TP7 : Intervalles de confiance

Objectifs : Calculer et étudier les propriétés des intervalles de confiances pour une probabilité inconnue ou pour la moyenne et la variance inconnues d'une variable aléatoire. Avec données simulées et données réelles `apnee.csv`.

1 Données simulées

Dans cette partie on considère un échantillon i.i.d. de X pour X de loi de Bernoulli $\mathcal{B}(p)$ où p est inconnu. Rappelons que si $X_1, ..., X_n$ est un échantillon aléatoire de la variable de Bernoulli $\mathcal{B}(p)$, l'intervalle suivant est un intervalle de confiance de niveau approximatif (pour n assez grand) $1-\alpha$ pour le paramètre inconnu p :

$$\left[\bar{X}_n - \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}u_{1-\frac{\alpha}{2}}, \bar{X}_n + \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}u_{1-\frac{\alpha}{2}}\right]$$

Dans ce modèle $E(X) = p$, donc pour estimer p, on utilise l'estimateur usuel $\bar{X}_n$. Cet estimateur est aussi noté F_n dans le cours et appelé fréquence empirique. C'est la fréquence d'apparition de l'évènement d'intérêt de probabilité inconnue p et codé par la valeur 1 dans l'échantillon $X_1, ..., X_n$, puisqu'une moyenne de valeurs qui sont soit 0 soit 1 est la fréquence d'apparition du 1.

Exercice 1 : Intervalle de confiance pour p

1. Définir $p = 0.3$ et $n = 500$. Générer un tirage d'un échantillon de taille n pour une variable de Bernoulli $\mathcal{B}(p)$ et l'affecter à `x`.

2. Pour chaque $k = 1, ..., n$, calculer les moyennes des k premières valeurs tirées de x (où $x = (x_1, ..., x_n)$) et les affecter à `moy.taille` (il y a n valeurs dans ce vecteur et on pourra utiliser `cumsum()` pour son calcul.

3. Calculer les bornes inférieures et supérieures des intervalles de confiance de niveau 95% obtenus pour chacun des k-échantillons précédents avec k = 1, ..., n. Affecter les résultats à `Binf.taille` et `Bsup.taille.`

4. Calculer les amplitudes de chacun des intervalles, à affecter à `amplitude.taille` et représenter graphiquement la suite des amplitudes obtenues. Que se passe-t-il lorsque la taille de l'échantillon augmente ? Ajouter la droite horizontale qui passe par l'ordonnée 0.1 et en déduire une lecture de la taille minimum d'échantillon qu'il faut prendre pour que l'amplitude obtenue soit inférieure ou égale à 0.1 (soit une précision de l'estimation de p à -0.05 ou $+0.05$ près).

5. On considère à présent l'échantillon (complet) des n valeurs tirées. Que retourne la commande `prop.test(sum(x),length(x))` ? Quelle option doit-on utiliser dans cette dernière fonction pour obtenir un niveau de confiance autre que le niveau 95% choisi par défaut ?

6. Calculer les intervalles obtenus pour des valeurs de $\alpha = 0.01, 0.02, ..., 0.30$ et leurs amplitudes. On pourra les affecter à `Binf.niveau, Bsup.niveau` et `amplitude.niveau`.

7. Représenter graphiquement la suite des amplitudes obtenues selon $1-\alpha$ et indiquer à partir de quel niveau de confiance on peut garantir une amplitude inférieure à 0.1 ?

8. Calculer exactement la valeur de niveau pour lequel l'intervalle est exactement d'amplitude 0.1.

2 Données réelles

Les données du fichier `apnee.csv` décrivent l'observation de 35 sujets souffrant d'apnée du sommeil (`apnee==1`)et de 65 sujets ne souffrant pas de ce problème (`apnee==0`). Pour chaque sujet ont été observés l'âge, le poids, la taille, le tabagisme (1 si fumeur), le sexe (0 pour les hommes) et le nombre de verres d'alcool consommés par jour. Ces données ont été collectées afin d'évaluer les effets potentiels du sexe, de la consommation d'alcool et de tabac ou du poids sur l'apnée du sommeil.

Exercice 2 : Intervalles de confiance pour μ et σ^2

1. Charger les données `apnee.csv` et affecter le data.frame à `data`.
2. Extraire du data.frame l'échantillon des mesures de la variable `poids` chez les hommes, avec la commande `data[data$sexe==0,"poids"])` et l'affecter à `poidsH`.
3. Cette variable semble-t-elle avoir une répartition gaussienne ?
4. Supposons que le poids des hommes suit une loi normale $\mathcal{N}(\mu, \sigma^2)$. Calculer l'estimation sans biais de la moyenne μ et celle de la variance σ^2.
5. On supposera ici $\sigma^2 = 19^2$ connu. Calculer l'intervalle de confiance de niveau 95% pour μ.
6. On suppose à présent σ inconnu et estimé par $\hat{\sigma}$, calculer l'intervalle de confiance pour μ de niveau 95%.
7. Que retourne la commande `t.test(poidsH)` ? En utilisant l'option `conf.level` dans les arguments de `t.test()`, calculer un intervalle de confiance de niveau 99%.
8. Calculer l'intervalle de confiance de niveau 90% pour la variance σ^2.
9. En quoi le fait que l'hypothèse gaussienne ne semble pas respectée (voir avec `qqnorm`) n'est pas ici une contre-indication pour la validité des intervalles de confiance, précédemment calculés ?

Exercice 3 : comparer des probabilités p_1 et p_2

1. Soit p_1 la probabilité de souffrir d'apnée du sommeil lorsque l'on est fumeur et p_2 la probabilité de souffrir d'apnée du sommeil lorsque l'on est non-fumeur.
2. Extraire l'échantillon des mesures d'apnée pour les fumeurs et pour les non-fumeurs et les affecter à `apnee.fumeurs` et à `apnee.non.fumeurs`.
3. Estimer p_1 et p_2 et calculer les intervalles de confiance de niveau 90%. Comparer ces intervalles et conclure.