

IFT 3245

Simulation et modèles

Fabian Bastin
DIRO
Université de Montréal

Automne 2012

Estimation séquentielle

Pour un intervalle de confiance de niveau $1 - \alpha$, si on fixe n , la largeur $I_2 - I_1$ est aléatoire.

Si on veut $I_2 - I_1 \leq w$ pour w fixé, la valeur minimale de n requise est une variable aléatoire N .

Comment prédir ce N ?

Pour $X_i \sim \text{binomiale}(1, p)$, avec $n = 1000$ on a obtenu $\bar{X}_n = 0.882$, $S_n^2 \approx 0.1042$, et la demi-largeur du intervalle de confiance à 95% était de 0.020. Combien de répétitions additionnelles faut-il pour réduire la demi-largeur à environ 0.005?

Estimation séquentielle

Nous voulons $1.96S_n/\sqrt{n} \leq 0.005$. En supposant que S_n ne changera pas significativement, cela donne
 $n \geq (1.96 \times S_n/0.005)^2 \approx 16011.8$. En conséquence, nous pouvons recommander de faire 15012 répétitions additionnelles.

Procédure à deux étapes

Cette approche est valable pour la loi de Student (ou normale).

Faire n_0 répétitions et calculer $S_{n_0}^2$; la prédiction du n requis est

$$\hat{N}^* = \min \left\{ n \mid (t_{n-1, 1-\alpha/2}) S_{n_0} / \sqrt{n} \leq r \right\}.$$

On fera $\max(0, \hat{N}^* - n_0)$ répétitions additionnelles.

Bien sûr, il se peut que ce soit insuffisant, ou trop.

Après n_0 , recalculer S_n^2 et la demi-largeur pour chaque n . On s'arrête dès que $I_2 - I_1 \leq w$.

Procédure à deux étapes

Cette procédure est biaisée, car on tend à s'arrêter à un N où S_N^2 sous-estime la variance.

Mais lorsque $w \rightarrow 0$, le bias disparaît, $N/n^* \rightarrow 1$ a.p.1 où n^* est la valeur optimale de N si on connaissait σ^2 , et

$$P[|\bar{X}_N - \mu| \leq w/2] \rightarrow 1 - \alpha.$$

IC pour des estimateurs de quantile

Si X est de répartition F , le q -quantile de F est

$$\xi_q = F^{-1}(q) = \inf\{x : F(x) \geq q\}.$$

Soit $X_{(1)}, \dots, X_{(n)}$ un échantillon i.i.d. de X , trié, et \hat{F}_n la fonction de répartition empirique. Un estimateur simple de ξ_q est le quantile empirique

$$\hat{\xi}_{q,n} = \hat{F}_n^{-1}(q) = \inf\{x : \hat{F}_n(x) \geq q\} = X_{(\lceil nq \rceil)}.$$

Il est biaisé mais fortement consistent et obéit au théorème de la limite centrale, comme le montre le théorème ci-après.

- (i) Pour chaque q , $\hat{\xi}_{q,n}$ to as ξ_q quand $n \rightarrow \infty$.
- (ii) Si X a une densité f strictement positive et continue dans un voisinage de ξ_q , alors

$$\frac{\sqrt{n}(\hat{\xi}_{q,n} - \xi_q)f(\xi_q)}{\sqrt{q(1-q)}} \xrightarrow{D} N(0, 1) \quad \text{quand } n \rightarrow \infty.$$

Ce TLC indique qu'il y a beaucoup de bruit (variance) si $f(\xi_q)$ est petit.

De plus, pour l'utiliser afin de construire un intervalle de confiance, il faut estimer $f(\xi_q)$, ce qui est difficile.

IC pour des estimateurs de quantile

Nous pouvons néanmoins construire une méthode non-asymptotique de calcul d'un intervalle de confiance pour ξ_q : supposons que F est continue en ξ_q .

Soit B le nombre d'observations $X_{(i)}$ inférieures à ξ_q .

Puisque $P[X < \xi_q] = q$, B est binomiale(n, q).

Si $1 \leq j < k \leq n$, $X_{(j)} < \xi_q \leq X_{(k)}$ ssi $j \leq B < k$. Alors

$$P[X_{(j)} < \xi_q \leq X_{(k)}] = P[j \leq B < k] = \sum_{i=j}^{k-1} \binom{n}{i} q^i (1-q)^{n-i}.$$

On choisit j et k pour que cette somme soit supérieure ou égale à $1 - \alpha$ (intervalle unilatéral ou bilatéral).

