Statistiques appliquées (L3)
Groupes 14 et 15

Corrigé de l'interrogation n° 1*bis*

**Vrai ou faux**

1- Faux : la variance de la somme de deux variables aléatoires est égale à la somme de leurs variances si et seulement si leur covariance est nulle. C'est en particulier le cas lorsque les deux variables aléatoires en question sont indépendantes.

2- Vrai

3- Vrai

4- Vrai

**Exercice 1**

**Première partie**

1- La variable aléatoire $X_i$ définie par $X_i = 0$ si l'individu $i$ ne fume pas et $X_i = 1$ si l'individu $i$ fume suit une loi de Bernouilli de paramètre $p$. Le tirage s'effectue dans une très large population (par rapport à la taille de l'échantillon). Dans ce cas, même si le tirage est fait sans remise, on peut considérer que les variables $X_i$ sont indépendantes et identiquement distribuées, comme s'il avait été fait avec remise.

2- $\hat{p}_n = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ est un estimateur convergent de $p$ car $E(X_i) = p$. L'estimateur $\hat{p}_n$ n'est autre que la proportion (ou la fréquence) de fumeurs observée dans l'échantillon. Pour déterminer la loi de $\hat{p}_n$, il suffit de remarquer que $n\hat{p}_n = \sum\limits_{i=1}^{n} X_i$ est la somme de $n$ variables de Bernouilli indépendantes et de même paramètre $p$ : $n\hat{p}_n$ suit donc une loi binômiale $B(n,p)$ (la connaissance de la loi de $n\hat{p}_n$ est équivalente à la connaissance de la loi de $\hat{p}_n$)

Pour déterminer l'espérance et la variance de $\hat{p}_n$, on peut procéder de deux façons.

La première consiste à utiliser le fait que $n\hat{p}_n$ suit une loi binômiale $B(n,p)$ ce qui implique que $E(n\hat{p}_n) = np$ et $V(n\hat{p}_n) = np(1-p)$ (cette méthode exige la connaissance préalable de l'espérance et de la variance d'une loi binômiale $B(n,p)$). Or $E(n\hat{p}_n) = nE(\hat{p}_n)$ et $V(n\hat{p}_n) = n^2V(\hat{p}_n)$ donc $E(\hat{p}_n) = \frac{np}{n} = p$ et $V(\hat{p}_n) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$.

La seconde méthode (plus générale) consiste à faire le calcul classique suivant:

$E(\hat{p}_n) = E\left(\frac{1}{n}\sum\limits_{i=1}^{n} X_i\right) = \frac{1}{n}\sum\limits_{i=1}^{n} E(X_i) = \frac{1}{n}np = p$

$V(\hat{p}_n) = V\left(\frac{1}{n}\sum\limits_{i=1}^{n} X_i\right) = \frac{1}{n^2}V\left(\sum\limits_{i=1}^{n} X_i\right)$ or $V\left(\sum\limits_{i=1}^{n} X_i\right) = \sum\limits_{i=1}^{n} V(X_i)$ car les $X_i$ sont supposés indépendants, donc

$V(\hat{p}_n) = \frac{1}{n^2}\sum\limits_{i=1}^{n} V(X_i) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$ (rappel: $V(X_i) = p(1-p)$ car $X_i$ suit une loi de Bernouilli de paramètre $p$).

3- Le nombre d'observations est $n = 200$ et la proportion de fumeurs observée dans l'échantillon est $\hat{p}_{200} = 0.35$. On a donc $n\hat{p}_n(1-\hat{p}_n) = 200 \times 0.35 \times 0.65 = 45.5 > 15$. Il paraît donc légitime de penser que la condition $np(1-p) > 15$ est vérifiée, ce qui permet d'utiliser l'approximation normale de la loi binômiale (rappelons qu'il n'est pas possible de calculer $np(1-p)$ car $p$ est un paramètre inconnu, d'où le recours à un estimateur de $p$, à savoir $\hat{p}_n$). Cette approximation permet d'écrire que:

$$n\hat{p}_n \rightsquigarrow N(np, np(1-p))$$

ce qui revient à dire que :

$$\hat{p}_n \rightsquigarrow N\left(p, \frac{p(1-p)}{n}\right)$$

ou encore :

$$U_n = \frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow N(0,1)$$

Il est demandé de déterminer un intervalle bilatéral de confiance à 95%. On a donc besoin de déterminer le nombre $a$ tel que $P(|U_n| \leq a) = 0.95$. Le caractère symétrique de la loi normale centrée réduite permet d'écrire que: $P(U_n \leq a) = P(|U_n| \leq a) + \frac{1-P(|U_n|\leq a)}{2} = 0.95 + 0.025 = 0.975$. Sur une table statistique, on trouve $P(U_n \leq 1.96) = 0.975$ c'est-à-dire $a = 1.96$. On a donc:

$$P\left(-1.96 \leq \frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \leq 1.96\right) = 0.95$$

égalité qu'on peut réécrire sous la forme

$$P\left(-1.96\sqrt{\frac{p(1-p)}{n}} \leq \hat{p}_n - p \leq 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

ou encore

$$P\left(\hat{p}_n - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p}_n + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

Les bornes de l'intervalle de confiance ne devant pas contenir de paramètre inconnu, on remplace $p$ par son estimation $\hat{p}_n$ (cette approximation est valable pour $n$ suffisamment grand; on suppose que c'est le cas ici). On obtient alors

$$P\left(\hat{p}_n - 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leq p \leq \hat{p}_n + 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right) \simeq 0.95$$

Application numérique : pour $n = 200$ et $\hat{p}_{200} = 0.35$ on trouve l'intervalle $[0.2839, 0.4161]$.

