TD 4 STATISTIQUE - 1SN

Exercice 1.
Afin de tester la satisfaction des clients à service donné, on effectue un sondage et on définit une variable aléatoire Y_i de la façon suivante :

$$\begin{aligned} Y_i &= 1 \text{ si le client } i \text{ est satisfait} \\ Y_i &= 0 \text{ si le client } i \text{ n'est pas satisfait} \end{aligned}$$

A l'aide d'un échantillon $(Y_1, ..., Y_n)$ de même loi de Bernoulli

$$\begin{aligned} P\left[Y_i = 0\right] &= \theta \\ P\left[Y_i = 1\right] &= 1 - \theta \end{aligned}$$

on désire tester les hypothèese $H_0 : \theta = \theta_0 = 0.52$ et $H_1 : \theta = \theta_1 = 0.48$.

1. Construire la vraisemblance des observations $y_1, ..., y_n$ et expliciter la région de rejet de H_0 du test de Neyman-Pearson (pour l'application numérique, on choisira un risque de première espèce $\alpha = 0.1$).

2. Déterminer la puissance de ce test.

Exercice 2. Soit $X_1, ..., X_n$ un échantillon d'une loi normale de moyenne m et de variance σ^2. On veut faire le test d'hypothèses binaires suivant :

$$\begin{aligned} H_0 &: m = m_0; \sigma^2 \text{ quelconque} \\ H_1 &: m \neq m_0; \sigma^2 \text{ quelconque} \end{aligned}$$

Pour construire le test, on retient le test du rapport des vraisemblances maximales ou test GLR (Generalized Likelihood Ratio).

1. On suppose $m = m_0$ connu. Rappeler l'estimateur du maximum de vraisemblance (EMV) de σ^2.

2. Lorsque m et σ^2 sont inconnus, rappeler leurs estimateurs du maximum de vraisemblance.

3. Donner la forme du test GLR.

4. En décomposant $\sum_{i=1}^{n} (x_i - m_0)^2$, montrer que l'on peut définir un test équivalent à l'aide de la statistique
$$T_n = \frac{\overline{X} - m_0}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}}$$

5. On rappelle que sous l'hypothèse H_0, les deux variables aléatoires
$$U = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}} \text{ et } V = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sigma^2}$$
ont des lois connues $U \sim \mathcal{N}(0, 1)$ et $V \sim \chi^2_{n-1}$. En déduire la loi de T_n. Soit $\alpha = 5\%$ le risque de première espèce. Donner la région critique du test effectué à l'aide de T_n.

Exercice 3.

On considère les observations $x_i, i = 1, ..., n$ (avec $n = 10$) définies par

$x_1 = 1$	$x_2 = 0$	$x_3 = 1$	$x_4 = 1$	$x_5 = 1$	$x_6 = 1$	$x_7 = 1$	$x_8 = 2$	$x_9 = 0$	$x_{10} = 0$

On suppose que les variables aléatoires associées à ces observations sont indépendantes et issues de la même loi de Poisson $P(\lambda)$. O*n rappelle que si X suit une une loi de Poisson de paramètre λ, on a* $E[X] = \text{var}[X] = \lambda$ *et* $\varphi_X(t) = E\left[e^{itX}\right] = \exp\left[\lambda\left(e^{it} - 1\right)\right]$. On désire tester les deux hypothèses

$$\begin{cases} H_0 : \lambda = \lambda_0 \text{ (absence de planète)} \\ H_1 : \lambda = \lambda_1 \text{ (présence de planète)} \end{cases}$$

avec $\lambda_1 < \lambda_0$.

1. Vérifier que la statistique du test de Neyman-Pearson peut s'écrire $T = \sum_{i=1}^{n} X_i$ et déterminer la région critique associée.

2. Déterminer la fonction caractéristique de T et en déduire que T suit une loi de Poisson que l'on précisera sous chaque hypothèse.

3. On suppose que n est suffisamment grand pour pouvoir utiliser les résultats du théorème de la limite centrale.

   - Donner la loi approchée de T issue de ce théorème.
   - Quelle est la valeur du seuil obtenue lorsqu'on confond la loi de T avec son approximation.
   - Déterminer les courbes COR (caractéristiques opérationnelles du récepteur) découlant de cette loi approchée. On posera

   $$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

   et on notera $F^{-1}(x)$ son inverse. En supposant que n est suffisamment grand pour faire les approximations nécessaires, déterminer les paramètres qui influent sur la performance asymptotique ($n \to \infty$) du test. De ces deux cas
   Premier Cas : $n = 100, \lambda_0 = 1, \lambda_1 = 0.1$
   Deuxième Cas : $n = 100, \lambda_0 = 2, \lambda_1 = 1.1$
   indiquer celui qui engendre la meilleure performance.

Exercice 4.

