

Préparation à l'agrégation externe de Sciences Sociales

Statistique inférentielle - Estimateurs et Intervalles de confiance

2022-2023

Exercice 1

1. (a) X suit une loi binômiale de paramètres 925 et 0,32.
- (b) D'après le théorème central limite, la loi de X peut être approchée par une loi normale de paramètres $925 \times 0,32$ et $925 \times 0,32 \times 0,68$.
- (c) Si l'on vous donne seulement Φ la fonction de répartition d'une loi gaussienne centrée réduite $\mathcal{N}(0, 1)$ et en utilisant l'approximation précédente¹,

$$\begin{aligned} & \mathbf{P}(0,30 \times 925 \leq X \leq 0,40 \times 925) \\ & \approx \Phi\left(\frac{0,30 \times 925 - 0,32 \times 925}{\sqrt{925 \times 0,32 \times 0,68}}\right) - \Phi\left(\frac{0,40 \times 925 - 0,32 \times 925}{\sqrt{925 \times 0,32 \times 0,68}}\right) \\ & \approx 0,903 \text{ à } 10^{-3} \text{ près.} \end{aligned}$$

2. (a) L'estimation de son score est de $\frac{46}{200} = 23\%$.
Appelons \hat{s} l'estimateur associé, qui compte le nombre de voix en faveur de la tête de liste et le divise par 200 soit

$$\hat{s} = \frac{\sum_{i=1}^{200} X_i}{200}$$

où les X_i sont indépendants et suivent la loi de X . Alors $200\hat{s}$ suit une loi binômiale de paramètres 200 et s , où s est le vrai score.

- (b) On approxime cette loi par une loi normale de paramètres $200s = \mathbf{E}(\hat{s})$ et $200s(1-s) = \mathbf{Var}(\hat{s})$.
Donc, avec une probabilité au moins égale à 95%, on a

$$|200\hat{s} - 200s| \leq 1,96 \times \sqrt{200s(1-s)} \leq 1,96 \times \sqrt{50},$$

car on a $p(1-p) \leq 1/4$ quelque soit p compris entre 0 et 1. Soit :

$$|\hat{s} - s| \leq 1,96 \times \sqrt{\frac{s(1-s)}{200}} \leq 1,96 \times \sqrt{\frac{1}{4 \times 200}} \approx 0,069$$

D'où l'intervalle de confiance au niveau de confiance 95% pour le score :

$$s \in [\hat{s} - 0,07; \hat{s} + 0,07] = [0,16; 0,30] = [16\%; 30\%].$$

Exercice 2

On appelle X_1, \dots, X_{n_1} les variables aléatoires observées dans le premier échantillon, la variable X_i prenant la valeur 1 si le i -ème foyer possède le bien et 0 sinon, et X'_1, \dots, X'_{n_2} les variables aléatoires construites de la même façon à partir des observations réalisées dans le deuxième échantillon.

1. Avec certaines calculatrices vous pouvez directement calculer $\mathbf{P}(0,30 \times 925 \leq X \leq 0,40 \times 925)$, avec $X \sim \mathcal{N}(925 \times 0,32, 925 \times 0,32 \times 0,68)$, vérifiez si c'est votre cas!

1. Les observations $X_1, \dots, X_{n_1}, X'_1, \dots, X'_{n_2}$ sont indépendantes et de même loi de Bernoulli de paramètre p . La somme

