# Applications du bootstrap

## 1 Comparaison des intervalles de confiance

On veut comparer les résultats donnés par les différentes méthodes de construction d'intervalle de confiance pour le problème suivant.

Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de loi gamma $\Gamma(p, \theta)$ et $Y_1, \ldots, Y_n$ un échantillon i.i.d. de loi gamma $\Gamma(q, \kappa)$. On veut construire un intervalle de confiance pour le paramètre

$$\tau = \frac{\mathbb{E}(X)}{\mathbb{E}(Y)} = \frac{p\kappa}{q\theta}.$$

On utilise pour cela la statistique

$$T(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = \frac{\bar{X}_n}{\bar{Y}_n} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n Y_i} = \hat{\tau}.$$

**Remarque :**

1. La densité de la loi gamma $\Gamma(p, \theta)$ est définie par :

$$f_{\Gamma(p,\theta)}(x) = \frac{\theta^p}{\Gamma(p)} x^{p-1} e^{-\theta x} \text{ pour } x \geq 0.$$

2. Le paramètre $p$ est le paramètre de forme (shape) de la loi gamma et le paramètre $\theta$ est l'inverse du paramètre d'échelle (scale) et est appelé taux (rate). Pour tirer un échantillon de taille $n$ d'une loi gamma, on écrit en R `rgamma(n, shape=, rate =)`.

### 1.1 Intervalle de confiance exact

1. Montrer que la v.a. $(\theta \sum_{i=1}^n X_i)/(\kappa \sum_{i=1}^n Y_i) = (\theta/\kappa)\hat{\tau}$ suit la loi beta de seconde espèce $\beta(np, nq)$. La densité de la loi beta est donnée par :

$$f_{\beta(np,nq)}(x) = \frac{\Gamma(np+nq)}{\Gamma(np)\Gamma(nq)} \frac{x^{np-1}}{(1-x)^{np+nq}} \text{ pour } x \geq 0.$$

2. En déduire que l'intervalle de confiance exact au niveau de confiance $1-\alpha$ pour $\tau$ s'écrit :

$$IC_{\text{exact}}(\alpha) = \left[ \frac{p\hat{\tau}}{qF^{-1}_{\beta(np,nq)}(1-\alpha/2)}; \frac{p\hat{\tau}}{qF^{-1}_{\beta(np,nq)}(\alpha/2)} \right],$$

où $F^{-1}_{\beta(np,nq)}(a)$ est le fractile d'ordre de la loi beta de seconde espèce.

**Remarque :** $F^{-1}_{\beta(np,nq)}(a)$ s'écrit en R `qbeta(a,np,nq)/(1-qbeta(a,np,nq))` car c'est la loi beta de première espèce qui est programmée dans le logiciel.

## 1.2 Intervalle de confiance par approximation normale

On sait, par le Théorème Central Limite, que :

$$\sqrt{n}\left(\bar{X}_n - p/\theta\right) \overset{Loi}{\rightsquigarrow} U \sim \mathcal{N}(0, p/\theta^2) \text{ et } \sqrt{n}\left(\bar{Y}_n - q/\kappa\right) \overset{Loi}{\rightsquigarrow} V \sim \mathcal{N}(0, q/\kappa^2).$$

Par ailleurs, la delta-méthode assure que si

$$\sqrt{n}\left((U_n, V_n) - (u, v)\right) \overset{Loi}{\rightsquigarrow} (U, V)$$

et si $\Psi$ est continûment différentiable, on a :

$$\sqrt{n}\left(\Psi(U_n, V_n) - \Psi(u, v)\right) \overset{Loi}{\rightsquigarrow} \Psi'(u, v) \cdot \begin{pmatrix} U \\ V \end{pmatrix}.$$

1. Montrer que :

$$\sqrt{n}\left(\frac{\bar{X}_n}{\bar{Y}_n} - \frac{p\kappa}{q\theta}\right) = \sqrt{n}\left(\hat{\tau} - \tau\right) \overset{Loi}{\rightsquigarrow} \frac{\kappa}{q} U - \frac{p}{\theta}\left(\frac{\kappa}{q}\right)^2 V.$$

2. Quelle est la loi limite de $\sqrt{n}\left(\hat{\tau} - \tau\right)$ ?
3. En déduire que l'intervalle de confiance asymptotique par approximation normale au niveau de confiance $1 - \alpha$ pour $\tau$ s'écrit :

$$IC_{\text{AN}}(\alpha) = \left[\hat{\tau} - \frac{\hat{\sigma}}{\sqrt{n}}\Phi^{-1}(1 - \alpha/2); \hat{\tau} + \frac{\hat{\sigma}}{\sqrt{n}}\Phi^{-1}(1 - \alpha/2)\right]$$

où

$$\hat{\sigma} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{\bar{Y}_n^2} + \frac{\bar{X}_n^2}{\bar{Y}_n^4}\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}.$$

## 1.3 Intervalle de confiance du bootstrap basique

Donner la construction de l'intervalle de confiance du bootstrap basique pour $\tau$ au niveau de confiance $1 - \alpha$.

**Remarques :**

1. Pour tirer un vecteur de longueur $l$ dans le vecteur $v$ avec remise, on écrit en R `sample(v,l,replace=T)`
2. Pour classer les composantes d'un vecteur $v$ par ordre croissant, on écrit en R `sort(v)`.
3. La partie entière supérieure du réel $r$ s'écrit en R `ceiling(r)`.

## 1.4 Intervalle de confiance percentile

Donner la construction de l'intervalle de confiance percentile pour $\tau$ au niveau de confiance $1 - \alpha$.

## 1.5 Intervalle de confiance t-bootstrap

Donner la construction de l'intervalle de confiance t-bootstrap pour $\tau$ au niveau de confiance $1 - \alpha$.

### 1.6 Comparaison des intervalles de confiance

**Premier Programme**

Écrire un programme dont les paramètres sont $n = 10$, $p = 0.7$, $\theta = 0.007$, $q = 1$, $\kappa = 0.02$, $\alpha = 0.05$ et qui calcule les bornes inférieures et supérieures des différents intervalles de confiance pour $\tau = 2$ avec $B = 1000$ tirages pour le bootstrap.

**Second Programme**

En utilisant le programme précédent, calculer sur $M = 1000$ échantillons (ce qui revient à répéter le premier programme 1000 fois.

1. Calculer la couverture moyenne des différents intervalles de confiance, où la couverture d'un intervalle de confiance $[ICinf, ICsup]$ vaut 1 si $ICinf \leq \tau \leq ICsup$ et 0 sinon.
2. Puis calculer la longueur moyenne des différents intervalles de confiance, où la longueur d'un intervalle de confiance $[ICinf, ICsup]$ vaut $ICsup - ICinf$.

Quelle est la couverture théorique ? Comparer les différents intervalles.

## 2 Application sur données réelles

Utiliser la technique du bootstrap pour donner un intervalle de confiance pour la médiane dans le jeu de données "galaxies" de la librairie MASS (`library(MASS)`). Construire également un histogramme calculé grâce au bootstrap pour estimer la loi de la médiane.

Le jeu de données "galaxies" contient les vitesses (en km/sec) de 82 galaxies appartenant à 6 sections coniques distinctes de la région de la Couronne Boréale. Dans une telle étude, la multimodalité met en évidence des vides (voids) et des grands regroupements (superclusters) de galaxies dans le lointain univers.

(`http://en.wikipedia.org/wiki/Corona_Borealis`)