Université des Sciences et Technologies de Lille
U.F.R. de Mathématiques Pures et Appliquées
Bât. M2, F-59655 Villeneuve d'Ascq Cedex

Agrégation externe **Année 2002-2003**

T.P. sur les tests

Avant de commencer ce T.P., il n'est pas inutile de regarder la documentation des fonctions Scilab `gsort`, `find`, `cdfnor`.

Ex 1. *Test de Neyman Pearson et détection radar* [2]

Un radar actif de surveillance aérienne a des caractéristiques telles qu'une éventuelle cible réfléchit $N = 20$ impulsions lors d'un balayage. À l'aide d'un traitement adapté, ces N impulsions réfléchies en cas de présence de la cible fournissent un vecteur d'observations $(z_i)_{1\leq i\leq N}$ avec

$$\begin{array}{ll} (\mathcal{H}_1) & z_i = A + b_i \quad \text{en présence de cible,} \\ (\mathcal{H}_0) & z_i = 0 + b_i \quad \text{en l'absence de cible,} \end{array}$$

où les b_i sont des variables aléatoires gaussiennes $\mathfrak{N}(0,\sigma)$ indépendantes modélisant les divers bruits (σ est connu).

1) Donner le test de Neyman Pearson de niveau α de $(\mathcal{H}_0)$ contre $(\mathcal{H}_1)$. Application numérique : $A = 1$, $\sigma = 0,6$, $\alpha = 10^{-6}$.

2) Écrire un script permettant de simuler n échantillons de taille N sous l'hypothèse nulle, auxquels on appliquera le test de Neyman Pearson et calculer la fréquence empirique des fausses détections. Construire de manière économique n échantillons sous $(\mathcal{H}_1)$ et calculer la fréquence empirique des cibles non détectées.

3) Utiliser ce script pour étudier l'influence de α, σ et A.

Ex 2. *Le test des longueurs* [3]

Soient X et Y deux variables aléatoires. On veut tester

$(\mathcal{H}_0)$ « X et Y ont même loi » contre $(\mathcal{H}_1)$ « X et Y n'ont pas même loi ».

On dispose pour cela d'un échantillon $(X_1,\ldots,X_m)$ de la loi de X et d'un échantillon $(Y_1,\ldots,Y_n)$ de celle de Y, toutes ces $m+n$ variables aléatoires étant indépendantes. On range par ordre croissant les $m+n$ valeurs effectivement observées et après effacement des indices on obtient un « mot » de la forme $XXYXYYYXYXX$. On compte alors le nombre R de blocs de lettres identiques (les longueurs ou « runs » ou composantes connexes) dans ce mot, (sur cet exemple, $R = 7$, soit 4 longueurs de X et 3 de Y). L'idée intuitive de ce test est que c'est sous $(\mathcal{H}_0)$ que le mélange entre les valeurs de X et celles de Y sera le plus intense et donc le nombre de longueurs le plus élévé.

Sous $(\mathcal{H}_0)$, la loi de R est donnée par les formules suivantes où l'on suppose $m \leq n$.

$$\begin{aligned}
\mathbf{P}(R = 2k) &= \frac{2C_{m-1}^{k-1}C_{n-1}^{k-1}}{C_{m+n}^{m}}, \qquad 1 \leq k \leq m; \\
\mathbf{P}(R = 2k+1) &= \frac{C_{m-1}^{k}C_{n-1}^{k-1} + C_{m-1}^{k-1}C_{n-1}^{k}}{C_{m+n}^{m}}, \quad 1 \leq k < m; \\
\mathbf{P}(R = 2m+1) &= \frac{C_{n-1}^{m}}{C_{m+n}^{m}} \quad \text{si } m < n \text{ et } 0 \text{ sinon.}
\end{aligned}$$

Lorsque m et n sont grands, R est (toujours sous $(\mathcal{H}_0)$) approximativement gaussienne avec

$$\mathrm{E}\, R = 1 + \frac{2mn}{m+n}, \qquad \mathrm{Var}\, R = \frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)}.$$

On définit alors un test de niveau au plus ε par la règle « rejet de $(\mathcal{H}_0)$ si $R \leq r_\varepsilon$ », où r_ε est le plus grand entier x tel que $\mathbf{P}(R \leq x) \leq \varepsilon$.

1) Écrire une fonction Scilab `nblong(X,Y)` qui calcule R pour les échantillons X et Y.

2) Écrire une fonction `testlong(X,Y,niveau)` qui réalise le test des longueurs au niveau asymptotique `niveau` pour m et n grands.

3) Expérimentez-la !

4) Écrire une fonction `fdrlong(m,n,x)` qui calcule la fonction de répartition de R. La comparer avec la f.d.r. gaussienne correspondant à l'approximation utilisée ci-dessus.

