CORRECTION EXAMEN STATISTIQUE - 1TR

Lundi 28 Novembre 2016

Partiel sans document (Une feuille A4 recto-verso autorisée)

Exercice 1: Test statistique

On considère n variables aléatoires $Y_1,...,Y_n$ indépendantes de même loi normale $\mathcal{N}(0,\sigma^2)$ (σ^2 étant un paramètre connu) et on définit une suite de variable aléatoire X_i comme suit

$$X_i = ra_i + Y_i, \quad i = 1, ..., n$$

où $a = (a_1, ..., a_n)^T$ est un vecteur de paramètres connu (avec $a \neq 0$) et r est un paramètre inconnu dont on cherche à tester la valeur. On considère le test d'hypothèses simples

$$H_0: r = r_0, \quad H_1: r = r_1 \quad (\text{avec } r_1 > r_0).$$

- 1. Quelle est la loi de la variable aléatoire X_i ? $R\acute{e}ponse: X_i$ étant obtenue par transformation affine de Y_i , cette variable suit une loi normale. Sa moyenne est $E[X_i] = ra_i$ et sa variance est $var[X_i] = \sigma^2$. On en déduit $X_i \sim \mathcal{N}(ra_i, \sigma^2)$.
- 2. Montrer que le test de Neyman Pearson conduit à la statistique de test

$$T_n = \sum_{i=1}^n a_i X_i$$

et indiquer la région critique de ce test. Déterminer loi de T_n sous les deux hypothèses H_0 et H_1 . Réponse: puisque les variables aléatoires X_i sont indépendantes, la la vraisemblance des observations $x_1, ..., x_n$ est

$$p(x_1, ..., x_n; r) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (x_i - ra_i)^2\right] = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - ra_i)^2\right].$$

Le test de Neyman Pearson est défini par

Rejet de
$$H_0$$
 si $\frac{p(x_1, ..., x_n; r_1)}{p(x_1, ..., x_n; r_0)} > S_{\alpha}$

où S_{α} est un seuil dépendant du risque de première espèce α . Mais

$$\frac{p(x_1, ..., x_n; r_1)}{p(x_1, ..., x_n; r_0)} > S_\alpha \Leftrightarrow \frac{\exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - r_1 a_i)^2\right]}{\exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - r_0 a_i)^2\right]} > S_\alpha \Leftrightarrow \sum_{i=1}^n a_i x_i > K_\alpha$$

où la dernière égalité a été obtenue en utilisant la condition $r_1 > r_0$. La région critique du test est l'ensemble des vecteurs $(x_1,...,x_n) \in \mathbb{R}^n$ tels que $\sum_{i=1}^n a_i x_i > K_\alpha$ et la statistique de test est

$$T_n = \sum_{i=1}^n a_i X_i.$$

La statistique de test peut s'écrire sous la forme $T_n = \mathbf{a}^T \mathbf{X}$, avec $\mathbf{X} = (X_1, ..., X_n)^T$ et où \mathbf{a}^T est une matrice de rang 1. Donc T_n suit une loi normale de moyenne

$$E[T_n] = \sum_{i=1}^{n} a_i E[X_i] = r \left(\sum_{i=1}^{n} a_i^2\right)$$

et de variance

$$\operatorname{var}[T_n] = \sum_{i=1}^n a_i^2 \operatorname{var}[X_i] = \sigma^2 \left(\sum_{i=1}^n a_i^2 \right)$$

Donc, sous H_0 on a

$$T_n \sim \mathcal{N}\left(r_0\left(\sum_{i=1}^n a_i^2\right), \sigma^2\left(\sum_{i=1}^n a_i^2\right)\right)$$

tandis que sous H_1

$$T_n \sim \mathcal{N}\left(r_1\left(\sum_{i=1}^n a_i^2\right), \sigma^2\left(\sum_{i=1}^n a_i^2\right)\right).$$