$$S_1 = \sum_{k=1}^{n_1} X_k$$

suit donc une loi binomiale de paramètres n_1 et p , tandis que la somme

$$S_2 = \sum_{k=1}^{n_2} X'_k$$

suit donc une loi binomiale de paramètres n_2 et p . Leur espérances sont

$$E(S_1) = n_1 p \text{ et } E(S_2) = n_2 p$$

et leur variances

$$V(S_1) = n_1 p(1-p) \text{ et } V(S_2) = n_2 p(1-p).$$

2. $E(F_1) = \frac{1}{n_1} E(S_1) = p$ et $E(F_2) = \frac{1}{n_2} E(S_2)$.
 $V(F_1) = \frac{1}{n_1^2} V(S_1) = \frac{p(1-p)}{n_1}$ et $V(F_2) = \frac{1}{n_2^2} V(S_2) = \frac{p(1-p)}{n_2}$.
3. F_1 et F_2 sont des estimateurs car ils sont définis uniquement en fonction des observations X_i et X'_i et peuvent être calculés à l'aide de ces observations sans connaître p . Le fait qu'ils soient sans biais résulte de la question précédente puisque l'on a vu que leurs espérances étaient toutes deux égales à p .
4. $E(G) = \frac{E(F_1) + E(F_2)}{2} = p$. En tant que fonctionnelle d'estimateurs de p , G est un estimateur de p ; on vient de plus de voir qu'il est sans biais.
5. F_1 et F_2 sont indépendantes, donc

$$\begin{aligned} V(G) &= \frac{1}{4} (V(F_1) + V(F_2)) \\ &= \frac{1}{4} \left(\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2} \right) \\ &= \frac{1}{4} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) p(1-p) \end{aligned}$$

6. G est obtenu en faisant la moyenne des deux estimateurs F_1 et F_2 , ce qui revient à leur donner la même importance. Pourtant, F_1 est obtenue à partir d'un échantillon plus grand que F_2 et devrait à ce titre être davantage pondérée pour obtenir un estimateur de meilleure qualité (c'est le sens de la question suivante!). Lorsque l'on calcule la moyenne de F_1 et F_2 , on prend certes en compte plus d'information que lors du calcul de F_1 , mais on donne à l'information issue du deuxième échantillon un poids démesuré. On s'attend donc à ce que G soit meilleur que F_1 (et donc que F_2) lorsque n_1 et n_2 ne sont pas trop éloignés : c'est la zone dans laquelle l'effet positif de l'ajout d'information prévaut. On s'attend par contre à perdre en précision du fait de la moyennisation avec F_2 lorsque F_1 est beaucoup plus précis que F_2 , c'est-à-dire lorsque n_1 est bien plus grand que n_2 . Le tout est de quantifier le seuil auquel ce phénomène arrive... La qualité d'un estimateur se mesure (pour nous) à l'aide de son risque quadratique. G étant sans biais en tant qu'estimateur de p , on a

$$\text{RQ}(G, p) = V(G) = \frac{1}{4} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) p(1-p)$$

Par ailleurs, F_1 et F_2 sont des estimateurs sans biais de p , donc leur risque quadratique est une fois encore égal à leur variance; comme $n_1 > n_2$, le risque quadratique le plus bas est celui de

F_1 d'après la question 2. G est donc un meilleur estimateur que F_1 et F_2 si et seulement s'il est meilleur que F_1 , c'est-à-dire si

$$\text{RQ}(G, p) < \text{RQ}(F_1, p).$$

Alors,

$$\begin{aligned} \text{RQ}(G, p) < \text{RQ}(F_1, p) &\Leftrightarrow \frac{1}{4} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) p(1-p) < \frac{1}{n_1} p(1-p) \\ &\Leftrightarrow \frac{1}{4} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) < \frac{1}{n_1} \\ &\Leftrightarrow \frac{1}{4} \left(1 + \frac{n_1}{n_2} \right) < 1 \\ &\Leftrightarrow 1 + \frac{n_1}{n_2} < 4 \\ &\Leftrightarrow \frac{n_1}{n_2} < 3. \end{aligned}$$

G est donc meilleur que F_1 et F_2 si et seulement si $n_1 < 3n_2$.

7. L'estimateur $uF_1 + vF_2$ est sans biais si et seulement si $u + v = 1$, et sa variance est alors égale à $u^2 \frac{p(1-p)}{n_1} + (1-u)^2 \frac{p(1-p)}{n_2}$.