Un statisticien pose la question suivante à un échantillon de 30 participants : "Préférez-vous boire du thé ou du café ?". Parmi cet échantillon, 10 préfèrent le thé et 20 préfèrent le café. Il désire effectuer un test du chi-deux pour déterminer s'il y a une véritable préférence pour le café dans cet échantillon. Pour cela, il définit l'hypothèse H_0 par "la probabilité de boire du thé est égale à la probabilité de boire du café", i.e., les deux classes $\{\text{Thé}\}$ et $\{\text{Café}\}$ sont équiprobables ($P[\text{Thé}] = P[\text{Café}] = \frac{1}{2}$).

1. Déterminer la statistique du test du chi-deux noté ϕ associée à ce problème.

2. Rappeler la loi de ϕ sous l'hypothèse H_0 (définie par "Il n'y a pas de préférence ni pour le thé, ni pour le café").

3. Expliquer comment déterminer le seuil de décision S_α du test du chi-deux à l'aide de la fonction de répartition de la loi déterminée à la question précédente et du risque α de ce test. Pour $\alpha = 0.05$, on trouve $S_{0.05} = 3.84$. Que conclut-on ?

Correction exercice 1

1) La vraisemblance de ce problème est

$$\begin{aligned} L(y_1,...,y_n;\theta) &= \prod_{i=1}^{n} P[Y_i = y_i] \\ &= \prod_{i=1}^{n} \theta^{1-y_i}(1-\theta)^{y_i} \\ &= \theta^{n-n\overline{y}}(1-\theta)^{n\overline{y}} \end{aligned}$$

avec

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

On rejette donc H_0 si

$$\frac{L(y_1,...,y_n;\theta_1)}{L(y_1,...,y_n;\theta_0)} > K_\alpha \Longleftrightarrow \overline{y}\ln\left(\frac{\theta_0}{\theta_1}\frac{1-\theta_1}{1-\theta_0}\right) > S_\alpha$$

Pour $\theta_0 = 0.52$ et $\theta_1 = 0.48$, on a

$$\frac{\theta_0}{\theta_1}\frac{1-\theta_1}{1-\theta_0} = \left(\frac{0.52}{0.48}\right)^2 > 1$$

donc on rejette H_0 si

$$\overline{y} > \nu_\alpha$$

où ν_α est un seuil dépendant du risque de première espèce α. Pour déterminer ce seuil, on se fixe une valeur de α

$$\begin{aligned} \alpha &= P\left[\text{Rejeter } H_0 \middle| H_0 \text{ vraie}\right] \\ &= P\left[\overline{Y} > \nu_\alpha \middle| \theta = \theta_0\right] \end{aligned}$$

En utilisant le théorème de la limite centrale, on peut approcher la loi de $\overline{Y}$ comme suit

$$\overline{Y} \sim \mathcal{N}\left(1-\theta, \frac{\theta(1-\theta)}{n}\right)$$

Donc

$$\begin{aligned} \alpha &= P\left[U = \frac{\overline{Y}-(1-\theta_0)}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} > \frac{\nu_\alpha-(1-\theta_0)}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} \middle| U \sim \mathcal{N}(0,1)\right] \\ &= 1 - F\left(\frac{\nu_\alpha-(1-\theta_0)}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}\right) \end{aligned}$$

On en déduit

$$\frac{\nu_\alpha-(1-\theta_0)}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} = F^{-1}(1-\alpha)$$

où F est la fonction de répartition d'une loi normale $\mathcal{N}(0,1)$, d'où

$$\nu_\alpha = \sqrt{\frac{\theta_0(1-\theta_0)}{n}}F^{-1}(1-\alpha) + (1-\theta_0)$$

2) La puissance du test est

$$\begin{aligned}\pi &= P\left[\text{Rejeter } H_0 \middle| H_1 \text{ vraie}\right] \\ &= P\left[\overline{Y} > \nu_\alpha \middle| \theta = \theta_1\right] \\ &= 1 - F\left(\frac{\nu_\alpha - (1-\theta_1)}{\sqrt{\frac{\theta_1(1-\theta_1)}{n}}}\right)\end{aligned}$$

Correction exercice 2

1)

$$\tilde{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^n (X_i - m_0)^2$$

2)

$$\hat{m}_{MV} = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i, \qquad \hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$$

3) Le test GLR est défini par

$$\text{Rejet de } H_0 \text{ si } \frac{L\left(X_1, ..., X_n; H_1\right)}{L\left(X_1, ..., X_n; H_1\right)} > S_\alpha$$

c'est-à-dire

$$\text{Rejet de } H_0 \text{ si } \frac{\left(2\pi\hat{\sigma}^2_{MV}\right)^{-n/2} \exp\left[-\frac{1}{2\hat{\sigma}^2_{MV}}\sum\left(X_i - \overline{X}\right)^2\right]}{\left(2\pi\tilde{\sigma}^2_{MV}\right)^{-n/2} \exp\left[-\frac{1}{2\tilde{\sigma}^2_{MV}}\sum\left(X_i - m_0\right)^2\right]} > K_\alpha$$

c'est-à-dire

$$\text{Rejet de } H_0 \text{ si } \frac{\tilde{\sigma}^2_{MV}}{\hat{\sigma}^2_{MV}} > S_\alpha \Leftrightarrow \frac{\sum (X_i - m_0)^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} > S_\alpha$$

