Chapitre 4

Intervalles de confiance

Dans une étude statistique, on peut se demander si les mesures observées peuvent correspondre à une certaine réalité. Par exemple, est-ce que notre série de jets d'une pièce de monnaie $(P; P; F; P; F; P)$ peut correspondre à une pièce bien équilibrée ?

Pour répondre à cette question, on cherche à estimer la probabilité p que la pièce montre pile. Si on arrive à établir que $p = \frac{1}{2}$, alors la pièce est parfaitement équilibrée. Malheureusement, il est extrêmement difficile, voire impossible, de montrer que $p = \frac{1}{2}$.

On peut tenter d'estimer p à l'aide d'un estimateur $\widehat{P}$, mais la probabilité $\mathbf{P}(\widehat{P} = p)$ que $\widehat{P}$ soit exactement égal au paramètre p est, en général, extrêmement petite, voire nulle[1]. Par conséquent, il est nécessaire de s'intéresser à la probabilité

$$\mathbf{P}\big(\widehat{P} \geqslant p - r \ \text{ et } \ \widehat{P} \leqslant p + r\big) = \mathbf{P}\big(\widehat{P} - r \leqslant p \leqslant \widehat{P} + r\big) = \mathbf{P}\big(\big[\widehat{P} - r, \widehat{P} + r\big] \ni p\big)$$

Cette probabilité s'approche de 1 au fur et à mesure que le nombre positif r devient grand. Néanmoins, plus r grandit, plus l'approximation de p par un intervalle perd en précision. Il faut donc trouver des compromis.

Définition

Un *intervalle de confiance* est un intervalle $[G, D]$ qui contient le paramètre θ à estimer avec une certaine probabilité β (G et D sont des variables aléatoires).

$$\mathbf{P}([G, D] \ni \theta) = \beta$$

On dit que β est le *seuil de confiance* ou encore la *probabilité de couverture*, tandis que $\alpha = 1 - \beta$ est appelé le *risque d'erreur*.

Construction d'un intervalle de confiance

Pour construire un intervalle de confiance pour le paramètre θ, on commence par choisir le seuil de confiance β désiré (ce qui est équivalent à choisir un risque d'erreur $\alpha = 1 - \beta$). Ensuite, il faut réussir à déterminer le nombre positif r correspondant en utilisant la distribution théorique de l'estimateur $\hat{\theta}$.

Comme d'habitude, on choisit un seuil de confiance de 95% dans une démarche infirmative, et un seuil de confiance de 99% dans une démarche confirmative.

1. De plus, la probabilité pour que deux estimations $\widehat{p}_1$ et $\widehat{p}_2$ de l'estimateur $\widehat{P}$ soient égales est aussi, en général, extrêmement faible.

4.1 L'intervalle de confiance sur une moyenne, variance connue

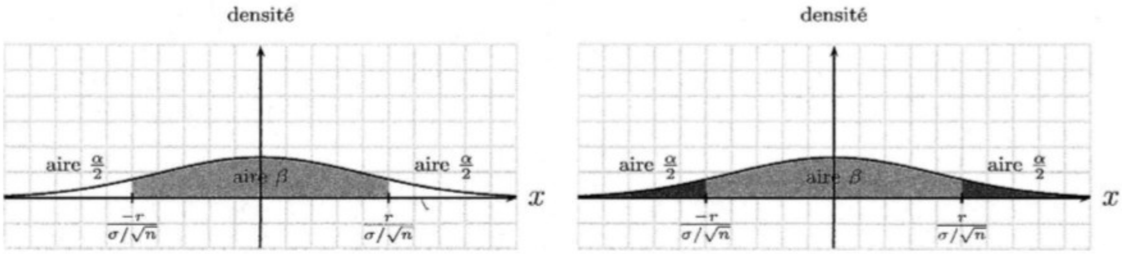
On considère n variables aléatoires indépendantes, notées X_1, X_2, $\dots$, X_n, qui suivent toutes une loi normale $\mathcal{N}(\mu, \sigma^2)$ où seul le paramètre σ est connu. Pour estimer μ, on prend le MLE $\overline{X}$ qui est aussi sans biais. Sous ces hypothèses, $\overline{X}$ suit une loi normale de paramètres $\mathcal{N}(\mu, \frac{\sigma^2}{n})$. Afin de déterminer l'intervalle de confiance, on va devoir centrer-réduire cet estimateur de sorte à pouvoir utiliser la table concernant la fonction de répartition ϕ de la loi normale $\mathcal{N}(0,1)$. On trouvera l'intervalle de confiance suivant :

$$\left[\overline{X} - \phi^{-1}(1-\tfrac{\alpha}{2})\cdot\tfrac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + \phi^{-1}(1-\tfrac{\alpha}{2})\cdot\tfrac{\sigma}{\sqrt{n}}\right]$$

