Université de Strasbourg
M1 - Mathématiques fondamentales
Statistique

Ségolen Geffray
geffray@math.unistra.fr
Année 2020/2021

---

**TP1: illustration des propriétés d'un intervalle de confiance**

---

Le but de ce TP est de proposer une illustration par simulations les propriétés **fréquentistes** d'un intervalle de confiance pour un paramètre réel noté $\theta$ lorsque cet intervalle est déterminé à partir d'un échantillon i.i.d. noté $(X_1, ..., X_n)$.
Donnons-nous un niveau de confiance noté $(1-\alpha)$ **choisi** dans $]0;1[$ idéalement le plus proche possible de 1. Le choix classique est $(1-\alpha) = 0.95$. Un intervalle de confiance pour $\theta$ au niveau de confiance $(1-\alpha)$ consiste en la donnée de deux statistiques (= fonctions mesurables de l'échantillon) ne dépendant pas de $\theta$, ni d'aucun paramètre inconnu, notées $\widehat{\theta}_{n,\text{inf}} = \varphi_1(X_1, ..., X_n)$ et $\widehat{\theta}_{n,\text{sup}} = \varphi_2(X_1, ..., X_n)$, telles que pour tout $\theta \in \Theta$, on a:

$$\mathbb{P}_\theta[\widehat{\theta}_{n,\text{inf}} \leq \theta \leq \widehat{\theta}_{n,\text{sup}}] = 1 - \alpha. \tag{1}$$

Lorsque l'égalité (1) est vraie pour tout $n$, on dit qu'il s'agit d'un **IC à distance finie**. Lorsque l'égalité (1) est vraie asymptotiquement, on dit qu'il s'agit d'un **IC asymptotique**.
NB: L'intervalle $[\varphi_1(x_1, ..., x_n), \varphi_2(x_1, ..., x_n)]$ est **déterministe**, il n'a donc pas une probabilité de $(1-\alpha)$ de contenir $\theta$. Seul l'intervalle aléatoire $[\varphi_1(X_1, ..., X_n), \varphi_2(X_1, ..., X_n)]$ possède cette propriété fréquentiste.

Notons

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S_n'^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 .$$

**• Intervalle de confiance bilatéral symétrique pour $\theta = \mathbb{E}[X]$ pour un grand échantillon i.i.d. de loi admettant un moment d'ordre 2:**

- L'**intervalle de confiance** asymptotique bilatéral pour $\theta$ au niveau de confiance $(1-\alpha)$ est issu de l'encadrement en probabilité suivant

$$\mathbb{P}\left[\overline{X}_n - F_N^{-1}(1-\alpha/2)\frac{\sqrt{S_n'^2}}{\sqrt{n}} \leq \theta \leq \overline{X}_n + F_N^{-1}(1-\alpha/2)\frac{\sqrt{S_n'^2}}{\sqrt{n}}\right] \approx 1-\alpha\,,$$

  l'approximation étant d'autant meilleure que $n$ est grand.
  La notation $F_N^{-1}(1-\alpha/2)$ désigne le quantile d'ordre $(1-\alpha/2)$ de la loi gaussienne standard notée $\mathcal{N}(0,1)$.

- Il y a une probabilité égale environ à $(1-\alpha)$ pour que l'intervalle de confiance aléatoire contienne la moyenne de la population, l'approximation étant d'autant meilleure que $n$ est grand.. De manière à peu près équivalente, sur 100 réalisations de l'intervalle de confiance, le paramètre inconnu sera effectivement dans la fourchette proposée dans environ $(1-\alpha)100\%$ des cas, l'approximation étant d'autant meilleure que $n$ est grand.

**• Intervalle de confiance bilatéral symétrique pour $\theta = \mathbb{E}[X]$ pour un échantillon i.i.d. gaussien de taille quelconque:**

- L'**intervalle de confiance** bilatéral pour $\theta$ au niveau de confiance $(1-\alpha)$ est issu de l'encadrement en probabilité suivant

$$\mathbb{P}\left[\overline{X}_n - F^{-1}_{T(n-1)}(1-\alpha/2)\frac{\sqrt{S'^2_n}}{\sqrt{n}} \leq \theta \leq \overline{X}_n + F^{-1}_{T(n-1)}(1-\alpha/2)\frac{\sqrt{S'^2_n}}{\sqrt{n}}\right] = 1-\alpha\,,$$

  où $F^{-1}_{T(n-1)}(1-\alpha/2)$ désigne le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à $(n-1)$ degrés de liberté notée $T(n-1)$.

