Statistiques appliquées (L3) **Antoine Bozio**
Correction **2005-2006**

Correction de l'interrogation écrite 2

Vrai ou faux ? (3 points)

1. FAUX : Un estimateur est sans biais si et seulement si son espérance est égale au paramètre estimé, $E(\theta) = \theta$. La notion de convergence nécessite une condition supplémentaire sur la variance asymtotique de l'estimateur (qui tend vers 0).
2. FAUX : La moyenne empirique d'une variable X_i qui suit une loi normale de paramètres m et σ suit une loi normale de paramètres m et $\frac{\sigma}{n}$.
3. FAUX : Avec un échantillon, la vraie valeur du paramètre est soit dans l'intervalle de confiance, soit n'y est pas. Un intervalle de confiance à 95 % signifie que si on dispose de 100 échantillons, dans 95 cas la vraie valeur du paramètre estimé se trouve dans l'intervalle de confiance. La probabilité de 95 % concerne donc les tirages d'échantillon.
4. FAUX : La variance d'une loi de Bernouilli de paramètre p est $p(1-p)$.
5. VRAI
6. FAUX : C'est l'inverse.

Le Paludisme

A - Test d'un vaccin

1. Soit l'objectif est de supprimer les signes de malaria, soit l'objectif est d'améliorer la survie des patients. Le taux de survie devrait être l'objectif du vaccin. En effet, certains patients peuvent déjà être immunisés contre la malaria naturellement ou présenter des signes de malaria mais non mortels. On devrait ainsi juger l'efficacité d'un vaccin contre le paludisme au taux de survie final des patients.
2. La survie suit une loi de Bernouilli. Soit le patient survit (égal à 1) soit il meurt (égal à 0). La probabilité de survie correspond donc à la fréquence de l'événement 1 (le patient survit). La densité d'une loi de Bernouilli s'écrit : $f(x) = p^x(1-p)^{1-x}$. On peut calculer l'espérance :

$$E(X) = \sum x\ f(x) = \sum_{x=0} x\ p^x(1-p)^{1-x} + \sum_{x=1} x\ p^x(1-p)^{1-x}$$

$$E(X) = \sum_{x=1} x\ p^x = p$$

   Puis calculer la variance :

$$V(X) = E(X^2) - [E(X)]^2 = \sum_{x=0} x^2\ p^x(1-p)^{1-x} + \sum_{x=1} x^2\ p^x(1-p)^{1-x} - p^2$$

$$V(X) = p - p^2 = p(1-p)$$

3. Le meilleur estimateur de la probabilité de survie est la moyenne empirique de la fréquence de survie : $\hat{F} = \frac{1}{n}\sum x_i$ On a vu que $x_i \rightsquigarrow B(1,p)$, du coup :

$$\sum x_i \rightsquigarrow B(1,n)$$

Pour des grands échantillons, on peut approximer la loi binomiale par une loi normale :

$$B(1,n) \approx N(np, np(1-p))$$

On en déduit classiquement :

$$\hat{F} \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

$$\frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow N(0,1)$$

On lit dans la table de la loi normale $P(|U| < 1,96) = 0,95$ d'où :

$$p\left(\left|\frac{F-p}{\sqrt{\frac{p(1-p)}{n}}}\right| < 1,96\right) = 0,95$$

Pour approcher l'inégalité nous remplaçons p au dénominateur par son estimation f et nous obtenons :

$$p\left(-1,96 < \frac{F-p}{\sqrt{\frac{f(1-f)}{n}}} < 1,96\right) = 0,95$$

$$p\left(F - 1,96\sqrt{\frac{f(1-f)}{n}} < p < F + 1,96\sqrt{\frac{f(1-f)}{n}}\right) = 0,95$$

$$f - 1,96\sqrt{\frac{f(1-f)}{n}} < p < f + 1,96\sqrt{\frac{f(1-f)}{n}}$$

La fréquence observée f est égale à 0,7 et n est égal à 1000. On en déduit que l'intervalle de confiance à 95 % de la probabilité de survie pour cet échantillon est : $[0,672\ ;\ 0,728]$

4. Ce test ne permet pas de vérifier l'efficacité du vaccin puisque on ne sait pas quelle aurait été la probabilité de survie en l'absence du vaccin.

