EISTI – DEPARTEMENT MATHEMATIQUES

RATTRAPAGE DE STATISTIQUE INFERENTIELLE

juillet 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

Exercice 1

Soit un échantillon $X_1,\ldots,X_n$ issu d'une variable aléatoire X définie par sa fonction de densité,

$$f_\theta(x)=\theta^2 x e^{-\theta x} 1_{]0;+\infty[}(x)$$

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

1) Montrez que $2/\bar{X}$ est l'estimateur du maximum de vraisemblance de θ.
2) Calculez l'information de Fisher.

On considère maintenant $T=\bar{X}/2$ comme estimateur de $1/\theta$.

3) Est-ce que l'estimateur T est sans biais ?
4) Est-ce qu'il converge en moyenne quadratique ?

1) Fonction de vraisemblance

$$L(x_1,\ldots,x_n;\theta)=\prod_{i=1}^n f(x_i)=\prod_{i=1}^n \theta^2 x_i e^{-\theta x_i}\ ,\ x_i>0$$

$$\Rightarrow \ln L(x_1,\ldots,x_n;\theta)=-2n\ln(\theta)+\ln\left(\prod_{i=1}^n x_i\right)-\theta\sum_{i=1}^n x_i^2$$

Fonction dérivable en θ sur R_+ avec,

$$\frac{\partial \ln L(x_1,\ldots,x_n;\theta)}{\partial\theta}=-\frac{2n}{\theta}+\sum_{i=1}^n x_i>0 \Leftrightarrow \frac{1}{\theta}>\frac{1}{2n}\sum_{i=1}^n x_i=\frac{\bar{x}}{2} \Leftrightarrow \theta<\frac{2}{\bar{x}}$$

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 $2/\bar{X}$.

2) Information de Fisher

Le support de X ne dépend pas de θ **(-0.5 si pas justifié)** donc l'information de Fisher s'obtient en dérivant une deuxième fois

$$\frac{\partial^2 \ln L(x_1,\ldots,x_n;\theta)}{\partial\theta^2}=\frac{2n}{\theta^2} \Rightarrow I_n(\theta)=E\left[-\frac{\partial^2 \ln L(X_1,\ldots,X_n;\theta)}{\partial\theta^2}\right]=\frac{2n}{\theta^2}$$

3) Biais

$E(T)=E(\bar{X})/2=E(X)/2=1/\theta$

Donc T est un estimateur sans biais de $1/\theta$.

4) Risque quadratique

$R_\theta(T)=V(T)$ car l'estimateur est sans biais **(-0.5 si pas justifié)**

$$=V(\bar{X})/4=V(X)/4n=(2/\theta^2)/4n=\frac{1}{2n\theta^2}\xrightarrow[n\to+\infty]{}0$$

T converge en m.q. vers θ.

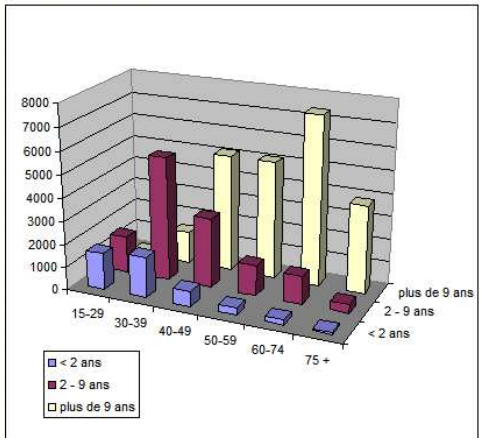
Exercice 2

Une étude de l'INSEE Alsace tente de mettre en relation la durée d'occupation d'un logement avec l'âge du locataire (de référence dans le foyer).

Age	Durée d'occupation			*total*
	< 2 ans	2 - 9 ans	plus de 9 ans	
15-29	1606	1592	140	3338
30-39	1761	5393	1393	8547
40-49	675	3053	5148	8876
50-59	341	1343	5132	6816
60-74	221	1227	7380	8828
75 +	95	393	3836	4324
Total	4699	13001	23029	40729

En fait, l'âge et la durée d'occupation sont des variables quantitatives, toutefois dans cet exercice elles peuvent être considérées comme des variables nominales.

