

**UNIVERSITE PARIS 1 PANTHEON - SORBONNE**  
**LICENCE DE SCIENCES ECONOMIQUE ET SOCIALE**  
**Statistiques - Interrogation du 16 mai 2001**

**Durée : 45 minutes**

Une maladie contagieuse inconnue touchant les bovins vient de se déclarer en France et vous êtes chargé(e) par le ministère de la santé d'organiser la campagne de dépistage de cette maladie. Cette maladie se caractérise par un taux moyen élevé d'un certain virus dans l'organisme de l'animal.

Le ministère vous a imposé d'effectuer 3 mesures du taux de ce virus par animal, et votre travail consiste à élaborer un test statistique permettant, en fonction de ces 3 mesures, de déterminer si un animal est malade ou non.

Une première étude sur ce virus a été réalisée. Elle indique que le taux observé lors d'une mesure peut être considéré comme une variable aléatoire  $X$  suivant une loi normale  $N(m, 4)$ , où l'espérance  $m$  inconnue est le taux moyen du virus dans l'organisme de l'animal ( $m$  est exprimée en pourcentage). L'étude indique également qu'on peut considérer différentes mesures effectuées sur un même animal comme indépendantes.

1) Etant données ces hypothèses, dans quel modèle statistique devez-vous travailler ?

L'étude précédente indique qu'il n'y a que deux valeurs possibles de  $m$  : soit  $m = 20$  et l'animal est sain, soit  $m = 30$  et l'animal est malade.

2) Quelle hypothèse de base choisissez-vous ? Pourquoi ?

3) Vous choisissez un seuil  $\alpha = 5\%$ . Que représente ce seuil ? Que représente la puissance d'un test ?

4) Qu'est-ce que le test le plus puissant de seuil  $\alpha$  ? Effectuez ce test, sans démontrer la forme de la région de rejet.

5) Déterminez la puissance de ce test.

Une seconde étude, plus poussée, sur ce virus vient d'être réalisée. Elle indique que chez un animal sain, le taux moyen de ce virus peut prendre toute valeur strictement inférieure à 30%. Cette étude indique également que lorsqu'un animal contracte la maladie, ce taux monte à 30%, puis reste constant jusqu'à la mort de l'animal.

6) La prise en compte de cette nouvelle étude vous conduit-elle à modifier votre test ? Pourquoi ?

**UNIVERSITE PARIS 1 PANTHEON - SORBONNE**  
**LICENCE DE SCIENCES ECONOMIQUE ET SOCIALE**  
**Statistiques - Corrigé Interrogation du 16 mai 2001**

1) Le modèle statistique est  $X_1, X_2, X_3 \stackrel{iid}{\sim} \mathcal{N}(m, 4)$ , où  $m$  est inconnu.

2) On veut tester l'hypothèse  $S$ ="l'animal est sain" contre l'hypothèse  $M$ ="l'animal est malade". D'après la première étude,  $S$ ="m = 20" et  $M$ ="m = 30". Il est clair que pour le ministère de la santé, l'erreur la plus grave consiste à "laisser passer" un animal malade. On prend donc  $M$  comme hypothèse de base et  $S$  comme hypothèse alternative.

3) Le seuil du test est la probabilité de rejeter l'hypothèse de base alors qu'elle est vraie. Ici, c'est donc la probabilité de laisser passer un animal sain.

L'autre erreur possible consiste à accepter  $M$  alors que  $S$  est vraie. La probabilité de commettre cette seconde erreur est  $1 - \pi$ , où  $\pi$  est la puissance du test. En effet, la puissance du test est par définition la probabilité de rejeter l'hypothèse de base alors que l'hypothèse alternative est vraie.

4) Le test le plus puissant de seuil  $\alpha$  a une puissance supérieure à celle de tout autre test de même seuil. On sait que ce test est le test de Neyman, dont la forme de la région de rejet est donnée par  $\frac{L(X_1, X_2, X_3 | M)}{L(X_1, X_2, X_3 | S)} \geq A$ .

Comme  $20 < 30$ , on sait d'après le formulaire que cette équation est équivalente à  $\bar{X} \leq B$ , où  $\bar{X}$  est la moyenne arithmétique des  $X_i$ . Le nombre  $B$  est déterminé par le seuil  $\alpha$  : on a  $\bar{X} \sim \mathcal{N}\left(m, \frac{4}{3}\right)$ , et donc  $\frac{\bar{X} - m}{2/\sqrt{3}} \sim \mathcal{N}(0, 1)$ . On lit sur une table  $P(\mathcal{N}(0, 1) \leq -1.645) = \alpha = 0.05$ . On en déduit que  $\frac{B - 30}{2/\sqrt{3}} = -1.645$ , et donc que  $B = 30 - 1.645 \frac{2}{\sqrt{3}} = 28.10$ .

On rejette donc  $M$  si la moyenne des taux observés est inférieure à 28,10%.

5)  $\pi = P(\bar{X} \leq B | S) = P\left(\mathcal{N}(0, 1) \leq \frac{B - 20}{2/\sqrt{3}}\right) = P(\mathcal{N}(0, 1) \leq 7.02) \simeq 1$ .

