Licence Economie Gestion troisième année 2009-2010

Corrigé de l'examen de Statistique du premier février 2010

Exercice 1. On observe un échantillon de 290 pingouins et 285 manchots ayant fait les mêmes études et classés par tranches de salaire. On veut savoir si à qualifications égales, le salaire dépend de l'espèce.

	1000-2000	2000-3000	3000-4000	4000-5000
Pingouins	50	70	110	60
Manchots	60	75	100	50

On utilisera les notations suivantes : soit $n_{1,j}$ et $n_{2,j}$ les nombre respectifs de pingouins et de manchots de la tranche de salaire j, $1 \leq j \leq 4$, n_{1+} le nombre total de pingouins, n_{2+} le nombre total de manchots et $n = n_{1+} + n_{2+}$.

(i) Donner le nom du test à appliquer.

Corrigé On va appliquer le test du χ^2 d'indépendance.

(ii) Définir l'hypothèse nulle H_0 qui convient.

Corrigé L'hypothèse H_0 est "Le salaire et l'espèce sont indépendants".

(iii) Pour $j = 1, \ldots, 4$, calculer n_{+j} le nombre total d'individus de la catégorie de salaire j. Calculer $\sum_{j=1}^{4} n_{+j}$.

Corrigé $n_{+1} = 110$, $n_{+2} = 145$, $n_{+3} = 210$, $n_{4+} = 110$. On a évidemment $\sum_{j=1}^{4} n_{+j} = n$.

(iv) Donner la définition de la statistique de test T_n. (Utiliser les notations indiquées)

Corrigé La statistique de test est

$$T_n = \sum_{i=1}^{2} \sum_{j=1}^{4} \frac{(n_{i,j} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n} \ .$$

(v) Quelle est sa loi approximative sous H_0 ? Quelle est la région de rejet de H_0 au seuil de risque 5% ?

Corrigé Sa loi approximative sous H_0 est la loi du χ^2 à $(2-1)(4-1) = 3$ degrés de liberté. La région de rejet de H_0 au seuil de risque 5% est donc $\{T_n > 7.81\}$.

(vi) La valeur observée de T_n est $t = 2.43$. Peut-on accepter au seuil 5% l'hypothèse que le salaire est indépendant de l'espèce ?

Corrigé le quantile d'ordre 95% de la loi du χ^2 à trois degrés de liberté est 7.81, donc on accepte l'hypothèse H_0.

(vii) Quel est le plus grand seuil auquel on peut accepter cette hypothèse ?

Corrigé On lit dans la table du χ^2 à trois degrés de liberté que le plus grand seuil auquel on peut accepter l'hypothèse est 50%.

Exercice 2. On considère un échantillon $(X_1, X_2, \ldots, X_n)$ de loi normale $\mathbf{N}(m, 1)$ avec m inconnue. On rappelle que la densité f de la loi $\mathbf{N}(01)$ est donnée par :

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-(x-m)^2/2\right\} \ .$$

(i) Déterminer la fonction de vraisemblance V de l'échantillon. (Préciser sur quel ensemble elle est définie).

Corrigé La vraisemblance de l'échantillon est la fonction V définie sur $\mathbf{R}^{n+1}$ par

$$V(m, x_1, \ldots, x_n) = (2\pi)^{-n/2} \prod_{i=1}^{n} \exp\left\{-(x_i - m)^2/2\right\}$$

$$a = (2\pi)^{-n/2} \exp\left\{ -\sum_{i=1}^{n} (x_i - m)^2/2 \right\} .$$

(ii) Calculer la log-vraisemblance L et ses dérivées première et seconde par rapport à m.

Corrigé On omet les variables x_i dans la notation.

$$L(m) = -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} (x_i - m)^2/2 ,$$

$$\frac{\partial L}{\partial m}(m) = \sum_{i=1}^{n} (x_i - m) = \sum_{i=1}^{n} x_i - nm , \qquad \frac{\partial^2 L}{\partial m^2}(m) = -n .$$

(iii) Calculer l'information de Fisher $I_n(m)$.

Corrigé

$$I_n(m) = -\mathbf{E}\left[\frac{\partial^2 L}{\partial m^2}(m)\right] = n .$$

(iv) Déterminer l'estimateur du maximum de vraisemblance de m.

Corrigé Cherchons tout d'abord les zéros de la dérivée première de la log-vraisemblance.

$$\frac{\partial L}{\partial m}(m) = 0 \;\Leftrightarrow m = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Ce point est un maximum car la log-vraisemblance est concave. L'estimateur du maximum de vraisemblance de m est donc la moyenne empirique.

(v) Cet estimateur est-il sans biais ?

Corrigé Oui, car la moyenne empirique est toujours un estimateur sans biais de l'espérance.

(vi) Cet estimateur est-il convergent ?

Corrigé Oui par la loi forte des grands nombres.

