IFT 3245
Simulation et modèles

Fabian Bastin
DIRO
Université de Montréal

Automne 2012

Estimation et intervalles de confiance

- Déterminer les quantités à mesurer déterminées.
- Construire des estimateurs de celles-ci.
- Mesurer leur précision.

Définition Un *estimateur* $\hat{\mu}$ d'une quantité fixe, mais inconnue, μ, est une variable ou un vecteur aléatoire qui associe aux données une valeur supposée approcher la véritable valeur μ.

Efficacité des estimateurs

Considérons un estimateur X d'une certaine quantité inconnue μ.

Le biais, la variance, l'erreur quadratique moyenne (MSE, pour mean square error), et l'erreur relative (RE, pour relative error) de X sont définis respectivement comme suit:

$$\begin{aligned} \beta &= E[X] - \mu; \\ \sigma^2 &= \mathrm{Var}(X) = E[(X - E[X])^2]; \\ \mathrm{MSE}[X] &= E[(X-\mu)^2] = \beta^2 + \sigma^2; \\ \mathrm{RE}[X] &= \sqrt{\mathrm{MSE}[X]}/|\mu|, \quad \text{pour } \mu \neq 0. \end{aligned}$$

Un estimateur sera dit non-biaisé si $\beta = 0$.

Efficacité des estimateurs

La racine carrée du $MSE[X]$ est appelée l'erreur absolue; c'est une mesure de la précision statistique de l'estimateur X, et $RE[X]$ est une mesure de cette prédiction relativement à l'ordre de grandeur de la moyenne μ.

Supposons de plus que l'effort numérique requis pour calculer X (par exemple en termes de temps CPU) est une variable aléatoire (typiquement corrélée avec X) et dénotons son espérance mathématique par $C(X)$.

L'efficacité de X est

$$\text{Eff}(X) = \frac{1}{C(X) \cdot \text{MSE}(X)}. \quad (1)$$

Un estimateur de X sera dit être plus efficace qu'un autre estimateur Y si $\text{Eff}(X) > \text{Eff}(Y)$.

Amélioration d'efficacité

Améliorer l'efficacité signifie trouver un estimateur Y qui est plus efficace que l'estimateur X actuellement utilisé.

Souvent, les deux estimateurs sont non biaisés, et sont supposés présenter des temps de calcul similaires. Par conséquent, améliorer l'efficacité revient dans ce cas à réduire la variance. Pour cette raison, nous parlerons souvent de techniques de réduction de variance.

Il est toutefois parfois possible d'améliorer l'efficacité en augmentant la variance tout en réduisant le coût de calcul.

Si le temps de calcul n'est pas pris en compte, nous appellerons $\mathrm{Var}[X]/\mathrm{Var}[Y]$ le facteur de réduction de variance de Y par rapport à X. Il représente le facteur par lequel la variance est réduite en utilisant Y au lieu de X.

Intervalle de confiance

Toute mesure de qualité est imparfaite ou incomplète. Ainsi, l'efficacité Eff[X] suppose que le coût de l'erreur est symétrique et proportionnel à son carré.

Un autre aspect important que Eff[X] ne mesure pas est la disponibilité d'une bonne façon d'évaluer l'erreur d'estimation.

Par exemple, si on estime cette erreur par la variance de X, il nous faut un bon estimateur de cette variance.

Intervalle de confiance

L'évaluation de l'erreur d'estimation est habituellement fournie en donnant un intervalle de confiance (IC), défini comme suit.

Intervalle de confiance.
Un intervalle de confiance $[l_1, l2]$ pour une quantité μ est un intervalle défini au moyen de deux variables aléatoires l_1 et l_2 satisfaisant $l_1 \leq l_2$, donnant une certaine probabilité de contenir μ.

$[l_1, l_2]$ est un intervalle de confiance de niveau $1 - \alpha$ (ou à $100(1 - \alpha)\%$) pour μ si $P[l_1 \leq \mu \leq l_2] = 1 - \alpha$.