3. On note

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$$

la fonction de répartition d'une loi normale $\mathcal{N}(0,1)$ et F^{-1} son inverse. Déterminer la valeur du seuil K_{α} du test de Neyman Pearson en fonction de r_0, α, σ , des paramètres a_i et de F^{-1} . Réponse : Le risque α est défini par

$$\alpha = P[\text{Rejeter } H_0 | H_0 \text{ vraie}] = P\left[T_n > K_\alpha | T_n \sim \mathcal{N}\left(r_0\left(\sum_{i=1}^n a_i^2\right), \sigma^2\left(\sum_{i=1}^n a_i^2\right)\right)\right].$$

On pose $||a||^2 = \sum_{i=1}^n a_i^2$ et on obtient

$$\alpha = P\left[\frac{T_n - r_0 ||a||^2}{\sigma \sqrt{||a||^2}} > \frac{K_\alpha - r_0 ||a||^2}{\sigma \sqrt{||a||^2}}\right] = 1 - F\left[\frac{K_\alpha - r_0 ||a||^2}{\sigma ||a||}\right].$$

On a donc

$$K_{\alpha} = \sigma ||a|| F^{-1} (1 - \alpha) + r_0 ||a||^2.$$

4. Déterminer la puissance du test en fonction du seuil K_{α} , r_1 , σ , des paramètres a_i et F. Déterminer les courbes COR du test étudié dans cet exercice et tracer la forme de ces courbes pour différentes valeurs du couple (r_0, r_1) et pour différentes valeurs du paramètre σ . Comment ces courbes COR évoluent-elles en fonction des paramètres a_i ? Expliquer.

 $\it R\'eponse$: la puissance du test est définie par $\pi=1-\beta$ et se calcule comme suit

$$\pi = 1 - \beta = P[\text{Rejeter } H_0 | H_1 \text{ vraie}] = P\left[T_n > K_\alpha | T_n \sim \mathcal{N}\left(r_1\left(\sum_{i=1}^n a_i^2\right), \sigma^2\left(\sum_{i=1}^n a_i^2\right)\right)\right].$$

D'où

$$\pi = 1 - F \left[\frac{K_{\alpha} - r_1 ||a||^2}{\sigma ||a||} \right].$$

On obtient la courbe COR en remplaçant K_{α} par son expression en fonction de α , soit

$$\pi = 1 - F \left[-\frac{(r_1 - r_0)}{\sigma} ||a|| + F^{-1} (1 - \alpha) \right].$$

On peut donc remarquer que le test est d'autant plus puissant que $r_1 - r_0$ est grand ou que σ est petit, ce qui est logique. Par ailleurs la performance de ce test est d'autant meilleure que ||a|| est grand.

Exercice 2 : Une vraisemblance capricieuse (8 points)

On considère n variables aléatoires $X_1, ..., X_n$ indépendantes suivant la même loi continue de loi normale $\mathcal{N}(am, a^2\sigma^2)$ où m est un paramètre connu et a est un paramètre inconnu.

1. En utilisant la valeur de $E[X_i]$, déterminer un estimateur des moments de a noté \widehat{a}_{Mo} . Déterminer le biais et la variance de cet estimateur. L'estimateur \widehat{a}_{Mo} est-il l'estimateur efficace de a? $R\acute{e}ponse$: on a $E[X_i] = am$, d'où $a = \frac{1}{m}E[X_i]$. On en déduit l'estimateur des moments

$$\widehat{a}_{\mathsf{Mo}} = \frac{1}{mn} \sum_{i=1}^{n} X_i.$$

La moyenne de cet estimateur est

$$E[\widehat{a}_{\mathsf{Mo}}] = \frac{1}{mn} \sum_{i=1}^{n} (am) = a$$

et donc il est non biaisé. La variance de \hat{a}_{Mo} est

$$\operatorname{var}[\widehat{a}_{\mathrm{Mo}}] = \frac{1}{m^2 n^2} \sum_{i=1}^{n} (a^2 \sigma^2) = \frac{a^2 \sigma^2}{n m^2}.$$

L'estimateur \hat{a}_{Mo} est donc convergent. Pour vérifier s'il est efficace, il faut déterminer la borne de Cramer-Rao d'un estimateur non biaisé de a. La vraisemblance de ce modèle statistique est définie par

$$p(x_1, ..., x_n; a) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi a^2 \sigma^2}} \exp\left[-\frac{1}{2a^2 \sigma^2} (x_i - ma)^2\right]$$
$$= \frac{1}{(2\pi a^2 \sigma^2)^{n/2}} \exp\left[-\frac{1}{2a^2 \sigma^2} \sum_{i=1}^{n} (x_i - am)^2\right].$$