4) On décompose $\sum (X_i - m_0)^2$ comme suit

$$\begin{aligned}\sum (X_i - m_0)^2 &= \sum \left(X_i - \overline{X} + \overline{X} - m_0\right)^2 \\ &= \sum \left(X_i - \overline{X}\right)^2 + n\left(\overline{X} - m_0\right)^2\end{aligned}$$

donc le test GLR est défini par

$$\begin{aligned}\text{Rejet de } H_0 \text{ si } \frac{\sum \left(X_i - \overline{X}\right)^2 + n\left(\overline{X} - m_0\right)^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} &> S_\alpha \Leftrightarrow T_n^2 > \mu_\alpha \\ &\Leftrightarrow T_n \in \left]-\infty, -\mu_\alpha\right[\cup \left]\mu_\alpha, \infty\right[\end{aligned}$$

5) La statistique T_n s'écrit sous la forme suivante :

$$T_n = \frac{\sigma}{\sqrt{n}}\frac{U}{\sigma\sqrt{V}} = \frac{1}{\sqrt{n(n-1)}}\frac{U}{\sqrt{\frac{V}{n-1}}}$$

où

$$W_n = \frac{U}{\sqrt{\frac{V}{n-1}}} \sim t_{n-1}$$

On en déduit

$$\text{Rejet de } H_0 \text{ si } W_n \in \left]-\infty, -c_\alpha\right[\cup \left]c_\alpha, \infty\right[$$

et

$$\begin{aligned}1-\alpha &= 1-P\left[\text{rejeter } H_0 \middle| H_0 \text{ vraie}\right]\\ &= P\left[\text{accepter } H_0 \middle| H_0 \text{ vraie}\right]\\ &= P\left[\left|W_n\right| < c_\alpha \middle| H_0 \text{ vraie}\right] = 0.95\end{aligned}$$

Les tables de la loi de Student donnent la valeur de c_α.

Correction exercice 3

1) Des calculs élémentaires donnent

$$\text{Rejet de } H_0 \text{ si } T = \sum_{i=1}^{n} X_i < S_\alpha$$

2) La fonction caractéristique de T est

$$E\left[e^{itT}\right] = \prod_{j=1}^{n} E\left[e^{itX_j}\right] = \exp\left[n\lambda\left(e^{it}-1\right)\right]$$

qui est la fonction caractéristique d'une loi de Poisson de paramètre $n\lambda$ donc $T \sim P(n\lambda)$. Sous H_0, on a $T \sim P(n\lambda_0)$ et sous H_1, on a $T \sim P(n\lambda_1)$.
3) a) Pour n grand, l'approximation normale est $\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\lambda, n\lambda)$.
b) On trouve $K_\alpha = n\lambda_0 + \sqrt{n\lambda_0}F^{-1}(\alpha)$.
c) Un calcul simple conduit à

$$\text{PD} = 1-\beta = F\left(\sqrt{n}\frac{\lambda_0-\lambda_1}{\sqrt{\lambda_1}} + \sqrt{\frac{\lambda_0}{\lambda_1}}F^{-1}(\alpha)\right)$$

c'est-à-dire asymptotiquement

$$\text{PD} = 1-\beta \sim F\left(\sqrt{n}\frac{\lambda_0-\lambda_1}{\sqrt{\lambda_1}}\right)$$

Le paramètre qui règle la performance asymptotique du test est donc $\sqrt{n}\frac{\lambda_0-\lambda_1}{\sqrt{\lambda_1}}$. Dans les deux cas proposés $\lambda_0-\lambda_1 = 0.9$ et $n = 100$. Le premier test est meilleur car PD est une fonction décroissante de λ_1 lorsque $\lambda_0-\lambda_1$ et n sont fixés.

Correction exercice 4

1. La statistique du test du chi-deux noté ϕ est définie par
$$\phi = \sum_{k=1}^{2} \frac{(Z_k - np_k)^2}{np_k}$$
avec $p_1 = p_2 = 0.5$, $n = 30$, $Z_1 = 10$ et $Z_2 = 20$. Une application numérique donne
$$\phi = \frac{(10-15)^2}{15} + \frac{(20-15)^2}{15} = \frac{10}{3}.$$
2. Sous l'hypothèse H_0, ϕ suit une loi du χ^2 à 1 degré de liberté, i.e., $\phi \sim \chi_1^2$.
3. On a
$$\alpha = P[\text{Rejeter} H_0|H_0 \text{ vraie}] = P[\phi > S_\alpha|\phi \sim \chi_1^2] = 1 - F_{\chi_1^2}(S_\alpha).$$
donc
$$S_\alpha = F_{\chi_1^2}^{-1}(1-\alpha).$$
Pour $\alpha = 0.05$, on trouve $S_{0.05} = 3.84 > \phi$ donc on rejette H_0 avec $\alpha = 0.05$.