Feuille de travaux dirigés 5 : Tests, Neyman-Pearson

Exercice 1 (Test de Neyman-Pearson : gaussiennes à moyenne connue)**:**

1. Soit Y un vecteur gaussien centré de taille n. On veut tester l'hypothèse $H_0 : Y \sim \mathcal{N}(0, \Sigma_0)$ versus $H_1 : Y \sim \mathcal{N}(0, \Sigma_1)$ où Σ_0, Σ_1 sont inversibles. Montrer que le test de Neyman-Pearson revient à comparer $y^T(\Sigma_1^{-1} - \Sigma_0^{-1})y$ à un seuil.
2. Soient X et V deux variables gaussiennes réelles, centrées, de variances respectives σ_X^2 et σ_V^2. La variable X est un signal utile et V est un bruit de mesure. L'observation est donnée par $Y = X + V$. On recoit n observations indépendantes.

   Proposer un test au niveau α permettant de détecter la présence du signal X.
3. Pour le test précédent, donner la valeur du seuil en fonction des quantiles de la loi du chi-deux. On précise que la loi du chi-deux à n degrés de libertés est la loi suivie par la somme de n variables normales centrées réduites indépendantes :

$$X \sim \chi_n^2 \iff X \overset{\text{loi}}{=} \sum_1^n U_i^2 \text{ où } U_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

Exercice 2 (Dosages)**:**

On souhaite tester si la concentration d'un produit est la même dans deux bassins différents A et B dans lesquels vivent des poissons d'élevage. A cet effet, des dosages sont effectuées dans les deux bassins et ont donné les résultats suivants :
Bassin A : 12, 14, 13, 13 (en mg/l)
Bassin B : 11, 13, 12 (en mg/l)

On admet que le résultat d'un dosage est une réalisation d'une variable aléatoire gaussienne dont l'espérance est la concentration du produit dans le bassin choisi (μ_A et μ_B respectivement pour les bassins A et B). On admet également que tous les dosages sont effectués de manière indépendante. On supposera que l'écart-type, pour la méthode de mesure utilisée, est égal à $\sigma = 1$ mg/l.

On cherche un test de l'hypothèse H_0 : "$\mu_A = \mu_B$" contre H_1 : "$\mu_A > \mu_B$".

1. On note n_A et n_B les tailles respectives des échantillons A et B et $\bar{X}_A$ et $\bar{X}_B$ les moyennes empiriques respectives des échantillons A et B. On considère la statistique de test

$$T(X) = \bar{X}_A - \bar{X}_B.$$

   Quelle est la loi de $T(X)$ sous l'hypothèse nulle ?

   Dans la suite on suppose que la seule observation disponible pour le statisticien est $T(X)$.
2. On considère provisoirement le test de l'hypothèse nulle H_0 contre

$$\tilde{H}_1 : \mu_A - \mu_B = \Delta$$

où $\Delta > 0$ est fixé. Montrer que le test de Neyman-Pearson de niveau $\alpha = 5\%$ revient à comparer $T(X)$ un seuil C que l'on précisera en fonction $\bar{n}_A$, $\bar{n}_B$, σ, et du quantile de niveau u de la loi gaussienne standard $\mathcal{N}(0,1)$, noté $q_{\mathcal{N}}(u)$, où u est à déterminer.

3. Quelle est la réponse de votre test pour les valeurs numériques données dans l'énoncé ? On donne $q_{\mathcal{N}}(0.95) \simeq 1.645, q_{\mathcal{N}}(0.975) \simeq 1.96$, $\sqrt{2} \simeq 1.414$.
4. Montrer que le test construit à la question 2 est aussi un test de niveau $\alpha = 5\%$ de H_0 contre H_1.
5. Montrer que le test construit à la question 2 est uniformément plus puissant de niveau α pour tester l'hypothèse H_0 contre H_1.

Exercice 3 (gestion de réseau)**:**
On reprend l'exemple du modèle de Pareto pour la modélisation du trafic internet, vu au TD 2. On rappelle que une variable aléatoire X_1 suit une loi de Pareto $\mathcal{P}ar(u,\theta)$, avec $u > 0$ et $\theta > 0$, si la fonction de répartition de X_1 est

$$F_\theta(x) = \begin{cases} 0 & \text{si } x \le u \\ 1 - \left(\frac{u}{x}\right)^\theta & \text{si } x > u. \end{cases}.$$

Dans la suite on fixe $u > 0$ supposé connu. On pourra utiliser les résultats suivant :