**Deuxième partie**

1- Le meilleur estimateur de $m$ est la moyenne empirique $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Cet estimateur est convergent d'après la loi des grands nombres. Il est sans biais car, par linéarité de l'espérance, on a: $E(\bar{Y}_n) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{nm}{n} = m$. Une autre façon d'établir que l'estimateur est convergent est la suivante: on commence par établir qu'il est sans biais : $E(\bar{Y}_n) = m$, puis on calcule sa variance : $V(\bar{Y}_n) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i)$ car les variables $Y_i$ sont indépendantes, d'où $V(\bar{Y}_n) = \frac{\sigma^2}{n}$; on peut donc affirmer que

$$E\left(\left(\bar{Y}_n - m\right)^2\right) = E\left(\left(\bar{Y}_n - E(\bar{Y}_n)\right)^2\right) = V(\bar{Y}_n) = \frac{\sigma^2}{n} \underset{n\to+\infty}{\longrightarrow} 0$$

c'est-à-dire que $\bar{Y}_n$ converge en moyenne quadratique vers $m$. Or la convergence en moyenne quadratique implique la convergence en probabilité, donc $\bar{Y}_n$ converge en probabilité vers $m$ (lorsqu'on parle de la convergence d'un estimateur sans précision supplémentaire, c'est de la convergence en probabilité qu'il s'agit).

L'estimateur $\bar{Y}_n$ est la moyenne de $n$ variables normales indépendantes. Il suit donc une loi normale, de paramètres $E(\bar{Y}_n) = m$ et $V(\bar{Y}_n) = \frac{\sigma^2}{n} = \frac{16}{200} = \frac{2}{25}$.

2- On sait que $\bar{Y}_n \rightsquigarrow N(m, \frac{\sigma^2}{n})$ donc, en centrant et en réduisant la variable $\bar{Y}_n$, on obtient:

$$T_n = \frac{\bar{Y}_n - m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0, 1)$$

Notons que cette relation est valable pour toute valeur de $n$ (y compris les petites valeurs de $n$).

Il est demandé de déterminer un intervalle de confiance bilatéral à 90%. Nous avons donc besoin de déterminer le nombre $b$ tel que $P(|T_n| \leq b) = 0.90$. Or $T_n$ est symétrique donc $P(T_n \leq b) = P(|T_n| \leq a) + \frac{1-P(|T_n|\leq a)}{2} = 0.90 + 0.05 = 0.95$.

**Remarque:** La valeur de $b$, à savoir 1.645, est donnée dans l'énoncé. Cette valeur ne se lit pas directement sur la table statistique de la loi normale centrée. En effet, sur la table statistique de $T_n \rightsquigarrow N(0,1)$, on trouve $P(T_n \leq 1.64) = 0.9495$ et $P(T_n \leq 1.65) = 0.9505$. On sait donc que $b$ se trouve entre 1.64 et 1.65. On procède alors par approximation (ou extrapolation) linéaire, c'est-à-dire qu'on fait l'approximation suivante: $P(T_n \leq 1.645) \simeq \frac{P(T_n \leq 1.64)+P(T_n \leq 1.65)}{2} = \frac{0.9495+0.9505}{2} = 0.95$

Ainsi, on a :

$$P\left(-1.645 \leq \frac{\bar{Y}_n - m}{\frac{\sigma}{\sqrt{n}}} \leq 1.645\right) = 0.90$$

c'est-à-dire:

$$P\left(-1.645\frac{\sigma}{\sqrt{n}} \leq \bar{Y}_n - m \leq 1.645\frac{\sigma}{\sqrt{n}}\right) = 0.90$$

qu'on peut réécrire sous la forme :

$$P\left(\bar{Y}_n - 1.645\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{Y}_n + 1.645\frac{\sigma}{\sqrt{n}}\right) = 0.90$$

Application numérique: pour $n = 200$, $\sigma = 4$ et $\bar{Y}_{200} = 12$, on trouve l'intervalle $[11.5347, 12.4653]$

3- La différence fondamentale entre l'estimation par intervalle de confiance de la proportion $p$ et l'estimation par intervalle de confiance de la consommation journalière moyenne $m$ est que la première n'est valable que si la taille de l'échantillon est suffisamment grande (en l'occurence $n > \frac{15}{p(1-p)}$) ce qui permet de recourir à une approximation de la loi de l'estimateur par une loi normale, alors que la seconde est valable quelle que soit la taille de l'échantillon.