Le test est donc très puissant.

6) D'après la seconde étude, on a toujours  $M$ ="m = 30", mais par contre  $S$  devient  $S'$ ="m < 30", qui est une hypothèse multiple. Dans la question 4, on voit que la forme de la région de rejet ne dépend pas de la valeur de  $m$  prise pour l'hypothèse simple  $S$ , pourvu que cette valeur soit inférieure à 30. On en déduit que le test précédent est le test uniformément le plus puissant de  $M$  contre  $S'$ . On ne modifie donc pas notre test.

**UNIVERSITE PARIS 1 PANTHEON - SORBONNE**  
**LICENCE DE SCIENCES ECONOMIQUE ET SOCIALE**  
**Statistiques - Interrogation du 10 avril 2002**

**30 minutes - Sans calculatrice**

On considère un  $n$ -échantillon  $(X_1, \dots, X_n)$  de loi  $\mathcal{N}(m, 1)$ , où  $m$  est inconnu. On estime  $m$  par la moyenne empirique du  $n$ -échantillon notée  $\hat{m} = \frac{1}{n} \sum_{i=1}^n X_i$ .

- 1) (2 points)** Déterminez la loi de  $\hat{m}$ .
- 2) (4 points)** Définissez puis calculez la vraisemblance et la log-vraisemblance du paramètre  $m$ . Montrez que  $\hat{m}$  est efficace.
- 3) (4 points)** Soit  $X$  une variable aléatoire de même loi que les  $X_i$  et indépendante des  $X_i$ . Déterminez un intervalle de prévision à 90% pour  $X$ .

**UNIVERSITE PARIS 1 PANTHEON - SORBONNE**  
**LICENCE DE SCIENCES ECONOMIQUE ET SOCIALE**  
**Statistiques - Corrigé Interrogation du 10 avril 2002**

1)  $\hat{m}$  est la moyenne empirique des  $X_i$ . On sait donc que son espérance est  $m$  et sa variance  $\frac{1}{n}$ . (si vous refaites les calculs, vous devez utiliser l'indépendance des  $X_i$  pour le calcul de la variance, mais pas pour celui de l'espérance.) Comme de plus  $\hat{m}$  est une combinaison linéaire de lois normales, on a  $\hat{m} \approx \mathcal{N}\left(m, \frac{1}{n}\right)$ .

2) La vraisemblance est :

$$\begin{aligned} l(x_1, \dots, x_n, m) &= P(X_1 = x_1, \dots, X_n = x_n) \\ &= \prod_{i=1}^n P(X_i = x_i) \text{ car les } X_i \text{ sont indépendantes} \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - m)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - m)^2\right) \end{aligned}$$

Donc la log-vraisemblance est :

$$\begin{aligned} L(x_1, \dots, x_n, m) &= \ln(l(x_1, \dots, x_n, m)) \\ &= -n \ln(\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^n (x_i - m)^2 \end{aligned}$$

Il ne vous était pas demandé de montrer que  $\hat{m}$  est l'estimateur du maximum de vraisemblance de  $m$ , mais seulement de montrer que  $\hat{m}$  est efficace. Comme le support de la loi normale ne dépend pas de  $m$ , on peut utiliser le théorème de Cramer-Rao :  $\hat{m}$  est efficace si sa variance est égale à  $\frac{1}{I(m)}$ , où  $I(m) = E\left(-\frac{\partial^2 L}{\partial m^2}\right)$ . On a :

$$\begin{aligned} \frac{\partial L}{\partial m} &= \sum_{i=1}^n (x_i - m) = \sum_{i=1}^n x_i - nm \\ \frac{\partial^2 L}{\partial m^2} &= -n \end{aligned}$$

Donc  $I(m) = n$ , donc  $\frac{1}{I(m)} = \frac{1}{n}$ , qui est bien la variance de  $\hat{m}$ .

3) Ceux qui ont mal lu l'énoncé et ont fait un intervalle de confiance pour  $m$  ont été notés sur 2 points. La meilleure prévision ponctuelle de  $X$  est  $\hat{m}$ . Soit  $\mathcal{E} = X - \hat{m}$  l'erreur de prévision.  $\mathcal{E}$  suit une loi normale car c'est une combinaison linéaire de lois normales. On calcule  $E(\mathcal{E}) = 0$  et  $V(\mathcal{E}) = 1 + \frac{1}{n}$  (en utilisant l'indépendance pour le calcul de la variance). Donc  $\mathcal{E} \approx \mathcal{N}\left(0, 1 + \frac{1}{n}\right)$ , donc

$\frac{\mathcal{E}}{\sqrt{1 + \frac{1}{n}}} \approx \mathcal{N}(0, 1)$ . On lit sur la table  $\mathcal{N}(0, 1)$  que  $P\left(\left|\frac{\mathcal{E}}{\sqrt{1 + \frac{1}{n}}}\right| \leq 1,64\right) = 0,9$ .  
 Donc l'intervalle de prévision à 90% est  $\left[\hat{m} - 1,64\sqrt{1 + \frac{1}{n}}, \hat{m} + 1,64\sqrt{1 + \frac{1}{n}}\right]$ .