Partiel TISD - Master Pro
Vendredi 13 novembre 2009

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 (Etude du débit du Nil)
Nous considérons les débits annuels du Nil, relevés entre 1915 et 1970. Le graphique est représenté ci-dessous.

(a) (b) (c)

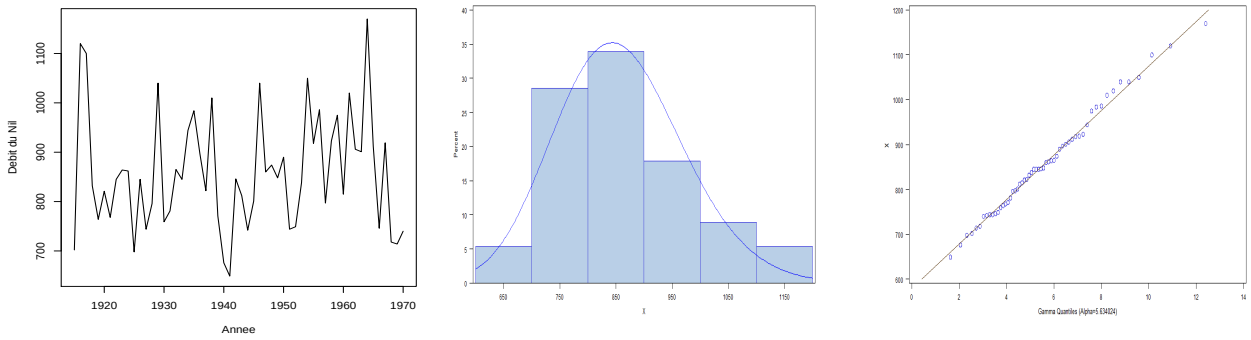

FIG. 1 – *(a) Débit annuel du Nil entre 1915 et 1970. (b) Histogramme des 56 observations du débit annuel du Nil avec densité de la loi Gamma superposée. (c) QQ-plot de la distribution empirique avec la loi Gamma ajustée sur les données.*

1. Nous notons X_t pour $t \in \{1, \ldots, 56\}$ les variables aléatoires représentant ces débits durant les années successives de 1915 à 1970. Au vu de la Figure 1 (a), que pensez-vous de l'hypothèse que ces variables aléatoires sont i.i.d. ? (hypothèse qui sera faite dans toute la suite de l'exercice)

2. On réalise la procédure `univariate` suivante :

```
proc univariate data=malib.nil plot;
var X;
histogram X / gamma (color=blue);
qqplot X / gamma(alpha=est sigma=est theta=est);
symbol v=circle;
run;
```

En vous reportant à l'annexe 1, quelle est le débit moyen ? son écart-type ? sa médiane ? l'intervalle inter-quartile ? commenter les coefficient d'asymétrie et d'aplatissement de la distribution.

3. On cherche à tester l'adéquation à une loi Gamma (voir annexe 1).
3.1. Qu'en pensez-vous au vu de l'histogramme et du graphique quantile-quantile ? (Fig. 1)

3.2. Commenter les tests d'adéquation donnés par SAS.

4. A une observation donnée, on associe deux variables supplémentaires Y_t et Z_t définies par le code suivant :

```
data malib.nil;
set malib.nil;
if X>845 then Y=1;
if X<=845 then Y=0;
if lag(X)>845 then Z=1;
if lag(X)<=845 then Z=0;
run;
```

4.1. Traduire ce code en donnant la définition "mathématique" des variables Y et Z.

On réalise un croisement de Y et Z :

```
proc freq data=malib.nil;
tables  Y*Z / chisq;
run;
```

Les résultats sont présentés à l'annexe 2, dans laquelle certains résultats ont été malencontreusement effacés.
4.2. Donner le tableau de fréquences correspondant et calculer les distributions marginales.

4.3. On cherche à tester l'indépendance de Y et Z. Pourquoi ?

4.4. Donner et calculer la statistique ξ pour le test dont on parle à la question **4.2**. Quel est son comportement quand le nombre d'observations n tend vers $+\infty$ sous H_0 ? sous l'hypothèse alternative ?

4.5. Pour commenter le test réalisé en **5.4**, on utilise **R** qui nous renvoie les résultats suivants :

```
> qchisq(0.95,1)
[1] 3.841459
> qchisq(0.95,2)
[1] 5.991465
> qchisq(0.95,4)
[1] 9.487729
```

Quelle en est la conclusion ? Commenter. Ceci concorde-t-il avec la p-valeur donnée par **SAS** dans l'annexe 2 ?