**Exercice 2**

1-a- La moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur convergent de $E(X)$. Or $E(X) = \theta + 1$, donc $\hat{\theta}_n = \bar{X}_n - 1$ est un estimateur convergent de $\theta$.

Cet estimateur est sans biais car:

$$E(\hat{\theta}_n) = E\left(\bar{X}_n - 1\right) = E\left(\bar{X}_n\right) - 1 = E(X) - 1 = \theta$$

Calculons la variance de l'estimateur $\hat{\theta}_n$ :

$$V(\hat{\theta}_n) = V\left(\bar{X}_n - 1\right) = E\left((\bar{X}_n - 1 - E\left(\bar{X}_n - 1\right))^2\right)$$

or $\bar{X}_n - 1 - E\left(\bar{X}_n - 1\right) = \bar{X}_n - 1 - \left(E\left(\bar{X}_n\right) - 1\right) = \bar{X}_n - E\left(\bar{X}_n\right)$ donc

$$V(\hat{\theta}_n) = E\left((\bar{X}_n - E\left(\bar{X}_n\right))^2\right) = V(\bar{X}_n)$$

Par ailleurs, $V(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$ car les $X_i$ sont indépendants et $V(X_i) = V(X)$ car les $X_i$ sont identiquement distribués donc $V(\bar{X}_n) = \frac{nV(X)}{n^2} = \frac{1}{n}$ d'où

$$V(\hat{\theta}_n) = \frac{1}{n}$$

1-b- D'après le TCL, on sait que, lorsque $n \to +\infty$

$$\frac{\bar{X}_n - E(X)}{\sqrt{\frac{V(X)}{n}}} \xrightarrow{loi} N(0,1)$$

or $\bar{X}_n = \hat{\theta}_n + 1$ , $E(X) = \theta + 1$ et $V(X) = 1$ donc, lorsque $n \to +\infty$

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\frac{1}{n}}} \xrightarrow{loi} N(0,1)$$

On dispose d'un échantillon de grande taille $n$ donc on peut faire l'approximation :

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\frac{1}{n}}} \rightsquigarrow N(0,1)$$

Il s'ensuit que

$$P(-1.96 \leq \frac{\hat{\theta}_n - \theta}{\sqrt{\frac{1}{n}}} \leq 1.96) \simeq 0.95$$

ce qu'on peut réécrire sous la forme :

$$P(\hat{\theta}_n - 1.96\frac{1}{\sqrt{n}} \leq \theta \leq \hat{\theta}_n + 1.96\frac{1}{\sqrt{n}}) \simeq 0.95$$

Ainsi, $\left[\hat{\theta}_n - 1.96\frac{1}{\sqrt{n}}, \hat{\theta}_n + 1.96\frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance à 95% du paramètre $\theta$.

2- La fonction de vraisemblance de l'échantillon $(x_1, x_2, ..., x_n)$ est

$$L(\theta, x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i)$$

donc

$$L(\theta, x_1, ..., x_n) = \begin{cases} e^{n\theta - \sum\limits_{i=1}^{n} x_i} & \text{si} \quad \forall i \ \ x_i \geq \theta \\ 0 & \text{sinon} \end{cases}$$

L'assertion $\forall i \ \ x_i \geq \theta$ étant équivalente à $\theta \leq \min(x_i)$, on peut réécrire la fonction de vraisemblance sous la forme:

$$L(\theta, x_1, ..., x_n) = \begin{cases} e^{n\theta - \sum\limits_{i=1}^{n} x_i} & \text{si} \quad \theta \leq \min(x_i) \\ 0 & \text{si} \quad \theta > \min(x_i) \end{cases}$$

On en déduit la log-vraisemblance

$$l(\theta, x_1, ..., x_n) = \begin{cases} n\theta - \sum\limits_{i=1}^{n} x_i & \text{si} \quad \theta \leq \min(x_i) \\ -\infty & \text{si} \quad \theta > \min(x_i) \end{cases}$$

**Question bonus**: l'estimateur du maximum de vraisemblance $\hat{\theta}_{MV}$ est donné par la valeur de $\theta$ qui maximise la fonction de vraisemblance (ou de log-vraisemblance, c'est la même chose car le logarithme est une fonction strictement croissante) à $(x_1, x_2, ..., x_n)$ donnés. La fonction $\theta \to L(\theta, x_1, ..., x_n)$ est strictement croissante et strictement positive sur $]-\infty, \min(x_i)]$ et est nulle sur $]\min(x_i), +\infty[$. Elle atteint donc son maximum au point $\theta = \min(x_i)$ (pour s'en convaincre tracer le graphe de $\theta \to L(\theta, x_1, ..., x_n)$ à $(x_1, x_2, ..., x_n)$ donnés). L'estimateur du maximu de vraisemblance est donc $\hat{\theta}_{MV} = \min(x_i)$.

**Remarque**: La fonction $\theta \to L(\theta, x_1, ..., x_n)$ est discontinue, et donc non dérivable, au point $\theta = \min(x_i)$. On ne peut donc pas procéder par dérivation pour déterminer le maximum de cette fonction.