Éléments de compréhension des statistiques

Jeffery P.

Doctorant au Laboratoire des Sciences du Numérique de Nantes (LS2N)

2019

Intervalle de confiance : préambule

Jusqu'ici on a calculé des probabilités sur une loi normale connue.

- Imaginons maintenant que l'on dispose d'un échantillon de N observations que l'on suppose issues d'une loi normale $\mathcal{N}(\mu, \sigma^2)$

Comment faire pour estimer la vraie moyenne μ avec les données dont on dispose ?

Intervalle de confiance : intérêt

Une première intuition peut consister à supposer que la vraie moyenne μ est « à peu près » égale à la moyenne de l'échantillon dite empirique i.e., $\bar{x}_{\mathrm{N}}$

$$\text{Estimation ponctuelle : } \mu \sim \bar{x}_{\mathrm{N}} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} x_i$$

Néanmoins, une question que l'on peut se poser est la certitude sur cette estimation. . . ?

Définition d'un intervalle de confiance

Un **intervalle de confiance** pour la moyenne μ, de niveau $1-\alpha$ où $\alpha \in]0;1[$, est un intervalle qui a une probabilité $1-\alpha$ de contenir la vraie valeur de μ.

Attention

- Cela ne veut pas dire que la vraie valeur μ à une probabilité $1-\alpha$ d'être dans l'intervalle. Ça n'aurait pas de sens car μ n'est pas aléatoire !

Intervalle de confiance dans le cas gaussien

Lorsqu'on dispose d'observation d'une distribution supposée gaussienne (disons $\mathcal{N}(\mu, \sigma^2)$), on dispose de la vraie loi de la moyenne empirique :

$$\bar{X}_N \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

Ainsi en normalisant (centrage+réduction), on peut déduire que :

$$\sqrt{N}\frac{\bar{X}_N - \mu}{\sigma} = \frac{\bar{X}_N - \mu}{\frac{\sigma}{\sqrt{N}}} \sim \mathcal{N}(0, 1)$$

Rappel sur les quantiles

En reprenant les notation des quantiles de la loi normale i.e. si $\mathrm{Z} \sim \mathcal{N}(0,1)$ et $\alpha \in]0.5;1[$, $z_{1-\alpha/2}$ vérifie

$$P(\mathrm{Z} \leq z_{(1-\alpha/2)}) = 1-\alpha/2 \quad \text{et} \quad P(\mathrm{Z} \leq -z_{(1-\alpha/2)}) = \alpha/2$$

D'où :

$$P(-z_{(1-\alpha/2)} \leq \mathrm{Z} \leq z_{(1-\alpha/2)}) = P(|\mathrm{Z}| \leq z_{(1-\alpha/2)}) = 1-\alpha$$

Si l'écart-type de la population est connu

On peut écrire :

$$P(-z_{(1-\alpha/2)} \leq \sqrt{\mathrm{N}}\frac{\bar{\mathrm{X}}_{\mathrm{N}} - \mu}{\sigma} \leq z_{(1-\alpha/2)}) = 1 - \alpha$$

Ce qui est équvalent à écrire que :

$$P(\bar{\mathrm{X}}_{\mathrm{N}} - \frac{\sigma}{\sqrt{\mathrm{N}}}z_{(1-\alpha/2)} \leq \mu \leq \bar{\mathrm{X}}_{\mathrm{N}} + \frac{\sigma}{\sqrt{\mathrm{N}}}z_{(1-\alpha/2)}) = 1 - \alpha$$

Formule de l'IC si l'écart-type de la population est connu

D'après les calculs précédents on peut écrire de manière équivalente :

$$P(\mu \in [\bar{X}_N - \frac{\sigma}{\sqrt{N}} z_{(1-\alpha/2)}; \bar{X}_N + \frac{\sigma}{\sqrt{N}} z_{(1-\alpha/2)}]) = 1 - \alpha$$

Ce qui nous donne un intervalle de **probabilité** pour la moyenne Eureka ! !

