TP de Modélisation Stochastique n^0 9 : Intervalles de confiance

On rappelle les théorèmes suivants qui permettent la construction d'intervalles de confiance pour l'estimation d'un ou plusieurs paramètres. Pour commnecer, on suppose que $(X_1, X_2, .., X_n)$ est un n-échantillon réel, d'espérance m et de variance σ^2. On note:

$$\bar{X}_n = \frac{S_n}{n} = \frac{X_1 + X_2 + ... + X_n}{n} \quad \text{et} \quad \widehat{\sigma_n}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

1. **Intervalles sur la moyenne (variance connue):**
   - On suppose que la variance σ^2 est connue. Alors:

     $$\sqrt{n}\frac{(\bar{X}_n - m)}{\sigma} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

     Ceci donne lorsque l'on a une réalisation de $(X_1, X_2, .., X_n)$, l'intervalle de confiance:

     $$I_\alpha = [\bar{X}_n - \frac{\sigma.q_{(1-\alpha/2)}}{\sqrt{n}}, \bar{X}_n + \frac{\sigma.q_{(1-\alpha/2)}}{\sqrt{n}}], \quad \text{pour le niveau de confiance } \ 1-\alpha,$$

     avec q_a le quantile d'ordre a pour la loi $\mathcal{N}(0,1)$ (on utilise en pratique ce résultat dès que $n \geq 30$).
   - Si on suppose que $(X_1, X_2, .., X_n)$ est un échantillon de loi gaussienne. Alors:

     $$\sqrt{n}\frac{(\bar{X}_n - m)}{\sigma} \overset{\mathcal{L}}{\simeq} \mathcal{N}(0,1),$$

     pour tout $n \geq 2$ et l'intervalle de confiance est le même que précédemment.

2. **Intervalles sur la moyenne (variance inconnue):**
   - On suppose que la variance σ^2 est inconnue. Alors:

     $$\sqrt{n}\frac{(\bar{X}_n - m)}{\widehat{\sigma_n}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

     Ceci donne lorsque l'on a une réalisation de $(X_1, X_2, .., X_n)$, l'intervalle de confiance:

     $$I_\alpha = [\bar{X}_n - \frac{\widehat{\sigma_n}.q_{(1-\alpha/2)}}{\sqrt{n}}, \bar{X}_n + \frac{\widehat{\sigma_n}.q_{(1-\alpha/2)}}{\sqrt{n}}], \quad \text{pour le niveau de confiance } \ 1-\alpha,$$

     avec q_a le quantile d'ordre a pour la loi $\mathcal{N}(0,1)$ (ce résultat est vrai avec tout autre estimation convergente de σ. On lm'utilise en pratique dès que $n \geq 30$).
   - **Student** Si on suppose que $(X_1, X_2, .., X_n)$ est un échantillon de loi gaussienne, le résultat précédent est toujours vrai, mais on peut l'améliorer pour les n petits en utilisant le théorème suivant:

     $$\sqrt{n}\frac{(\bar{X}_n - m)}{\widehat{\sigma_n}} \overset{\mathcal{L}}{\simeq} t(n-1),$$

     pour tout $n \geq 2$ avec t la loi de Student à $(n-1)$ degrés de liberté. Ceci donne lorsque l'on a une réalisation de $(X_1, X_2, .., X_n)$, l'intervalle de confiance:

     $$I_\alpha = [\bar{X}_n - \frac{\widehat{\sigma_n}.t_{(1-\alpha/2)}(n-1)}{\sqrt{n}}, \bar{X}_n + \frac{\widehat{\sigma_n}.t_{(1-\alpha/2)}(n-1)}{\sqrt{n}}], \quad \text{pour le niveau de confiance } \ 1-\alpha,$$

     avec $t_a(n)$ le quantile d'ordre a pour la loi $t(n)$. Ce résultat est vrai pour tout $n \geq 2$.

- **Cas particulier de la binomiale** Si $(X_1, X_2, .., X_n)$ est un échantillon de loi de Bernoulli de paramètre p, alors:
$$\sqrt{n}\frac{(\bar{X}_n - p)}{\sqrt{n.\bar{X}_n.(1-\bar{X}_n)}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$
Ce résultat est untilisé en pratique pour $np(1-p) \geq 20$.