Intervalle de confiance

Habituellement, on construit un intervalle de confiance pour un niveau visé ou nominal $1 - \alpha$, mais la véritable probabilité de couverture est différente et inconnue. La différence est l'erreur de couverture.

La largeur de l'intervalle est $I_2 - I_1$ (une variable aléatoire). Idéalement, nous voudrions assurer la bonne couverture, tout en conservant $E[I_2 - I_1]$ et $\mathrm{Var}[I_2 - I_1]$ petits.

Par abus de notation, nous dénoterons parfois une suite d'estimateurs $\{Y_n,\ n \geq 1\}$ par Y_n. Deux exemples classiques sont $\overline{X}_n$ et S_n^2.

Intervalle de confiance

Lorsque $n \to \infty$, Y_n est dit asymptotiquement sans biais si $E[Y_n - \mu] \to 0$, consistant si $Y_n \to \mu$ en probabilité, i.e. $P[|Y_n - \mu| > \epsilon] \to 0$ pour tout $\epsilon > 0$, et fortement consistant si $Y_n \to \mu$ avec probabilité 1 (ou presque sûrement).

$\overline{X}_n$ et S_n^2 pour sont ainsi fortement consistants par rapport à $\mu = E[X_i]$.

Un intervalle de confiance $(I_{n,1}, I_{n,2})$ est asymptotiquement valide si son erreur de couverture converge vers 0.

Erreur non prise en compte

Les intervalles de confiance considérés ici prennent en compte l'erreur due aux aléas de la simulation, mais pas l'erreur dans l'estimation des paramètres du modèle.

Supposons par exemple que dans un certain système, les durées de service sont indépendantes et suivent la loi gamma de paramètres (α, β) inconnus. Supposons de plus que nous disposions de 200 observations de durées de service et que l'on estime (α, β) par $(\hat{\alpha}, \hat{\beta})$ à partir de ces 200 observations.

On suppose pour simplifier que l'on identifie la bonne loi.

On utilise ensuite la loi estimée dans un modèle de simulation et on calcule un intervalle de confiance pour une mesure de performance quelconque (par exemple, la durée d'attente moyenne) en faisant *n* répétitions de la simulation avec la loi estimée.

Erreur non prise en compte

Si n tend vers l'infini, la largeur de l'intervalle de confiance tend vers 0, mais l'estimateur converge vers la valeur exacte du modèle avec $(\hat{\alpha}, \hat{\beta})$, qui diffère de celle du modèle avec $(\alpha\, \beta)$.

Il y a donc deux sources d'erreur:

1. l'une due au fait que n est fini,
2. l'autre due à l'erreur dans les paramètres du modèle.

Souvent, il y a plusieurs sources d'erreur de ce second type et elles dominent lorsque n est grand.

Horizon fini

Supposons que nous observons $X_1, \ldots, X_n$, des copies i.i.d. de X obtenues en faisant n répétitions de la simulation, et que nous voulons estimer $\mu = E[X]$.

Nous estimons μ par $\overline{X}_n$ et $\sigma^2 = \text{var}[X]$ par S_n^2.

Théorème
Si $X_1, \ldots, X_n$ sont i.i.d. $N(\mu, \sigma^2)$, alors

(i) $\overline{X}_n$ et S_n^2 sont indépendants;

(ii) $(n-1)S_n^2/\sigma^2 \sim \chi^2(n-1)$;

(iii) $\sqrt{n}(\overline{X}_n - \mu)/S_n \sim$ Student-$t(n-1)$.

Ce théorème permet de calculer un intervalle de confiance pour μ au niveau $1-\alpha$:

$$(\overline{X}_n \pm t_{n-1,1-\alpha/2} S_n/\sqrt{n}),$$

où $P[T_{n-1} \leq t_{n-1,1-\alpha/2}] = 1-\alpha/2$.

Lorsque n est grand, nous pouvons approximer la loi Student-t à $n-1$ degrés de liberté au moyen d'une $N(0,1)$.

