EILCO **2023-2024**
Cycle Ingénieur $1^{\text{ère}}$ année

TP4 : Estimation, intervalles de confiance

Exercice 1

1. On tente de répondre à un questionnaire en donnant des réponses au hasard.
   (a) Pour chaque question, on a 20% de chances de répondre correctement. Simuler les réponses (correcte/incorrecte) pour un test comportant 20 questions.
   (b) Ecrire une fonction, dépendant de n et p, simulant les réponses à un questionnaire constitué de n questions lorsque la probabilité de bonne réponse, de chaque question, est égale à p.
2. Supposons qu'une classe de 100 étudiants passe un test "vrai ou faux" comportant 20 questions et que tous les étudiants répondent au hasard à chaque question (une réponse correcte rapporte 1 point ; et une réponse incorrecte 0).
   (a) Simuler l'ensemble des notes obtenues par les étudiants.
   (b) Donner la moyenne et l'écart-type des notes obtenus par la classe. Comparer les résultats obtenus aux valeurs théoriques.
   (c) Donner la proportion d'étudiants qui ont obtenu un pourcentage de bonnes réponses supérieur (ou égal) à 40%.

Exercice 2 Soit $X_1, \ldots, X_n$ un échantillon tiré d'une population distribuée selon une loi uniforme dans l'intervalle $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. On considère trois estimateurs sans biais du paramètre inconnu θ :

— la moyenne arithmétique $\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$;

— la médiane empirique

$$\hat{\theta}_2 = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ impair} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right) & \text{si } n \text{ pair;} \end{cases}$$

— la mi-étendue $\hat{\theta}_3 = \frac{X_{(1)} + X_{(n)}}{2}$.

1. Montrer, théoriquement, que les trois estimateurs sont consistants et sans biais.

2. A l'aide de simulations, vérifier que les trois estimateurs sont consistants et sans biais. *Indication : simuler un grand nombre N d'échantillons aléatoires de taille n. Puis, pour chaque échantillon, calculer les trois estimateurs, ainsi que la moyenne et la variance, par type d'estimateur, de tous les estimateurs obtenus.*
3. Déterminer lequel a la plus faible variance.

Exercice 3 Soit $X_1, \ldots, X_n$ un échantillon de loi exponentielle de paramètre λ. On pose

$$T = \frac{n}{X_1 + \cdots + X_n}.$$

1. Montrer que T est un estimateur consistant de λ.
2. Justifier que la variable aléatoire $\frac{n\lambda}{T}$ suit la loi Gamma [1] de paramètre $(n, 1)$.
3. En déduire un intervalle de confiance pour λ au niveau $1-\alpha$, pour tout $\alpha \in]0,1[$.
4. Un modèle théorique suggère que la durée des appels téléphoniques suit une distribution exponentielle de paramètre λ. Les données (en minutes), pour $n = 10$ appels, sont les suivantes :

   2.84, 2.37, 7.52, 2.76, 3.83, 1.32, 8.43, 2.25, 1.63, 0.27.

   (a) Donner une estimation ponctuelle de λ.
   (b) Calculer l'intervalle de confiance au seuil de 95% issu de ces données.

Exercice 4 Dans cet exercice, on suppose que X suive une loi gaussienne de variance connue mais d'espérance inconnue.

1. Simuler un échantillon de taille $n = 50$ d'espérance $\mu = 1$ et d'écart-type $\sigma = 1$. Calculer la moyenne empirique puis l'intervalle de confiance de μ au seuil de 95%.
2. Ecrire une fonction `confint.mean` qui prend comme arguments, l'échantillon, l'écart-type (connu) et le niveau de confiance, et retourne les deux bornes de l'intervalle de confiance. Tester cette fonction avec l'échantillon simulé avec des niveaux de confiance 0.90, 0.95 et 0.99.
3. (a) Simuler 100 échantillons de taille $n = 50$ avec une espérance $\mu = 1$ et un écart-type $\sigma = 1$, puis calculer les 100 intervalles de confiance au niveau 0.95 en utilisant la fonction `confint.mean`.
   (b) Compter le nombre d'intervalles de confiance ne contenant pas la vraie valeur μ.

1. On rappelle que la densité de la loi Gamma de paramètre (α, β) est donnée par $f_{\alpha,\beta}(x) = \frac{x^{\alpha-1}\beta^\alpha e^{-\beta x}}{\Gamma(\alpha)}$ pour tout $x > 0$.

Exercice 5 On considère des pairs de points (x_i, y_i), $i = 1, \ldots, n$, et on modélise leur relation par le modèle linéaire simple :

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

où les $(\varepsilon_i)_{i=1,\ldots,n}$ sont des erreurs iid, indépendantes des X_i, et telles que $\mathbb{E}[\varepsilon_i] = 0$ et $\mathbb{V}[\varepsilon_i] = \sigma^2$ pour tout i. En particulier, on a $\mathbb{E}[Y_i | X_i = x] = \alpha + \beta x$ et $\mathbb{V}[Y_i] = \sigma^2$.

On peut estimer α, β par la méthode des moindres carrés. Si on fait l'hypothèse que les ε_i suivent une loi normale $\mathcal{N}(0, \sigma^2)$, on peut montrer que les estimateurs des moindres carrés $\hat{\alpha}$, $\hat{\beta}$ possèdent de bonnes propriétés. Par exemple, pour β, on a :

— $\hat{\beta}_n \to \beta$ quand n tend vers l'infini (consistance) ;

— $\frac{\hat{\beta}_n - \beta}{\widehat{\mathrm{se}}} \sim \mathcal{N}(0,1)$ quand n tend vers l'infini (normalité asymptotique), où $\widehat{\mathrm{se}} = \sqrt{V(\hat{\beta}_n)}$.

Dans cet exercice, on veut vérifier empiriquement ces deux propriétés.

1. Consistance.
   (a) Simuler $n = 1000$ réalisations iid x_i selon une loi $\mathcal{N}(165, 100)$.
   (b) Simuler n réalisations y_i selon les lois $\mathcal{N}(\alpha + \beta x_i, \sigma^2)$, avec $\alpha = 165$, $\beta = 0.1$ et $\sigma = 0.5$ (cela revient à simuler $Y_i = \alpha + \beta x_i + \varepsilon_i$, avec $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$).
   (c) Fitter le modèle linéaire expliquant y par x et déterminer $\hat{\alpha}_n$ et $\hat{\beta}_n$.
   (d) Visualiser le nuage de points (x_i, y_i) et tracer la droite de régression.
2. Normalité asymptotique.
   (a) Simuler 500000 réalisations de (X_i, Y_i) comme ci-dessus.
   (b) Pour étudier la distribution de $\hat{\beta}_n$, on échantillonne aléatoirement les observations et, pour chaque échantillon, on estime le modèle (c'est l'idée à la base du *bootstrap*). Prendre 500 échantillons de taille $n = 1000$: pour chaque échantillon, fitter le modèle et déterminer $\hat{\beta}_n(1), \ldots, \hat{\beta}_n(500)$.
   (c) Visualiser l'histogramme de $\frac{\hat{\beta} - \beta}{\widehat{\mathrm{se}}}$.
   (d) Superposer la densité de $\mathcal{N}(0,1)$ à l'histogramme.