IFT 3245
Simulation et modèles

Fabian Bastin
DIRO
Université de Montréal

Automne 2012

Intervalle de confiance pour une fonction de plusieurs moyennes

Dans le cas déterministe, nous savons que si $\boldsymbol{Y}_n = (Y_{1n}, \ldots, Y_{dn})$ converge vers un certain vecteur $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ et si $g : \mathcal{R}^d \to \mathcal{R}$ est continue, alors $g(\boldsymbol{Y}_n) \to g(\boldsymbol{\mu})$.

Supposons à présent que les $\boldsymbol{Y}_n$ sont des vecteurs aléatoires et que $r(n)(\boldsymbol{Y}_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{Y}$.

Par exemple, si $\boldsymbol{Y}_n$ est une moyenne de n vecteurs, nous savons que $\sqrt{n}(\boldsymbol{Y}_n - \boldsymbol{\mu}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_y)$.

Avons-nous encore la convergence de $r(n)(g(\boldsymbol{Y}_n) - g(\boldsymbol{\mu}))$, et le cas échéant, vers quelle distribution?

Théorème Delta

Theorème (Théorème Delta)

Soit $g : \mathcal{R}^d \to \mathcal{R}$ *continûment différentiable dans un voisinage de* $\boldsymbol{\mu}$*, et* ∇g *son gradient. Si* $r(n)(\boldsymbol{Y}_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{Y}$ *quand* $n \to \infty$*, alors*

$$r(n)(g(\boldsymbol{Y}_n) - g(\boldsymbol{\mu})) \xrightarrow{D} (\nabla g(\boldsymbol{\mu}))^T \boldsymbol{Y} \quad \text{quand } n \to \infty.$$

Corollaire (Corollaire)

Si $\sqrt{n}(\boldsymbol{Y}_n - \boldsymbol{\mu}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_y)$ *quand* $n \to \infty$*, alors on a le TLC:*

$$\sqrt{n}(g(\boldsymbol{Y}_n) - g(\boldsymbol{\mu}))/\sigma_g \xrightarrow{D} N(0, 1) \quad \text{quand } n \to \infty,$$

où ${\sigma_g}^2 = (\nabla g(\boldsymbol{\mu}))^T \boldsymbol{\Sigma}_y \nabla g(\boldsymbol{\mu})$.

Quotient de deux espérances

Soient $(X_1, Y_1), \ldots, (X_n, Y_n)$ des copies i.i.d. de (X, Y) et supposons que l'on estime $\nu = E[X]/E[Y]$ par

$$\hat{\nu}_n = \frac{\overline{X}_n}{\overline{Y}_n} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n Y_i}.$$

Cet estimateur est biaisé mais fortement consistant.

Posons $\mu_1 = E[X]$, $\mu_2 = E[Y]$, $g(\mu_1, \mu_2) = \mu_1/\mu_2$, $\sigma_1^2 = \mathrm{Var}[X]$, $\sigma_2^2 = \mathrm{Var}[Y]$, et $\sigma_{12} = \mathrm{Cov}[X, Y]$. Supposons que ces quantités soient finies et que $\mu_2 \neq 0$, $\sigma_1^2 > 0$, et $\sigma_2^2 > 0$.

Le gradient de g est

$$\nabla g(\mu_1, \mu_2) = (1/\mu_2, -\mu_1/\mu_2^2)^T.$$

Quotient de deux espérances

En vertu du théorème de la limite centrale,

$$\sqrt{n}(\overline{X}_n - \mu_1, \overline{Y}_n - \mu_2)^T \xrightarrow{D} (W_1, W_2)^T \sim N(\mathbf{0}, \mathbf{\Sigma})$$

où

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

Puis, par le théorème delta (ou son corollaire), nous avons

$$\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{D} (W_1, W_2) \cdot \nabla g(\mu_1, \mu_2) = W_1/\mu_2 - W_2\mu_1/\mu_2^2 \sim N(0, \sigma_g^2)$$

où

$$\begin{aligned} \sigma_g^2 &= (\nabla g(\boldsymbol{\mu}))^T \mathbf{\Sigma}_y \nabla g(\boldsymbol{\mu}) \\ &= \sigma_1^2/\mu_2^2 + \sigma_2^2\mu_1^2/\mu_2^4 - 2\sigma_{12}\mu_1/\mu_2^3, \end{aligned}$$

ou encore

$$\sigma_g^2 = \frac{\sigma_1^2 + \sigma_2^2\nu^2 - 2\sigma_{12}\nu}{\mu_2^2}.$$

Quotient de deux espérances

Nous pouvons calculer un intervalle de confiance en utilisant ce dernier théorème de la limite centrale si nous disposons d'un bon estimateur de σ_g^2. Un candidat évident est:

$$\hat{\sigma}_{g,n}^2 = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 \hat{\nu}_n^2 - 2\hat{\sigma}_{12}\hat{\nu}_n}{\overline{Y}_n^2},$$

où

$$\begin{aligned}
\hat{\sigma}_1^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2, \\
\hat{\sigma}_2^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y}_n)^2, \\
\hat{\sigma}_{12} &= \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)(Y_j - \overline{Y}_n).
\end{aligned}$$

