EISTI	**ING1-GM** **EXAMEN DE STATISTIQUES INFERENTIELLES 2018-2019**
Durée : 2h	Calculatrice EISTI autorisée 3 feuilles manuscrites R/V autorisées

Consignes :

- *Chacune des quatre parties du sujet peut être traitée séparément des deux autres.*
- *Le barème est donné à titre indicatif*

Afin de satisfaire au mieux sa clientèle, le directeur d'un grand magasin souhaite étudier le temps d'attente en caisse des clients dans la tranche 12h-14h. Le temps d'attente d'un client est naturellement modélisé par une variable aléatoire X de loi exponentielle de paramètre θ **(rappels et résultats sur la loi exponentielle à la fin du sujet)**. Afin de déterminer le paramètre θ, on veut estimer la quantité

$$q = \frac{1}{\theta^2}$$

sur un échantillon de *n* clients. On note $X_1,\ldots,X_n$ le temps d'attente de chaque client.

PREMIERE PARTIE : L'ESTIMATEUR ET SES PROPRIETES [6 POINTS]

On considère l'estimateur suivant

$$Q = \frac{1}{2n}\sum_{i=1}^{n} X_i^2$$

On note que $Q = \frac{1}{2}\overline{Y}$ où $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ avec $Y = X^2$.

(a) Montrer que Q est un estimateur sans biais de q.

(b) Montrer que le risque quadratique est

$$R_q(Q) = \frac{5}{n}q^2.$$

(c) Etudier la loi asymptotique de Q.

(d) Cet estimateur est-il efficace ?

DEUXIEME PARTIE : ESTIMATION ET INTERVALLE DE CONFIANCE [4 POINTS]

Le directeur dispose d'un échantillon de 100 clients sur lequel il a observé une valeur de $\hat{q} = 25$.

(a) On constate qu'un client est mécontent à partir du moment où il attend plus de 7 minutes en caisse. Quelle est la probabilité estimée qu'un client soit mécontent ?

(b) D'après la première partie, on peut supposer que l'estimateur Q suit une loi normale

$$N(q, \frac{5\times 25^2}{n}).$$

Donner un intervalle de confiance de niveau 5% pour q.

(c) En déduire un intervalle de confiance de niveau 5% pour le temps moyen d'attente en caisse.

TROISIEME PARTIE : TEST D'HYPOTHESES : CONSTRUCTION [5 POINTS]

Le directeur envisage d'ouvrir des caisses supplémentaires entre 12h et 14h si le temps d'attente s'avère trop long. Il décide donc de tester les hypothèses

H_0 : « le temps d'attente moyen est acceptable » : $\mu=\mu_0=4$

H_1 : « le temps d'attente moyen est trop long » : $\mu=\mu_1>\mu_0$

On note que $\mu=E(X)=1/\theta$, donc les hypothèses sont équivalentes à

$H_0 : q=q_0=4^2$

$H_1 : q=q_1>q_0$

Le directeur dispose toujours de l'échantillon de 100 clients sur lequel il a observé une valeur de l'estimateur de q de $\hat{q}=25$. On choisit donc l'estimateur Q comme variable de décision et, d'après la première partie, on peut supposer que l'estimateur Q suit une loi normale

$$N(q,\frac{\sigma^2}{n}) \text{ où } \sigma^2=5q^2.$$

(a) Déterminer graphiquement l'allure de la région critique.
(b) On suppose un risque de 1ère espèce de 5%. Calculer alors le seuil de la région critique.
(c) Quelle décision doit prendre le directeur et quelle est la probabilité qu'il se trompe ?

QUATRIEME PARTIE : TEST D'HYPOTHESES : P-VALEUR [6 POINTS]

Le directeur gère deux magasins. Il souhaite faire une étude comparative entre les deux magasins.

1) Il souhaite savoir si le temps d'attente est le même pour les deux magasins. Il dispose d'un échantillon de 50 clients par magasin. Le résultat du test calculé par le logiciel R est le suivant.

```
        Welch Two Sample t-test

data:  Temps by Magasin
t = -0.76504, df = 93.074, p-value = 0.4462
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.736739  1.658268
sample estimates:
mean in group M1 mean in group M2
        5.197092         6.236327
```

(a) Quelle est l'hypothèse nulle H_0 du test ?
(b) Quelle est l'hypothèse alternative est H_1 du test ?
(c) Quelle est la conclusion du test ? Justifier votre réponse.
(d) A votre avis, ce test est-il pertinent ?

2) Il effectue une enquête de satisfaction sur 100 clients et obtient les résultats suivants.

	Très Satisfait	Satisfait	Peu satisfait	Insatisfait
Magasin 1	15	20	10	5
Magasin 2	9	21	12	8

Le résultat du test calculé par le logiciel R est le suivant.

```
        Pearson's Chi-squared test

data:  Mydata$Opinion and Mydata$Magasin
X-squared = 2.3985, df = 3, p-value = 0.4939
```

(a) Quelle est l'hypothèse nulle H_0 du test ?
(b) Quelle est la statistique du test (variable de décision) ? Expliquer comment on la calcule. Quelle est sa valeur sur l'échantillon ?
(c) Quelle est la conclusion du test ? Justifier votre réponse.
(d) A votre avis, ce test est-il pertinent ?

