

IFT 3245

Simulation et modèles

Fabian Bastin
DIRO
Université de Montréal

Automne 2012

Amélioration de l'Efficacité

Rappelons que l'efficacité $\text{Eff}[X]$ d'un estimateur X se définit comme suit:

$$\text{Eff}[X] = \frac{1}{\text{MSE}[X]C(X)}.$$

Est-il possible, étant donné X , de construire un nouvel estimateur Y plus efficace que X ?

Nous allons introduire les principaux concepts d'amélioration de l'efficacité au moyen d'un exemple introductif sur les centres d'appels téléphoniques.

Exemple introductif

Posons B , le facteur d'achalandage pour la journée, et supposons que $P[B = b_t] = q_t$, où

t	1	2	3	4
b_t	0.8	1.0	1.2	1.4
q_t	0.25	0.55	0.15	0.05

Il est facile de vérifier que $E[B] = 1$.

Les arrivées des appels suivent processus de Poisson de taux $B\lambda_j$ durant l'heure j .

Notons $G_i(s)$ = nombre d'appels dont le service a débuté après moins de s secondes d'attente, le jour i .

On veut estimer $\mu = E[G_i(s)]$, disons pour $s = 20$.

Exemple introductif

Nous supposons de plus que les durées de service des appels suivent la loi $\Gamma(\alpha, \gamma)$, dont la moyenne est $\alpha\gamma$.

Dans notre exemple, on a $\alpha = 1$ et $\gamma = \gamma_1 = 100$.

Nombre d'agents n_j et taux d'arrivée λ_j (par heure) pour 13 périodes d'une heure dans le centre d'appel:

j	0	1	2	3	4	5	6	7	8	9	10	11	12
n_j	4	6	8	8	8	7	8	8	6	6	4	4	4
λ_j	100	150	150	180	200	150	150	150	120	100	80	70	60

On simule n jours, indépendamment.

Exemple introductif

Soit $X_i = G_i(s)$ pour le jour i , et

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

On a $E[\bar{X}_n] = \mu$ et $\text{Var}[\bar{X}_n] = \text{Var}[X_i]/n$.

Une expérience avec $n = 1000$ donne $\bar{X}_n = 1518.3$ et $S_n^2 = 21615$. La variance estimée de \bar{X}_n est alors $\widehat{\text{Var}}[\bar{X}_n] = 21.6$.

Exemple introductif

Nous souhaitons construire un intervalle de confiance pour $\sigma^2 = n\text{Var}[\bar{X}_n]$, sous l'hypothèse que $(n-1)S_n^2/\sigma^2$ suit approximativement une χ^2 à $n-1$ degrés de liberté. Ceci permet de construire l'intervalle à 90%: $[0.930S_n^2, 1.077S_n^2]$.

En d'autres termes, l'erreur relative de cet estimateur est inférieure à 8% avec une "confiance" d'environ 90%.

Voyons comment améliorer cet estimateur \bar{X}_n , en réduisant sa variance. Pour chaque méthode proposée, nous donnerons des résultats numériques pour $n = 1000$.

Estimation indirecte.

Soit A_i le nombre total d'arrivées au jour i ; posons $D_i = A_i - X_i$.

On sait que $a = E[A_i] = \sum_{j=1}^m \lambda_j = 1660$.

Nous pouvons dès lors écrire

$\mu = E[X_i] = E[A_i - D_i] = a - E[D_i]$, que l'on peut estimer par

$$\bar{X}_{i,n} = E[A_i] - \bar{D}_n = a - \frac{1}{n} \sum_{i=1}^n D_i.$$

Cet estimateur a moins de variance que \bar{X}_n ssi $\text{Var}[D_i] < \text{Var}[X_i]$. Variance estimée: $\widehat{\text{Var}}[X_{i,i}] = 18389$.

Variable de contrôle (VC)

Idée: exploiter l'information auxiliaire.

Par exemple, si A_i est plus grand que d'habitude ($A_i > E[A_i] = 1660$), on s'attend à ce que ce jour là, X_i et D_i surestiment $E[X_i]$ et $E[D_i]$.