Quotient de deux espérances

Puisque $\hat{\sigma}^2_{g,n}$ est fortement consistant, on obtient le théorème de la limite centrale

$$\frac{\sqrt{n}(\hat{\nu}_n - \nu)}{\hat{\sigma}_{g,n}} \xrightarrow{D} \frac{\sqrt{n}(\hat{\nu}_n - \nu)}{\sigma_g} \xrightarrow{D} N(0,1) \qquad \text{quand } n \to \infty.$$

L'intervalle de confiance <u>classique</u> pour ν au niveau nominal $1-\alpha$ est $(\hat{\nu}_n - r,\ \hat{\nu}_n + r)$ où $r = z_{1-\alpha/2}\hat{\sigma}_{g,n}/\sqrt{n}$.

Son erreur de couverture est parfois grande lorsque n n'est pas très grand, ou lorsque la convergence vers $N(0,1)$ est lente.

Dans ce cas, on recommande d'utiliser le <u>bootstrap-*t* non-paramétrique</u>, en prenant $\hat{\nu}_n$ et $\hat{\sigma}_{g,n}$ comme estimateurs de la moyenne et de la variance.

Quotient de deux espérances

Pour le cas particulier d'un rapport de deux espérances, la dérivation suivante est plus directe.

Les variables aléatoires

$$Z_j = X_j - \nu Y_j,$$

sont i.i.d. de moyenne 0 et de variance

$$\begin{aligned}\sigma_z^2 = \mathrm{Var}[Z_j] &= \mathrm{Var}[X_j] + \nu^2 \mathrm{Var}[Y_j] - 2\nu \mathrm{Cov}(X_j, Y_j)\\ &= \sigma_1^2 + \sigma_2^2 \nu^2 - 2\sigma_{12}\nu.\end{aligned}$$

En appliquant le TLC aux Z_j, on obtient

$$\frac{\sqrt{n}\bar{Y}_n(\hat{\nu}_n - \nu)}{\sigma_z} = \frac{\sqrt{n}\bar{Z}_n}{\sigma_z} \xrightarrow{D} N(0,1) \qquad \text{quand } n \to \infty.$$

Quotient de deux espérances

C'est équivalent, car $\sigma_z/\overline{Y}_n$ to as $\sigma_z/\mu_2 = \sigma_g$ quand $n \to \infty$.

Remarque importante: on préfère $\text{Cov}(X_j, Y_j) > 0$!

Différence entre deux moyennes

On a n_1 observations i.i.d. $X_{11}, \ldots, X_{1,n_1}$, de moyenne μ_1, et n_2 observations i.i.d. $X_{21}, \ldots, X_{2,n_2}$, de moyenne μ_2. On veut un intervalle de confiance pour $\mu_1 - \mu_2$.

Les deux méthodes suivantes supposent que les X_{ji} suivent la loi normale.
(Pas toujours valide!)

Dans la seconde (Welch), les deux échantillons doivent être indépendants mais on peut avoir $n_1 \neq n_2$. Dans la première, il faut $n_1 = n_2$ mais X_{1i} et X_{2i} peuvent être corrélés.

Soit $n_1 = n_2 = n$.

Posons $Z_i = X_{1i} - X_{2i}$ pour $1 \leq i \leq n$,

$$\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i, \qquad \text{et} \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2.$$

Puisque les Z_i sont i.i.d. normales,

$$\sqrt{n}[\overline{Z}_n - (\mu_1 - \mu_2)]/S_n \sim \underline{\text{Student}}(n-1).$$

On utilise cela pour calculer l'intervalle de confiance. Puisque

$$\text{Var}[Z_i] = \text{Var}[X_{1i}] + \text{Var}[X_{2i}] - 2\text{Cov}[X_{1i}, X_{2i}],$$

il est avantageux d'avoir $\underline{\text{Cov}[X_{1i}, X_{2i}] > 0}$.

Première approche: méthode de Welch

On suppose que X_{1i} et X_{2i} sont indépendants. Soit

$$\overline{X}_{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ki} \qquad \text{et} \qquad S^2_{(k)} = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(X_{ki} - \overline{X}_{(k)})^2,$$

pour $k = 1, 2$.

Alors,

$$\frac{\overline{X}_{(1)} - \overline{X}_{(2)} - (\mu_1 - \mu_2)}{[S^2_{(1)}/n_1 + S^2_{(2)}/n_2]^{1/2}} \approx \underline{\text{Student}}(\hat{\ell})$$

où

$$\hat{\ell} = \frac{[S^2_{(1)}/n_1 + S^2_{(2)}/n_2]^2}{[S^2_{(1)}/n_1]^2/(n_1 - 1) + [S^2_{(2)}/n_2]^2/(n_2 - 1)}.$$