Fig. 1 La durée d'occupation d'un logement selon l'âge de la personne de référence



Avec le graphique Figure 1, il semble que les deux variables soient liées. On va mettre en place un test d'indépendance du chi-deux pour le confirmer ?

1) Enoncez « en français » les hypothèses nulle et alternative du test, puis exprimez les à l'aide de la distribution conjointe p_{ij} des variables et des distributions marginales $p_{i.}$, $p_{.j}$
2) Comment calcule-t-on les effectifs théoriques de la distribution conjointe des deux variables ?
3) Si les effectifs observés sur l'échantillon sont égaux aux effectifs théoriques, qu'est-ce que cela signifie sur le lien entre les deux variables ?
4) Comment calcule-t-on la statistique du test ?
5) Sous l'hypothèse H_0, quelle est la loi de la statistique du test (sans oublier les degrés de liberté)?
6) Le calcul du chi-deux sur l'échantillon donne une valeur 17 033 pour une valeur critique de 18,30704 pour un seuil critique de 5%. Quelle est la conclusion ? Connait-on le risque d'erreur pour cette décision ?
7) Etant donné la valeur et la valeur critique du chi-deux, pensez-vous que la p-valeur du test est plus proche de 50% ou de 0,5% ?

1)
H_0 : la durée d'occupation et l'âge du locataire sont indépendantes : $p_{ij}=p_{i.}\times p_{.j}$
H_1 : la durée d'occupation et l'âge du locataire sont liées : $p_{ij}\neq p_{i.}\times p_{.j}$
2) On calcule la fréquence marginale de la $i^{ème}$ ligne $f_{i.}$ et on multiplie par la fréquence marginale de la $j^{ème}$ colonne. On obtient ainsi la fréquence conjointe f_{ij}. Il suffit de multiplier par l'effectif total pour avoir l'effectif théorique n_{ij}.
3) Cela signifie que $p_{ij}=p_{i.}\times p_{.j}$ donc que les variables sont indépendantes.
4) La distance du chi-deux est la somme des (eff. théorique – eff. observé)/eff. théorique.
5) Sous l'hypothèse H_0, la statistique suit une loi du chi-deux à (3-1)×(6-1) d.d.l.
6) $F>F_{5\%}$ donc on accepte H_1, c-à-d que les variables sont liées avec un risque de 5% **(-0.5 si pas le risque)**

7) On constate que $F>>F_{5\%}$ donc H_1 est très largement acceptée. On peut donc en conclure que la p-valeur est très petite donc plus proche de 0,5% que de 50%.

Exercice 3

Dans le cadre d'une étude sur la santé au travail, on a interrogé au hasard 500 salariés de différents secteurs et de différentes régions de France. 145 d'entre eux déclarent avoir déjà subi un harcèlement moral au travail. On souhaite estimer la proportion de salariés ayant déjà subi un harcèlement moral au travail.

1) Quel est la variable aléatoire observée sur l'échantillon ? Quelle est sa loi ?
2) Quel est l'estimateur de la proportion ? Quelle est sa loi (justifier) ?
3) Donner une estimation de cette proportion par un intervalle de confiance à 90%.
4) Si avec les mêmes données on calculait un intervalle de confiance à 95%, serait-il plus grand ou plus petit que celui trouvé à la question précédente? (justifier sans calcul.)

1) On observe la variable aléatoire X=1 si la personne est harcelée et 0 sinon. Elle suit une loi de Bernoulli B(p) où p est la proportion de gens harcelés au travail.
2) On utilise la fréquence (moyenne des Xi) comme estimateur de p. 2tant donnée que l'échantillon est très grand, on utilise le TCL pour approcher la loi de l'estimateur par une loi normale N(p,p(1-p)/n).
3) [0,257 ;0,323]
4) Si on augmente le niveau de confiance alors nécessairement cela agrandit l'intervalle de confiance pour qu'il y ait plus de chance que la vraie valeur soit à l'intérieur.

Exercice 4

Un négociant en vin s'intéresse à la contenance des bouteilles d'un producteur soupçonné par certains clients de frauder. Il souhaite s'assurer que cette contenance respecte bien en moyenne la limite légale de 75 cl. À cet effet, il mesure le contenu de 10 bouteilles prises au hasard et obtient une moyenne de 74,43cl et un écart-type estimé de 0,5cl.

