Corrigé de l'interrogation n°2

Question [4]**:**

Supposons qu'on dispose d'un échantillon $(X_1, X_2, ..., X_n)$ tiré de façon i.i.d. dans une loi $N(m, \sigma^2)$. On cherche à déterminer un intervalle bilatéral de confiance au niveau $1-\alpha$ pour l'espérance m (on se restreint aux intervalles bilatéraux symétriques à cause du caractère symétrique de la loi normale). On note $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ la moyenne empirique, qui est un estimateur convergent de m.

1^{er} cas : l'écart-type σ est connu.

Il est aisé de montrer que $\bar{X}_n \rightsquigarrow N(m, \frac{\sigma^2}{n})$, donc la statistique U_n définie par $U_n = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$ suit la loi $N(0,1)$. Pour déterminer un intervalle bilatéral de confiance au niveau $1-\alpha$ pour le paramètre m, on commence par chercher le nombre u tel que $P(-u \leq U_n \leq u) = 1-\alpha$. Le caractère symétrique (par rapport à 0) de la densité de la loi $N(0,1)$ permet d'affirmer que u n'est autre que le fractile d'ordre $1-\frac{\alpha}{2}$ de la loi $N(0,1)$ (c'est-à-dire que $\Phi(u) = 1-\frac{\alpha}{2}$ où Φ est la fonction de répartition de $N(0,1)$). Une fois que u est déterminé en utilisant une table statistique de la loi $N(0,1)$, il suffit d'exploiter le fait que $P\left(-u \leq \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \leq u\right) = P\left(\bar{X}_n - u\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X}_n + u\frac{\sigma}{\sqrt{n}}\right)$ pour pouvoir affirmer que $P\left(\bar{X}_n - u\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X}_n + u\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$ c'est-à-dire que $\left[\bar{X}_n - u\frac{\sigma}{\sqrt{n}}, \bar{X}_n + u\frac{\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance au niveau $1-\alpha$ pour l'espérance m.

2^{nd} cas: l'écart-type σ est inconnu.

Dans ce cas, on ne peut plus utiliser la statistique U_n qui contient, en plus du paramètre inconnu m qu'on cherche à estimer, un autre paramètre inconnu, à savoir σ.Il va donc falloir utiliser un estimateur convergent de σ.Pour cela, on introduit la variance empirique modifiée $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$ qui se trouve être un estimateur convergent (et sans biais) de la variance σ^2. On peut alors déterminer un intervalle de confiance pour m car on connaît la loi de la statistique $Z_n = \frac{\bar{X}_n - m}{\frac{S_n}{\sqrt{n}}}$. En effet, on sait que : $Z_n = \frac{\bar{X}_n - m}{\frac{S_n}{\sqrt{n}}} \rightsquigarrow T_{n-1}$où T_{n-1} désigne la loi de Student à $n-1$ degrés de liberté. Le même argument de symétrie que celui invoqué pour $N(0,1)$ permet d'affirmer que le nombre t tel que $P(-t \leq Z_n \leq t) = 1-\alpha$ est le fractile d'ordre $1-\frac{\alpha}{2}$ de la loi T_{n-1}. Une fois que t est déterminé en utilisant une table statistique de la loi T_{n-1} , il suffit d'exploiter le fait que

$$P\left(-t \leq \frac{\bar{X}_n - m}{\frac{S_n}{\sqrt{n}}} \leq t\right) = P\left(\bar{X}_n - t\frac{S_n}{\sqrt{n}} \leq m \leq \bar{X}_n + t\frac{S_n}{\sqrt{n}}\right)$$

pour pouvoir affirmer que $\left[\bar{X}_n - t\frac{S_n}{\sqrt{n}}, \bar{X}_n + t\frac{S_n}{\sqrt{n}}\right]$ est un intervalle de confiance au niveau $1-\alpha$ pour l'espérance m.

