ENSIMAG 1ère année 23 mai 2013

Principes et Méthodes Statistiques

Durée : 3 heures. Tous documents autorisés.

Les deux parties sont indépendantes.

Les résultats vus en cours ou en TD peuvent être utilisés sans être redémontrés.

Il sera grandement tenu compte de la qualité de la rédaction (présentation et justification des réponses) dans la notation.

Barème indicatif - Partie 1 : 6 pts, Partie 2 : 14 pts.

Première partie

A l'automne 2012, une équipe de biologistes français a publié une étude visant à étudier l'impact de la nourriture à base d'Organismes Génétiquement Modifiés (OGM) sur le développement de tumeurs chez les rats. Pour cela, un échantillon de 140 rats, 70 mâles et 70 femelles, a été divisé en 4 groupes. Pour chaque sexe, un groupe témoin de 10 rats a été nourri sans OGM et un groupe de 60 rats a été nourri avec du maïs génétiquement modifié à différentes doses.

Le tableau suivant donne le nombre de rats ayant développé des tumeurs avant la fin de l'étude :

	sans OGM	avec OGM	total
mâles	2	30	32
femelles	5	44	49
total	7	74	81

1. Peut-on en déduire que la prise d'OGM augmente le risque de développer des tumeurs ?

2. Le risque de développer une tumeur est-il plus fort pour les femelles que pour les mâles ?

Les réponses à ces questions doivent être données à l'aide de tests d'hypothèses : décrire les tests utilisés en donnant les hypothèses de modélisation, donner une approximation des p-valeurs, donner les commandes `R` permettant de mettre en œuvre les tests et répondre aux questions posées. Pour la question 1, interpréter les conséquences des erreurs de première et deuxième espèce.

Deuxième partie

Une variable aléatoire X positive est dite de loi de Pareto $\mathcal{P}a(a,b)$, avec $a \geq 2$ et $b > 0$, si et seulement sa densité est :

$$f(x) = \frac{a\,b^a}{(b+x)^{a+1}},\ \forall x \geq 0.$$

1. Montrer que $E(X) = \dfrac{b}{a-1}$ et $Var(X) = \dfrac{a\,b^2}{(a-1)^2\,(a-2)}$. Calculer la fonction de répartition de X.

2. Soient $x_1, \ldots, x_n$ n réalisations de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi $\mathcal{P}a(a,b)$. Calculer les estimateurs de a et b par la méthode des moments.

3. Ecrire les équations donnant les estimateurs de a et b par la méthode du maximum de vraisemblance. Exprimer l'estimateur de a en fonction de celui de b et montrer que l'estimateur de b est solution d'une équation implicite.

   Dans toute la suite, on supposera que b est connu, égal à 1.

4. Calculer l'estimateur des moments $\tilde{a}_n$ de a.

5. Calculer l'estimateur de maximum de vraisemblance $\hat{a}_n$ de a.

6. Donner la loi de probabilité de $Y = \ln(1+X)$.

7. Montrer que $\hat{a}_n$ est biaisé. En déduire un estimateur sans biais $\hat{a}'_n$. Montrer que cet estimateur est convergent.

8. Donner l'expression du graphe de probabilités pour la loi $\mathcal{P}a(a,1)$. Expliquer comment calculer un estimateur graphique a_g de a.

9. On a relevé les durées en milliers d'heures entre les défaillances successives d'un système informatique, supposées indépendantes et de même loi :

   0.252 1.017 0.094 0.980 0.046 0.449

   On donne les résultats suivants en `R` :

```
> x<-c(0.252,1.017,0.094,0.980,0.046,0.449)
> y<-log(1+x)

> mean(x)
[1] 0.473
> mean(y)
[1] 0.352
```

   La régression linéaire sur le graphe de probabilités de la question 8 donne :

```
> summary(reg)

Residuals:
        1         2         3         4         5
 0.066891 -0.045836 -0.001531 -0.047373  0.027848

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.13854    0.04005  -3.459 0.040670 *
sort(y)[1:5] -2.46096    0.10984 -22.405 0.000195 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05656 on 3 degrees of freedom
Multiple R-Squared: 0.9941,     Adjusted R-squared: 0.9921
F-statistic:   502 on 1 and 3 DF,  p-value: 0.0001947
```

Justifier le fait que l'on admette que cet échantillon est issu de la loi $\mathcal{P}a(a,1)$. Donner les valeurs des estimations $\hat{a}'_n$, $\tilde{a}_n$ et a_g. Laquelle doit-on retenir ?

10. Donner une fonction pivotale pour a. En déduire l'expression d'un intervalle de confiance bilatéral de seuil α pour a. Pour les données de l'exemple, donner cet intervalle au seuil 5%.