(vii) Cet estimateur est-il efficace ?

Corrigé On a $\mathrm{var}(\bar{X}_n) = 1/n = 1/I_n(m)$ donc l'estimateur du maximum de vraisemblance est efficace.

(viii) Quelle est la loi de la moyenne empirique $\bar{X}_n$?

Corrigé La moyenne empirique $\bar{X}_n$ suit la loi $\mathbf{N}(m, 1/n)$.

(ix) Donner un intervalle de confiance pour m de niveau 95%, pour $n = 16$ et la valeur observée $\bar{x}_n = 5$ de $\bar{X}_n$.

Corrigé L'intervalle de confiance de niveau 95% pour m est

$$[\bar{x}_n - 1.96/\sqrt{n}, \bar{x}_n + 1.96\sqrt{n}] = [4.51, 5.49] \ .$$

Exercice 3. Les notes d'examen de deux groupes d'étudiants sont modélisées par des variables gaussiennes indépendantes d'espérances et de variances respectives m_1, m_2, σ_1^2 et σ_2^2. Le groupe 1 a $n_1 = 21$ étudiants et le groupe 2 a $n_2 = 31$ étudiants. Pour chaque groupe on mesure la moyenne empirique et la variance empirique (sans biais) des notes.

	$\bar{x}_{n,i}$	$\hat{\sigma}^2_{n,i}$
groupe 1	14.12	3.52
groupe 2	13.25	3.43

On veut tester l'hypothèse que les deux groupes ont le même niveau moyen.

(i) On teste tout d'abord l'égalité des variances. Quelle est la statistique de test et quelle est sa loi sous l'hypohtèse nulle ? Quelle est la région de rejet de H_0 au seuil de risque 5% ?

Corrigé La statistique de test est $\hat{\sigma}^2_{n,1}/\hat{\sigma}^2_{n,1}$. Sa loi sous H_0 est la loi de Fisher $F_{20,30}$. La région de rejet de H_0 au seuil de risque H_0 est $\mathcal{R} = [0, 0.43[\cup]2.19, \infty[$.

(ii) La valeur de la statistique de test est 1.05. L'hypothèse d'égalité des variances peut-elle être acceptée au seuil 5% ?

Corrigé La valeur observée de la statistique de test n'est pas dans la région de rejet, donc on accepte H_0 au seuil de risque 5%.

(iii) Sous l'hypothèse d'égalité des variances, donner un nouvel estimateur sans biais $\tilde{\sigma}^2_n$ de la variance commune. (Justifier la réponse)

Corrigé On définit

$$\tilde{\sigma}^2_n = \frac{(n_1 - 1)\sigma^2_{n,1} + (n_2 - 1)\sigma^2_{n,2}}{n_1 + n_2 - 2} \ .$$

Sous l'hypothèse d'égalité des variances $\tilde{\sigma}^2_n$ est bien un estimateur sans biais de σ^2 car c'est un barycentre de deux estimateurs sans biais.

(iv) On teste maintenant l'égalité des moyennes. Donner l'expression de la statistique de test T_n à utiliser. Quelle est sa loi sous H_0 ? Quelle approximation peut-on utiliser ? Quelle est la région de rejet de H_0 au seuil de risque 5% ?

Corrigé La statistique de test est

$$T_n = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{(\bar{X}_{n,1} - \bar{X}_{n,2})}{\tilde{\sigma}_n} \ .$$

Sa loi sous H_0 est la loi de Student à $n_1 + n_2 - 2 = 50$ degrés de liberté. On peut l'approcher par la loi normale centrée réduite. La région de rejet de H_0 au seuil de risque H_0 est donc $]-\infty, -1.96[\cup]1.96, +\infty[$.

(v) La valeur observée de la statistique de test est $t = 1.65$. L'hypothèse d'égalité des moyennes peut-elle être acceptée au seuil 5% ?

Corrigé La valeur observée de la statistique de test n'est pas dans la région de rejet, donc on accepte H_0 au seuil de risque 5%.

(vi) Quel est le plus grand seuil auquel on accepte l'hypothèse d'égalité des moyennes ?

Corrigé Le plus grand seuil auquel on accepte H_0 est 10% car $\Phi(1.65) = 95\%$.

Valeurs numériques

– Soit Φ la fonction de répartition de la loi normale centrée réduite.

$$\Phi(1.65) = 95\% \ , \quad \Phi(1.96) = 97.5\% \ .$$

– Soit Q_ν la fonction de répartition de la loi du χ^2 à ν degrés de liberté.

$$Q_3(2.37) = 50\% \ , \quad Q_3(7.81) = 95\% \ .$$

– Soit $F_{n,m}$ la fonction de répartition de la loi de Fisher à n et m degrés de liberté.

$$F_{20,30}(0.43) = 2.5\% \ , \quad F_{20,30}(2.19) = 97.5\% \ .$$