Statistique Mathématique

Correction du partiel — février 2023

Intervalles de confiance pour une Bernoulli

Partie I : Wald ou l'approximation gaussienne

I.1. (1 point). On a $\mathbb{E}[X] = p$ et $\mathbb{E}\left[X^2\right] = p$, d'où l'on déduit facilement $\mathrm{Var}(X) = p(1-p)$.

I.2. (1 point). Il n'y a qu'un paramètre inconnu, p. On sait que $\mathbb{E}[X] = p$ d'après la question suivante. On exprime p en fonction des moments: $p = \mathbb{E}[X]$, puis on remplace l'espérance par sa version empirique pour obtenir $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

I.3. (1 point). Les X_i sont i.i.d. de loi commune X par hypothèse. De plus, ils sont bornés, donc intégrables. D'après la loi forte des grands nombres,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mathbb{E}[X_1] = p\,.$$

Ainsi $\hat{p}$ est un estimateur fortement consistant de p.

I.4. (1 point). Comme à la question précédente, on utilise le fait que les X_i soient i.i.d. De plus, ils sont bornés, donc possèdent un moment d'ordre deux. D'après le théorème central limite,

$$\sqrt{n}(\hat{p} - p) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}(X_1)) = \mathcal{N}(0, p(1-p))\ .$$

I.5. (2 points). On sait que $\hat{p} \xrightarrow{\text{a.s.}} p$ d'après la question I.3., et $p \in]0,1[$ par hypothèse. De plus, la fonction $x \mapsto 1/\sqrt{x(1-x)}$ est continue sur $]0,1[$. Par le théorème de continuité, on en déduit que

$$\frac{1}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow{\text{a.s.}} \frac{1}{\sqrt{p(1-p)}}\,.$$

Par le lemme de Slutsky, en utilisant la convergence prouvée dans la question I.4., on obtient que

$$\sqrt{\frac{n}{\hat{p}(1-\hat{p})}}(\hat{p} - p) \xrightarrow{\mathcal{L}} \frac{1}{\sqrt{p(1-p)}} \cdot \mathcal{N}(0, p(1-p)) = \mathcal{N}(0,1)\ .$$

I.6. (2 points). Par définition de z, on a que $\mathbb{P}(\mathcal{N}(0,1) \leq z) \leq 1-\alpha/2$. De plus, $\mathbb{P}(\mathcal{N}(0,1) \geq -z) = 1 - \mathbb{P}(\mathcal{N}(0,1) \leq z) = \alpha/2$, puisque la gaussienne est symétrique par rapport à l'origine. On en déduit

$$\mathbb{P}(\mathcal{N}(0,1) \in [-z, z]) = 1 - \alpha\,.$$

En utilisant la question I.5., on a donc

$$\mathbb{P}\left(\sqrt{\frac{n}{\hat{p}(1-\hat{p})}}(\hat{p} - p) \in [-z, z]\right) \longrightarrow 1 - \alpha$$

lorsque $n \to +\infty$. En réécrivant l'événement duquel on prend la probabilité, on obtient

$$\mathbb{P}\left(\hat{p} - p \in \left[\pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]\right) \longrightarrow 1 - \alpha\,,$$

soit

$$\mathbb{P}\left(p \in \left[\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]\right) \longrightarrow 1 - \alpha\,,$$

ce qu'il fallait démontrer.

I.7. (1 point). A la lecture de la Figure 1, on se rend compte que

- la probabilité pour p d'appartenir à l'intervalle de confiance C_A, à n fixé, peut être bien **inférieure au seuil de 95%**. Pour $25 \leq n \leq 100$, c'est même la situation typique : ce seuil n'est franchi que six fois;
- la probabilité pour p d'appartenir à l'intervalle de confiance C_A **n'est pas croissante avec** n : ajouter de nouveaux points d'observation n'est pas une garantie d'amélioration de l'intervalle.

Le lecteur intéressé peut consulter Brown et al. [2001] pour approfondir le sujet.

Partie II : Wilson

II.1. (1 point). En gardant à l'esprit le fait que $z > 0$, on écrit

$$\begin{aligned}\mathbb{P}\left(\frac{n}{p(1-p)}(\hat{p}-p)^2 \leq z^2\right) &= \mathbb{P}\left(\sqrt{\frac{n}{p(1-p)}}\,|\hat{p}-p| \leq z\right) \\ &= \mathbb{P}\left(\sqrt{\frac{n}{p(1-p)}}(\hat{p}-p) \leq z \text{ et } -\sqrt{\frac{n}{p(1-p)}}(\hat{p}-p) \leq z\right) \\ &= \mathbb{P}\left(\sqrt{\frac{n}{p(1-p)}}(\hat{p}-p) \in [-z,z]\right).\end{aligned}$$

Cette quantité tend vers $1-\alpha$ d'après la question I.4.

II.2. (3 points). Plaçons nous sur l'événement

$$\frac{n}{p(1-p)}(\hat{p}-p)^2 \leq z^2\,.$$

Comme $p \in]0,1[$, on peut multiplier les deux côtés de l'inégalité par $p(1-p)/n$. On obtient

$$(\hat{p}-p)^2 \leq \frac{z^2}{n}p(1-p)\,,$$

soit

$$\left(1+\frac{z^2}{n}\right)p^2 - \left(2\hat{p}+\frac{z^2}{n}\right)p + \hat{p}^2 \leq 0\,.$$

Le déterminant de l'équation du second degré auquel correspond le terme de gauche est

$$\Delta = \left(2\hat{p}+\frac{z^2}{n}\right)^2 - 4\left(1+\frac{z^2}{n}\right)\hat{p}^2 = 4\hat{p}(1-\hat{p})\frac{z^2}{n} + \frac{z^4}{n^2}\,.$$

On remarque que $\Delta > 0$. De plus, le coefficient du terme p^2 est également positif, donc l'inégalité du second degré est satisfaite pour tout p compris entre les deux racines données par

$$p_\pm := \frac{-\left(2\hat{p}+\frac{z^2}{n}\right) \pm \sqrt{\Delta}}{2\left(1+\frac{z^2}{n}\right)}\,.$$

On remarque que l'on peut réécrire ces solutions comme

$$p_\pm = \frac{1}{1+\frac{z^2}{n}}\left(\hat{p}+\frac{z^2}{2n}\right) \pm \frac{z}{1+\frac{z^2}{n}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2}{4n^2}}\,,$$

d'où l'on déduit l'intervalle de confiance attendu.

II.3. (1 point). Lorsque n est grand, on a

$$\frac{1}{1+\frac{z^2}{n}} = 1 + o\left(\frac{1}{n}\right) \qquad \text{et} \qquad \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + o\left(\frac{1}{n}\right) ,$$

où l'on a utilisé le fait que $\hat{p} \xrightarrow{\text{a.s.}} p$. On en déduit que

$$p_\pm = \hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + o\left(\frac{1}{n}\right) ,$$

ce qu'il fallait démontrer.

II.4. (1 point). A première vue, on préfère l'utilisation de C_B : la probabilité pour C_B de contenir p à n fixé apparaît bien supérieure à $1-\alpha$. Cependant, cette amélioration a un coût, et C_B **est typiquement beaucoup plus large que** C_A.

References

L. D. Brown, T. T. Cai, and A. DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133, 2001.