IFT 3245
Simulation et modèles

Fabian Bastin
DIRO
Université de Montréal

Automne 2016

Intervalle de confiance par rééchantillonnage (“bootstrap”)

Il s’agit de de techniques de simulation appliquées en statistique; l’idée est d’estimer la distribution (inconnue et quelconque) de l’estimateur en rééchantillonnant des échantillons de taille n en tirant avec remplacement dans l’échantillon de taille n original.

Pour chaque échantillon ainsi construit, nous recalculons l’estimateur en cours d’étude.

Principe de plug-in

Considérons un échantillon i.i.d. $\boldsymbol{X} = X_1, \ldots, X_n$, issu d'une loi de fonction de répartition F, et un estimateur $Y = g(X_1, \ldots, X_n)$ d'une valeur réelle inconnue θ.

Exemple:

$$Y = \overline{X}_n$$

avec $\theta = \mu$, ou

$$Y = S_n^2$$

avec $\theta = \sigma^2$.

Y peut être biaisé (i.e. $E[Y] \neq \theta$), mais nous supposons que g ne dépend pas de l'ordre des X_i's.

Principe de plug-in

Si nous ne connaissons pas la ditribution exacte F, nous pouvons toujours nous diriger vers la distribution empirique construite à partir de l'échantillon $\boldsymbol{X}$.

L'estimateur plug-in d'un paramètre $\theta = g(F)$ est défini comme

$$\hat{\theta} = g(\hat{F}).$$

En général, l'estimateur plug-in d'une espérance $\theta = E_F(x)$ est

$$E_{\hat{F}} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x},$$

autrement dit nous retrouvons l'estimateur de moyenne classique.

Principe de plug-in

Le principe de plug-in est en général assez bon, si la seule source d'information disponible à propos de F vient de l'échantillon $\boldsymbol{X}$.

Sous cette circonstance, $\hat{\theta}_n = g(\hat{F}_n)$ ne peut pas être amélioré comme estimateur de $\theta = g(F)$, du moins pas dans le sens asymptotique habituel en théorie statistique. Par exemple, si $\hat{f}_k$ est l'estimateur de fréquence plug-in $\#\{x_i = k\}/n$, alors

$$\hat{f}_k \sim \text{Bi}(n, f_k)/n.$$

Dans ce cas, l'estimateur $\hat{f}_k$ est non-biaisé pour f_k, $E[\hat{f}_k] = f_k$, de variance $f_k(1 - f_k)/n$. Il s'agit de la plus petite variance possible pour un estimateur sans biais de f_k.

Bootstrap non-paramétrique

Considérons à présent $K_n(F, z) = P[Y - \theta \leq z]$ pour $z \in \mathcal{R}$. Un intervalle de confiance exact pour θ, au niveau $1 - \alpha_1 - \alpha_2$, est

$$(I_1, I_2) = (Y - K_n^{-1}(F, 1 - \alpha_1),\ Y - K_n^{-1}(F, \alpha_2)),$$

où $K_n^{-1}(F, q)$ est le q-quantile de $K_n(F, \cdot)$.

En effet,

$$\begin{aligned} P[I_1 > \theta] &= P[Y - \theta > K_n^{-1}(F, 1 - \alpha_1)] \\ &= 1 - K_n(F,\ K_n^{-1}(F, 1 - \alpha_1)) \\ &= \alpha_1. \end{aligned}$$

De même, nous avons $P[I_2 < \theta] = \alpha_2$.

Bootstrap non-paramétrique

Toutefois, il est rare de connaître $K_n(F, \cdot)$.

Une première idée serait de répéter l'expérience m fois afin obtenir m copies i.i.d. de Y pour estimer sa distribution. Si $E[Y] = \theta$, on peut estimer ainsi la distribution de $Y - \theta$. Mais cela ferait mn simulations!

Souvent, il est très coûteux, voire même impossible, d'avoir de nouvelles copies de Y. L'idée du bootstrap consiste à remplacer F par $\hat{F}_n$ et θ par y dans $K_n(F, z)$.

Soient $x_1, \ldots, x_n$ les valeurs de $X_1, \ldots, X_n$ et $y = g(x_1, \ldots, x_n)$. Tirons $X_1^*, \ldots, X_n^*$ au hasard avec remplacement de l'échantillon de départ $\{x_1, \ldots, x_n\}$ (i.e., de $\hat{F}_n$) et calculons $Y^* = g(X_1^*, \ldots, X_n^*)$.