Remarque importante : Lorsque la taille de l'échantillon est **suffisamment grande** (en pratique, on considère que c'est le cas pour $n > 30$), on peut approcher la loi T_{n-1} par la loi $N(0,1)$. Dans ces conditions, on a approximativement :

$$Z_n = \frac{\bar{X}_n - m}{\frac{S_n}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

ce qui implique que le nombre t tel que $P(-t \leq Z_n \leq t) = 1-\alpha$ est le fractile d'ordre $1-\frac{\alpha}{2}$ de la loi $N(0,1)$, autrement dit ce n'est autre que le nombre qu'on a appelé u dans le traitement du cas d'un écart-type connu. Ceci permet d'affirmer que, dans ce cas particulier, $\left[\bar{X}_n - u\frac{S_n}{\sqrt{n}}, \bar{X}_n + u\frac{S_n}{\sqrt{n}}\right]$ est un intervalle de confiance au niveau $1-\alpha$ pour l'espérance m.

Une autre façon (plus générale) d'aboutir à ce résultat est de dire que pour n "suffisamment grand" on peut approcher la loi de $\frac{\bar{X}_n - m}{\frac{S_n}{\sqrt{n}}}$ par celle de $\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$ (car S_n est un estimateur convergent de σ). Ceci revient à dire que lorsque n est suffisamment grand, on peut remplacer σ par S_n dans la relation $\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1)$, obtenant ainsi que $\frac{\bar{X}_n - m}{\frac{S_n}{\sqrt{n}}} \rightsquigarrow N(0,1)$.

Exercice 1 [11]**:**

1.a. Pour tout $i \in \{1, ..., N\}$, on a $P(X_i = 1) = p$ et $P(X_i = 0) = 1 - p$. On peut regrouper ces deux égalités, en écrivant que pour $x_i \in \{0,1\}$, $P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$. La fonction de vraisemblance associée à la loi de Bernouilli de paramètre p est donc:

$$L(p; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

La log-vraisemblance d'un échantillon $(x_1, x_2, ..., x_n)$ est donc égale à :

$$\begin{aligned} \ln L(p; x_1, x_2, ..., x_n) &= \sum_{i=1}^{n} [x_i \ln p + (1 - x_i) \ln(1-p)] \\ &= \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p) \end{aligned}$$

1.b. Afin de déterminer l'estimateur du maximum de vraisemblance, il faut maximiser la log-vraisemblance par rapport à la variable p à $(x_1, x_2, ..., x_n)$ donnés. La première étape consiste à résoudre l'équation correspondant à la condition du premier ordre :

$$\frac{\partial \ln L}{\partial p}(p, x_1, x_2, ..., x_n) = 0$$

or

$$\frac{\partial \ln L}{\partial p}(p, x_1, x_2, ..., x_n) = \frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} x_i\right)$$

donc : $\frac{\partial \ln L}{\partial p}(p, x_1, x_2, ..., x_n) = 0 \Longleftrightarrow \frac{1}{p}\sum_{i=1}^{n} x_i = \frac{1}{1-p}\left(n - \sum_{i=1}^{n} x_i\right) \Longleftrightarrow (1-p)\sum_{i=1}^{n} x_i = p\left(n - \sum_{i=1}^{n} x_i\right)$

d'où $\frac{\partial \ln L}{\partial p}(p, x_1, x_2, ..., x_n) = 0 \Longleftrightarrow \sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i = pn - p\sum_{i=1}^{n} x_i \Longleftrightarrow \sum_{i=1}^{n} x_i = pn \Longleftrightarrow p = \frac{1}{n}\sum_{i=1}^{n} x_i$

