HMMA 102

<u>*Statistique mathématique*</u> Feuille 9

Table de contingence et modèle log-linéaire

76) Considérez un tableau de contingence des probabilités R x R dont on dénote les probabilités par p_{jk}, $j, k = 1,\ldots, R$. On a $0 < p_{jk} < 1$ avec $\sum_{j=1}^{R}\sum_{k=1}^{R} p_{jk} = 1$. Dans un premier temps, on veut tester l'hypothèse nulle

$$H_0\text{: } p_{jj}=p^* \text{ et } p_{jk} = p_{kj} \text{ pour tout } j \neq k$$

où p^* est inconnu ainsi que les valeurs des p_{jk}.

a) Exprimez H_0 sous forme d'un jeu de contraintes.

b) Si l'on effectue un test du rapport de vraisemblance maximale de H_0, déterminez le degré de liberté de la loi khi-deux qui doit être utilisée.

c) On observe $X_1,\ldots, X_n$, des matrices aléatoires *i.i.d.* R x R composées de 0 partout sauf à un endroit où il y a un "1". On note les éléments de ces matrices $x_{jk,i}$ avec $\sum_{j=1}^{R}\sum_{k=1}^{R} x_{jk,i} = 1$ et $\sum_{i=1}^{n}\sum_{j=1}^{R}\sum_{k=1}^{R} x_{jk,i} = n$. La probabilité que le "1" soit en position (j, k) est p_{jk}. (c'est-à-dire $P[x_{jk,i} = 1] = p_k$).

> *i*) Déterminez la fonction de vraisemblance de l'échantillon sous l'hypothèse H_0.
>
> *ii*) Déterminez les EVM contraints de p^* et des p_{jk} $(j \neq k)$ (Attention, ils doivent respecter la contrainte $\sum_{j=1}^{R}\sum_{k=1}^{R} p_{jk} = Rp^* + 2\sum_{j=1}^{R-1}\sum_{k=j+1}^{R} p_{jk} = 1$).

Supposons que H_0 ait été accepté. On veut maintenant tester H_0^*: $p^* = 1/(2R)$ et $p_{jk} = 1/(2(R^2- R))$ contre $H_1^* = H_0$. Écrivez la statistique du test RVM de H_0^* contre H_1^*. Quel est son degré de liberté?

77) Soit $X_1,\ldots, X_n$, des variables aléatoires *iid* pouvant prendre les valeurs (1,0,0,0,0), (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0), (0,0,0,0,1) et (0,0,0,0,0) avec probabilités p_{11}, p_{12}, p_{13},

p_{21}, p_{22} et p_{23} respectivement. On suppose que $0 < p_{ij} < 1$ et que $\sum_{i=1}^{2}\sum_{j=1}^{3} p_{ij} = 1$. Dans ce contexte, on veut tester

$$H_0: p_{ij} = q_i\, r_j, \quad \text{pour tout } i = 1, 2 \text{ et } j = 1, 2, 3$$

contre

$$H_1: p_{ij} \neq q_i\, r_j, \text{ pour au moins un couple } (i, j).$$

Dans l'expression de H_0, les q_i et les r_j sont inconnus. Déterminez le test RVM pour ce problème. Précisez la loi limite qui doit être utilisée pour ce test.

78) Considérez un tableau de contingence 5 x 6 x 3. On veut tester l'hypothèse que la variable W est indépendante de Y et Z. Écrire sous H_0 les p_{rcl} en terme de p_{ij+}, p_{++l} etc. Déterminer les EVM contraints des p_{rcl}. Trouvez le degré de liberté de la loi khi-deux asymptotique.

79) Les données suivantes portent sur la réponse à la chimiothérapie pour des sujets atteints d'un certain type de cancer.

	Réponse		Total
Sexe	+	–	
Mâle	13	5	18
Femelle	5	7	12
Total	18	12	30

Déterminez, en utilisant R et au niveau $\alpha = 10\%$ si la variable « sexe » est indépendante de la variable « réponse ».

80) Les données suivantes sont issues d'une étude sur la relation entre quelques maladies gastriques et le groupe sanguin.

Type Sangun	Ulcère peptique	Cancer Gastrique	Contrôle	Total
O	983	383	2892	4258
A	679	416	2625	3720
B	134	84	570	788
Total	1796	883	6087	8766

Déterminez en utilisant R et au niveau $\alpha = 5\%$ si le groupe sanguin est indépendant du type de maladie.

81) Les données suivantes proviennent d'une étude liant la survie de prématurés selon le niveau de soins reçus et selon la clinique où ils ont été admis.

		Survie	
	Niveau de soin	Décès	Survie
Clinique A	Faible	3	1776
	Élevé	4	293
Clinique B	Faible	17	197
	Élevé	2	23

Étudiez, par modèle log-linéaire et en utilisant R, la structure de dépendance entre les variables « clinique», « niveau de soin» et « survie».

82) Les données suivantes proviennent du recensement de 1959 de la population soviétique habitant en milieu urbain.

		Scolarisation			
Sexe	âge	< 4 années	4 – 7	8 – 10	> 10
Homme	16 – 29	631.571	4.381.454	7.557.356	915.307
	30 – 49	1.176.623	4.174.797	5.384.310	1.055.949
	50 et +	2.376.214	1.807.879	1.187.537	501.741
Femme	16 – 29	664.308	2.843.925	9.180.042	1.274.867
	30 – 49	3.825.831	4.481.894	6.873.450	1.170.910
	50 et +	7.432.953	1.888.644	1.312.793	275.948

Étudiez, par modèle log-linéaire et en utilisant R, la structure de dépendance entre les variables « sexe », « âge » et « scolarisation ».

Réponse : 1b) dl = $R(R+1)/2 - 1$; 1c) dl = $R(R-1)/2$; 2) dl = 2; 3) dl = 58