Bootstrap non-paramétrique de base

L'opération est répétée m fois, de sorte que nous obtenions m copies i.i.d. de Y^*, à savoir $Y_1^*, \ldots, Y_m^*$.

Cela revient à répéter l'expérience m fois avec $\hat{F}_n$ au lieu de F.

La notation étoile indique que $\boldsymbol{x}^*$ n'est pas l'ensemble de données réel $\boldsymbol{x}$, mais plutôt une version randomisée, ou rééchantillonnée, de $\boldsymbol{x}$.

Ex: algorithme bootstrap d'estimation d'écarts-type

1. Tirer m échantillons bootstrap indépendants $\boldsymbol{x}_1^*$, $\boldsymbol{x}_2$,..., $\boldsymbol{x}_m^*$, chacun consistant de n valeurs de données tirées avec remplacement de $\boldsymbol{x}$.
2. Evaluer la réplication bootstrap correspondante à chaque échantillon bootstrap,

$$\hat{\theta}^*(i) = g(x_i^*), \ i = 1, 2, \ldots, m.$$

3. Estimer l'erreur standard $\mathrm{se}_F(\hat{\theta})$ par l'écart-type échantillonnal des m réplications:

$$\hat{\mathrm{se}}_m = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}^*(i) - \hat{\theta}^*(\cdot))^2},$$

où $\hat{\theta}^*(\cdot) = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}^*(i)$.

$$\lim_{m\to\infty} \hat{\mathrm{se}}_m = \mathrm{se}_{\hat{F}} = \mathrm{se}_{\hat{F}}(\hat{\theta}^*).$$

Ex: algorithme bootstrap d'estimation d'écarts-type

L'estimateur de bootstrap idéal $\text{se}_{\hat{F}}(\hat{\theta}^*)$ et son approximation $\hat{\text{se}}_m$ sont parfois appelés estimateurs bootstrap non-paramétriques car ils sont basés sur $\hat{F}$, l'estimateur non-paramétrique de la population F.
Soit $\hat{K}_{n,m}$ la fonction de répartition empirique de $Y_1^* - y, \ldots, Y_m^* - y$. Pour $m \to \infty$, elle converge vers la fonction de répartition de $Y^* - y$, qui est $K_n(\hat{F}_n, \cdot)$. L'intervalle de confiance retourné est:

$$(y - \hat{K}_{n,m}^{-1}(1-\alpha_1),\ y - \hat{K}_{n,m}^{-1}(\alpha_2)) = (2y - Y^*_{(\lceil m(1-\alpha_1) \rceil)},\ 2y - Y^*_{(\lceil m\alpha_2 \rceil)}).$$

Cela revient à remplacer F par $\hat{F}_n$ puis à approximer $K_n(\hat{F}_n, \cdot)$ par $\hat{K}_{n,m}$. Il y a donc deux sources d'erreur, qui sont cependant la plupart du temps inévitables.

Bootstrap-t non-paramétrique

Supposons que nous disposons également d'un estimateur de la variance de Y, disons $S^2 = h^2(X_1, \ldots, X_n)$.

Soit $J_n(F, \cdot)$ la fonction de répartition de la statistique studentisée $(Y - \theta)/S$.

Un intervalle de confiance exact de niveau $(1 - \alpha_1 - \alpha_2)$:

$$(I_1, I_2) = (Y - J_n^{-1}(F, 1 - \alpha_1)S,\ Y - J_n^{-1}(F, \alpha_2)S).$$

L'algorithme du bootstrap-t non-paramétrique consiste, pour chacune des m répétitions bootstrap, à générer n observations $X_1^*, \ldots, X_n^*$ comme avant, puis à calculer $Y^* = g(X_1^*, \ldots, X_n^*)$, $S^* = h(X_1^*, \ldots, X_n^*)$, et $Z^* = (Y^* - y)/S^*$.

Bootstrap-t non-paramétrique

Soient $Z_1^*, \ldots, Z_m^*$ les m copies i.i.d. de Z^* et $\hat{J}_{n,m}$ leur fonction de répartition empirique. Pour calculer l'intervalle de confiance, on remplace $J_n(F, \cdot)$ par $\hat{J}_{n,m}(\cdot)$:

$$\begin{aligned}(I_1, I_2) &= (y - \hat{J}_{n,m}^{-1}(1-\alpha_1)S,\ y - \hat{J}_{n,m}^{-1}(\alpha_2)S) \\ &= (y - Z^*_{(\lceil m(1-\alpha_1)\rceil)}S,\ y - Z^*_{(\lceil m\alpha_2\rceil)}S).\end{aligned}$$

Empiriquement, le bootstrap-t performe souvent le mieux.

