6 juin 2025

---

1 Estimateur du nombre de boules dans une urne

On considère une urne contenant n boules, numérotées de 1 à n. On ne connaît pas n et on souhaite l'estimer. Pour cela, on procède à m tirages avec remise.

1. Donner un estimateur simple de n.
2. Calculer l'estimateur par maximum de vraisemblance de n.
3. Montrer que le biais de cet estimateur est

$$-\sum_{k=1}^{n-1}\left(\frac{k}{n}\right)^m .$$

Que vaut-il quand $m = 1$? Quand $m \to +\infty$?

Solution

Modélisation : Échantillon aléatoire $(X_1,X_2,\ldots,X_m)$ où X_i est le numéro de la boule tirée au i-ème tirage.

La loi de X est donnée par :

$$\mathbb{P}_X(X = k|n) = \begin{cases} \frac{1}{n} & \text{si } k \in [1,\ldots,n] \\ 0 & \text{sinon} \end{cases} = \frac{1}{n}\, 1_{[1,\ldots,n]}(k).$$

Question 1. Un estimateur naturel de n est

$$N_m = \max_{i=1,\ldots,m}(X_i).$$

Une autre possiblité est d'observer que X étant uniforme sur $[1,n]$, la moyenne empirique des X_i s'approche de $\frac{n+1}{2}$, et de suggérer

$$N_m = 2\frac{1}{m}\sum_{i=1}^{m} X_i - 1.$$

Question 2. La vraisemblance d'un échantillon $(x_1,x_2,\ldots,x_m)$ pour l'estimation η de n vaut

$$L(x_1,x_2,\ldots,x_m|\eta) = \prod_{i=1}^{m}\frac{1}{\eta}\, 1_{[1,\ldots,\eta]}(x_i).$$

L'estimation par maximum de vraisemblance de n est donc

$$\widehat{n}_{\text{MLE}} = \underset{\eta\in\mathbb{N}^*}{\arg\max}\prod_{i=1}^{m}\frac{1}{\eta}\, 1_{[1,\ldots,\eta]}(x_i).$$

La quantité $\prod_{i=1}^m \frac{1}{\eta}\, 1_{[1,\dots,\eta]}(x_i)$ est positive, et non nulle dès que toutes les indicatrices valent 1, c'est-à-dire que $x_i \leq \eta$ pour tout $i = 1, \dots, m$. Enfin, quand elle est non-nulle, elle est d'autant plus petite que η est grand. Ainsi $\widehat{n}_{\text{MLE}}$ doit être aussi petit que possible tout en majorant tous les x_i. On a donc

$$\widehat{n}_{\text{MLE}} = \max_{i=1,\dots,m}(x_i),$$

et l'estimateur par maximum de vraisemblance de n est la variable aléatoire réelle

$$\widehat{N}_{\text{MLE}} = \max_{i=1,\dots,m}(X_i).$$

Le premier des estimateurs que nous avons proposé à la question précédente est en fait l'estimateur par maximum de vraisemblance de N.

Question 3. Le biais de l'estimateur que nous avons proposé est, par définiition :

$$\mathrm{B}(N_m) = \mathbb{E}(N_m) - n.$$

Calculons $\mathbb{E}(N_m)$:

$$\begin{aligned}
\mathbb{E}(N_m) &= \sum_{k=1}^{n} k\mathbb{P}(N_m = k) \text{ par définition} \\
&= \sum_{k=1}^{n} \mathbb{P}(N_m \geq k) \\
&= \sum_{k=1}^{n} 1 - \mathbb{P}(N_m \leq (k-1)) \\
&= n - \sum_{k=1}^{n} \left(\frac{(k-1)}{n}\right)^m \\
&= n - \sum_{k=1}^{n-1} \left(\frac{k}{n}\right)^m .
\end{aligned}$$

La deuxième ligne s'obtient en observant que

$$\begin{aligned}
\sum_{k=1}^{n} \mathbb{P}(N_m \geq k) &= \mathbb{P}(N_m \geq n) + \mathbb{P}(N_m \geq n-1) + \cdots + \mathbb{P}(N_m \geq 1) \\
&= \mathbb{P}(N_m = n) + (\mathbb{P}(N_m = n) + \mathbb{P}(N_m = n-1)) + \cdots + \\
&\quad (\mathbb{P}(N_m = n) + \mathbb{P}(N_m = n-1) + \cdots + \mathbb{P}(N_m = 1)),
\end{aligned}$$

puis en regroupant ensemble les n termes $\mathbb{P}(N_m = n)$, les $(n-1)$ termes $\mathbb{P}(N_m = n-1)$, etc.