On suppose que la contenance des bouteilles (en cl) suit une loi gaussienne.

1) Construisez un intervalle de confiance avec un niveau de risque de 10% pour la contenance moyenne des bouteilles de ce producteur.
2) Pour que 75cl appartient à cet intervalle, faut-il augmenter ou diminuer le niveau de risque ?
3) Si à la place d'avoir un écart-type estimé à 0,5cl, on suppose que 0,5cl est la valeur exacte de l'écart-type, que se passe-t-il pour l'intervalle confiance ? (Justifiez sans faire de calcul)

1) Soit X la contenance d'une bouteille. Par hypothèse on suppose que $X \sim N(\mu, 0{,}5^2)$ où μ est la contenance moyenne.
On cherche $a>0$ et $b>0$ tels que $P(a \leq \mu \leq b)=0.90$.
L'estimateur usuel de μ est $\overline{X}$ la moyenne de l'échantillon. L'échantillon gaussien de variance inconnue **(-0.5 si pas correctement justifié)** donc

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

suit une loi de Student à (10-1)=9 d.d.l.
D'où

$$P(a \le \mu \le b) = P(\sqrt{n}\frac{\bar{x}-b}{s^*} \le \sqrt{n}\frac{\bar{X}-\mu}{S*} \le \sqrt{n}\frac{\bar{x}-a}{s^*}) = P(b' \le Z \le a') = 0.90$$

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

$$\begin{cases} P(Z \le a') = 1-\alpha/2 = 0.95 \\ b' = -a' \end{cases} \Rightarrow \begin{cases} a' = 1.83 \\ b' = -1.83 \end{cases} \Rightarrow \begin{cases} a = 74.43 - 1.83 * 0.5/\sqrt{10} = 74.14 \\ b = 74.43 + 1.83 * 0.5/\sqrt{10} = 74.72 \end{cases}$$

2) Il faut un intervalle plus grand. Pour cela, on augmente le niveau de confiance donc on diminue le niveau de risque.
3) Si on connait l'écart-type alors on a moins d'incertitude, il est donc logique que l'intervalle de confiance diminue. Effectivement, dans ce cas l'estimateur suit une loi normale et le quantile associé à un risque de 10% est 1.64 à la place de 1.83.

Le négociant décide de tester l'hypothèse nulle (H_0) : μ=75 contre l'alternative (H_1) : μ <75.

4) Quel point de vue le négociant adopte-t-il en choisissant ces hypothèses ? Justifiez précisément la réponse.
5) Construisez, à l'aide d'une règle de décision intuitive basée sur la moyenne empirique, un test de niveau 10% de (H_0) contre (H_1). Quelle est la conclusion de ce test ?
6) Sur un graphique, dessinez les erreurs de première et deuxième espèces.

4) H_0 : La contenance moyenne est correcte / H_1 : la contenance moyenne est en dessous
Risque de 1ère espèce : Dire que la contenance moyenne est en dessous de ce qui est affiché alors que c'est faux.
Le test est du point de vue du producteur qui ne veut pas qu'on l'accuse à tord.
5) La statistique du test est toujours la moyenne. La région critique est de la forme W={$\bar{X}$<C}. Pour trouver C, on résout l'équation : $\alpha = P(W|H_0)$. Sous l'hypothèse H_0,

$$T = \sqrt{n}\frac{\bar{X}-75}{S^*}$$

suit une loi de Student à (10-1)=9 d.d.l., d'où

$$0.10 = P(\bar{X} < C) = P(\frac{\bar{X}-75}{0.5}\sqrt{10} < \frac{C-75}{0.5}\sqrt{10}) = P(T < C') \Rightarrow \text{(table) } C' = -1.383$$

$$\Rightarrow C = 75 - 1.383 * \frac{0.5}{\sqrt{10}} = 74.78.$$

Sur l'échantillon, on a calculé $\bar{x} = 74.43 < 74.78$. Donc on peut dire que la contenance moyenne est en-dessous de ce qui est affiché avec un risque de 10% de se tromper **(-0.5 si pas le risque).**
6)…