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. Préciser le test de puissance maximale tel que le risque de première espèce α vérifie $\alpha \leq 0.05$. On précisera le risque maximal α, la décision prise au vu des données $x_i, i = 1, ..., 10$ et la puissance de ce test. Pour les applications numériques, on prendra $\lambda_0 = 1$ et $\lambda_1 = 0.1$.

4. 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. En comparant avec la valeur obtenue précédemment, dire ce que vous pensez de cette approximation pour $n = 10$.
   - Déterminer les courbes COR (caractéristiques opérationnelles du récepteur) découlant de cette loi approchée. On posera

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

     et on notera $\Phi^{-1}(x)$ 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.

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

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

c'est-à-dire

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

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

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

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 $T \sim P(n\lambda_0) = P(10)$ et sous H_1, on a $T \sim P(n\lambda_1) = P(1)$.
3) On a $\alpha = P\left[\text{rejeter } H_0 \middle| H_0 \text{ vraie}\right] = P\left[T < S_\alpha \middle| T \sim P(n\lambda_0) = P(10)\right]$. En analysant les tables de la loi de Poisson $P(10)$, on trouve

$$S_\alpha = 5 \Rightarrow \alpha = 0.0293 \text{ et } S_\alpha = 6 \Rightarrow \alpha > 0.05.$$

Donc le test est défini par

$$\text{Rejet de } H_0 \text{ si } T < 5$$

et le risque de première espèce associé est $\alpha = 0.0293 < 0.05$. Les données sont telles que $\sum_{i=1}^{n} x_i = 8$ et donc on accepte l'hypothèse H_0 avec $\alpha = 0.0293$. Le calcul de la puissance du test conduit à

$$\begin{aligned} \pi &= 1-\beta = P\left[\text{rejeter } H_0 \middle| H_1 \text{ vraie}\right] \\ &= P\left[T < 5 \middle| T \sim P(1)\right] \\ &= \sum_{i=0}^{4} p_i \sim 0.9963 \end{aligned}$$

La puissance du test est donc excellente.
4) a) Pour n grand, l'approximation normale est $\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\lambda, n\lambda)$.
b) On trouve $K_\alpha = n\lambda_0 + \sqrt{n\lambda_0}\Phi^{-1}(\alpha) \sim 4.8$. On trouve 4.8 au lieu de 5 et donc l'approximation est satisfaisante pour $n = 10$ (puisque T prend des valeurs discrètes avoir $T < 5$ ou $T < 4.8$, c'est la même chose).
c) Un calcul simple conduit à

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

c'est-à-dire asymptotiquement

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

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.

k \ λ	**0.8**	**0.9**	**1.0**	**2.0**	**3.0**	**4.0**	**5.0**	**6.0**
0	0.4493	0.4066	0.3679	0.1353	0.0498	0.0183	0.0067	0.0025
1	0.3595	0.3659	0.3679	0.2707	0.1494	0.0733	0.0337	0.0149
2	0.1438	0.1647	0.1839	0.2707	0.2240	0.1465	0.0842	0.0446
3	0.0383	0.0494	0.0613	0.1804	0.2240	0.1954	0.1404	0.0892
4	0.0077	0.0111	0.0153	0.0902	0.1680	0.1954	0.1755	0.1339
5	0.0012	0.0020	0.0031	0.0361	0.1008	0.1563	0.1755	0.1606
6	0.0002	0.0003	0.0005	0.0120	0.0504	0.1042	0.1462	0.1606
7			0.0001	0.0034	0.0216	0.0595	0.1044	0.1377
8				0.0009	0.0081	0.0298	0.0653	0.1033
9				0.0002	0.0027	0.0132	0.0363	0.0688
10					0.0008	0.0053	0.0181	0.0413
11					0.0002	0.0019	0.0082	0.0225
12					0.0001	0.0006	0.0034	0.0113
13						0.0002	0.0013	0.0052
14						0.0001	0.0005	0.0022
15							0.0002	0.0009
16								0.0003
17								0.0001

k \ λ	**7.0**	**8.0**	**9.0**	**10.0**	**11.0**	**12.0**	**13.0**
0	0.0009	0.0003	0.0001				
1	0.0064	0.0027	0.0011	0.0005	0.0002	0.0001	
2	0.0223	0.0107	0.0050	0.0023	0.0010	0.0004	0.0002
3	0.0521	0.0286	0.0150	0.0076	0.0037	0.0018	0.0008
4	0.0912	0.0573	0.0337	0.0189	0.0102	0.0053	0.0027
5	0.1277	0.0916	0.0607	0.0378	0.0224	0.0127	0.0070
6	0.1490	0.1221	0.0911	0.0631	0.0411	0.0255	0.0152
7	0.1490	0.1396	0.1171	0.0901	0.0646	0.0437	0.0281
8	0.1304	0.1396	0.1318	0.1126	0.0888	0.0655	0.0457
9	0.1014	0.1241	0.1318	0.1251	0.1085	0.0874	0.0661
10	0.0710	0.0993	0.1186	0.1251	0.1194	0.1048	0.0859
11	0.0452	0.0722	0.0970	0.1137	0.1194	0.1144	0.1015
12	0.0263	0.0481	0.0728	0.0948	0.1094	0.1144	0.1099
13	0.0142	0.0296	0.0504	0.0729	0.0926	0.1056	0.1099
14	0.0071	0.0169	0.0324	0.0521	0.0728	0.0905	0.1021
15	0.0033	0.0090	0.0194	0.0347	0.0534	0.0724	0.0885
16	0.0014	0.0045	0.0109	0.0217	0.0367	0.0543	0.0719
17	0.0006	0.0021	0.0058	0.0128	0.0237	0.0383	0.0550
18	0.0002	0.0009	0.0029	0.0071	0.0145	0.0255	0.0397
19	0.0001	0.0004	0.0014	0.0037	0.0084	0.0161	0.0272
20		0.0002	0.0006	0.0019	0.0046	0.0097	0.0177
21		0.0001	0.0003	0.0009	0.0024	0.0055	0.0109
22			0.0001	0.0004	0.0012	0.0030	0.0065
23				0.0002	0.0006	0.0016	0.0037
24				0.0001	0.0003	0.0008	0.0020
25					0.0001	0.0004	0.0010
26						0.0002	0.0005
27						0.0001	0.0002
28							0.0001
29							0.0001
30							