- Il y a une probabilité d'exactement $(1-\alpha)$ pour que l'intervalle de confiance aléatoire contienne la moyenne de la population. De manière à peu près équivalente, sur 100 réalisations de l'intervalle de confiance, le paramètre inconnu sera effectivement dans la fourchette proposée dans $(1-\alpha)100\%$ des cas.

**• Intervalle de confiance bilatéral symétrique pour $\theta = \mathbb{E}[X]$ pour un grand échantillon i.i.d. de loi de Poisson:**

- L'**intervalle de confiance** asymptotique bilatéral pour $\theta$ au niveau de confiance $(1-\alpha)$ est issu de l'encadrement en probabilité suivant

$$\mathbb{P}\left[\overline{X}_n - F^{-1}_{N}(1-\alpha/2)\frac{\sqrt{\overline{X}_n}}{\sqrt{n}} \leq \theta \leq \overline{X}_n + F^{-1}_{N}(1-\alpha/2)\frac{\sqrt{\overline{X}_n}}{\sqrt{n}}\right] \approx 1-\alpha\,,$$

  l'approximation étant d'autant meilleure que $n$ est grand.
  La notation $F^{-1}_N(1-\alpha/2)$ désigne le quantile d'ordre $(1-\alpha/2)$ de la loi gaussienne standard notée $\mathcal{N}(0,1)$.

- Il y a une probabilité égale environ à $(1-\alpha)$ pour que l'intervalle de confiance aléatoire contienne la moyenne de la population. De manière à peu près équivalente, sur 100 réalisations de l'intervalle de confiance, le paramètre inconnu sera effectivement dans la fourchette proposée dans $(1-\alpha)100\%$ des cas.

**• Intervalle de confiance bilatéral symétrique pour $\theta = \mathbb{E}[X]$ pour un échantillon i.i.d. de loi de Poisson de taille quelconque:**

- L'**intervalle de confiance** bilatéral pour $\theta$ au niveau de confiance $(1-\alpha)$ est issu de l'encadrement en probabilité suivant

$$\mathbb{P}\left[\left(-\frac{1}{\sqrt{4n\alpha}} + \sqrt{\frac{1}{4n\alpha} + \overline{X}_n}\right)^2 \leq \theta \leq \left(\frac{1}{\sqrt{4n\alpha}} + \sqrt{\frac{1}{4n\alpha} + \overline{X}_n}\right)^2\right] \geq 1-\alpha\,.$$

- Il y a une probabilité d'au moins $(1-\alpha)$ pour que l'intervalle de confiance aléatoire contienne la moyenne de la population, l'approximation étant d'autant meilleure que $n$ est grand. De manière à peu près équivalente, sur 100 réalisations de l'intervalle de confiance, le paramètre inconnu sera effectivement dans la fourchette proposée dans $(1-\alpha)100\%$ des cas, l'approximation étant d'autant meilleure que $n$ est grand.

• **Illustration numérique:**
Elaborer un plan de simulations explicitant le nombre d'échantillons générés noté $M$, la taille requise pour l'échantillon que l'on a noté $n$ et les différentes lois utilisées pour simuler les échantillons. Vous étudierez la probabilité de couverture (définie comme étant la proportion de fois où le paramètre inconnu à estimer appartient effectivement à l'intervalle de confiance proposé) et la largeur de l'intervalle proposé de façon à illustrer les éléments présentés ci-dessus.

On utilisera les valeurs $M = 1000$ et $n = 20, 50, 100, 200, 500$.

Le travail à effectuer peut se mettre sous la forme de l'algorithme suivant:

• choisir une loi de probabilité paramétrée par $\theta$ (et éventuellement d'autres paramètres), notée $P_\theta$, pour une valeur fixée de $\theta$ parmi l'ensemble des valeurs admissibles,
• pour $n = 20, 50, 100, 200, 500$,

- pour $m = 1, ..., M$,
  - simuler un échantillon i.i.d. $\left(X_1^{\{m\}}, ..., X_n^{\{m\}}\right)$ selon $P_\theta$,
  - calculer les bornes inférieure $\widehat{\theta}_{n,\text{inf}}^{\{m\}}$ et supérieure $\widehat{\theta}_{n,\text{sup}}^{\{m\}}$ de l'intervalle de confiance considéré,
- déterminer
  - la largeur moyenne empirique de l'intervalle de confiance considéré,
  - la probabilité de couverture (définie comme étant la proportion de fois où le paramètre inconnu à estimer appartient effectivement à l'intervalle de confiance proposé).

• recommencer en simulant les données selon une autre loi.