Pour obtenir un intervalle de confiance pour σ^2, on choisira x_1 et x_2 tels que

$$P[x_1 < \chi^2_{n-1} < x_2] = 1-\alpha,$$

ce qui permet de poser

$$[l_1, l_2] = [(n-1)S_n^2/x_2, (n-1)S_n^2/x_1].$$

Nous avons alors

$$\begin{aligned} P[I_1 \leq \sigma^2 \leq I_2] &= P[(n-1)S_n^2/x_2 \leq \sigma^2 \leq (n-1)S_n^2/x_1] \\ &= P[x_1 \leq (n-1)S_n^2/\sigma^2 \leq x_2] \\ &= 1-\alpha. \end{aligned}$$

Ceci n'est valide que si les X_i suivent la loi normale.

Le tableau ci-après explicite les bornes $(n-1)/x_1$ et $(n-1)/x_2$ d'un intervalle de confiance sur σ^2/S_n^2. Par exemple, pour $n = 1000$, un intervalle de confiance à 90% pour σ^2 est donné par

$$[0.930\, S_n^2,\ 1.077\, S_n^2]$$

Horizon fini

Bornes $(n-1)/x_1$ et $(n-1)/x_2$ d'un intervalle de confiance sur σ^2/S_n^2:

	$\alpha = 0.02$		$\alpha = 0.10$	
n	$(n-1)/x_1$	$(n-1)/x_2$	$(n-1)/x_1$	$(n-1)/x_2$
10	0.388	3.518	0.492	2.284
30	0.570	1.939	0.663	1.568
100	0.729	1.413	0.796	1.270
300	0.831	1.216	0.876	1.146
1000	0.902	1.111	0.930	1.077

Si on a deux échantillons indépendants, $X_1, \ldots, X_m$ i.i.d. normales de variance σ_x^2 et $Y_1, \ldots, Y_n$ i.i.d. normales de variance σ_y^2, on peut calculer un intervalle de confiance sur le rapport des deux variances, en utilisant le fait que

$$F = \frac{S_{x,m}^2/\sigma_x^2}{S_{y,n}^2/\sigma_y^2} = \frac{S_{x,m}^2\sigma_y^2}{S_{y,n}^2\sigma_x^2} \sim F(m-1, n-1),$$

où $S_{x,m}^2$ et $S_{y,n}^2$ sont les variances échantillonnales.

Si $P[x_1 < F < x_2] = 1 - \alpha$, l'intervalle est

$$[I_1, I_2] = \left[\frac{1}{x_2}\frac{S_{x,m}^2}{S_{y,n}^2}, \frac{1}{x_1}\frac{S_{x,m}^2}{S_{y,n}^2}\right].$$

Ce type d'intervalles est potentiellement utile lorsqu'on estime le facteur de réduction de variance entre deux estimateurs.

Approximation normale

Lorsque n est grand, $\overline{X}_n$ est approximativement normale même si X ne l'est pas, en vertu du théorème de la limite centrale (TLC).

Il existe plusieurs versions du TLC: X_i de lois différentes, dépendance, TLCs multivariés, TLC fonctionnels, etc. Nous citerons le résultat suivant.

Approximation normale

Théorème.
Soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes, avec $E[X_i] = \mu_i$ et $\mathrm{Var}[X_i] = \sigma_i^2$. Posons $s_n^2 = \sigma_1^2 + \cdots + \sigma_n^2$,

$$Y_n = \frac{(X_1 - \mu_1) + \cdots + (X_n - \mu_n)}{s_n},$$

et $F_n(x) = P[Y_n \leq x]$. Alors, $E[Y_n] = 0$, $\mathrm{Var}[Y_n] = 1$, et

$$\sup_{n \geq 1,\, x \in \mathcal{R}} |F_n(x) - \Phi(x)| \leq \kappa \, \frac{E(|X_1 - \mu_1|^3) + \cdots + E(|X_n - \mu_n|^3)}{s_n^3}$$

où $\kappa = 3$ si les X_i sont i.i.d. et $\kappa = 6$ sinon.