Ex 3. *Le test de Wilcoxon (observations non couplées)* [3]

La variable aléatoire Y est dite *stochastiquement* supérieure à X si

$$\forall x \in \mathbb{R}, \quad \mathbf{P}(Y \leq x) \leq \mathbf{P}(X \leq x),$$

avec inégalité stricte pour au moins une valeur de x. On se propose de tester

$$\begin{gathered}
(\mathcal{H}_0) \quad \text{« } X \text{ et } Y \text{ ont même loi »} \\
\text{contre} \\
(\mathcal{H}_1) \quad \text{« } Y \text{ est stochastiquement supérieure à } X \text{ ».}
\end{gathered}$$

On conserve les notations de l'exercice 2 et on traite les données de la même façon pour obtenir un mot formé des seules lettres X et Y. On utilise cette fois la statistique T égale à la somme des rangs des X. Par exemple avec le mot $XXYXYYYXYXX$, les rangs des X sont : 1, 2, 4, 8, 10, 11 et $T = 36$. Intuitivement, si Y est stochastiquement supérieure à X, on s'attend à ce que la répartition des X dans le mot privilégie sa partie gauche et donc que la somme des rangs soit plus petite que sous $(\mathcal{H}_0)$. Ceci conduit à définir le test (dit de Wilcoxon) par une zone de rejet de la forme $T \leq t_\alpha$. Lorsque m et n ne sont pas trop grands, on peut calculer la f.d.r. de T sous $(\mathcal{H}_0)$ en dénombrant le nombre de configurations réalisant l'évènement $\{T \leq t\}$ et en divisant par le nombre

total de configurations possibles, soit C_{m+n}^n. Lorsque m et n sont grands, T est, sous $(\mathcal{H}_0)$, approximativement gaussienne avec

$$\mathrm{E}\,T = \frac{m(m+n+1)}{2}, \quad \mathrm{Var}\,T = \frac{mn(m+n+1)}{12}.$$

1) Écrire une fonction Scilab `wilcox(X,Y)` qui calcule T pour les échantillons X et Y.

2) Écrire une fonction `testwlcx(X,Y,niveau)` qui réalise le test de Wilcoxon au niveau asymptotique `niveau` pour m et n grands.

3) Expérimentez-la. (Attention au fait que l'hypothèse alternative n'est pas la négation de $(\mathcal{H}_0)$).

Ex 4. *Test des signes (observations couplées)* [1, p. 112]

Pour essayer un nouvel engrais B devant améliorer le rendement en blé, nous disposons de n champs d'une station expérimentale (n petit de l'ordre de 10) ; chacun de ces champs étant divisé en deux parcelles. Affectons avec la probabilité 1/2 l'engrais B à une parcelle d'un champ et l'engrais usuel A à l'autre parcelle. Pour le champ numéro i, on note (X_i, Y_i) les rendements respectifs des parcelles ayant reçu A et B. Insistons sur un problème important en agronomie. Le lot de champs à la disposition de l'expérimentateur est trés réduit. Il n'est pas raisonnable de supposer que $\big((X_i, Y_i)\big)_{1\le i\le n}$ est un n échantillon d'une loi sur $\mathbb{R}^2$: cela reviendrait à dire que les n champs sont de même type et indépendants.

On se propose de tester

$$\begin{array}{c}(\mathcal{H}_0) \quad \text{« } B \text{ et } A \text{ donnent le même rendement »}\\ \text{contre}\\ (\mathcal{H}_1) \quad \text{« } B \text{ améliore le rendement »}.\end{array}$$

Considérons alors sous $(\mathcal{H}_0)$, la probabilité (due au seul tirage au sort) d'avoir $\{X_i < Y_i\}$. Comme les engrais ont été affectés avec la probabilité 1/2 à chacune des parcelles, $\mathbf{P}(X_i > Y_i) = \mathbf{P}(X_i < Y_i)$. Supposons que $\mathbf{P}(X_i = Y_i)$ est nulle : alors les v.a. $U_i = \mathbf{1}_{\{X_i<Y_i\}}$ sont des v.a. de Bernoulli indépendantes de paramètre 1/2 et $S = U_1 + \cdots + U_n$ suit la loi $\mathrm{Bin}(n, 1/2)$. Sous l'hypothèse $(\mathcal{H}_1)$, on observera plus souvent $\{X_i < Y_i\}$ et S sera plus grand. D'où la région de rejet du *test du signe* : on définit s_α comme le plus petit entier s tel que

$$\sum_{k=s+1}^{n} \frac{C_n^k}{2^n} \le \alpha,$$

et on prend comme région de rejet $D_\alpha = \{S > s_\alpha\}$.

1) Déterminer les valeurs de s_α en fonction de α pour $n = 10$. On inversera directement la f.d.r. de la binomiale, sans faire confiance à `cdfbin`.

2) Simuler les rendements sous $(\mathcal{H}_0)$ et sous $(\mathcal{H}_1)$ et mettre en œuvre le test.

3) Construire un test *randomisé* permettant d'avoir exactement le niveau α.

Références

[1] D. Dacunha-Castelle et M. Duflo, *Probabilités et statistiques, 1. Problèmes à temps fixe*. Masson 1982.

[2] J.-P. Delmas *Probabilités et télécommunications*. Exercices et problèmes commentés. Masson 1987.

[3] P. Jaffard, *Statistique*, Résumé de cours-Exercices-Problèmes. Masson 1990.

[4] X. Milhaud, *Statistique*. Editions espaces 34, Belin 2001.