La quatrième ligne s'obtient en observant que l'événement "$N_m \leq k-1$" est équivalent à l'événement "$X_i \leq k-1$ pour $i = 1, \dots, m$", que $\mathbb{P}(X_i \leq k-1) = \frac{k-1}{n}$, et que les X_i sont indépendants.

Ainsi notre estimateur est biaisé, est son biais vaut

$$\mathrm{B}(N_m) = -\sum_{k=1}^{n-1} \left(\frac{k}{n}\right)^m .$$

Ce biais est négatif : la valeur estimée est en moyenne plus faible que le nombre de boules. Il est peu probable de tirer la boule numéro n, sauf à faire un très grand nombre de tirages.

Tirer une seule boule ($m = 1$) est équivalent à tirer une valeur uniformément entre 1 et n, on s'attend donc à être plus proche de $\frac{n+1}{2}$ (espérance d'une variable aléatoire réelle uniformément distribuée sur $[1,n]$) que de n. On a bien

$$\mathrm{B}(N_1) = -\sum_{k=1}^{n-1} \left(\frac{k}{n}\right) = -\frac{1}{n}\frac{n(n-1)}{2} = -\frac{n-1}{2} = \frac{n+1}{2} - n.$$

Par ailleurs, le biais tend vers 0 quand $m \to +\infty$. N_m est asymptotiquement non biaisé.

2 Estimation de densité

Considérons une variable aléatoire X suivant une loi normale de paramètres μ et σ^2.

1. Étant donné un échantillon de $n \in \mathbb{N}^*$ observations de X, calculer l'estimateur par maximum de vraisemblance de μ et σ.
2. Supposons maintenant que μ est la réalisation d'un variable aléatoire réelle M qui suit une loi normale de moyenne m et de variance τ^2. Calculer l'estimateur de Bayes de μ.
3. Décomposer l'estimateur de Bayes de μ en la somme d'un terme fonction de la moyenne empirique de l'échantillon et un terme dépendant de la moyenne a priori. Que se passe-t-il quand n augmente ?

Solution

Question 1 Appelons $(x_1,x_2,\dots,x_n)$ notre échantillon. Il s'agit d'une réalisation de l'échantillon aléatoire $(X_1,X_2,\dots,X_n)$ composé de variables i.i.d. de même loi que X. La vraisemblance de cet échantillon est

$$L(x_1,x_2,\dots,x_n;\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right),$$

et sa log-vraisemblance est donc

$$\ell(x_1,x_2,\dots,x_n;\mu,\sigma) = \sum_{i=1}^{n} -\ln(\sigma\sqrt{2\pi}) - \frac{(x_i-\mu)^2}{2\sigma^2}.$$

Nous cherchons

$$\widehat{\mu}_{\text{MLE}},\, \widehat{\sigma}_{\text{MLE}} \in \operatorname*{arg\,max}_{\widehat{\mu},\widehat{\sigma}\in\mathbb{R}^2} \ell(x_1,x_2,\dots,x_n;\widehat{\mu},\widehat{\sigma}).$$

La fonction $\mu \mapsto \sum_{i=1}^{n} -\ln(\sigma\sqrt{2\pi}) - \frac{(x_i-\mu)^2}{2\sigma^2}$ est une fonction concave, dont le gradient en μ vaut $\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\left(n\mu - \sum_{i=1}^{n} x_i\right)$. En annulant ce gradient, on obtient $\widehat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$. L'estimateur par maximum de vraisemblance de l'espérance de X est sa moyenne empirique.

En remplaçant μ par $\widehat{\mu}_{\text{MLE}}$ et en appelant $\alpha = \frac{1}{\sigma^2}$, la fonction $\alpha \mapsto \frac{1}{2}\sum_{i=1}^{n} -\ln(2\pi) + \ln(\alpha) - (x_i-\mu)^2\alpha$ est concave (sa dérivée seconde est proportionnelle à $-\frac{1}{\alpha^2} < 0$). En annulant son gradient, on obtient $\widehat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{\text{MLE}})^2$ et $\widehat{\sigma}_{\text{MLE}}$ est donc l'écart-type empirique.

Question 2 Nous supposons maintenant que μ est la réalisation d'une variable aléatoire réelle $M \sim \mathcal{N}(m,\tau^2)$. Son estimateur de Bayes est $\widehat{\mu}_{\text{Bayes}} = \mathbb{E}(M|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$.

