TP 4 – SY02
Intervalle de confiance

Les questions/sections marquées par un sont des questions qui sont prévues pour être traitées en autonomie en dehors de la séance de TP.

1 Fonctions pivotales

Dans cette section, nous allons vérifier expérimentalement les lois suivies par les 2 principales fonctions pivotales utilisées pour la construction d'intervalles de confiance dans le cadre d'un échantillon iid Gaussien, c'est à dire lorsque $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

Commençons par le théorème de Fisher qui stipule que :

$$\frac{(n-1)S^{*2}}{\sigma^2} \sim \chi^2_{n-1}. \tag{1}$$

1 On suppose fixés μ, σ et n. Définir une fonction `chisq1` qui renvoie une seule observation de la variable aléatoire

$$\frac{(n-1)S^{*2}}{\sigma^2},$$

puis illustrer graphiquement la relation (1) (cf TP3).

2 Réaliser le même travail pour illustrer le fait que,

$$\frac{\overline{X} - \mu}{S^*/\sqrt{n}} \sim \mathcal{T}_{n-1}.$$

2 Intervalles de confiance

Dans cette section, nous allons illustrer la notion de "niveau de confiance" d'un intervalle de confiance. Comme dans la section précédente, notre modèle est un n-échantillon iid Gaussien, i.e., $X_1, \dots, X_n \overset{\text{iid}}{\sim} X$, avec $X \sim \mathcal{N}(\mu, \sigma^2)$ et nous considérons les cas σ^2 connu et σ^2 inconnu.

3 (σ^2 connu) Après avoir rappelé l'expression de l'intervalle de confiance bilatéral au niveau $1-\alpha$ sur l'espérance μ d'une variable aléatoire qui suit une loi normale de variance connue σ^2, générer un échantillon de taille n selon une loi normale avec les paramètres de votre choix et donner une réalisation de cet intervalle de confiance ; vérifier la cohérence de votre calcul par rapport au paramètre choisi μ.

On suppose dans la suite de la section que le paramètre σ n'est plus connu. On ne peut donc pas s'en servir dans l'expression de l'intervalle de confiance.

(4) Donner l'expression de l'intervalle de confiance lorsque σ n'est pas connu et calculer une réalisation de l'intervalle de confiance avec les observations précédentes. Retrouver cet intervalle en se servant de la fonction de test que l'on verra prochainement :

```r
t.test(x, conf.level = 1 - alpha)$conf.int
```

(5) Créer une fonction `gen_IC` qui prend en argument un échantillon `x` de taille quelconque et un niveau de signification α et renvoie l'intervalle de confiance sur l'espérance sous forme d'un vecteur de longueur 2 contenant la borne inférieure et la borne supérieure de l'intervalle.

(6) Utiliser la fonction `replicate` pour créer plusieurs intervalles de confiance. Que contient le résultat de `replicate` ?

Pour visualiser ces intervalles de confiance, la fonction `plot_ICs` est mise à votre disposition dans le fichier `utils.R` présent dans le sous-dossier `src/`. Pour l'utiliser, il faut charger les définitions présentes de le fichier avec la commande

```r
source("src/utils.R")
```

(7) Visualiser les intervalles de confiance avec la fonction `plot_ICs`. Quelle est la relation entre le niveau de l'intervalle de confiance (95%), le nombre d'intervalles verts et le nombre total d'intervalles ?

(8) Étudier l'influence de n sur la largeur moyenne de l'intervalle de confiance. Pour fixer l'échelle des abscisses et éviter que `R` décide, on pourra utiliser le paramètre optionnel `xlim` de la fonction `plot_ICs`.

(9) Étudier l'influence de la dispersion de l'échantillon sur la largeur moyenne de l'intervalle de confiance. Pour fixer l'échelle des abscisses, on pourra utiliser le paramètre optionnel `xlim` de la fonction `plot_ICs`.

Pour illustrer plus précisément la relation établie à la question 7 avec un nombre quelconque d'intervalles de confiance, on va créer une fonction qui génère un échantillon, calcule l'intervalle de confiance associé et renvoie `TRUE` si l'intervalle contient le paramètre.

(10) Créer cette fonction et calculer la proportion d'intervalles qui contient le paramètre (appelé le taux de recouvrement). Commenter le résultat.

3 Lemme de Slutsky

Le but de cette section est d'illustrer l'application du lemme de Slutsky lors de la recherche d'un intervalle de confiance asymptotique sur la proportion b dans un modèle Binomial $B(n, p)$. D'après le polycopié de cours, on a la convergence en loi suivante,

$$\frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \tag{2}$$

Il est difficile d'extraire p de l'expression précédente. On choisit donc d'utiliser le lemme de Slutsky. On trouve alors l'intervalle asymptotique suivant,

$$IC = \left[\widehat{p} - u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}; \widehat{p} + u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right]. \tag{3}$$

(11) Écrire une fonction qui prend en argument la proportion recherchée p, la longueur de l'échantillon n, le nombre de fois k où on réitère l'expérience et le niveau $1 - \alpha$ des intervalles de confiance et renvoie la proportion de réalisations des intervalles (3) qui contiennent le paramètre p parmi les k expériences.

(12) Tracer cette proportion en fonction de n. On pourra utiliser une échelle logarithmique pour n.

Le calcul de l'intervalle de confiance directement issu de (2) sans utiliser le lemme de Slutsky est possible. On trouve

$$IC = \left[\frac{2n\widehat{p} + u^2_{1-\alpha/2} \pm u_{1-\alpha/2}\sqrt{u^2_{1-\alpha/2} + 4n\widehat{p}(1-\widehat{p})}}{2n + 2u^2_{1-\alpha/2}}\right].$$

(13) Écrire la même fonction que précédemment avec ce nouvel intervalle de confiance et comparer. Qu'en concluez-vous ?