Notons $f(u) = \frac{u^2}{n_1} + \frac{(1-u)^2}{n_2} = u^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) - \frac{2u}{n_2} + \frac{1}{n_2}$ pour tout réel u . f est une fonction polynômiale du second degré, et son coefficient dominant est strictement positif, donc elle admet un minimum pour $u = u_0 = \frac{n_1}{n_1 + n_2}$. En effet sa dérivée $f'(u) = \frac{2u}{n_1} + \frac{-2(1-u)}{n_2}$ s'annule si et seulement si

$$\frac{2u}{n_1} + \frac{-2(1-u)}{n_2} = 0 \Leftrightarrow \frac{u}{n_1} = \frac{1-u}{n_2} \Leftrightarrow u = u_0$$

Donc le meilleur estimateur non-biaisé de la forme $uF_1 + vF_2$ est

$$\frac{n_1}{n_1 + n_2} F_1 + \frac{n_2}{n_1 + n_2} F_2.$$

Exercice 3

- Une personne répond «oui» si :
 - elle est d'accord (probabilité p) et dit la vérité (probabilité α) ou
 - elle n'est pas d'accord (probabilité $(1-p)$) et ment (probabilité $(1-\alpha)$).
 Donc, la probabilité de réponses positives est :

$$q = \alpha p + (1-\alpha)(1-p).$$

- Soient X_1, \dots, X_n des variables aléatoires modélisant les réponses des personnes interrogées. Les X_i sont indépendantes et identiquement distribuées, elles suivent toutes une loi $Ber(q)$. Leur somme $X = \sum_{i=1}^n X_i$ suit donc une $Bin(n, q)$. La fréquence observée par l'enquêteur est $F_n = \frac{X}{n}$. Pour que F_n soit un estimateur sans biais de q il faut que $E(F_n) = q$. Donc, il nous faut calculer $E(F_n)$:

$$E(F_n) = E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{nq}{n} = q.$$

La fréquence observée par l'enquêteur est donc un estimateur sans biais de q .

Et son risque quadratique est :

$$\begin{aligned}
 RQ(F_n) &= E[(F_n - q)^2] \\
 &= V(F_n) \\
 &= V\left(\frac{X}{n}\right) \\
 &= \frac{V(X)}{n^2} \\
 &= \frac{nq(1-q)}{n^2} \\
 &= \frac{q(1-q)}{n}
 \end{aligned}$$

On observe que $F_n \xrightarrow[n \rightarrow \infty]{} 0$.

3. Pour $\alpha \neq 1/2$,

$$\begin{aligned}
 q &= \alpha p + (1 - \alpha)(1 - p) \\
 \iff q &= \alpha p + 1 - \alpha - p + \alpha p \\
 \iff p &= \frac{q + \alpha - 1}{2\alpha - 1}.
 \end{aligned}$$

4. Supposons $\alpha \neq 1/2$. L'espérance de T_n est

$$E(T_n) = E\left(\frac{F_n - 1 + \alpha}{2\alpha - 1}\right) = \frac{E(F_n) - 1 + \alpha}{2\alpha - 1} = \frac{q - 1 + \alpha}{2\alpha - 1} = p.$$

Son risque quadratique est

$$\begin{aligned}
 RQ(T_n) &= E[(T_n - p)^2] = V(T_n) \\
 &= V\left(\frac{F_n - 1 + \alpha}{2\alpha - 1}\right) = \frac{V(F_n)}{(2\alpha - 1)^2} \\
 &= \frac{q(1-q)}{n(2\alpha - 1)^2}
 \end{aligned}$$

5. Pour n fixé, minimiser le risque quadratique revient à maximiser sur l'intervalle $[0, 1]$ la fonction $f(\alpha) = (2\alpha - 1)^2$ dont la dérivée est $f'(\alpha) = 4(2\alpha - 1)$. Nous obtenons le tableau de variation suivant :

α	0	1/2	1
$4(2\alpha - 1)$	−	0	+
$(2\alpha - 1)^2$	1	0	1
$RQ(T_n)$	$\frac{q(1-q)}{n}$	$+\infty$	$\frac{q(1-q)}{n}$

Le risque quadratique est donc minimum pour $\alpha = 0$ ou $\alpha = 1$, c'est-à-dire lorsque les personnes interrogées répondent toujours la vérité ou mentent toujours. Dans ce cas, l'enquêteur connaît l'opinion de celui qu'il interroge et nous retrouvons le problème de départ.

Le risque quadratique est maximum, même infini, pour $\alpha = 1/2$. En effet dans ce cas-là $q = 1/2$. La probabilité de réponses «oui» q ne dépend plus de la probabilité d'opinion positive p .

Avec le dé, $\alpha = 1/6$, le risque quadratique est égal à $\frac{9q(1-q)}{4n}$.