Le choix de m influence peu l'erreur de couverture, mais un trop petit m donne des intervalle de confiance dont la largeur varie beaucoup.

Un choix populaire consiste à prendre $m = 1000$.

Estimation du biais

Une application particulièrement intéressante du bootstrap est la possibilité d'estimer le biais d'un estimateur quelconque.

Sous la distribution F, le biais d'un estimateur $\hat{\theta} = g(\boldsymbol{X})$ d'une quantité inconnue $\theta = t(F)$ est défini comme

$$B_F(\hat{\theta}, \theta) = E_F[g(\boldsymbol{X})] - t(F).$$

L'estimateur bootstrap de biais est défini comme

$$B_{\hat{F}}(\hat{\theta}, \theta) = E_{\hat{F}}[g(\boldsymbol{X}^*)] - t(\hat{F}).$$

L'estimateur plug-in $t(\hat{F})$ de θ peut différer de $\hat{\theta} = g(x)$. En d'autres termes, $B_{\hat{F}}(\hat{\theta}, \theta)$ est l'estimateur plug-in de $B_F(\hat{\theta}, \theta)$, que $\hat{\theta}$ soit ou non l'estimateur plug-in de θ.

Estimation du biais

Dans la plupart des cas, $E_{\hat{F}}[g(\boldsymbol{X}^*)]$ devra être approximé par simulation Monte-Carlo:

$$\hat{\theta}^* = \frac{1}{m}\sum_{i=1}^{m}\theta^*(i) = \frac{1}{m}\sum_{i=1}^{m} g(\boldsymbol{x}_i^*).$$

L'estimateur de bootstrap de biais basé sur les m réplications bootstrap est

$$\hat{B}_m = \hat{\theta}^* - t(\hat{F}).$$

Estimation du biais: version améliorée

Il est possible d'améliorer cet estimateur quand $\hat{\theta}$ est l'estimateur plug-in $t(\hat{F})$ de $\theta = t(F)$.

Soit P_j^* la proportion du j^e point de données originales dans l'échantillon bootstrap $\boldsymbol{x}^* = \{x_1^*, x_2^*, \ldots, x_n^*\}$:

$$P_j^* = \frac{\#\{x_i^* = x_j\}}{n}, \; j = 1, 2, \ldots, n.$$

Le vecteur de rééchantillonnage

$$\boldsymbol{P}^* = (P_1^*, P_2^*, \ldots, P_n^*)$$

a des composantes non-négatives dont la somme est égale à 1.

Estimation du biais: version améliorée

Une réplication bootstrap $\hat{\theta}^*$ peut être vue comme une fonction du vecteur de rééchantillonnage $\boldsymbol{P}^*$. Pour $\hat{\theta} = t(\hat{F})$, l'estimateur plug-in de θ, nous écrivons

$$\hat{\theta}^* = T(\boldsymbol{P}^*)$$

pour indiquer que $\hat{\theta}^*$ est une fonction du vecteur de rééchantillonnage.

Les m échantillons bootstrap $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_m^*$ donnent lieu aux vecteurs de rééchantillonnage correspondants $P_1^*, P_2^*, \ldots, P_m^*$.

Définissons $\overline{P}^*$ comme la moyenne de ces vecteurs:

$$\overline{\boldsymbol{P}}^* = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{P}_i^*.$$

Estimation du biais: version améliorée

En écrivant

$$\boldsymbol{P}_0 = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right),$$

l'estimateur de biais bootstrap devient

$$\hat{B}_m = \hat{\theta}^* - T(\boldsymbol{P}_0).$$

L'estimateur de bootstrap amélioré est défini comme

$$\overline{B}_m = \hat{\theta}^* - T(\overline{\boldsymbol{P}}^*).$$

$\hat{B}_m$ et $\overline{B}_m$ convergent vers $B_{\hat{F}}$, toutefois il est possible de montrer que la convergence est plus rapide pour $\overline{B}_m$.
Il est toutefois dangereux d'utiliser ces estimations de biais pour corriger l'estimateur $\hat{\theta}$, car ils ajoutent de la variance à ce dernier.