1. **Loi Gamma** Une variable aléatoire Y suit une loi Gamma de paramètres α et λ ($\alpha > 0$ et $\lambda > 0$), notée $\mathcal{G}amma(\alpha, \lambda)$, si elle admet une densité par rapport à la mesure de Lebesgue donnée par
$$f^{\mathcal{G}}_{(\alpha,\lambda)}(y) = \mathbb{1}_{y>0} \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}.$$
On rappelle que pour $\alpha > 0$, $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. Si $Y \sim \mathcal{G}amma(\alpha, \lambda)$, on a
$$\mathbb{E}_{\alpha,\lambda}(Y) = \frac{\alpha}{\lambda} \quad ; \quad \mathbb{V}\mathrm{ar}_{\alpha,\beta}(Y) = \frac{\alpha}{\lambda^2}.$$
2. **Loi Inverse Gamma** Si $Y \sim \mathcal{G}amma(\alpha, \lambda)$, alors $T := \frac{1}{Y}$ suit une loi dite 'inverse gamma' $\mathcal{IG}(\alpha, \lambda)$, de densité
$$f^{\mathcal{IG}}_{\alpha,\lambda}(t) = \mathbb{1}_{t>0} \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{1}{t^{\alpha+1}} e^{-\lambda/t},$$
et l'on a, lorsque $\alpha > 1$ (*resp.* $\alpha > 2$) :
$$\mathbb{E}_{\alpha,\lambda}(T) = \frac{\lambda}{\alpha-1} \quad ; \quad (\text{resp. } \mathbb{V}\mathrm{ar}_{\alpha,\lambda}(T) = \frac{\lambda^2}{(\alpha-1)^2(\alpha-2)}.)$$
3. **Somme d'exponentielles** Si $(Z_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{E}(\lambda)$, alors $Y := \sum_{i=1}^n Z_i$ suit une loi Gamma de paramètres (n, λ) :
$$\sum_{i=1}^n Z_i \sim \mathcal{G}amma(n, \lambda).$$

Le gestionnaire du réseau s'intéresse à la probabilité que la réseau sature (pour une observation X_1). En l'état actuel, son réseau sature lorsque $X_1 > s$, où $s > u$ est connu. Soit $g(\theta) = \mathbb{P}_\theta(X_1 > s)$ (où θ est inconnu). La réglementation autorise le réseau à saturer “rarement”, c'est-à-dire, elle impose que $g(\theta) \leq \rho_0$ où $0 < \rho_0 < 1$ est petit. La question est de savoir si le gestionnaire doit redimensionner son installation ou non. L'hypothèse nulle est que tout va bien :

$$H_0 : \{g(\theta) \leq \rho_0\}.$$

(l'hypothèse alternative est donc $H_1 : \{g(\theta) > \rho_0\}$).

1. Donner l'expression de $g(\theta)$ en fonction de θ, s et u.
2. Montrer que H_0 est vérifiée si et seulement si

$$\theta \in \Theta_0 = [\theta_0, +\infty[\quad \text{où } \theta_0 = \frac{\log(\rho_0)}{\log(u/s)}$$

3. Considérons pour commencer le test d'hypothèses simples $\tilde{H}_0 : \{\theta = \theta_0\}$ contre $\tilde{H}_1 : \{\theta = \theta_1\}$, où $\theta_1 < \theta_0$. Écrire la statistique du rapport de vraisemblance et montrer que le test de Neyman Pearson revient à comparer la variable aléatoire

$$W = \sum_{i=1}^{n} \log(X_i/u)$$

à un seuil c (qu'on ne calculera pas pour l'instant).

4. Soit $\alpha \in]0, 1[$. Déterminer le seuil c tel que le test

$$\delta(X) = \begin{cases} 1 & \text{si } W \geq c \\ 0 & \text{si } W < c \end{cases}$$

soit un test uniformément plus puissant (U.P.P) au niveau α (c'est-à-dire, de risque de première espèce égal à α) pour l'hypothèse $\tilde{H}_0$ contre $\tilde{H}_1$. On exprimera c en fonction des quantiles de la loi $\mathcal{G}amma(n, 1)$.

5. Soit $t > \theta_0$ considérons le test de l'hypothèse $H_0(t) : \{\theta = t\}$ contre $\tilde{H}_1$. Montrer que le risque de première espèce du test δ construit à la question 4 est strictement inférieur à α.
6. En déduire que δ est U.P.P. de niveau α pour tester $H_0 : \{\theta \geq \theta_0\}$ contre $\tilde{H}_1$.
7. En déduire que δ est U.P.P. de niveau α pour H_0 contre H_1.