   Le test complémentaire indispensable est un test témoin sur un échantillon de 1000 souris à qui on injecte aussi le parasite mais sans vaccin. Il s'agit d'un groupe de contrôle qui permet véritablement de tester l'efficacité du vaccin.

5. Pour tester la différence entre deux fréquences F_1 et F_2, il suffit de tester que la différence $F_1 - F_2$ est différente de 0. Ainsi, on a :

   $$F_1 \rightsquigarrow \mathcal{N}\left(p_1; \frac{p_1(1-p_1)}{n_1}\right)$$

   $$F_2 \rightsquigarrow \mathcal{N}\left(p_2; \frac{p_2(1-p_2)}{n_2}\right)$$

   avec F_1 la fréquence du premier échantillon et F_2 la fréquence du second échantillon. On a alors :

   $$F_1 - F_2 \rightsquigarrow \mathcal{N}\left(p_1 - p_2; \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

   Sous H_0, c'est-à-dire sous l'hypothèse d'absence de différence, $p_1 = p_2 = p$. Sur les deux échantillons le taux de survie moyen est de 0,5 ($\frac{0,7+0,3}{2} = 0,5$). On a donc :

   $$F_1 - F_2 \rightsquigarrow_{H_0} \mathcal{N}\left(0; p(1-p)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]\right)$$

La région critique associée au test est donc :

$$\left\{\left|\frac{F_1 - F_2}{\sqrt{p(1-p)\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}}\right| > 1,96\right\}$$

$$\left\{\left|\frac{0,7 - 0,3}{\sqrt{0,5(1-0,5)\left[\frac{1}{1000}+\frac{1}{1000}\right]}}\right| > 1,96\right\}$$

$$\left\{\left|\frac{0,4}{0,0005}\right| > 1,96\right\}$$

$$\left\{\left|800\right| > 1,96\right\}$$

Du coup, on rejète l'hypothèse H_0 d'identité des taux de survie dans les deux échantillons. En conclusion, les personnes vaccinées ont un taux de survie nettement plus élevé. Le vaccin semble fonctionner, même s'il ne garantit pas une protection absolue.

B - Mise sur le marché d'un vaccin

1. Soit on teste l'hypothèse le vaccin n'est pas dangereux (p=0,02) et l'erreur est d'accepter à tort cette hypothèse (c'est-à-dire que le vaccin n'est pas dangereux alors qu'il l'est), soit on test l'hypothèse le vaccin est dangereux (p=0,05) et l'erreur est d'accepter à tort cette hypothèse (c'est-à-dire dire que le vaccin est dangereux alors qu'il ne l'est pas).

   Les experts de l'AFSSAPS souhaitent éviter de mettre sur le marché un vaccin dangereux en soi. Ils vont donc prendre comme hypothèse $H_0\{p = 0,05\}$.

2. On a donc une variable aléatoire X_i qui prend la valeur 1 si les globules blancs augmentent et 0 sinon. p est la probabilité que les globules blancs augmentent. Le test mis en œuvre par les experts de l'AFSSAPS est le test optimal de Neyman de $H_0 = \{p = 0,05\}$ contre $H_1 = \{p = 0,02\}$. On sait que la statistique de test est $F_n = \frac{1}{n}\sum X_i$. D'après l'énoncé $F_n = 0,032$. La région de rejet de p = 0,05 est de la forme $F_n < A$, puisque l'autre terme de l'alternative est $p = 0,02 < 0,05$. La valeur A est définie par le seuil du test, donc pour un seuil de 1% :

   $$P\{F_N < A; p = 0,05\} = 0,01$$

   Sous H_0, $\sum X_i$ suit une loi $B(n\ ; 0,05)$. Ici, n = 2000 :

   $$np(1-p) = 2000 \times 0,05 \times (1 - 0,05) = 95 > 15$$

   On peut donc faire une approximation de la loi Binomiale par une loi Normale d'espérance np, et de variance $np(1-p)$ sans faire de correction de continuité. On utilisera donc pour F_n la loi normale $\mathcal{N}\left(p; \frac{p(1-p)}{N}\right)$.