La fonction $p \to \ln L(p, x_1, x_2, ..., x_n)$ atteint donc un extrémum (c'est-à-dire un maximum ou un minimum) au point $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Il s'agit bien d'un maximum car la fonction $p \to \ln L(x_1, x_2, ..., x_n, p)$ est concave. En effet, pour tout $(p; x_1, x_2, ..., x_n)$, on a :

$$\frac{\partial^2 \ln L}{\partial p^2}(p, x_1, x_2, ..., x_n) = -\frac{1}{p^2}\sum_{i=1}^{n} x_i - \frac{1}{(1-p)^2}\left(n - \sum_{i=1}^{n} x_i\right) < 0$$

car $\sum_{i=1}^{n} x_i \geq 0$ et $n - \sum_{i=1}^{n} x_i \geq 0$ (du fait que $x_i \in \{0,1\}$ pour tout i) et ces deux quantités ne peuvent pas être simultanément nulles. En particulier, la condition du second ordre $\frac{\partial^2 \ln L}{\partial p^2}(\hat{p}_n, x_1, x_2, ..., x_n) < 0$ est satisfaite.

Conclusion: L'estimateur du maximum de vraisemblance du paramètre p est $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1.c. $E(\hat{p}_n) = \frac{1}{n}\sum\limits_{i=1}^{n} E(X_i) = \frac{np}{n} = p$ (l'estimateur $\hat{p}_n$ est sans biais).

$V(\hat{p}_n) = V\left(\frac{1}{n}\sum\limits_{i=1}^{n} X_i\right) = \frac{1}{n^2}V\left(\sum\limits_{i=1}^{n} X_i\right)$ 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).

Il est clair que $V(\hat{p}_n) \underset{n\longrightarrow+\infty}{\longrightarrow} 0$ donc $\hat{p}_n$ est un estimateur sans biais dont la variance tend vers 0 lorsque n tend vers $+\infty$, ce qui permet d'affirmer que $\hat{p}_n$ est un estimateur convergent.

1.d. L'expression de la quantité d'information de Fisher est :

$$I_n(p) = E\left(\left[\frac{\partial \ln L(p, X_1, ..., X_n)}{\partial p}\right]^2\right) = -E\left(\frac{\partial^2 \ln L(p, X_1, ..., X_n)}{\partial p^2}\right)$$

La deuxième égalité est valable sous les conditions de Cramer-Rao : pour les lois usuelles vues dans ce cours, il suffit de vérifier que le support de la loi ne dépend pas du paramètre à estimer. Cest le cas pour la loi de Bernouilli puisque son support est $\{0, 1\}$ et ne dépend donc pas du paramètre p.

D'après ce qui précède, on a :

$$\frac{\partial^2 \ln L(p, X_1, ..., X_n)}{\partial p^2} = -\frac{1}{p^2}\sum_{i=1}^{n} X_i - \frac{1}{(1-p)^2}\left(n - \sum_{i=1}^{n} X_i\right)$$

donc

$$E\left(\frac{\partial^2 \ln L(p, X_1, ..., X_n)}{\partial p^2}\right) = -\frac{1}{p^2}\sum_{i=1}^{n} E(X_i) - \frac{1}{(1-p)^2}\left(n - \sum_{i=1}^{n} E(X_i)\right)$$

d'où

$$E\left(\frac{\partial^2 \ln L(p, X_1, ..., X_n)}{\partial p^2}\right) = -\frac{np}{p^2} - \frac{n(1-p)}{(1-p)^2} = -\frac{n}{p} - \frac{n}{1-p} = \frac{-n + np - np}{p(1-p)} = -\frac{n}{p(1-p)}$$

donc

$$I_n(p) = \frac{n}{p(1-p)}$$

L'estimateur $\hat{p}_n$ est un estimateur sans biais et sa variance vérifie

$$V(\hat{p}_n) = \frac{1}{I_n(p)}$$

Il s'agit donc d'un estimateur efficace.