NOTE :
Si X suit une loi exponentielle de paramètre θ alors la fonction de densité est

$$f(x)=\begin{cases}\theta e^{-\theta x} & \text{si } x>0\\ 0 & \text{si } x\le 0\end{cases}.$$

et la fonction de répartition est

$$F(t)=\begin{cases}1-e^{-\theta t} & \text{si } t\ge 0\\ 0 & \text{si } t<0\end{cases}$$

De plus pour tout entier $k\ge 0$,

$$E\left[X^k\right]=\frac{k!}{\theta^k}.$$

Correction

Barême
Première partie (6 points)
(a) = 1pt (b) = 2pt (c)= 1pt (d)=2pt
Deuxième partie (4 points)
(a)= 1pt (b)= 2pt (c)=1pt
Troisième partie (5 points)
(a)=1.5 pt (b)= 2pt (c)=1pt pour décision + 0.5pt pour risque
Quatrième partie (6 points)

1) **(a) =1pt (b)=0.5pt (c)=1pt (d)=0.5pt**
2) **(a) =0.5pt (b)=0.5pt pour statistique+calcul + 0.5pt pour valeur (c)=1pt (d)=0.5pt**

Première partie

(a) $E(Q)=\frac{1}{2}E(\bar{Y})=\frac{1}{2}E(Y)=\frac{1}{2}E(X^2)=\frac{1}{2}\times\frac{2!}{\theta^2}=\frac{1}{\theta^2}=q$. Donc Q est sans biais

(b) $E(Y_i)=E(X_i^2)=\frac{2!}{\theta^2}=2q$ et $V(Y_i)=V(X_i^2)=E(X_i^4)-4q^2=\frac{4!}{\theta^4}-4q^2=20q^2$

$R_q(Q)=V(Q)$ car b=0

$$=\frac{1}{4}V(\bar{Y})=\frac{1}{4}\frac{V(Y)}{n}=\frac{5}{n}q^2 \underset{n\to+\infty}{\to} 0$$

donc Q converge en m.q vers q.

(c) Les Y_i étant indépendantes, d'après le T.C.L., $\bar{Y}$ converge en loi vers une loi normale N((E(Y),V(Y)/n) telle que

$$E(Y)=2q \text{ et } V(Y)=20q^2.$$

Donc Q converge en loi vers $N(q,\frac{5}{n}q^2)$.

(d) Soit on calcule la fonction de vraisemblance en fonction q et on dérive par rapport à q et non θ ; soit on travaille avec θ et on utilise la propriété :

Soit T_n un estimateur de $h(\theta)$ avec h dérivable, alors dans le cas où le support des X_i ne dépend pas de θ,

$$\text{var}(T_n)\geq\frac{(h'(\theta))^2}{I_n(\theta)}.$$

Le terme de droite est appelé la *borne de Cramer-Rao*.

Deuxième partie

(a) On cherche $P[X>7]=1-P[X\leq 7]=1-F(7)=e^{-7\times\theta}$. Or q est estimé par $\hat{q}=25$ donc θ est approché par $1/\sqrt{\hat{q}}=0.2$. Donc $P[X>7]=e^{-7\times 0.2}\approx 0.25$. Il y a donc 25% de clients mécontents.

(b) On cherche a et b tels que $P[a\leq q\leq b]=0.95$. On sait que

$$Q \sim N(q, \frac{\sigma^2}{n}) \Leftrightarrow Z = \sqrt{n}\frac{Q-q}{\sigma} \sim N(0,1)$$

d'où

$$P[a \le q \le b] = P[-b \le -q \le -a] = P[\sqrt{n}\frac{\hat{q}-b}{\sigma} \le \sqrt{n}\frac{Q-q}{\sigma} \le \sqrt{n}\frac{\hat{q}-a}{\sigma}] = 0.95$$

$$\Leftrightarrow P[b' \le Z \le a'] = 0.95$$

On suppose un risque symétrique. De plus la loi N(0,1) est symétrique par rapport à Oy donc

$$\Rightarrow \begin{cases} b' = -a' \\ P[Z \ge a'] = 0.025 \end{cases} \Rightarrow \begin{cases} a' = 1.96 \\ b' = -1.96 \end{cases} \Rightarrow \begin{cases} a = \hat{q} - 1.96\frac{\sigma}{\sqrt{n}} \\ b = \hat{q} + 1.96\frac{\sigma}{\sqrt{n}} \end{cases}$$

On remarque que σ est inconnu et dépend de q puisque $\sigma = \sqrt{5}q$. L'échantillon est grand et on peut donc approcher la valeur de σ par $\sigma \approx \sqrt{5}\hat{q} = 55.9$. Finalement, on obtient

