

# IFT 3245

## Simulation et modèles

Fabian Bastin  
DIRO  
Université de Montréal

Automne 2016

# Intervalle de confiance pour une fonction de plusieurs moyennes

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

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

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

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

# Théorème Delta

## Theorème (Théorème Delta)

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

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

## Corollaire (Corollaire)

Si  $\sqrt{n}(\mathbf{Y}_n - \mu) \xrightarrow{D} N(\mathbf{0}, \Sigma_y)$  quand  $n \rightarrow \infty$ , alors on a le TLC:

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

où  $\sigma_g^2 = (\nabla g(\mu))^T \Sigma_y \nabla g(\mu)$ .

# Quotient de deux espérances

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

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

Cet estimateur est biaisé mais fortement consistant.

Posons  $\mu_1 = E[X]$ ,  $\mu_2 = E[Y]$ ,  $g(\mu_1, \mu_2) = \mu_1/\mu_2$ ,  $\sigma_1^2 = \text{Var}[X]$ ,  $\sigma_2^2 = \text{Var}[Y]$ , et  $\sigma_{12} = \text{Cov}[X, Y]$ . 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,

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

où

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

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

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

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  $\sigma_g^2$ . Un candidat évident est:

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

où

$$\hat{\sigma}_1^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X}_n)^2,$$

$$\hat{\sigma}_2^2 = \frac{1}{n-1} \sum_{j=1}^n (Y_j - \bar{Y}_n)^2,$$

$$\hat{\sigma}_{12} = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X}_n)(Y_j - \bar{Y}_n).$$

# Quotient de deux espérances

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

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

L'intervalle de confiance classique pour  $\nu$  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 bootstrap-t non-paramétrique, 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

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

En appliquant le TLC aux  $Z_j$ , on obtient

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

# Quotient de deux espérances

C'est équivalent, car  $\sigma_z/\bar{Y}_n$  converge vers  $\sigma_z/\mu_2 = \sigma_g$  quand  $n \rightarrow \infty$ , presque sûrement.

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

# Différence entre deux moyennes

On a  $n_1$  observations i.i.d.  $X_{11}, \dots, X_{1,n_1}$ , de moyenne  $\mu_1$ , et  $n_2$  observations i.i.d.  $X_{21}, \dots, X_{2,n_2}$ , de moyenne  $\mu_2$ . 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.

# Première approche: observations couplées

Soit  $n_1 = n_2 = n$ .

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

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

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

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

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  $\underline{\text{Cov}[X_{1i}, X_{2i}] > 0}$ .



## Deuxième approche: méthode de Welch

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

$$\bar{X}_{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ki} \quad \text{et} \quad S_{(k)}^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (X_{ki} - \bar{X}_{(k)})^2,$$

pour  $k = 1, 2$ .

Alors,

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

où

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