2.a. On sait que $n\hat{p}_n = \sum\limits_{i=1}^{n} X_i \rightsquigarrow B(n, p)$. Sous l'hypothèse de validité de l'approximation normale de la loi binômiale, on approche $B(n, p)$ par $N(np, np(1-p))$ ce qui permet d'affirmer qu'on a approximativement : $n\hat{p}_n \rightsquigarrow N(np, np(1-p))$ d'où $\hat{p}_n \rightsquigarrow N(p, \frac{p(1-p)}{n})$ et donc :$\frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow N(0, 1)$

Remarque :l'approximation normale de la loi binômiale n'est rien d'autre qu'un cas particulier de l'approximation issue du théorème central limite pour n "suffisamment grand", appliquée à la loi de Bernouilli (la notion de "suffisamment grand" est dans ce cas plus précise puisqu'on estime qu'on peut faire l'approximation lorsque $np(1-p) > 15$)).

$\hat{p}_n$ est un estimateur convergent de p donc pour n suffisamment grand (on estime que c'est le cas ici), les lois de $\frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}}$ et de $\frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}}$ sont "proches", ce qui permet de dire qu'on a approximativement :

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

Posons $U_n = \frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}}$. $P\left(-u \leq U_n \leq u\right) = 0.90 = 2F(u) - 1$ où F est la fonction de répartition de la loi $N(0,1)$, d'où $F(u) = 0.95$. Sur la table statistique de $N(0,1)$, on trouve : $F(1.64) = 0.9495$ et $F(1.65) = 0.9505$, or il est clair que $0.95 = \frac{1}{2} \times 0.9495 + \frac{1}{2} \times 0.9505$ donc par interpolation linéaire $u \simeq \frac{1}{2} \times 1.64 + \frac{1}{2} \times 1.65 = 1.645$. (Vous pouvez utiliser $u = 1,64$)

$$P\left(-1.645 \leq U_n \leq 1.645\right) = P\left(-1.645 \leq \frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}} \leq 1.645\right) \simeq 0.90$$

donc :

$$P\left(\hat{p}_n - 1.645\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leq p \leq \hat{p}_n + 1.645\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right) \simeq 0.90$$

Pour les valeurs observées $\hat{p}_n = 0,4$ et $n = 400$, on obtient l'intervalle de confiance pour p au seuil de 90% suivant : $\left[0.36\ ,\quad 0.44\right]$.

2.b. Minorer p (c'est-à-dire chercher une borne inférieure pour p) revient à majorer U_n puisque U_n est clairement décroissante en p. On commence donc par chercher le nombre α tel que $P(U_n \leq \alpha) = 0.95$ c'est-à-dire tels que $F(\alpha) = 0.95$ (Cf. question précédente) $\alpha = 1.645$. On a donc

$$P\left(\frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}} \leq 1.645\right) \simeq 0.95$$

qu'on peut réécrire sous la forme

$$P\left(-p \leq -\hat{p}_n + 1.645\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right) \simeq 0.95$$

d'où :

$$P\left(p \geq \hat{p}_n - 1.645\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right) \simeq 0.95$$

donc $\hat{p}_n - 1.645\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$ est une borne inférieure de confiance à 95% pour le paramètre p. Pour une taille d'échantillon $n = 400$ et une valeur observée de 0.4 pour $\hat{p}_n$, cette borne prend la valeur 0.36.

Remarque : une borne inférieure de confiance correspond à un intervalle unilatéral *à droite*.

2.b. Lorsqu'on veut déterminer une borne supérieure de confiance pour p, on cherche à majorer p, or majorer p revient à minorer U_n. On commence donc par chercher le nombre β tel que $P(U_n \geq \beta) = 0,95$. Or $P(U_n \geq \beta) = 1 - F(\beta) = F(-\beta)$ donc $F(-\beta) = 0.95$; par conséquent, $-\beta = \alpha = 1.645$ (Cf. question précédente) d'où $\beta = -1.645$. On a donc:

$$P\left(\frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}} \geq -1.645\right) \simeq 0.95$$

qu'on peut réécrire sous la forme :

$$P\left(-p \geq -\hat{p}_n - 1.645\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right) \simeq 0.95$$

d'où :

$$P\left(p \leq \hat{p}_n + 1.645\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right) \simeq 0.95$$

donc $\hat{p}_n + 1.645\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$ est une borne supérieure de confiance à 95% pour le paralètre p.Pour une taille d'échantillon $n = 400$ et une valeur observée de 0.4 pour $\hat{p}_n$, cette borne prend la valeur 0.44.