IC pour des estimateurs de quantile

Si n est grand et q n'est pas trop proche de 0 ou 1, on peut approximer la loi binomiale par la loi normale:

$$\frac{B - nq}{\sqrt{nq(1 - q)}} \approx N(0, 1).$$

On obtient alors $j = \lfloor nq + 1 - \delta \rfloor$ et $k = \lfloor nq + 1 + \delta \rfloor$, où
 $\delta = \sqrt{nq(1 - q)}\Phi^{-1}(1 - \alpha/2)$.

Exemple: valeur à risque

Soit L la perte nette de valeur d'un porte-feuille d'actifs pour une période de temps donnée $[0, T]$. La valeur à risque (VAR) (au temps 0) est la valeur de x_p telle que $P[L > x_p] = p$. C'est le $(1 - p)$ -quantile de L .

Valeurs courantes: $p = 0.01$, $T = 2$ semaines (banques), $T =$ mois ou années (assurance, fonds de pension).

On peut critiquer l'utilisation de la VAR, vu qu'elle donne une information très limitée.

Par exemple si $x_{0.01} = 10^7$ dollars, que sait-on sur l'importance réelle de la perte?

Une mesure complémentaire pourrait être $E[L \mid L > x_p]$, par exemple.

Exemple: valeur à risque

Modèles pour estimer la VAR: on doit modéliser l'évolution du prix des actifs (souvent plusieurs milliers, dépendants).

Souvent: modèles à facteurs.

On peut remplacer les actifs par des prêts, comptes à payer, etc. Sauf dans les cas simples, on estime la VAR par simulation. Quand p est petit: "importance sampling".

Intervalle de confiance par rééchantillonnage (“bootstrap”)

Il s'agit 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. $\mathbf{X} = X_1, \dots, X_n$, issu d'une loi de fonction de répartition F , et un estimateur $Y = g(X_1, \dots, X_n)$ d'une valeur réelle inconnue θ .

Exemple:

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

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 distribution exacte F , nous pouvons toujours nous diriger vers la distribution empirique construite à partir de l'échantillon \mathbf{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

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

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 \mathbf{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 .

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 $P[I_2 < \theta] = \alpha_2$.

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

Une première idée serait de répéter l'expérience m fois afin d'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, \dots, x_n les valeurs de X_1, \dots, X_n et $y = g(x_1, \dots, x_n)$. Tirons X_1^*, \dots, X_n^* au hasard avec remplacement de l'échantillon de départ $\{x_1, \dots, x_n\}$ (i.e., de \hat{F}_n) et calculons $Y^* = g(X_1^*, \dots, X_n^*)$.

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^*, \dots, Y_m^* .

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

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

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

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

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

- ➌ Estimer l'erreur standard $\text{se}_F(\hat{\theta})$ par l'écart-type échantillonnaux des m réplications:

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

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

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

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

L'estimateur de bootstrap idéal $\hat{s}_{\hat{F}}(\hat{\theta}^*)$ et son approximation $\hat{s}_{\hat{e}_m}$ 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

$Y_1^* - y, \dots, Y_m^* - y$. Pour $m \rightarrow \infty$, 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 $S^2 = h^2(X_1, \dots, X_n)$.

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^*, \dots, X_n^* comme avant, puis à calculer $Y^* = g(X_1^*, \dots, X_n^*)$, $S^* = h(X_1^*, \dots, X_n^*)$, et $Z^* = (Y^* - y)/S^*$.

Bootstrap-t non-paramétrique

Soient Z_1^*, \dots, 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 $m = 1000$.

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 $\hat{\theta} = g(\mathbf{X})$ d'une quantité inconnue $\theta = t(F)$ est défini comme

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

L'estimateur bootstrap de biais est défini comme

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

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, $E_{\hat{F}}[g(\mathbf{X}^*)]$ devra être approximé par simulation Monte-Carlo:

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

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 $\mathbf{x}^* = \{x_1^*, x_2^*, \dots, x_n^*\}$:

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

Le vecteur de rééchantillonnage

$$\mathbf{P}^* = (P_1^*, P_2^*, \dots, P_n^*)$$

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 \mathbf{P}^* . Pour $\hat{\theta} = t(\hat{F})$, l'estimateur plug-in de θ , nous écrivons

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

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

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

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

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

Estimation du biais: version améliorée

En écrivant

$$\mathbf{P}_0 = \left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n} \right),$$

l'estimateur de biais bootstrap devient

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

L'estimateur de bootstrap amélioré est défini comme

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

\hat{B}_m et \bar{B}_m convergent vers $B_{\hat{F}}$, toutefois il est possible de montrer que la convergence est plus rapide pour \bar{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.