Pour déterminer cette espérance, calculons la densité correspondante :

$$\mathbb{P}(M = \mu|X_1,X_2,\ldots,X_n) = \frac{\mathbb{P}(X_1,X_2,\ldots,X_n|M=\mu)\mathbb{P}(M=\mu)}{\mathbb{P}(X_1,X_2,\ldots,X_n)} \text{ (Bayes)}$$
$$= \frac{1}{\mathbb{P}(X_1,X_2,\ldots,X_n)}\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2}\right)\frac{1}{\tau\sqrt{2\pi}}\exp\left(-\frac{(\mu-m)^2}{2\tau^2}\right)$$
$$= \mathcal{K}_1 \exp\left(-\sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2} - \frac{(\mu-m)^2}{2\tau^2}\right) \text{ où } \mathcal{K}_1 \text{ ne dépend pas de } \mu.$$

Ainsi,

$$\mathbb{P}(M=\mu|X_1,X_2,\ldots,X_n) = \mathcal{K}_1 \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{i=1}^n (X_i^2+\mu^2-2X_i\mu) + \frac{1}{\tau^2}(\mu^2+m^2-2\mu m)\right]\right)$$
$$= \mathcal{K}_1 \exp\left(-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\mu^2 - 2\left(\frac{\sum_{i=1}^n X_i}{\sigma^2}+\frac{m}{\tau^2}\right)\mu + \frac{\sum_{i=1}^n X_i^2}{\sigma^2}+\frac{m^2}{\tau^2}\right]\right)$$
$$= \mathcal{K}_1 \exp\left(-\frac{1}{2}\frac{(\mu-a)^2}{b^2}\right)\exp\left(-\frac{1}{2}\mathcal{K}_2\right),$$

où $\mathcal{K}_1$ et $\mathcal{K}_2$ ne dépendent pas de μ,

$$a = \frac{\sum_{i=1}^n X_i\tau^2 + m\sigma^2}{n\tau^2+\sigma^2} \text{ et } b = \sqrt{\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}}.$$

L'intégrale d'une densité de probabilité vaut 1. Cela vaut pour la loi normale centrée en a et d'écart-type b, comme pour $\mathbb{P}(M|X_1,X_2,\ldots,X_n)$, et donc $\mathcal{K}_1 \exp\left(-\frac{1}{2}\mathcal{K}_2\right) = \frac{1}{b\sqrt{2\pi}}$.

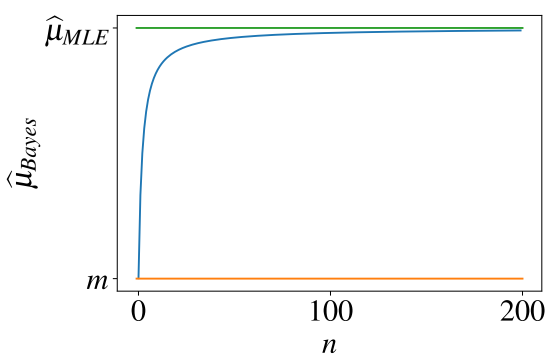
Ainsi, $M|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ suit une loi normale centrée en a. Son espérance vaut donc a, et ainsi

$$\widehat{\mu}_{\text{Bayes}} = \frac{\sum_{i=1}^n x_i\tau^2 + m\sigma^2}{n\tau^2+\sigma^2}.$$

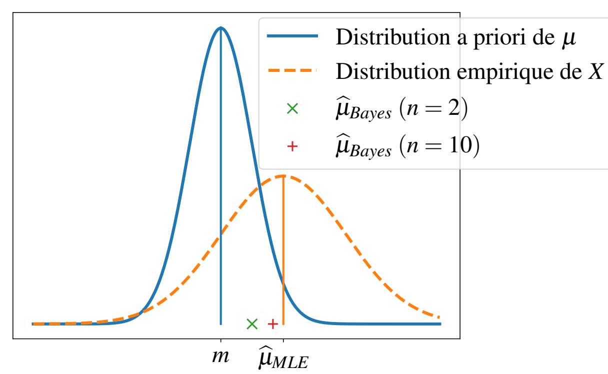
Question 3

$$\widehat{\mu}_{\text{Bayes}} = \frac{n/\sigma^2}{n/\sigma^2+1/\tau^2}\widehat{\mu}_{\text{MLE}} + \frac{1/\tau^2}{n/\sigma^2+1/\tau^2}m.$$

Plus il y a de données et plus l'estimateur de Bayes est proche de la moyenne empirique. Quand il y a peu de données, l'estimateur de Bayes est plus proche de la valeur a priori du paramètre. Ce phénomène est illustré sur la figure 1.



(A) Estimation de Bayes en fonction de la taille de l'échantillon.



(B) Distribution empirique de X et distribution a priori de μ.

FIGURE 1 – Estimation de Bayes de la moyenne d'une gaussienne. Ici $\widehat{\mu}^2_{\text{MLE}} = 2$ et $\widehat{\sigma}^2_{\text{MLE}} = 2$, tandis que $m = 0$ et $\tau^2 = 1$. Plus n est grand, plus on s'éloigne de la valeur a priori de μ pour se rapprocher de son estimation empirique.