4.6. Retrouver le coefficient Φ^2. Pourquoi le Φ^2 et le V de Cramer sont-il égaux ?

5. On rappelle que si X suit une loi $\Gamma(k,\theta)$, $\mathbb{E}(X^p) = k \times \cdots \times (p+k-1) \times \theta^p$ pour tout $p \geq 1$. On suppose connu un estimateur fortement convergent $\widehat{k}$ de k. On considère l'estimateur suivant du paramètre d'échelle θ :

$$\widehat{\theta} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n X_i}.$$

5.1. Exprimer $\widehat{\theta}$ en fonction de $\bar{X}_n$ et $\overline{X_n^2} = \sum_{i=1}^n X_i^2/n$.

5.2. En utilisant le théorème central limite, obtenir un intervalle de confiance pour θ.

5.3. Faire l'application numérique en utilisant les sorties de l'annexe 1 ("échelle" se dit "scale" en anglais).

Exercice 2 (Eléphants (d'après Dussaix et Grosbras))
Un directeur de cirque possède $N = 100$ éléphants. Il souhaite estimer le poids total de son troupeau afin de les emmener en tournée en bateau. On s'intéresse à la variable poids Y et on notera Y_i le poids du $i^{\text{ème}}$ éléphant ($i \in \{1, \ldots, 100\}$). Les éléphants sont regroupés en deux classes : les mâles ($h = 1$) et les femelles ($h = 2$).

L'année précédente, le directeur avait déjà fait peser ses éléphants et trouvait les résultats suivants :

Classe h	Effectif N_h	Poids moyen Y_h	Variance corrigée S_h^2
Mâles : $h = 1$	60	6	4
Femelles : $h = 2$	40	4	2.25

1. On commence par recalculer les statistiques de l'année précédente :
1.1. Quel était le poids total P' du troupeau l'année dernière ?

1.2. Calculer le poids moyen $\bar{Y}$ d'un éléphant du troupeau.

1.3. Rappeler la définition de la variance corrigée, puis calculer la variance totale S^2 (non corrigée) de Y (attention aux renormalisations).

1.4. Pour l'étude de Y, est-il nécessaire de distinguer les éléphants mâles et femelles ?

2. Le directeur suppose que les variances de Y par classe restent celles de l'année dernière. Pour estimer le poids total $P = \sum_{i=1}^N Y_i$, il procède à un sondage aléatoire simple (SAS) de n éléphants tirés sans remise. Il constitue ainsi un échantillon s.
2.1. Recalculer les probabilités d'inclusion π_i.

2.2. Rappeler l'expression de l'estimateur d'Horvitz-Thompson $\widetilde{P}$ de P. Quel est l'estimateur associé $\widetilde{Y}$ de $\mathbb{E}(Y)$?

2.3. On rappelle que la variance de $\widetilde{P}$ est

$$\mathrm{Var}(\widetilde{P}) = N^2\Big(1 - \frac{n}{N}\Big)\frac{\bar{S}^2}{n}$$

où $\bar{S}^2$ est la variance corrigée de Y. Faire l'application numérique si l'on tire 10 éléphants pour le SAS.

3. Compte-tenu des résultats de la question **1.4**, le directeur choisit de faire un sondage stratifié. Soient n_1 et n_2 le nombre d'éléphants mâles et femelles à tirer. On pose $n = n_1 + n_2$.

3.1. Quelles sont les probabilités d'inclusion ?

3.2. Donner la définition de l'estimateur de Horvitz-Thompson $\widehat{P}$ dans ce cas. Quel est l'estimateur associé $\widehat{Y}$ pour $\mathbb{E}(Y)$?

3.3. En utilisant la question **2.3**, donner en justifiant la variance de $\widehat{P}$ en fonction de n_1, n_2, N_1, N_2, $\bar{S}_1^2$ et $\bar{S}_2^2$?

4. On considère un sondage stratifié proportionnel : le taux de sondage $f = 1/10$ est le même dans chacune des strates.

4.1. Calculer n_1 et n_2. Que devient $\widehat{Y}$ dans ce cas ?

4.2. En utilisant la question **3.3**, montrer que

$$\mathrm{Var}(\widetilde{Y}) = \frac{1-f}{n}\frac{N_1\bar{S}_1^2 + N_2\bar{S}_2^2}{N}.$$

Calculer numériquement cette variance.

4.3. Justifier pourquoi on a :

$$\mathrm{Var}(\widetilde{Y}) = \frac{1-f}{n}\Big(\frac{N_1\big(\bar{Y}_1 - \mathbb{E}(Y)\big)^2 + N_2\big(\bar{Y}_2 - \mathbb{E}(Y)\big)^2}{N-1}\Big) + \frac{N}{N-1}\mathrm{Var}(\widehat{Y}) - \frac{1-f}{n}\frac{\bar{S}_1^2 + \bar{S}_2^2}{N-1}.$$

Commenter.

4.4. Quelle est la part de réduction de la variance par rapport au SAS : $1 - \mathrm{Var}(\widehat{Y})/\mathrm{Var}(\widetilde{Y})$. Commenter.

5. On considère un sondage stratifié avec répartition de Neyman. Il consiste à choisi n_1 et n_2 suivant la règle d'allocation optimale suivante :

$$\frac{n_h}{N_h\sqrt{S_h^2}} = \text{constante} = \frac{n}{N_1\sqrt{S_1^2} + N_2\sqrt{S_2^2}}, \quad h \in \{1,2\}.$$

5.1. Calculer n_1 et n_2.

5.2. Calculer numériquement la variance de $\widehat{P}$ en utilisant la question **3.3**. Commenter.