EISTI – DEPARTEMENT MATHEMATIQUES

EXAMEN DE STATISTIQUE INFERENTIELLE

6 juin 2017 – Durée 2h00

La consultation et l'échange de documents sont interdits
Les calculatrices sont autorisées
L'utilisation de 3 feuilles manuscrites recto-verso format A4 est autorisée

!!!!!! Ne pas oublier de rendre la feuille jointe avec votre nom et prénom !!!!!!!

Barème sur 30 donné à titre indicatif

Exercice 1 (5 points)
A compléter directement sur la feuille

Exercice 2 (8 points)
Soit un échantillon $X_1,\ldots,X_n$ issu d'une variable aléatoire X définie par sa fonction de densité,

$$f_\theta(x)=\frac{1}{\theta}e^{-\frac{1}{\theta}x}1_{]0;+\infty[}(x)$$

où $\theta>0$. On sait que $E(X)=\theta$ et $V(X)=\theta^2$.

1) Montrez que $\overline{X}$ la moyenne de l'échantillon est l'estimateur du maximum de vraisemblance de θ.
2) Est-ce que l'estimateur est sans biais ?
3) Quel est son risque quadratique ?
4) Est-il efficace ?

Barème
1) 3 pt / 2) 1pt / 3) 2pt / 4) 2pt
Même correction que loi exponentielle de paramètre 1/θ.
1) Fonction de vraisemblance

$$L(x_1,\ldots,x_n;\theta)=\prod_{i=1}^{n}f(x_i)=\prod_{i=1}^{n}\frac{1}{\theta}e^{-\frac{x_i}{\theta}}=\frac{1}{\theta^n}e^{-\frac{\sum_{i=1}^{n}x_i}{\theta}}$$

$$\Rightarrow \ln L(x_1,\ldots,x_n;\theta)=-n\ln(\theta)-\frac{1}{\theta}\sum_{i=1}^{n}x_i$$

D'où $$\frac{\partial \ln L(x_1,\ldots,x_n;\theta)}{\partial\theta}=-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}x_i=\frac{1}{\theta}\left[\frac{1}{\theta}\sum_{i=1}^{n}x_i-n\right]>0$$

$$\Leftrightarrow \left[\frac{1}{\theta}\sum_{i=1}^{n}x_i-n\right]>0 \text{ car } \theta>0$$

$$\Leftrightarrow \theta<\frac{1}{n}\sum_{i=1}^{n}x_i$$

D'après le tableau de variations, on constate la valeur ci-dessus est un maximum de $\ln L(x_1,\ldots,x_n;.)$. Donc l'EMV est

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ .$$

2) $E(\overline{X}) = E(X) = \theta$. Donc l'EMV est sans biais.

3) D'après la loi des grands nombres, la moyenne converge en probabilité vers E(X)=θ.

4) $R_{\overline{X}}(\theta) = \mathrm{var}(\overline{X})$ car l'EMV est sans biais

$$= \frac{\mathrm{var}(X)}{n} = \frac{\theta^2}{n}$$

On remarque que le risque quadratique de l'EMV tend vers 0, donc l'EMV converge en m.q. vers θ.

5) D'après le T.C.L., $\overline{X}$ converge en loi vers une loi normale de paramètre μ=θ et $\sigma^2=\theta^2/n$.

6) Le support de la loi ne dépend pas de θ, d'où

$$\frac{\partial^2 \ln L(x_1,\ldots,x_n;\theta)}{\partial\theta^2} = \frac{\partial}{\partial\theta}\left[-\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i\right] = \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n} x_i$$

$$\Rightarrow I_n(\theta) = -E\left[\frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n} X_i\right] = -\frac{n}{\theta^2} + \frac{2}{\theta^3} nE(X) = -\frac{n}{\theta^2} + \frac{2}{\theta^3} n\theta = \frac{n}{\theta^2}$$

On remarque que

$$\mathrm{var}(\overline{X}) = \frac{1}{I_n(\theta)} = \frac{\theta^2}{n}$$

Donc l'EMV est efficace.

Exercice 3 (5 points)

Le poids moyen d'un échantillon de 49 enfants nés au mois de janvier 2004 dans l'hôpital de Charleville-Mézière a été de 3,6kg avec un écart-type estimé de 0,5kg.

Barème

a) 3pt / b) 2pt

a) Déterminez un intervalle de confiance à 95% pour le poids moyen des nouveaux nés de cet hôpital.