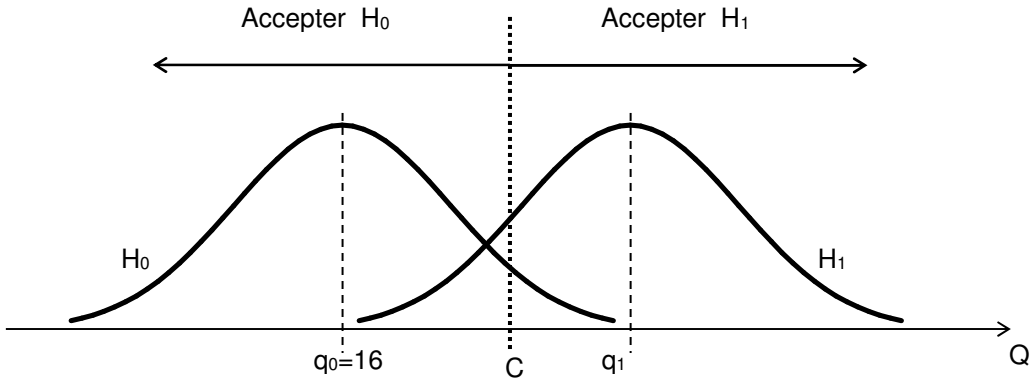
$$\begin{cases} a = 25 - 1.96\frac{55.9}{10} \approx 14.04 \\ \qquad\qquad 55.9 \qquad\qquad \Rightarrow P[14.04 \le q \le 35.96] = 0.95 \\ b = 25 + 1.96\frac{55.9}{10} \approx 35.96 \end{cases}$$

(c) Le temps moyen d'attente est donné par : $E(X) = \frac{1}{\theta} = \sqrt{q}$. D'où l'intervalle de confiance de niveau 5% est défini par

$$P[\sqrt{14.04} \le \sqrt{q} \le \sqrt{35.96}] \approx P[3.75 \le \sqrt{q} \le 6] = 0.95$$

Troisième partie

(a) Sous l'hypothèse H_0, Q suit une loi normale centrée en q_0=16 et sous l'hypothèse H_1, Q suit une loi normale centrée en $q_1 > q_0$, d'où le graphique



Donc la région critique est de la forme

$$W = \{Q > C\}.$$

(c) $\alpha = P[W \mid H_0 \text{ vraie}]$.

Supposons H_0 vraie alors Q suit une loi $N(16, \frac{5 \times 4^4}{n})$, d'où

$\alpha = P[Q > C] = P[\sqrt{n}\frac{Q-16}{16\sqrt{5}} > \sqrt{n}\frac{C-16}{16\sqrt{5}}] = P[Z > C']$, où Z suit une loi N(0,1).

$\Rightarrow C'=1.64 \Rightarrow C = 16 + \frac{1.64\times 16\times\sqrt{5}}{\sqrt{n}} \approx 22$

(e) L'échantillon de 100 clients fournit $\hat{q}=25>22$ donc le directeur accepte H_1, *i.e* qu'il décide d'ouvrir des caisses supplémentaires, avec 5% de chance de se tromper, *i.e* 5% de chance que cela soit inutile.

Quatrième partie

1) Il s'agit d'un test paramétrique de comparaison de moyennes.
   (a) Si on note μ_1 le temps moyen d'attente dans le magasin 1 et μ_2 celui du magasin 2, alors l'hypothèse nulle est $H_0 : \mu_1=\mu_2$.
   (b) L'hypothèse alternative est $H_0 : \mu_1\neq\mu_2$.
   (c) La p-valeur du test est 0.4462>>α donc on accepte H_0, c'est-à-dire qu'il n'y a pas de différence significative entre les magasins.
   (d) Le temps t'attente suit une loi exponentielle. La moyenne n'est donc pas représentative de la population à cause des valeurs extrêmes. Le test n'est donc pas adapté.
2) Il s'agit du test d'indépendance du chi-deux.
   (a) H_0 : Les variables sont indépendantes $\Leftrightarrow$ loi conjointe= produit des lois marginales
   (b) La statistique du test est la distance du chi-deux entre le tableau de effectifs observés et celui des effectifs théoriques
   $$\frac{(\text{eff. th.} - \text{eff. obs.})^2}{\text{eff. th.}}.$$
   Les effectifs théoriques sont calculés par $n\times\hat{p}_{i.}\times\hat{p}_{.j}$, où n est la taille de l'échantillon, $\hat{p}_{i.}$ est la fréquence de la modalité i de la première variable et $\hat{p}_{.j}$ est la fréquence de la modalité j de la deuxième variable. Sa valeur calculée sur l'échantillon est 2.3985.
   (c) La p-valeur du test est 0.49>>α donc on accepte H_0, c'est-à-dire que la satisfaction d'un client est indépendante du magasin dans lequel il va.
   (d) Le logiciel n'indique pas de warning. On peut donc supposer que tous les effectifs théoriques sont supérieurs à 5.