On peut donc calculer les dérivées de la log-vraisemblance. La première dérivée s'écrit

$$\frac{\partial \ln p(x_1, ..., x_n; a)}{\partial a} = -\frac{n}{a} + \frac{1}{a^3 \sigma^2} \sum_{i=1}^n (x_i - am)^2 + \frac{m}{a^2 \sigma^2} \sum_{i=1}^n (x_i - am)$$
$$= -\frac{n}{a} + \frac{nS_n^2}{a^3 \sigma^2} - \frac{nm\overline{x}}{a^2 \sigma^2}$$

avec

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$
, et $S_n^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2$.

La seconde dérivée de la log-vraisemblance a pour expression

$$\frac{\partial^2 \ln p(x_1, ..., x_n; a)}{\partial a^2} = \frac{n}{a^2} - \frac{3n(\sigma^2 + 1)}{a^2 \sigma^2} + \frac{2n}{a^2 \sigma^2}.$$

Puisque $E[\overline{X}] = a$ et $E[S_n^2] = a^2(\sigma^2 + 1)$, on en déduit

$$E\left[\frac{\partial^2 \ln p(X_1,...,X_n;a)}{\partial a^2}\right] = E\left[\frac{n}{a^2} - \frac{3nS_n^2}{a^4\sigma^2} + \frac{2nm\overline{x}}{a^3\sigma^2}\right] = \frac{-n(2\sigma^2 + m^2)}{a^2\sigma^2}.$$

La borne de Cramer-Rao d'un estimateur non biaisé de a est donc

$$BCR(a) = \frac{a^2 \sigma^2}{n(2\sigma^2 + m^2)}.$$

Comme la variance de \widehat{a}_{Mo} est différente de BCR(a), l'estimateur \widehat{a}_{Mo} n'est pas l'estimateur efficace du paramètre a.

2. Déterminer la vraisemblance associée à l'échantillon $(X_1,...,X_n)$ notée $L(x_1,...,x_n;a)$. Montrer que

$$\frac{\partial \ln L(x_1, ..., x_n; a)}{\partial a} = -\frac{n}{a^3 \sigma^2} \left[a^2 \sigma^2 + m a \overline{x} - S_n^2 \right]$$

avec

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$
, et $S_n^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2$.

Etudier les variations de la fonction $\frac{\partial \ln L(x_1,...,x_n;a)}{\partial a}$ et montrer que la vraisemblance admet deux extrema locaux que l'on déterminera. Expliquer la méthode à utiliser pour déterminer l'estimateur du maximum de vraisemblance du paramètre a (on ne demande pas de le faire).

Réponse : on a trouvé à la question précédente

$$\frac{\partial \ln L(x_1, ..., x_n; a)}{\partial a} = -\frac{n}{a} + \frac{nS_n^2}{a^3 \sigma^2} - \frac{nm\overline{x}}{a^2 \sigma^2}$$

qui peut bien se mettre sous la forme demandée. Le sens de variation de la vraisemblance en fonction des valeurs de a s'obtient en étudiant le signe de

$$g(a) = a^2 \sigma^2 + ma\overline{x} - S_n^2.$$

C'est une équation du second degré de déterminant

$$\Delta = m^2 \overline{x}^2 + 4S_n^2 \sigma^2$$

qui admet les deux racines suivantes

$$a_1 = \frac{1}{2\sigma^2} \left(-m\overline{x} + \sqrt{m^2\overline{x}^2 + 4S_n^2\sigma^2} \right) \text{ et } a_2 = \frac{1}{2\sigma^2} \left(-m\overline{x} - \sqrt{m^2\overline{x}^2 + 4S_n^2\sigma^2} \right).$$

La dérivée de la log-vraisemblance s'écrit donc

$$\frac{\partial \ln L(x_1, ..., x_n; a)}{\partial a} = -\frac{n(a - a_1)(a - a_2)}{a^3}.$$