Soit μ le poids moyen des nouveaux nés de cet hôpital. On cherche a>0 et b>0 tels que P(a≤μ≤b)=0.95.

L'estimateur usuel de μ est $\overline{X}$ la moyenne de l'échantillon. L'échantillon est assez grand pour utiliser le TCL et approcher la loi de

$$Z = \sqrt{n}\,\frac{\overline{X} - \mu}{\sigma}$$

par la loi normale N(0,1) et aussi pour remplacer σ par son estimation.

ou bien

On suppose que le poids d'un nouveau né est une variable gaussienne. La variance étant inconnue, la statistique

$$T = \sqrt{n}\,\frac{\overline{X} - \mu}{S^*}$$

suit une loi de Student à n=49-1=49 d.d.l.

D'où

$$P(a \le \mu \le b) = P\left(\sqrt{n}\,\frac{\overline{x} - b}{s^*} \le \sqrt{n}\,\frac{\overline{X} - \mu}{s^*} \le \sqrt{n}\,\frac{\overline{x} - a}{s^*}\right) = P(b' \le Z \le a') = 0.95$$

On suppose que l'intervalle est symétrique, d'où

$$\begin{cases} P(Z \le a') = 1 - \alpha/2 = 0.975 \\ b' = -a' \end{cases} \Rightarrow \begin{cases} a' = 1.96 \\ b' = -1.96 \end{cases} \Rightarrow \begin{cases} a = \bar{x} - 1.96 * s^* / \sqrt{n} = 3.46 \\ b = \bar{x} + 1.96 * s^* / \sqrt{n} = 3.74 \end{cases}$$

b) Quel serait le niveau de confiance d'un intervalle de longueur 0,1kg centré en 3,6kg pour ce poids moyen ?

On cherche α tel que

$$P(3.55 \le \mu \le 3.65) = 1-\alpha$$

De la même façon, on arrive à

$$\begin{cases} P(Z \le a') = 1 - \alpha/2 \\ a' = \sqrt{n}\,\dfrac{\bar{x} - a}{s^*} = 7 * \dfrac{3.6 - 3.55}{0.5} = 0.7 \end{cases}$$

Or d'après la table de la loi N(0,1), $P(Z \le 0.7) = 0.75804$. Donc $1-\alpha/2 = 0.75804 \Rightarrow \alpha = 2*(1-0.75804) = 0.48$

Exercice 4 (7 points)

Des études en psychologie du développement ont montré qu'à l'âge de 1 mois, 50% des bébés marchent. On souhaite mener une étude sur les retards de développement des bébés prématurés. On souhaite savoir si les bébés prématurés marchent plus tardivement. Pour cela, on observe un échantillon de 80 bébés prématurés.

Barème

1) 1pt / 2) 2pt / 3) 1pt / 4) 2 pt / 5) 1 pt

1) Etablissez les hypothèses du test.

Soit p la proportion de bébés prématurés marchant à 12 mois. On souhaite tester

$$H_0 : p = 0{,}5 \text{ contre } H_1 : p < 0{,}5$$

2) Quelle est la statistique du test (variable de décision) et sa loi ? Justifiez.

L'estimateur de p est la fréquence

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

où $X_i = 1$ si le bébé marche et 0 sinon. Les X_i sont indépendantes de loi de Bernoulli B(p). Comme l'échantillon est grand, on peut utiliser le TCL pour approcher la loi de la fréquence par une loi normale de paramètres : $\mu = p$ et $\sigma^2 = p(1-p)/n$.

3) Déterminez la région critique en fonction d'un ou deux seuils.

La région critique est de la forme : $W = \{\bar{X} < C\}$ (faire dessin si besoin)

4) Calculez la ou les valeurs du ou des seuils.

On résout l'équation : $\alpha = P(W|H_0)$. Sous l'hypothèse H_0, $\bar{X} \sim N(0.5 ; 0.5^2/80)$, d'où

$$0.05 = P(\bar{X} < C) = P\left(\frac{\bar{X} - 0.5}{0.5}\sqrt{80} < \frac{C - 0.5}{0.5}\sqrt{80}\right) = P(Z < C') \Rightarrow \text{(table) } C' = -1.6449$$

$$\Rightarrow C = 0.5 - 1.6449 * \frac{0.5}{\sqrt{80}} = 0.408$$

5) A 12 mois, 35 de ces 80 bébés marchent. Quelle conclusion peut-on en tirer ?