Remarque : une borne supérieure de confiance correspond à un intervalle unilatéral de confiance *à gauche*.

3.a. D'après ce qui précède, on sait que $\hat{p}_n \rightsquigarrow N(p, \frac{p(1-p)}{n})$. Par analogie, en posant $\hat{q}_m = \frac{1}{m}\sum_{j=1}^{m} Y_j$, on a : $\hat{q}_m \rightsquigarrow N(q, \frac{q(1-q)}{m})$. Par ailleurs $\hat{p}_n$ est une combinaison linéaire de $X_1, X_2, ..., X_n$ et $\hat{q}_m$ est une combinaison linéaire de $Y_1, Y_2, ..., Y_m$ or les variables $X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m$ sont indépendantes donc les estimateurs $\hat{p}_n$ et $\hat{q}_m$ sont indépendants. Il s'avère donc que $\hat{p}_n$ et $\hat{q}_m$ sont des variables aléatoires normales indépendantes. Ceci implique que $\hat{p}_n - \hat{q}_m$ suit une loi normale d'espérance

$$E(\hat{p}_n - \hat{q}_m) = E(\hat{p}_n) - E(\hat{q}_m) = p - q$$

et de variance

$$\begin{aligned} V(\hat{p}_n - \hat{q}_m) &= V(\hat{p}_n) + V(\hat{q}_m) - \underbrace{2cov(\hat{p}_n, \hat{q}_m)}_{\substack{= 0 \text{ car } \hat{p}_n \text{ et } \hat{q}_m \\ \text{sont indépendants}}} \\ &= \frac{p(1-p)}{n} + \frac{q(1-q)}{m} \end{aligned}$$

ce qu'on peut résumer par :

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

donc en centrant et en réduisant cette variable, on obtient :

$$\frac{\hat{p}_n - \hat{q}_m - (p-q)}{\sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{m}}} \rightsquigarrow N(0,1)$$

Or on sait que si une variable aléatoire U suit la loi $N(0,1)$ alors $P\left(|U| \leq 1.96\right) = 0.95$ (on obtient classiquement la valeur 1.96 en utilisant la formule $P\left(|U| \leq u\right) = 2F(u) - 1$ ou un raisonnement graphique exploitant le caractère symétrique de la densité de la loi $N(0,1)$), donc

$$P\left(\left|\frac{\hat{p}_n - \hat{q}_m - (p-q)}{\sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{m}}}\right| \leq 1,96\right) = 0.95$$

d'où

$$P\left(-1,96\sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{m}} \leq \hat{p}_n - \hat{q}_m - (p-q) \leq 1,96\sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{m}}\right) = 0,95$$

donc

$$P\left(-(\hat{p}_n - \hat{q}_m) - 1,96\sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{m}} \leq -(p-q) \leq -(\hat{p}_n - \hat{q}_m) + 1,96\sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{m}}\right) = 0.95$$

et par conséquent

$$P\left(\hat{p}_n - \hat{q}_m - 1,96\sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{m}} \leq p - q \leq \hat{p}_n - \hat{q}_m + 1,96\sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{m}}\right) = 0.95$$