La formule de l'intervalle de **confiance** de **niveau** $1 - \alpha$ pour la moyenne est donc une réalisation de cet intervalle de probabilité compte tenu des données dont on dispose :

$$[\bar{x}_N - \frac{\sigma}{\sqrt{N}} z_{(1-\alpha/2)}; \bar{x}_N + \frac{\sigma}{\sqrt{N}} z_{(1-\alpha/2)}]$$

Si l'écart-type de la population est inconnu

La formule précédente ne fonctionne que si on connaît la valeur σ... ce qui est rarement le cas ! Dans la pratique, on remplace souvent σ par son estimation :

$$\sigma \quad \text{estimé par} \quad s = \sqrt{\frac{1}{\mathrm{N}-1}\sum_{i=1}^{\mathrm{N}}(x_i - \bar{x}_{\mathrm{N}})^2} = \sqrt{\frac{\mathrm{N}}{\mathrm{N}-1}\sigma^2_{\mathrm{obs}}}$$

Cela change légèrement la loi utilisée car :

$$\sqrt{\mathrm{N}}\frac{\bar{\mathrm{X}}_{\mathrm{N}} - \mu}{s} \sim \mathrm{Student}(\mathrm{N}-1)$$

On se contente de l'admettre mais cela peut se démontrer avec un peu de théorie statistique...ça ne sort pas d'un chapeau !

Formule de l'IC si l'écart-type de la population est inconnu

Quand on ne connaît pas l'écart-type de la distribution étudiée, l'intervalle de **confiance** de **niveau** $1 - \alpha$ pour la moyenne est :

$$[\bar{x}_{\mathrm{N}} - \frac{\mathrm{S}}{\sqrt{\mathrm{N}}} t_{(1-\alpha/2)} ; \bar{x}_{\mathrm{N}} + \frac{\mathrm{S}}{\sqrt{\mathrm{N}}} t_{(1-\alpha/2)}]$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $\mathrm{N} - 1$ degrés de liberté

Convergence de la loi de Student : approximation gaussienne

Dès que $\mathrm{N} - 1 \geq 30$ on considère que la loi de Student peut être approximée par la loi normale centrée réduite donc :

$$t_{(1-\alpha/2)} \simeq z_{(1-\alpha/2)}$$

et par conséquent :

$$\begin{aligned} & [\bar{x}_{\mathrm{N}} - \frac{\sigma}{\sqrt{\mathrm{N}}} t_{(1-\alpha/2)}; \bar{x}_{\mathrm{N}} + \frac{\sigma}{\sqrt{\mathrm{N}}} t_{(1-\alpha/2)}] \\ \simeq \quad & [\bar{x}_{\mathrm{N}} - \frac{\sigma}{\sqrt{\mathrm{N}}} z_{(1-\alpha/2)}; \bar{x}_{\mathrm{N}} + \frac{\sigma}{\sqrt{\mathrm{N}}} z_{(1-\alpha/2)}] \end{aligned}$$

Marge d'erreur

Remarque

L'intervalle de confiance $[\bar{x}_N - \frac{\sigma}{\sqrt{N}} z_{(1-\alpha/2)}; \bar{x}_N + \frac{\sigma}{\sqrt{N}} z_{(1-\alpha/2)}]$ est donc symétrique, centré sur $\bar{x}_N$. On écrit d'ailleurs souvent par abus de notation :

$$[\bar{x}_N \pm \frac{\sigma}{\sqrt{N}} z_{(1-\alpha/2)}] \quad \text{ou} \quad [\bar{x}_N \pm \frac{s}{\sqrt{N}} t_{(1-\alpha/2)}]$$

On dit souvent que la marge d'erreur est :

$$\varepsilon = \frac{\sigma}{\sqrt{N}} z_{(1-\alpha/2)} \quad \text{ou} \quad \varepsilon = \frac{s}{\sqrt{N}} t_{(1-\alpha/2)}$$

Intervalle de confiance pour une proportion

Lorsqu'on cherche à estimer une proportion p_0, on utilise des formules comparables. Néanmoins, on ne peut formuler un intervalle que lorsque n est suffisamment grand par rapport à p car on doit se contenter de la convergence vers la loi normale :

$$P \underset{n\to\infty}{\longrightarrow} \mathcal{N}(\mu = p_0, \sigma^2 = \frac{p_0(1-p_0)}{\mathrm{N}})$$

Formule de l'IC pour une proportion

On pratique, on vérifie les conditions suivantes :

- $\mathrm{N} \geq 3$,
- $\mathrm{N} \times p \geq 5$
- $\mathrm{N} \times (1 - p) \geq 5$

Et on estime $\frac{p_0(1-p_0)}{N}$ par $\frac{p(1-p)}{N}$

Ainsi on en déduit l'intervalle de confiance de niveau $1 - \alpha$ suivant :

$$[p - \sqrt{\frac{p(1-p)}{\mathrm{N}}} z_{(1-\alpha/2)}; p + \sqrt{\frac{p(1-p)}{\mathrm{N}}} z_{(1-\alpha/2)}]$$