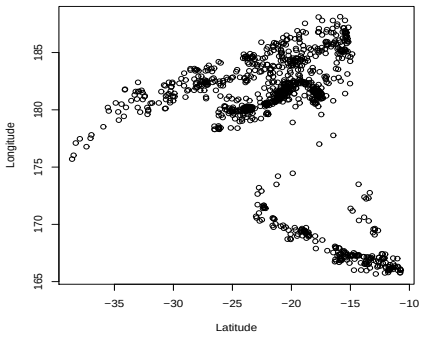


FIG. 1 – *Localisation des séismes*

En regardant la localisation géographique des séismes, on s'aperçoit qu'il y a clairement deux régions d'activité sismique. La première correspond à une jonction de plaque, la seconde à la tranchée de Tonga en Nouvelle Zélande. Nous nous demandons si la magnitude du séisme est liée à la région géographique où il se produit.

Le code et les sorties **SAS** sont fournis en annexe. La table **SAS** s'appelle `exam.quakes`.

1. Nous considérons la variable `mag` sans distinguer la localisation géographique dans un premier temps (voir annexe 1 et code en annexe 5).

1.1. Quelle est la magnitude moyenne, l'écart-type ?

1.2. Calculer l'étendue et l'intervalle inter-quartile.

1.3. Dessiner la *boîte à moustache*.

2. Nous créons une variable `groupe` qui vaut 1 si le séisme se produit à une longitude supérieure à 175 et 2 s'il se produit à une longitude inférieure à 175.

2.1. Nous effectuons une `proc univariate` avec la commande `by groupe` pour étudier séparément les séismes des deux groupes (voir annexe 2 et code en annexe 5). Donner et comparer les

moyennes et les variances pour chaque groupe.

2.2. Les graphiques renvoyés par la procédure sont donnés à la Figure 2. Nous superposons à l'histogramme la densité de lois Gamma bien choisies. Commenter l'histogramme et les graphiques quantiles-quantiles (pour lesquels les quantiles empiriques sont en ordonnée, et les quantiles de la loi Gamma théorique en abscisse).

