TP 9 : Estimation par intervalles de confiance, introduction aux tests statistiques

Dans ce TP, nous abordons pour la première fois la notion d'intervalle de confiance, en s'exercant sur des exemples simples où l'on peut écrire de tels intervalles, puis s'en servir pour un test statistique, ou pour de l'estimation.

Exercice 1 : Tester l'effet du maïs OGM En laboratoire, on a testé sur un échantillon de souris l'effet de la consommation de maïs OGM. Un groupe "test" de 15 souris a été nourri normalement, tandis qu'un autre groupe "témoin" de même taille a été nourri au maïs OGM pendant un an. A l'issue de l'expérience, on a pesé les souris (en grammes) et on a récolté les données. *Disclaimer: pour ce TP, les données sont factices et ne correspondent pas à une expérience réelle.*

1. Télécharger le fichier `souris30.csv` dans le même dossier que votre notebook, puis l'ouvrir et afficher les données. L'hypothèse d'une distribution normale vous paraît-elle cohérente ?

On cherche à répondre à la question suivante : y a-t-il une différence significative entre les poids des souris nourries par du maïs OGM, par comparaison avec les souris du groupe témoin ?

On rappelle pour cela que pour un échantillon i.i.d. $X_1, \dots, X_n$ de loi $\mathcal{N}(\mu, \sigma^2)$,

$$\sqrt{n}\frac{m_n - \mu}{\sqrt{v_n}} \sim t(n-1),$$

où m_n est la moyenne empirique, v_n l'estimateur sans biais de la variance, et $t(n-1)$ est la loi de Student à $n-1$ degrés de liberté.

2. On fait l'hypothèse d'une distribution normale. Déduire du précédent résultat un intervalle de confiance (non asymptotique) à 95% pour la moyenne des poids des souris test et des souris témoin. On pourra noter $t_\alpha^{(d)}$ le quantile d'ordre α de la loi de Student à d degré de libertés. Calculer en pratique les bornes de ces deux intervalles *(on pourra utiliser la fonction* `qt`*)*.

3. Retrouver ces mêmes intervalles avec la commande `t.test` appliquée aux deux jeux de données, consécutivement. Peut-on affirmer qu'il est probable que les moyennes sont différentes ? Que suggérez-vous pour améliorer la procédure de test ?

4. Répéter la même procédure pour le nouveau jeu de données `souris300.csv`, où le même test a été fait sur deux groupes de 150 souris. Calculer les intervalles de confiance asymptotiques. La réponse à la question "y a-t-il différence significative entre les deux échantillons" est-elle différente ?

Exercice 2 : Intervalle de confiance asymptotique pour le paramètre d'une loi de Poisson. On suppose que l'on dispose d'un échantillon $X_1, \dots, X_n$ d'une loi de Poisson de paramètre $\lambda > 0$.

1. Ecrire l'inégalité de Bienaymé-Tchebychev pour la moyenne empirique m_n. En déduire un intervalle de confiance *par excès et non asymptotique* de niveau $1-\alpha$. On pourra utiliser le raccourci suivant :

$$\forall x \geq 0,\ \forall \alpha > 0, \quad \lambda - \sqrt{\frac{\lambda}{\alpha n}} \leq x \leq \lambda + \sqrt{\frac{\lambda}{\alpha n}} \iff x - \frac{\sqrt{4\alpha n + 1} - 1}{2\alpha n} \leq \lambda \leq x + \frac{\sqrt{4\alpha n + 1} + 1}{2\alpha n}.$$

2. Ecrire une fonction `IC1(ech, alpha)` qui prend en argument un échantillon `ech` et un niveau d'erreur `alpha`, et qui renvoie l'intervalle de confiance établi en question 1 pour le paramètre λ. On pourra tester cette fonction sur un échantillon simulé.

3. Ecrire le théorème central limite pour la moyenne m_n. Pourquoi peut-on dire que

$$\sqrt{n}\frac{m_n - \lambda}{\sqrt{m_n}} \xrightarrow{(d)} \mathcal{N}(0,1)?$$

En déduire un nouvel intervalle de confiance, cette fois-ci *exact mais asymptotique*, pour le paramètre λ. On pourra noter q_α le quantile d'ordre α de la loi normale centrée réduite.

4. Ecrire une fonction `IC2(ech, alpha)` qui prend en argument un échantillon `ech` et un niveau d'erreur `alpha`, et qui renvoie l'intervalle de confiance établi en question 3 pour le paramètre λ *(on pourra utiliser la fonction `qnorm`)*. Pour le même échantillon simulé, comparer les intervalles de confiance `IC1` et `IC2`, plusieurs fois, en jouant sur n, λ et α. Commenter.

5. Dans cette question, on suppose que $\lambda = 3$ et on prendra $\alpha = 0.05$. Simuler $N = 10000$ échantillons de taille $n = 5000$, et estimer ainsi la probabilité que λ appartienne effectivement à l'échantillon `IC1`, ainsi que cette même probabilité pour `IC2`. Commenter les résultats obtenus.

6. Pour aller plus loin Pour λ et α fixés à des valeurs de votre choix, simuler un échatillon de taille $n = 4000$. Pour k variant de 1 à n, calculer les bornes des intervalles IC1 et IC2 *en ne prenant en compte que les k premières variables dans l'échantillon.* Représenter sur un graphique les bornes des intervalles IC1 et IC2 en fonction de n. On pourra aussi superposer une ligne horizontale correspondant au “vrai” λ. Commenter. On pourra jouer avec α.