On pourrait faire une "correction" à ces estimateurs: remplacer X_i par

$$X_{c,i} = X_i - \beta(A_i - 1660)$$

où β est une constante appropriée. Alors

$$\bar{X}_{c,n} = \bar{X}_n - \beta(\bar{A}_n - 1660).$$

Variable de contrôle (VC)

On a $E[\bar{X}_{c,n}] = E[X_i]$ et

$$\text{Var}[\bar{X}_{c,n}] = \frac{\text{Var}[X_i] + \beta^2 \text{Var}[A_i] - 2\beta \text{Cov}[A_i, X_i]}{n}.$$

Cette variance est une fonction quadratique en β , que l'on minimise en prenant

$$\beta = \beta^* = \text{Cov}[A_i, X_i] / \text{Var}[A_i].$$

La variance minimale est

$$\text{Var}[\bar{X}_{c,n}] = \frac{\text{Var}[X_i] - (\beta^*)^2 \text{Var}[A_i]}{n} = \text{Var}[\bar{X}_n] (1 - \rho^2[A_i, X_i])$$

où $\rho[A_i, X_i]$ est le coeff. de corrélation entre A_i et X_i .

Variable de contrôle (VC)

On ne connaît pas $\text{Cov}[A_i, X_i]$, mais:

- (a) On peut l'estimer par des expériences pilotes.
- (b) On peut l'estimer par les mêmes n simulations que \bar{X}_n .

Avec (b) on obtient l'estimateur (légèrement biaisé):

$$\bar{X}_{\text{ce},n} = \bar{X}_n - \frac{1}{n-1} \left[\sum_{i=1}^n (A_i - \bar{A}_n)(X_i - \bar{X}_n) \right] \frac{\bar{A}_n - a}{\text{Var}[A_i]}.$$

Conditionnellement à B_i , $A_i \sim \text{Poisson}(1660B_i)$. On a donc

$$\begin{aligned} \text{Var}[A_i] &= \text{Var}[E[A_i|B_i]] + E[\text{Var}[A_i|B_i]] \\ &= \text{Var}[1660B_i] + E[1660B_i] \\ &= 1660^2 \text{Var}[B_i] + 1660E[B_i] \\ &= 67794.4. \end{aligned}$$

Variance empirique obtenue ici: 3310.

Variable de contrôle (VC)

En prenant $\beta = 1$, on retrouve l'estimateur indirect:

$$\bar{X}_{i,n} = \bar{X}_n - (\bar{A}_n - 1660) = 1660 - \bar{D}_n.$$

Si on combine VC + indirect, on obtient:

$$\begin{aligned}\bar{X}_{i,c,n} &= a - \bar{D}_n - \beta_2(\bar{A}_n - a) \\ &= \bar{A}_n - \bar{D}_n - (1 + \beta_2)(\bar{A}_n - a) \\ &= \bar{X}_n - (1 + \beta_2)(\bar{A}_n - a),\end{aligned}$$

i.e., $\bar{X}_{i,c,n}$ est équivalent à $\bar{X}_{c,n}$ avec $\beta = 1 + \beta_2$. Par conséquent, en présence de la variable de contrôle, l'estimation indirecte n'apporte rien.

Nous pourrions aussi considérer d'autres variables de contrôle, comme B_j , la moyenne des durées de service, etc.

Le facteur d'achalandage B_i est une source importante de variabilité importante dans le cas présent. Essayons de la contrôler.

En posant $\mu_t = E[X_i | B_i = b_t]$, on a

$$\begin{aligned}\mu &= E[X_i] = \sum_{t=1}^4 P[B_i = b_t] \cdot E[X_i | B_i = b_t] \\ &= .25 E[X_i | B_i = 0.8] + .55 E[X_i | B_i = 1.0] \\ &\quad + .15 E[X_i | B_i = 1.2] + .05 E[X_i | B_i = 1.4] \\ &= .25 \mu_1 + .55 \mu_2 + .15 \mu_3 + .05 \mu_4.\end{aligned}$$

Idée: estimer μ_t séparément pour chaque t .

Stratification

Supposons qu'il y a N_t jours où $B_i = b_t$ et soient $X_{t,1}, \dots, X_{t,N_t}$ les valeurs de X_i pour ces jours.