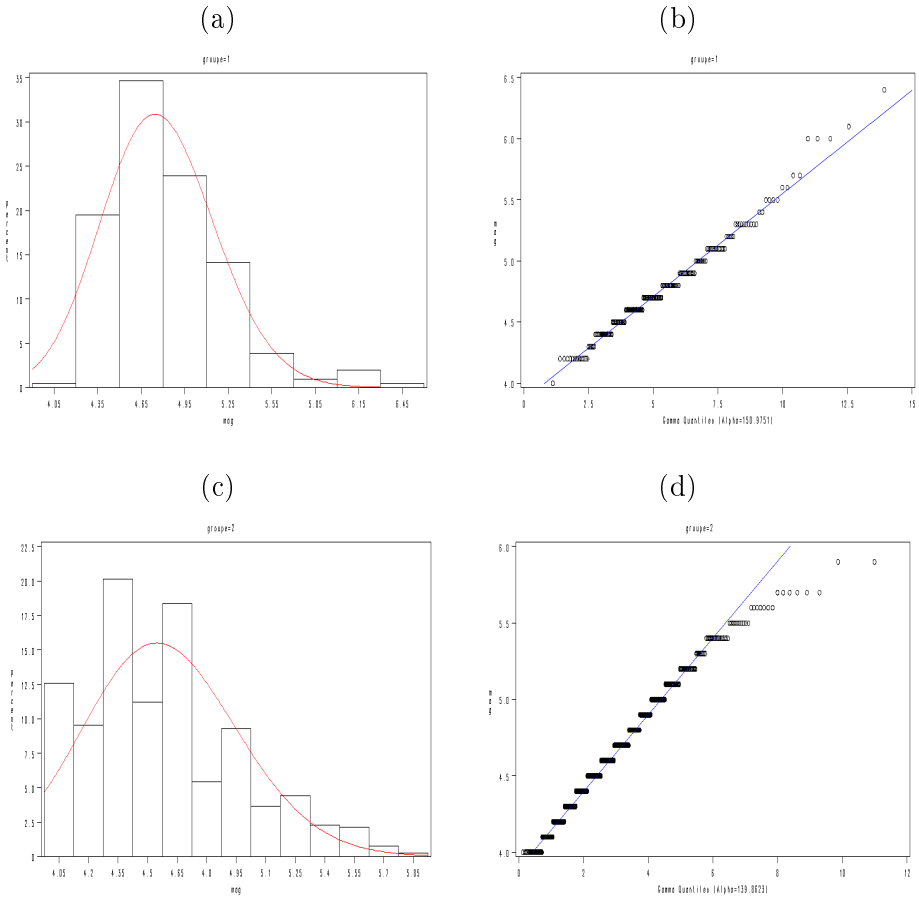


FIG. 2 – *Histogramme de la variable mag et qq-plot d'adéquation avec une loi Gamma, pour le groupe 1 (figures (a) et (b)) et pour le groupe 2 (figures (c) et (d)).*

2.3. A partir des sorties **SAS** de l'annexe 2, lire les tests d'adéquation à une loi Gamma pour les groupes 1 et 2 (on se fixe le seuil $\alpha = 5\%$). Quel est l'intérêt de tester l'adéquation de la distribution des magnitudes à une loi paramétrique (en l'occurrence une loi Gamma) ?

3. Nous étudions maintenant la décomposition de la variance pour ce découpage suivant les groupes 1 et 2.

3.1. Etablir la formule de décomposition de la variance vue en cours, et redonner l'interprétation de la variance inter et de la variance intra.

3.2. A partir de la *Seconde proc univariate*, en utilisant la commande `output` (voir annexe 5), nous obtenons la table `exam.sortiequakes` (voir ci-dessous) à partir de laquelle nous réalisons les *Troisième et quatrième proc univariate* (voir code et sorties en annexes 3, 4 et 5).

groupe	taille	moyenne	variance
1	205	4.7770731707	0.1592266858
2	795	4.58	0.1552292191

TAB. 1 – Table exam.sortiequakes

Expliquer pourquoi ces procédures nous donnent les variances inter et intra. Préciser leurs valeurs

numériques à partir des sorties **SAS** données en annexes 3 et 4.

3.3. Calculer la part de la variance expliquée par la variable `groupe` et conclure.

Exercice 3 (Intervalle de confiance)
Nous considérons n variables aléatoires *i.i.d.* X_1,...X_n de loi exponentielle $\mathcal{E}(\lambda)$ où $\lambda > 0$.

1. Rappeler ou calculer l'espérance et la variance de X_1.

2. Quelle est la loi de $S_n = \sum_{i=1}^{n} X_i$? Donner son espérance et sa variance.

3. Un estimateur naturel pour λ est $1/\bar{X}_n$. A partir du TCL, donner un théorème de normalité asymptotique pour $1/\bar{X}_n$. Construire un intervalle de confiance à 95% pour λ.

Annexe 5 : Code SAS pour l'exercice 2

```
libname exam "D:\Enseignements\TISD Lille\sujet exams";

data exam.quakes;
set exam.quakes;
if long<175 then groupe=1; else groupe=2;
run;

proc sort data=exam.quakes;
by groupe; run;

/*Premiere proc univariate*/

proc univariate data=exam.quakes;
var mag; run;

/*Seconde proc univariate*/

proc univariate data=exam.quakes;
var mag;
by groupe;
histogram mag /gamma (color=red);
qqplot mag / gamma(alpha=est sigma=est theta=est); symbol v=circle;
output out=exam.sortiequakes mean=moyenne var=variance n=taille;
run;

/*Troisieme proc univariate*/

proc univariate data=exam.sortiequakes;
var moyenne;
freq taille;
run;

/*Quatrieme proc univariate*/

proc univariate data=exam.sortiequakes;
var variance;
freq taille;
run;
```