Approximation normale

La borne sur l'erreur dépend donc de l'assymétrie des distributions. Sous l'hypothèse où n est suffisamment grand que pour pouvoir approximer la distribution de $\overline{X}_n$, nous pourrons choisir pour un intervalle de confiance

$$\left[\overline{X}_n - z_{1-\alpha/2} S_n/\sqrt{n}, \overline{X}_n + z_{1-\alpha/2} S_n/\sqrt{n}\right],$$

où $z_{1-\alpha/2}$ est le quantile $1-\alpha/2$ d'une normale $N(0,1)$.

Nous pouvons raisonablement recourir au TLC pour calculer un intervalle de confiance, sauf si un des cas suivant se présente:

- n est trop petit,
- α est proche de 0,
- les X_i ont une loi très asymétrique,
- il existe des moments supérieurs très élevés.

Exemple: binomiale

Supposons $n = 1000$, $X_i \sim \underline{\text{Binomiale}}(1, p)$; on veut estimer p. Si on a 882 succès, $\overline{X}_n = 0.882$.

On a alors $S_n^2 = \overline{X}_n(1 - \overline{X}_n)n/(n-1) \approx 0.1042$ et un intervalle de confiance à 95% (approximativement) est $(\overline{X}_n \pm 1.96 S_n/\sqrt{n}) \approx (0.862, 0.902)$.

L'intervalle de confiance ainsi construit nous donne aussi une idée des chiffres significatifs de l'estimateur.

Mais si $\overline{X}_n = 0.998$, alors on voit que p est trop proche de 1 et l'approx. normale sera très mauvaise. Dans ce cas, on va plutôt utiliser: $Y = \sum_{i=1}^{n}(1 - X_i) \approx \underline{\text{Poisson}(n(1-p))}$.

Exemple: durée de vie d'un système

Soit $X = \min(G_1, max(G_2, G_3))$. Les G_j sont i.i.d. Weibull ($\alpha = 0.5$, $\beta = 1$). On simule n fois, avec X_i la valeur de X pour la répétition i. On calcule un intervalle de confiance à 90% pour $E[X]$ via le théorème de la limite centrale.

n	Prob. couverture	Estim. $E[I_2 - I_1]/\mu$
5	0.708 ± 0.03	1.16
10	0.750 ± 0.03	0.82
20	0.800 ± 0.03	0.60
40	0.840 ± 0.03	0.44

Il y a dégradation significative de la couverture. Les G_j (et les X_i) suivent en effet une loi très éloignée de la normale, et on se trompe beaucoup en calculant un intervalle de confiance basé sur la loi normale.

Intervalle de confiance pour une loi discrète

Soit Y une variable aléatoire prenant ses valeurs dans $\{0, 1, 2, \dots\}$ et suivant une loi de paramètre μ, telle que $P_\mu[Y \geq y]$ est croissant en μ, où P_μ dénote la probabilité quand la valeur du paramètre est μ. (Le cas décroissant se traite de manière symétrique.)

Des exemples de telles distribution comptent les lois binomiale, géométrique, de Poisson,...

On veut un intervalle de confiance $[I_1, I_2]$ de niveau (approximatif) $1 - \alpha$ pour μ. Posons $\alpha = \alpha_1 + \alpha_2$, avec $\alpha_1 > 0$ et $\alpha_2 > 0$.

Nous voudrions $P[\mu < I_1] \approx \alpha_1$ et $P[\mu > I_2] \approx \alpha_2$. Si on observe $Y = y$, l'intervalle sera $[I_1(y), I_2(y)]$.

Intervalle de confiance pour une loi discrète

Algorithme: Prendre pour $I_1(y)$ et $I_2(y)$ les solutions de

$$\alpha_1 = P_{I_1}[Y \geq y] \quad \text{et} \quad \alpha_2 = P_{I_2}[Y \leq y]. \tag{2}$$

Ceci revient à fixer la probabilité que la variable Y soit supérieure (respectivement inférieure) à l'observation y, si le paramètre inconnu était de valeur I_1 (respectivement I_2).