On obtient sur cet échantillon, $\vec{x} = 35/80 = 0.4375 > C$. Donc on garde H_0, c-à-d que le développement est normal, sans connaitre le risque.

Exercice 5 (5 points)
On souhaite vérifier que toutes les couleurs sont équitablement représentées dans un paquet de M&M's. Sur 100 bonbons, on a la répartition suivante

Couleur	Jaune	Rouge	Orange	Bleu	Marron	Vert
Effectif	15	12	23	23	12	15

Barème

1) 2pt / 2) 3pt

1. Quel test doit-on mettre en place pour répondre à la question ? Précisez
   - les hypothèses
   - la statistique du test (variable de décision) et sa loi sous H_0 ainsi que
   - la région critique en fonction d'un seuil C
   - la (les) condition(s) d'application du test

Il s'agit du test d'ajustement du chi-deux. On test si la différence entre la distribution observée et la distribution théorique (ici uniforme) est nulle (H_0) ou strictement positive (H_1).
La statistique du test est la distance du chi-deux d_n sous l'hypothèse H_0 elle suit une loi du chi-deux à 6-1=5 ddl. La région critique est de la forme : $W=\{d_n>C\}$.
Pour appliquer le test, il faut que les effectifs théoriques soient ≥ 5

2. Etablissez le test pour un risque α=5% et répondez à la question.

Dans la table du chi-deux avec 5ddl on obtient un seuil C=11,07.

Couleur	Jaune	Rouge	Orange	Bleu	Marron	Vert
Eff. Obs	15	12	23	23	12	15
Eff. Th	16,67	16,67	16,67	16,67	16,67	16,67
chi-deux	0,17	1,31	2,41	2,41	1,31	0,17

D'où d_n=7,76<C. On peut donc conclure que la répartition des couleurs est uniforme

NOM :
PRENOM :

Question 1

On effectue le test

$H_0 : \mu=0$
$H_1 : \mu>0$

La statistique du test suit une loi normale. Si on augmente la valeur du seuil de décision, alors l'erreur

β augmente ~~β diminue~~ ~~α augmente~~ α diminue

0.5 pt par bonne réponse et -0.5 pt par mauvaise réponse

Question 2

0.5 pt par question

On souhaite savoir si la consommation de crème solaire dépend de la région où on habite. On a obtenu les résultats suivants

Source des variations	*Somme des carrés*	*Degré de liberté*	*Moyenne des carrés*	*F*	*Probabilité*	*Valeur critique pour F*
Expliquée	182,46	3	60,82	70,35	9,7234E-31	2,65
Résiduelle	165,98	192	0,86			
Total	348,45	195				

1) Quel est le nom du test : ANOVA
2) Combien de régions on été étudiées : ddl « Expliquée » =3=p-1 donc 4 régions
3) Combien de personnes ont été interrogées : ddl « Total »=195=n-1 donc 196 personnes
4) Quelle est l'hypothèse H_0 : $\mu_1=\mu_2=\mu_3=\mu_4$ où μ_i consommation moyenne de la région i
5) Quelle est l'hypothèse H_1 : au moins 2 moyennes sont différentes
6) Quelle est la conclusion du test : F>F critique ou p-valeur tres petite donc on accepte H_1, c-a-d que la région a un impact significatif sur la consommation avec un risque de 5% de se tromper.

Question 3

0.5 pt par question

Voici une liste de tests statistiques

Test de Student, Test de l'ANOVA, Test de Fisher, Test du chi-deux, Test de Shapiro, Test de Wilcoxon

Pour chaque situation, précisez le test correspondant.

1) On souhaite savoir si le PIB d'un pays dépend du continent : Si on suppose que le PIB par continent est une variable aléatoire gaussienne et que la variance est la même pour tous les continents, alors on applique le test de l'ANOVA

2) On souhaite savoir si la variabilité de la production deux chaines de fabrication d'un même produit est la même : Test de comparaison de variance. Si on suppose que les échantillons sont issus d'une loi gaussienne alors on fait le test de Fisher

3) Est-ce que la taille des marsupilami suit une loi normale : Test de Shapiro ou test de Kolmogorov ou chi-deux après découpage en classes

4) On souhaite comparer le temps d'exécution de deux programmes. On lance chaque programme 20 fois : Il s'agit de comparer deux échantillons. Le temps d'exécution d'un programme suit très certainement une loi exponentielle, on préférera le test de Wilcoxon à celui de Student