Elle est donc du signe de $(a-a_1)(a-a_2)$ pour a<0 et du signe opposé à $(a-a_1)(a-a_2)$ pour a>0. On peut alors faire une tableau de variations indiquant le signe de la dérivée de la log-vraisemblance et les variations de la vraisemblance. On en déduit alors facilement que la vraisemblance admet les deux extrema locaux $a=a_1$ et $a=a_2$. Pour déterminer l'estimateur du maximum de vraisemblance, il suffit donc de remplacer les valeurs de a_1 et a_2 dans la vraisemblance et de déterminer laquelle donne la valeur plus grande. On en déduirait les résultats suivants (ce n'était pas demandé ici)

Pour
$$\overline{x} > 0$$
, $\widehat{a}_{MV} = \frac{1}{2\sigma^2} \left(-m\overline{X} + \sqrt{m^2 \overline{X}^2 + 4S_n^2 \sigma^2} \right)$

et

Pour
$$\overline{x} < 0$$
, $\widehat{a}_{\text{MV}} = \frac{1}{2\sigma^2} \left(-m\overline{X} - \sqrt{m^2 \overline{X}^2 + 4S_n^2 \sigma^2} \right)$.

Exercice 3: Test d'adéquation (4 points)

Pour tester si un dé est truqué ou pas, on le lance 60 fois et on observe le résultat du dé à chaque lancer. On regroupe ces observations dans le tableau suivant

Face du dé	1	2	3	4	5	6
Nombre d'observations	10	8	10	15	5	12

Pour décider si le dé est truqué ou pas, on effectue un test du χ^2 .

1. Calculer la statistique du test du χ^2 .

Réponse : le test du χ^2 considéré pour résoudre se problème correspond aux 6 classes suivantes

$$C_1 = \{1\}, C_2 = \{2\}, C_3 = \{3\}, C_4 = \{4\}, C_5 = \{5\}, C_6 = \{6\}$$

qui sont équiprobables sous l'hypothèse H_0 (dé non truqué). On a donc

$$P(C_i) = p_i = \frac{1}{6}, \quad \forall i = 1, ..., 6.$$

La statistique du test du χ^2 est

$$\phi = \sum_{i=1}^{6} \frac{(N_i - np_i)^2}{np_i}$$

$$= \frac{1}{10} \sum_{i=1}^{6} (N_i - 10)^2$$

$$= \frac{1}{10} [0 + 4 + 0 + 25 + 25 + 4]$$

$$= \frac{58}{10} = 5.8.$$

2. Déterminer le seuil du test du χ^2 en fonction de α et de l'inverse de la fonction de répartition d'une loi du χ^2 dont on précisera le nombre de degrés de liberté.

Réponse : La régle de décision du test du χ^2 est

Rejet de l'hypothèse H_0 si $\phi > s_{\alpha}$,

avec

$$\alpha = P \left[\text{Rejeter } H_0 | H_0 \text{vraie} \right]$$

$$= P \left[\phi > K_{\alpha} | H_0 \text{ vraie} \right]$$

$$= P \left[\phi > K_{\alpha} | \phi \sim \chi_5^2 \right]$$

$$= 1 - F_5 \left(K_{\alpha} \right).$$

où F_5 est la fonction de répartition de la loi du χ_5^2 , d'où

$$K_{\alpha} = F_5^{-1}(1 - \alpha).$$

3. Pourquoi a-t-on $K_{0.05} < K_{0.01}$, où K_{α} est le seuil du test du χ^2 associé à un risque de première espèce α ?

Réponse : Comme α est la probabilité de rejeter H_0 sachant que l'hypothèse H_0 est vraie et qu'on rejette H_0 quand $\phi > K_{\alpha}$, quand α augmente, on rejette plus souvent H_0 et donc le seuil K_{α} diminue, ce qui se traduit par $K_{0.05} < K_{0.01}$.

4. Peut-on calculer la puissance du test? Pourquoi?

Réponse : on ne peut pas calculer la puissance du test définie par

$$\pi = 1 - \beta = P$$
 [Rejeter $H_0|H_1$ vraie]

car la loi de la statistique de test ϕ n'est pas connue sous l'hypothèse H_1 .