Dans chacune de ces deux configurations, on s'attend à ce qu'un événement ait une faible probabilité, vu qu'une faible (forte) valeur de μ défavorise l'évenement considéré, et ce en raison de la monotonie de $P_\mu[Y \geq y]$ par rapport à μ.

Intervalle de confiance pour une loi discrète

Nous pouvons résoudre par recherche binaire, par exemple.

Pour le cas où Y est décroissant avec μ, il suffit de permuter les signes $\geq$ et $\leq$.
La probabilité de couverture de cet intervalle est d'au moins $1 - \alpha$.

Proof.

Soit $y^*(\mu) = \min\{y \in \mathcal{N} : I_1(y) \geq \mu\}$ et $\nu = I_1(y^*(\mu)) \geq \mu$. Par conséquent, en vertu de la croissante de P_μ avec μ,

$$P_\mu[I_1(Y) \geq \mu] \leq P_\nu[I_1(Y) \geq \mu] = P_\nu[Y \geq y^*(\mu)] = \alpha_1.$$

On montre de même que $P_\mu[I_2(Y) \leq \mu] \leq \alpha_2$.
Nous avons dès lors

$$\begin{aligned} P_\mu[I_1(Y) \leq \mu \leq I_2(Y)] &= 1 - P[\mu < I_1(Y) \cup I_2(Y) < \mu] \\ &= 1 - P_\mu[\mu < I_1(Y)] - P_\mu[I_2(Y) < \mu] \\ &\geq 1 - \alpha_1 - \alpha_2 = 1 - \alpha. \end{aligned}$$

Intervalle de confiance pour une loi discrète

La probabilité de couverture exacte dépend de F_μ et est généralement inconnue.

Reprenons l'exemple de la binomiale.

Supposons que $X_1, \ldots, X_n$ sont i.i.d. avec $P[X_i = 1] = 1 - P[X_i = 0] = p$, de sorte que $Y = n\overline{X_n} = \sum_{i=1}^{n} X_i$ suit une binomiale(n, p).

Intervalle de confiance sur p basé sur l'observation de Y?

Intervalle de confiance pour une loi discrète

Pour n'importe quelles valeurs de p et de y, les probabilités dans l'intervalle de confiance peuvent être calculées en sommant les probabilités binomiales exactes si y est petit.

Si n est grand et p est petit, Y suit approximativement une variable aléatoire de Poisson de moyenne np, aussi peut-on approximer les probabilités dans en additionnant les probabilités de Poisson appropriées.

Pour p proche de 1, nous pouvons simplement remplacer p et X_i par $1 - p$ et $1 - X_i$.

Estimation séquentielle

Pour un intervalle de confiance de niveau $1 - \alpha$, si on fixe n, la largeur $l_2 - l_1$ est aléatoire.

Si on veut $l_2 - l_1 \leq w$ pour w fixé, la valeur minimale de n requise est une variable aléatoire N.

Comment prédire ce N?

Pour $X_i \sim$ binomiale$(1, p)$, avec $n = 1000$ on a obtenu $\overline{X}_n = 0.882$, $S_n^2 \approx 0.1042$, 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 $1.96 S_n/\sqrt{n} \leq 0.005$. 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 $l_2 - l_1 \leq w$.

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 $w \to 0$, le bias disparaît, $N/n^* \to 1$ a.p.1 où n^* est la valeur optimale de N si on connaissait σ^2, et $P[|\overline{X}_N - \mu| \leq w/2] \to 1 - \alpha$.

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 $X_{(1)}, \ldots, X_{(n)}$ 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.

IC pour des estimateurs de quantile: théorème

(i) *Pour chaque q, $\hat{\xi}_{q,n}$ to as ξ_q quand $n \to \infty$.*

(ii) *Si X a une densité f strictement positive et continue dans un voisinage de ξ_q, alors*

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

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".