En effet, on cherche $r > 0$ tel que :

$$\begin{aligned}
1-\alpha \;=\; \beta \;&=\; \mathbf{P}\big([\overline{X}-r, \overline{X}+r] \ni \mu\big) \;=\; \mathbf{P}\big(\overline{X}-r \leqslant \mu \leqslant \overline{X}+r\big)\\
&=\; \mathbf{P}\big(\overline{X} \geqslant \mu - r \ \text{ et } \ \overline{X} \leqslant \mu + r\big) \;=\; \mathbf{P}\big(\mu - r \leqslant \overline{X} \leqslant \mu + r\big)\\
&=\; \mathbf{P}\big(-r \leqslant \overline{X} - \mu \leqslant r\big) \;=\; \mathbf{P}\Big(\tfrac{-r}{\frac{\sigma}{\sqrt{n}}} \leqslant \tfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leqslant \tfrac{r}{\frac{\sigma}{\sqrt{n}}}\Big)
\end{aligned}$$

On s'est maintenant ramené à la loi normale centrée réduite Z.



On peut ensuite utiliser la fonction de répartition ϕ de la loi normale centrée réduite.

$$\begin{aligned}
1-\alpha \;&=\; \mathbf{P}\big(\tfrac{-r}{\sigma/\sqrt{n}} \leqslant Z \leqslant \tfrac{r}{\sigma/\sqrt{n}}\big) \;=\; \mathbf{P}\big(Z \leqslant \tfrac{r}{\sigma/\sqrt{n}}\big) - \mathbf{P}\big(Z < \tfrac{-r}{\sigma/\sqrt{n}}\big)\\
&=\; \mathbf{P}\big(Z \leqslant \tfrac{r}{\sigma/\sqrt{n}}\big) - \underbrace{\Big(1 - \mathbf{P}\big(Z \leqslant \tfrac{r}{\sigma/\sqrt{n}}\big)\Big)} \;=\; 2\mathbf{P}\big(Z \leqslant \tfrac{r}{\sigma/\sqrt{n}}\big) - 1\\
&=\; 2\phi\big(\tfrac{r}{\sigma/\sqrt{n}}\big) - 1 \qquad \text{car } \mathbf{P}\big(Z < \tfrac{-r}{\sigma/\sqrt{n}}\big) = \mathbf{P}\big(Z > \tfrac{r}{\sigma/\sqrt{n}}\big) \text{ par symétrie}
\end{aligned}$$

Finalement, on isole r :

$$\begin{aligned}
&1-\alpha \;=\; 2\phi\big(\tfrac{r}{\sigma/\sqrt{n}}\big) - 1 \iff 2\phi\big(\tfrac{r}{\sigma/\sqrt{n}}\big) \;=\; 2-\alpha \iff \phi\big(\tfrac{r}{\sigma/\sqrt{n}}\big) = 1 - \tfrac{\alpha}{2}\\
&\iff \tfrac{r}{\sigma/\sqrt{n}} = \phi^{-1}(1-\tfrac{\alpha}{2}) \iff r = \phi^{-1}(1-\tfrac{\alpha}{2})\cdot\tfrac{\sigma}{\sqrt{n}}
\end{aligned}$$

4.2 L'intervalle de confiance sur une moyenne, variance inconnue

On considère n variables aléatoires indépendantes, notées X_1, X_2, ..., X_n, qui suivent toutes une loi normale $\mathcal{N}(\mu, \sigma^2)$ où le paramètre σ est aussi inconnu. Pour estimer μ, on prend le MLE $\overline{X}$ qui est aussi sans biais. Sous ces hypothèses, $\overline{X}$ suit une loi normale de paramètres $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Malheureusement, comme on ne connaît pas σ, ce n'est pas très utile. Mais, en utilisant l'estimateur sans biais de σ, noté S, on peut utiliser le résultat de William Gosset : la variable aléatoire[2] $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ suit une distribution de Student avec $n-1$ degrés de liberté.

Afin de déterminer l'intervalle de confiance, on procède comme ci-dessus de sorte à pouvoir utiliser la table concernant la fonction de répartition ϕ_{n-1} de la loi de Student à $n-1$ degrés de liberté.

On trouvera l'intervalle de confiance suivant :

$$\left[\overline{X} - \phi_{n-1}^{-1}(1-\tfrac{\alpha}{2}) \cdot \tfrac{S}{\sqrt{n}} \ , \ \overline{X} + \phi_{n-1}^{-1}(1-\tfrac{\alpha}{2}) \cdot \tfrac{S}{\sqrt{n}}\right]$$

2. Ce n'est pas la variable centrée réduite de $\overline{X}$, car on a remplacé σ par son estimateur sans biais S.