Les bornes d'un intervalle de confiance ne devant pas contenir de paramètre inconnu, on y remplace p et q par leurs estimateurs respectifs $\hat{p}_n$ et $\hat{q}_m$ (cette approximation est valable pour n et m "suffisamment grands" car $\hat{p}_n$ et $\hat{q}_m$ sont des estimateurs convergents de p et q). Ainsi :

$$P\left(\hat{p}_n - \hat{q}_m - 1,96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{\hat{q}_m(1-\hat{q}_m)}{m}} \leq p - q \leq \hat{p}_n - \hat{q}_m + 1,96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{\hat{q}_m(1-\hat{q}_m)}{m}}\right) \simeq 0.95$$

donc $\left[\hat{p}_n - \hat{q}_m - 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{\hat{q}_m(1-\hat{q}_m)}{m}}, \hat{p}_n - \hat{q}_m + 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{\hat{q}_m(1-\hat{q}_m)}{m}}\right]$ est un intervalle bilatéral de confiance à 95% pour $p - q$.

3.b. Notons n la taille de l'échantillon masculin, m la taille de l'échantillon féminin, $\hat{p}_n$ la proportion empirique de fumeurs dans l'échantillon masculin et $\hat{q}_m$ la proportion empirique de fumeuses dans l'échantillon féminin. Les tailles des échantillons sont $n = 300$ et $m = 200$, la valeur observée de $\hat{p}_n$ est $\frac{105}{300} = 0.35$ et la valeur observée de $\hat{q}_m$ est $\frac{48}{200} = 0.24$. Il est aisé de vérifier que $300 \times 0.35 \times 0.65 = 68.25 > 15$ et $200 \times 0.24 \times 0.76 = 36.48 > 15$, ce qui permet de considérer que l'approximation normale de la loi binômiale est raisonnable pour les deux échantillons. On peut donc appliquer le résultat de la question précédente: l'intervalle $\left[0.35 - 0.24 - 1.96\sqrt{\frac{0.35\times0.65}{300} + \frac{0.24\times0.76}{200}}, 0.35 - 0.24 + 1.96\sqrt{\frac{0.35\times0.65}{300} + \frac{0.24\times0.76}{200}}\right]$, c'est-à-dire $[0.03, 0.19]$, est un intervalle de confiance à à 95% pour $p - q$.

Exercice 2 [10]**:**

1. L'estimation par la méthode des moindres carrés ordinaires se fait à partir de la minimisation de la somme des carrés des résidus $\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_i)^2$. Les équations normales s'obtiennent à partir des conditions du premier ordre :

$$\frac{\partial\left(\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2\right)}{\partial \beta_2} = 0$$

$$\frac{\partial\left(\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2\right)}{\partial \beta_1} = 0$$

ces deux conditions étant vérifiées pour $\beta_2 = \widehat{\beta_2}$ et $\beta_1 = \widehat{\beta_1}$.

Or $\frac{\partial\left(\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2\right)}{\partial \beta_2} = -2\sum_{i=1}^{N} x_i(y_i - \beta_1 - \beta_2 x_i) = -2\sum_{i=1}^{N}\left(x_i y_i - \beta_1 x_i - \beta_2 a x_i^2\right) = -2\left[\sum_{i=1}^{N} x_i y_i - \beta_2 \sum_{i=1}^{N} x_i^2 - \beta_1 \sum_{i=1}^{N} x_i\right]$

et $\frac{\partial\left(\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2\right)}{\partial \beta_1} = -2\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i) = -2\left[\sum_{i=1}^{N} y_i - \beta_1 N - \beta_{2i}\sum_{i=1}^{N} x_i\right]$ donc les estimateurs $\widehat{\beta_1}$ et $\widehat{\beta_2}$ vérifient les deux équations suivantes (dites équations normales) :

$$\sum_{i=1}^{N} x_i y_i - \widehat{\beta_2}\sum_{i=1}^{N} x_i^2 - \widehat{\beta_1}\sum_{i=1}^{N} x_i = 0 \tag{1}$$

$$\sum_{i=1}^{N} y_i - \widehat{\beta_2}\sum_{i=1}^{N} x_i - \widehat{\beta_1} N = 0 \tag{2}$$