   On centre et on réduit et on obtient :

   $$\frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow \mathcal{N}(0,1)$$

   $$P\{F_n < A; p = 0,05\} = P\left\{(F_n - p)\frac{\sqrt{n}}{\sqrt{p(1-p)}} < (A - p)\frac{\sqrt{n}}{\sqrt{p(1-p)}}\right\} = 0,01$$

On lit dans la table $\mathcal{N}(0;1)$: $P\{U < -2,33\} = 0,01$. On en déduit $(A - 0,05)\frac{\sqrt{2000}}{\sqrt{0,05\times0,95}} = -2,33$ d'où on tire $A = 0,05 - 2,33\frac{\sqrt{0,05\times0,95}}{\sqrt{2000}} = 0,038$. Au seuil de 1%, les experts de l'AFSSAPS ne peuvent pas refuser la mise sur le marché du vaccin si $F_{2000} < 0,038$. Or d'après l'énoncé les tests donnent une moyenne de 3,2 %, ce qui valide le test exigé.

3. En testant l'hypothèse que le vaccin est dangereux on prend le risque de rejeter la validité du vaccin alors qu'il n'est pas en réalité dangereux. Pour faire un test plus favorable à la mise sur le marché du vaccin, il faut tester l'hypothèse H_0 que le vaccin n'est pas dangereux, c'est-à-dire que $\{p = 0,02\}$.

C - Analyse économique du vaccin contre le paludisme

1. Une bonne approximation des revenus est la loi log-normale de densité :

$$f(lw) = \frac{1}{\sqrt{2\pi V(lw)}}e^{-\frac{(lw-E(lw))^2}{2V(lw)}}$$

On calcule rapidement sa vraisemblance :

$$L = \prod f(lw) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\sum \frac{(lw-mlw)^2}{2\sigma^2}}$$

où mlw est l'espérance de lw et σ^2 représente sa variance.
On passe en log :

$$ln(L) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(lw_i - mlw)^2$$

2. On dérive par rapport à mlw et on annule la dérivée :

$$\frac{\partial \ln(L)}{\partial mlw} = \frac{1}{\sigma^2}\sum(lw_i - mlw) = 0$$

On en tire :

$$\widehat{mlw} = \frac{1}{n}\sum lw_i = \overline{lw}$$

3. Il faut 400 Euros au moins pour pouvoir se payer le vaccin. On connaît la distribution des revenus de la population, on cherche donc la probabilité que les revenus de la population en question soit supérieurs ou égaux à 400 Euros.
$p(w \geqslant 400) = 1 - p(w \leqslant 400)$
$= 1 - p(\ lw \leqslant \ln 400)$
$= 1 - p(\frac{lw-mlw}{\sigma} \leqslant \frac{\ln(400)-mlw}{\sigma})$
$= 1 - p(u \leqslant \frac{\ln(400)-mlw}{\sigma})$ où u suit une loi normale centrée réduite.
$= 1 - F(\frac{\ln(400)-mlw}{\sigma})$ où F est la fonction de répartition de la loi normale évaluée

en $\frac{\ln(400)-mlw}{\sigma} = \frac{5,99-5,7}{\sqrt{0,5}} = 0,41$. Cette quantité $F(0,41)$ vaut 0,6591.
Par conséquent,
$p(w \geqslant 400) = 0,3409$
Seuls 34,09 % des individus concernés seront en état de payer le vaccin.

4. Cela donne 1318,2 millions de personnes qui ne pourront pas payer le vaccin s'il était disponible. Il faudrait donc 15,8 milliards d'Euros pour financer le vaccin à toutes les personnes ne pouvant pas le payer.