TP1 : Découverte de la fonction `boot`

ENSAE 3ème année, 2017

Applications du bootstrap et autres techniques de rééchantillonnage (cours de M. Roquain)

Exercice 0 (Création de votre fichier)

- Créer votre propre fichier `nom_prenom_TP1.R`
- Il est aussi possible de fournir un *pdf* assez propre à peu de frais à partir d'un document de style `.R` A cette fin, on pourra consuler le fichier `Template.R` dans *Rstudio* (voir le site web http://etienne.roquain.free.fr/teaching.html). Regarder attentivement le code puis le pdf généré.

Exercice 1 (Echauffement avec le théorème central limite)

- Faire une **première section** s'intitulant "*Illustration du théorème central limite*".
- Présenter des graphiques illustrant le théorème central limite pour n variables aléatoires i.i.d. suivant la distribution P de votre choix.

Pour les simulations, on pourra utiliser différentes fonctions de `R` comme `rbinom`, `rgamma`. On pourra faire par exemple 1000 simulations. Pour les graphiques, on pourra utiliser `hist`, `curve`, `dnorm` ou encore `qqnorm`. Si vous ne connaissez pas ces fonctions, vous avez la possibilité de consulter les aides de `R`, qui sont généralement assez claires. N'oubliez pas aussi de *privilégier les produits de matrices plutôt que les boucles "for"* (avec par exemple la fonction `sapply`). Notez que si vous connaissez un peu le `Latex`, vous pouvez aussi écrire l'énoncé du TCL dans votre fichier en vous inspirant de la formule du template.

Exercice 2 (Un échantillon bootstrap)

- Faire une **deuxième section** s'intitulant "*Génération d'un échantillon bootstrap*".
- Générer $\mathcal{X}_n = (X_1, \dots, X_n)$ un échantillon de variables i.i.d. suivant la distribution P de votre choix ($n = 50$).
- Représenter la fonction de répartition empirique de $\mathcal{X}_n$ à l'aide de la fonction `ecdf`. Quelle est la loi $\hat{P}_n$ associée à cette fonction de répartition ?
- Construire **un seul** échantillon bootstrap $(X_1^\star, \dots, X_n^\star)$, c'est-à-dire un n-échantillon i.i.d. selon la distribution $\hat{P}_n$. On pourra utiliser la fonction `sample`.
- Calculer le nombre d'éléments différents dans $(X_1^\star, \dots, X_n^\star)$ à l'aide des fonctions `unique` et `length`. Commenter.

Exercice 3 (Approximation bootstrap pour la moyenne)

Soit $\mathcal{X}_n = (X_1, \dots, X_n)$ un échantillon de variables i.i.d. suivant la distribution Bernoulli de paramètre $\theta = 1/2$. Nous cherchons à approcher la loi

$$L_n(P) = \mathcal{L}\left(n^{1/2}(\bar{X}_n - \theta)\right)$$

par

$$L_n(\hat{P}_n) = \mathcal{L}\left(n^{1/2}(\bar{X}_n^\star - \bar{X}_n)\middle|\mathcal{X}_n\right)$$

Pour se faire, il convient d'approcher $L_n(\hat{P}_n)$ par Monte-Carlo. Ceci peut se faire en appelant B fois la fonction `sample`, mais `R` a tout prévu et la librairie `boot` simplifie grandement les choses.

- Faire une **troisième section** s'intitulant "*Approximation bootstrap pour la moyenne*".
- Générer les $n = 100$ variables aléatoires de l'échantillon $\mathcal{X}_n$, et stocker les dans un vecteur `data`.
- Appeler la librairie `boot` et construire $B = 1000$ échantillon bootstrap à l'aide de la fonction `boot`. Voici un exemple :

```
library(boot)
n=100
theta=1/2
data = rbinom(n,1,theta)
thetachap=mean(data)
B=1000
bootdata = boot(data,function(y,i) mean(y[i]),R=B)
```

- A l'aide d'un graphique, représenter (l'approximation de) la distribution $L_n(\hat{P}_n)$ par un histogramme à l'aide de la fonction `hist`.
- Lorsque n est grand, quelle doit être $L(P)$ la limite de $L_n(\hat{P}_n)$? Ajouter la densité de la loi $L(P)$ sur l'histogramme précédent. Commenter.

Exercice 4 (Approximation bootstrap pour la médiane)

Adapter l'exercice précédent pour approcher cette fois

$$L_n(P) = \mathcal{L}\left(n^{1/2}(F_n^{-1}(1/2) - F^{-1}(1/2))\right)$$

lorsque $\mathcal{X}_n = (X_1, \ldots, X_n)$ est un échantillon de variables i.i.d. de loi P de Cauchy paramètre $\theta = 1$.

On rappelle que la densité associée à la loi de Cauchy de paramètre θ est

$$f_\theta(x) = \frac{1}{\pi(1 + (x - \theta)^2)}.$$

Aussi, on pourra utiliser la fonction `quantile` pour calculer la médiane mais attention au `type` utilisé ! Essayer par exemple `quantile(c(1,2),1/2)` et `quantile(c(1,2),1/2,type=1)` pour voir...