On peut estimer $\mu_t = E[X_i | B_i = b_t]$ par

$$\hat{\mu}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} X_{t,i}$$

et μ par

$$\bar{X}_{s,n} = \sum_{t=1}^4 q_t \hat{\mu}_t = .25\hat{\mu}_1 + .55\hat{\mu}_2 + .15\hat{\mu}_3 + .05\hat{\mu}_4.$$

On a

$$\begin{aligned}\text{Var}[\bar{X}_{s,n} \mid N_1, N_2, N_3, N_4] \\ &= \sum_{t=1}^4 q_t^2 \text{Var}[\hat{\mu}_t \mid N_t] = \sum_{t=1}^4 q_t^2 \sigma_t^2 / N_t \\ &= .25^2 \sigma_1^2 / N_1 + .55^2 \sigma_2^2 / N_2 + .15^2 \sigma_3^2 / N_3 + .05^2 \sigma_4^2 / N_4.\end{aligned}$$

où $\sigma_t^2 = \text{Var}[X_i \mid B_i = b_t]$.

La variance est réduite si les σ_t^2 sont inférieurs à $\text{Var}[X_i]$.

Si les B_i sont générés normalement: post-stratification.

Pour estimer μ par stratification, on peut aussi fixer les $N_t = n_t$ à l'avance, c'est-à-dire choisir à l'avance combien de jours on aura $B_i = b_t$ pour chaque valeur de t .

- Allocation proportionnelle: prendre $n_t = nq_t$.
Avec $n = 1000$, cela donne $n_1 = 250$, $n_2 = 550$, $n_3 = 150$, $n_4 = 50$.
- Allocation optimale: choisir les n_t pour minimiser $\text{Var}[\bar{X}_{s,n}]$ sous la contrainte $n_1 + n_2 + n_3 + n_4 = n$. On obtient:

$$\frac{n_t}{n} = \frac{\sigma_t P[B_i = t]}{\sum_{k=1}^4 \sigma_k P[B_i = k]} = \frac{\sigma_k q_k}{\sum_{k=1}^4 \sigma_k q_k}.$$

On ne connaît pas ces σ_k , mais on peut les estimer par des essais pilotes. Avec $n_0 = 800$ essais pilotes, 200 par valeur de t , on obtient par exemple $(n_1, n_2, n_3, n_4) = (219, 512, 182, 87)$ (après arrondi).

Stratification combinée avec VC:

$$\hat{\mu}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} X_{c,t,i} = \frac{1}{n_t} \sum_{i=1}^{n_t} [X_{t,i} - \beta_t(A_{t,i} - a b_t)].$$

On minimise $\sigma_t^2 = \text{Var}[X_{c,t,i}]$ en prenant

$$\beta_t = \beta_t^* = \text{Cov}[A_{t,i}, X_{t,i}] / \text{Var}[A_{t,i}] = \text{Cov}[A_i, X_i | B_i = b_t] / (a b_t).$$

L'ajout d'une VC change les σ_t^2 : l'allocation optimale n'est plus la même.

Avec $n_0 = 800$ essais pilotes, on obtient

$(\beta_1, \beta_2, \beta_3, \beta_4) = (1.020, 0.648, 0.224, -0.202)$ et

$(n_1, n_2, n_3, n_4) = (131, 503, 247, 119)$ comme estimation des valeurs optimales.

Stratégies combinées.

En répétant l'expérience avec $n = 100000$, on peut trouver les estimations suivantes pour la variance ($\pm 1\%$):

$$\text{Var}[X_n] = 21998; \quad \text{Var}[X_{i,n}] = 17996; \quad \text{Var}[X_{c,n}] = 3043;$$

$$\text{Var}[X_{\text{so},c,n}] = 885.$$

Résultats numériques pour $n = 1000$

Méthode	Estimateur	Mean	$S_n^2(\pm 9\%)$	Ratio
Crude estimator	\bar{X}_n	1518.2	21615	1.000
Indirect	$\bar{X}_{i,n}$	1502.5	18389	0.851
CV A_i , with pilot runs	$\bar{X}_{c,n}$	1510.1	3305	0.153
CV A_i , no pilot runs	$\bar{X}_{ce,n}$	1510.2	3310	0.153
Indirect + CV, no pilot runs	$\bar{X}_{i,c,n}$	1510.1	3309	0.153
Stratification (propor.)	$\bar{X}_{sp,n}$	1509.5	1778	0.082
Stratification (optimal)	$\bar{X}_{so,n}$	1509.4	1568	0.073
Strat. (propor.) + CV	$\bar{X}_{sp,c,n}$	1509.2	1140	0.053
Strat. (optimal) + CV	$\bar{X}_{so,c,n}$	1508.3	900	0.042