En divisant par N les deux membres de l'équation (2) on obtient l'expression de $\widehat{\beta_1}$ en fonction de $\widehat{\beta_2}$:

$$\widehat{\beta_1} = \bar{y} - \widehat{\beta_2}\bar{x}$$

et en remplaçant $\widehat{\beta_1}$dans l'équation (1) par son expression, on obtient :

$$\sum_{i=1}^{N} x_i y_i - \widehat{\beta_2} \sum_{i=1}^{N} x_i^2 - \left(\bar{y} - \widehat{\beta_2}\bar{x}\right) \sum_{i=1}^{N} x_i = 0$$

or $\sum\limits_{i=1}^{N} x_i = N\bar{x}$ donc

$$\sum_{i=1}^{N} x_i y_i - \widehat{\beta_2} \sum_{i=1}^{N} x_i^2 - N\bar{x}\bar{y} + \widehat{\beta_2} N\bar{x}^2 = 0$$

d'où

$$\widehat{\beta_2} \sum_{i=1}^{N} x_i^2 - \widehat{\beta_2} N\bar{x}^2 = \sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}$$

et par conséquent :

$$\widehat{\beta_2} = \frac{\sum\limits_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{\sum\limits_{i=1}^{N} x_i^2 - N\bar{x}^2}$$

2. Les estimateurs des MCO $\widehat{\beta_1}$ et $\widehat{\beta_2}$ sont sans biais, linéaires par rapport aux variables dépendantes y_i et de variance minimale dans la classe des estimateurs linéaires sans biais. Cette dernière propriété porte sur l'efficacité *relative* des estimateurs des MCO par rapport aux autres estimateurs linéaires sans biais (Théorème de Gauss-Markov). En aucun cas, ceci n'implique que les estimateurs $\widehat{\beta_1}$ et $\widehat{\beta_2}$ sont efficaces au sens où leur variance atteint la borne FDCR.

Le modèle $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ pour $i = 1, ..., N$ peut se réecrire sous la forme vectorielle : $y = X\beta + \varepsilon$ (voir TD), les estimateurs de MCO : $\widehat{\beta} = (X'X)^{-1}\ X'y = (X'X)^{-1}\ [X'X\beta + \varepsilon] = (X'X)^{-1} X'X\beta + (X'X)^{-1}\varepsilon = \beta + (X'X)^{-1}\varepsilon$

$E\left(\widehat{\beta}\right) = E\left(\beta + (X'X)^{-1}\varepsilon\right) = E(\beta) + E\left((X'X)^{-1}\varepsilon\right) = \beta + (X'X)^{-1} E(\varepsilon)$ or $E(\varepsilon) = 0$

$E\left(\widehat{\beta}\right) = \begin{bmatrix} E\left(\widehat{\beta_1}\right) \\ E\left(\widehat{\beta_2}\right) \end{bmatrix} = \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

3. La variance du résidu σ^2 peut être estimée sans biais par $\hat{\sigma}^2 = \frac{1}{N-2}\sum\limits_{i=1}^{N}\hat{\varepsilon}_i^2$ où $\hat{\varepsilon}_i$ est défini par $\hat{\varepsilon}_i = y_i - \widehat{\beta_2} x_i - \widehat{\beta_1}$.

4.a. On sait que $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ pour $i = 1, ..., N$ donc $\sum\limits_{i=1}^{N} y_i = \beta_1 N + \beta_2 \sum\limits_{i=1}^{N} x_i + \sum\limits_{i=1}^{N}\varepsilon_i$. En divisant par N les deux membres de cette équation, on obtient

$$\bar{y} = \beta_1 + \beta_2\bar{x} + \bar{\varepsilon}$$

Ainsi :

$$y_i - \bar{y} = \beta_1 + \beta_2 x_i + \varepsilon_i - (\beta_1 + \beta_2\bar{x} + \bar{\varepsilon}) = \beta_2 (x_i - \bar{x}) + \varepsilon_i - \bar{\varepsilon}$$

autrement dit :

$$q_i = a p_i + u_i$$

4.b. Par linéarité de l'espérance on a :

$$\begin{aligned} E(u_i) &= E(\varepsilon_i) - E(\bar{\varepsilon}) \\ &= E(\varepsilon_i) - \frac{1}{N}\sum_{j=1}^{N} E(\varepsilon_j) \\ &= 0 \end{aligned}$$

car par hypothèse $E(\varepsilon_1) = E(\varepsilon_2) = ... = E(\varepsilon_N) = 0$