3. **Intervalles sur la variance:**

- **moyenne connue** On suppose $(X_1, X_2, .., X_n)$ est un échantillon de **loi gaussienne**, de moyenne m connue. Alors:
$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - m)^2 \overset{\mathcal{L}}{\simeq} \chi^2(n),$$
pour tout $n \geq 2$. On obtient donc l'intervalle de confiance sur la variance:
$$I_\alpha = [\frac{1}{c_{(1-\alpha/2)}(n)}\sum_{i=1}^{n}(X_i - m)^2, \frac{1}{c_{\alpha/2}(n)}\sum_{i=1}^{n}(X_i - m)^2], \text{ pour le niveau de confiance } 1-\alpha,$$
avec $c_a(n)$ le quantile d'ordre a pour la loi $\chi^2(n)$.
- **moyenne inconnue** On suppose $(X_1, X_2, .., X_n)$ est un échantillon de **loi gaussienne**. Alors:
$$\frac{(n-1)}{\sigma^2}\widehat{\sigma_n}^2 \overset{\mathcal{L}}{\simeq} \chi^2(n-1),$$
pour tout $n \geq 2$. On obtient donc l'intervalle de confiance sur la variance:
$$I_\alpha = [\frac{(n-1)}{c_{(1-\alpha/2)}(n)}\widehat{\sigma_n}^2, \frac{(n-1)}{c_{\alpha/2}(n)}\widehat{\sigma_n}^2], \text{ pour le niveau de confiance } 1-\alpha,$$
avec $c_a(n)$ le quantile d'ordre a pour la loi $\chi^2(n)$.

4. **Intervalles sur la différence de deux moyennes:**
On suppose que $(X_1, X_2, .., X_n)$ et $(Y_1, Y_2, .., Y_m)$ sont deux échantillons indépendants de **loi gaussienne de même variance** (mais pas forcément de même moyenne). On note d la différence entre les moyennes de X et de Y, à savoir $d = m_X - m_Y$. Alors:
$$\sqrt{\frac{nm(n+m-2)}{n+m}}\frac{(\bar{X}_n - \bar{Y}_m) - d}{\sqrt{(n-1)\widehat{\sigma_n(X)}^2 + (m-1)\widehat{\sigma_n(Y)}^2}} \overset{\mathcal{L}}{\simeq} t(n+m-2),$$
pour tout $n \geq 2$ et $m \geq 2$, avec $t(n+m-2)$ la loi de Student à $(n+m-2)$ degrés de liberté. On en déduit l'intervalle de confiance sur d:
$$I_\alpha = \left[(\bar{X}_n - \bar{Y}_m) - t_{(1-\alpha/2)}(n+m-2).\sqrt{\frac{(n+m)((n-1)\widehat{\sigma_n(X)}^2 + (m-1)\widehat{\sigma_n(Y)}^2)}{nm(n+m-2)}},\right.$$
$$\left.(\bar{X}_n - \bar{Y}_m) + t_{(1-\alpha/2)}(n+m-2).\sqrt{\frac{(n+m)((n-1)\widehat{\sigma_n(X)}^2 + (m-1)\widehat{\sigma_n(Y)}^2)}{nm(n+m-2)}}\right]$$
avec $t_a(n)$ le quantile d'ordre a pour la loi $t(n)$.

Trois outils de StixboxA111

On a déjà rencontré des intervalles de confiance calculés automatiquement par Matlab dans le cas du modèle linéaire. On peut ajouter aussi:

- La commande *test1n* qui permet dans le cas du modèle gaussien d'obtenir des intervalles de confiance sur la moyenne et la variance pour un échantillon simple (Student...). Pour n grand, on obtient aussi l'intervalle de confiance sur la moyenne dans le cas non-gaussien.
- La commande *test1n* qui permet dans le cas de deux échantillons gaussiens de même variance d'obtenir des intervalles de confiance sur la différence des moyennes et sur la variance. Pour n grand, on obtient aussi l'intervalle de confiance sur la différence des moyennes dans le cas non-gaussien.
- La commande *ciquant* qui fournit un intervalle de confiance sur le quantile d'ordre p de la loi d'un échantillon. La méthode utilisée se réfère au test de signe...

Utiliser ces différentes commandes à partir d'échantillons de loi uniforme ou normale simulés.

Exercices sur les intervalles de confiance

1. Simuler un n-échantillon de variables de Bernoulli de paramètre p. Donner un intervalle de confiance à 95% pour p. Reproduire N fois la simulation précédente et déterminer le nombre de fois où l'intervalle de confiance proposé contient bien le véritable paramètre p. Reprendre cette étude avec des échantillons gaussiens (on calculera alors des intervalles de confiance pour la moyenne et pour la variance de ces variables).
2. Dans le cadre de l'exercice précédent, comparer les intervalles de confiances exactes (obtenus avec la loi binomiale) et ceux donner par le TCL.
3. Générer deux n-échantillons de Gaussiennes dont les moyennes sont donnée par deux réalisations de v.a. de loi uniforme sur $[-1, 1]$. Donner des intervelles de confiance aux niveaux $(1-\alpha)\%$ sur la différence des moyennes. Recommencer 100 fois et donner des intervalles de confiance sur la moyenne des différences à la moyenne.
4. On considère 10 réalisations indépendantes de loi quelconque (par exemple, exponentielle, Poisson et Weibull). Donner des intervalles de confiance à $(1-\alpha)\%$ sur la moyenne et sur la variance en utilisant des quantiles empiriques calculés à partir des estimations des lois de la moyenne et de la variance.
5. Générer un n-échantillon de loi gaussienne centré. Comparer les intervalles de confiance sur la variance obtenus avec la loi de la variance empirique et avec un TLC portant sur les X_i^2. Utiliser d'autres lois. Le deuxième intervalle de confiance est-il utilisable si la moyenne est inconnue?