MAP433 Statistique

PC6: Tests asymptotiques. Test de rangs

3 octobre 2014

1 Cancer et tabac

Voici les chiffres (fictifs) du suivi d'une population de 100 personnes (50 fumeurs, 50 non-fumeurs) pendant 20 ans.

	fumeur	non-fumeur
cancer diagnostiqué	11	5
pas de cancer	39	45

On s'interroge : la différence du nombre de cancers entre fumeurs et non-fumeurs est-elle statistiquement significative ? On note X_i la variable qui vaut 1 si le fumeur i a été atteint d'un cancer et 0 sinon. De même, on note Y_i la variable qui vaut 1 si le non-fumeur i a été atteint d'un cancer et 0 sinon. On suppose que les X_i sont i.i.d. de loi be Bernoulli $\mathcal{B}(\theta_f)$, les Y_i sont i.i.d. de loi $\mathcal{B}(\theta_{nf})$ et les X_i sont indépendants des Y_i.

1. Si $\theta_f \neq \theta_{nf}$, quelle est la limite de $\sqrt{n}|\bar{X}_n - \bar{Y}_n|$?
2. On suppose que $\theta_f = \theta_{nf} = \theta$ et on note $\hat{\theta} = (\bar{X}_n + \bar{Y}_n)/2$. Montrez que
$$\sqrt{\frac{n}{2\hat{\theta}(1-\hat{\theta})}}(\bar{X}_n - \bar{Y}_n) \overset{loi}{\to} \mathcal{N}(0,1).$$
3. Proposez un test de niveau asymptotique 5% de $\mathbf{H}_0$: "le taux de cancer n'est pas différent" ($\theta_f = \theta_{nf}$) contre $\mathbf{H}_1$: "le taux de cancer est différent" ($\theta_f \neq \theta_{nf}$).
4. Supposons maintenant qu'une étude supplémentaire permet d'avoir le suivi de 300 personnes et que les proportions sont les mêmes :

	fumeur	non-fumeur
cancer diagnostiqué	33	15
pas de cancer	117	135

   Quelle est la conclusion du test avec ces données ?
5. Revenons aux chiffres de la première étude : proposez un test de niveau asymptotique 5% de $\mathbf{H}_0$: "fumer n'a pas d'impact sur le taux de cancer" ($\theta_f = \theta_{nf}$) contre $\mathbf{H}_1$: "fumer augmente le taux de cancer" ($\theta_f > \theta_{nf}$) ? Quelle est sa conclusion ? Quelle est la p-value associée aux observations ?

2 Test de Wilcoxon

Une firme pharmaceutique a mis au point une nouvelle molécule pour faire chuter le taux de sucre dans le sang. Pour tester l'efficacité de cette molécule, elle le compare à un placebo. Elle réunit $n+m$ patients. A un premier groupe de m individus, elle administre un placebo (sans leur dire !). Au second groupe elle donne sa nouvelle molécule. Après un délai approprié, on mesure les taux de glycémie $\{X_i : i = 1, \dots, n\}$ et $\{Y_i : i = 1, \dots, m\}$ chez les deux groupes.

placebo : X	médicament : Y
1.0	1.4
1.41	0.94
0.61	3.1
0.22	0.54
5.9	1.2
0.84	0.043
0.49	3.0
	0.40
	0.075
	1.1

On suppose que les $X_1, \dots, X_n$ (resp. $Y_1, \dots, Y_m$) sont i.i.d. de loi de fonction de répartition F_X (resp. F_Y). On supposera que les fonctions F_X et F_Y sont continues. On veut tester si les lois des X et des Y sont les mêmes ou si les Y_i sont stochastiquement plus petits que les X_i, c'est à dire $F_X < F_Y$. On va donc tester $H_0 : F_X = F_Y$ contre $H_1 : F_X < F_Y$.
On pose $Z_i = X_i$ pour $i = 1, \dots, n$ et $Z_{n+i} = Y_i$ pour $i = 1, \dots, m$. On note $R(i)$ le rang de Z_i dans la suite $(Z_1, \dots, Z_{n+m})$, à savoir, $R(i) = 1$ si Z_i est la plus petite valeur, $R(i) = 2$ si Z_i est la seconde plus petite valeur, etc. La statistique de Wilcoxon est définie par

$$W_{n,m} = \sum_{i=1}^{n} R(i).$$

L'idée est que sous H_1 la statistique $W_{n,m}$ sera plus grande que sous H_0.

1. Soit $N = n + m$. Montrez que la *statistique de rang* $R \triangleq (R(1), \dots, R(N))$ suit sous H_0 la loi uniforme sur l'ensemble $\mathcal{S}_N$ de toutes les permutations de $\{1, \dots, N\}$. En déduire que sous H_0 la statistique de Wilcoxon est *libre*, i.e., la loi de $W_{n,m}$ ne dépend pas de F_X. Quelle est la loi de $R(i)$ sous H_0 ?
2. Montrez que sous H_0 on a $\mathbb{E}(W_{n,m}) = n(n+m+1)/2$.
3. Montrez que sous H_0 on a $\mathrm{Var}(W_{n,m}) = n\mathrm{Var}(R(1)) + n(n-1)\mathrm{Cov}(R(1), R(2))$ et

$$0 = \mathrm{Var}\left(\sum_{i=1}^{n+m} R(i)\right) = (n+m)\mathrm{Var}(R(1)) + (n+m)(n+m-1)\mathrm{Cov}(R(1), R(2)).$$

   En déduire que $\mathrm{Var}(W_{n,m}) = nm(n+m+1)/12$ sous H_0.
4. En admettant que $T_{n,m} = (W_{n,m} - \mathbb{E}(W_{n,m}))/\sqrt{\mathrm{Var}(W_{n,m})}$ converge en loi sous H_0 vers une loi normale $\mathcal{N}(0,1)$ lorsque $n \to \infty$, testez au niveau asymptotique 5% si la nouvelle molécule est plus efficace que le placebo.

3 Peut-on retarder sa mort ?

On prétend couramment que les mourants peuvent retarder leur décès jusqu'à certains événements importants. Pour tester cette théorie, Philips et King (1988, article paru dans *The Lancet*, prestigieux journal médical) ont collecté des données de décès aux environs d'une fête religieuse juive. Sur 1919 décès, 922 (resp. 997) ont eu lieu la semaine précédente (resp. suivante). Comment utiliser de telles données pour tester cette théorie grâce à un test asymptotique ?