4.c. On a :

$$\begin{aligned} V(u_i) &= V(\varepsilon_i - \bar{\varepsilon}) \\ &= V(\varepsilon_i) + V(\bar{\varepsilon}) - 2cov(\varepsilon_i, \bar{\varepsilon}) \end{aligned}$$

or pour tous i, j tels que $i \neq j$ on a par hypothèse $cov(\varepsilon_i, \varepsilon_j) = 0$ donc $V(\bar{\varepsilon}) = \frac{1}{N^2} V\left(\sum_{j=1}^{N} \varepsilon_j\right) = \frac{1}{N^2} \sum_{j=1}^{N} V(\varepsilon_j) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$.Par ailleurs,

$$\begin{aligned} cov(\varepsilon_i, \bar{\varepsilon}) &= cov(\varepsilon_i, \frac{1}{N}\sum_{j=1}^{N}\varepsilon_j) \\ &= \frac{1}{N}\sum_{j=1}^{N} cov(\varepsilon_i, \varepsilon_j) \\ &= \frac{1}{N}\left(cov(\varepsilon_i, \varepsilon_i) + \sum_{j \neq i} cov(\varepsilon_i, \varepsilon_j)\right) \end{aligned}$$

or pour tous i, j tels que $i \neq j$ on a par hypothèse $cov(\varepsilon_i, \varepsilon_j) = 0$ donc $cov(\varepsilon_i, \bar{\varepsilon}) = \frac{1}{N} cov(\varepsilon_i, \varepsilon_i) = \frac{1}{N} V(\varepsilon_i) = \frac{\sigma^2}{N}$. Finalement, on obtient :

$$\begin{aligned} V(u_i) &= \sigma^2 + \frac{N\sigma^2}{N^2} - 2\frac{\sigma^2}{N} \\ &= \sigma^2\left(1 - \frac{1}{N}\right) \end{aligned}$$

On a pour tous i, j tels que $i \neq j$:

$$\begin{aligned} cov(u_i, u_j) &= cov(\varepsilon_i - \bar{\varepsilon}, \varepsilon_j - \bar{\varepsilon}) \\ &= cov(\varepsilon_i, \varepsilon_j) - cov(\varepsilon_i, \bar{\varepsilon}) - cov(\bar{\varepsilon}, \varepsilon_j) + cov(\bar{\varepsilon}, \bar{\varepsilon}) \end{aligned}$$

Tout d'abord, on sait par hypothèse que $cov(\varepsilon_i, \varepsilon_j) = 0$ car $i \neq j$. Ensuite, d'après ce qui précède, on a $cov(\varepsilon_i, \bar{\varepsilon}) = \frac{\sigma^2}{N}$, et de même $cov(\bar{\varepsilon}, \varepsilon_j) = cov(\varepsilon_j, \bar{\varepsilon}) = \frac{\sigma^2}{N}$. Enfin, $cov(\bar{\varepsilon}, \bar{\varepsilon}) = V(\bar{\varepsilon}) = \frac{\sigma^2}{N}$ toujours d'après ce qui précède. Par conséquent, pour tous i, j tels que $i \neq j$:

$$cov(u_i, u_j) = -\frac{\sigma^2}{N} - \frac{\sigma^2}{N} + \frac{\sigma^2}{N}$$

soit :

$$cov(u_i, u_j) = -\frac{\sigma^2}{N}$$