Section 5: Méthodes de Monte Carlo

Pierre Duchesne

November 29, 2018

Introduction

- On cherche à apporter une solution statistique, typiquement computationnelle, reposant sur des simulations, à un problème complexe.
- Les résultats théoriques reposent sur des hypothèses. Avec des simulations (reposant sur des répétitions d'expériences dans les mêmes conditions), il est possible d'effectuer des analyses de sensibilités, afin de vérifier l'importance des hypothèses.

Exemple 1: Théorème limite central

- Le théorème limite central (TLC) stipule que la moyenne échantillonnale est approximativement de loi normale sous certaines conditions.
- Les conditions précises sont: $X_1, \ldots, X_n$, variables aléatoires indépendantes et identiquement distribuées (iid), de même moyenne μ et de variance finie σ^2.
- On a alors:

$$n^{1/2}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0,1),$$

avec $\bar{X}_n = n^{-1}\sum_{t=1}^{n} X_t$.

Exemple 2: Étude du biais exact d'un estimateur

- Comparaison d'estimateurs en échantillonnage.
- Selon la technique de linéarisation, certaines statistiques non-linéaires sont approximativement sans biais. On pourrait vérifier cette affirmation.
- Il faudra choisir une populations U de taille N jugée réaliste, choisir certains plans de sondages, pour certaines tailles d'échantillons.
- Exemple de statistique: $\hat{R} = \hat{t}_{y\pi}/\hat{t}_{x\pi}$, $\hat{t}_{y\pi} = \sum_s y_k/\pi_k$, $\hat{t}_{x\pi} = \sum_s x_k/\pi_k$.

Exemple 3: Variance approximative

- Lors de l'étude de l'estimateur par la régression, ou encore lors de l'étude de la technique de linéarisation, on a considéré les variances approximatives, notées AV.
- La variance approximative (calculée sur U) devrait s'approcher de la vraie variance de l'estimateur (calculée sur U).
- La variance (approximative) de l'estimateur par régression, pour le tirage SI, a été montrée égale à:

$$AV(\hat{t}_{y,reg}) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{EU},$$

$$\hat{t}_{y,reg} = N\{\bar{y}_s + (\bar{x}_U - \bar{x}_s)\hat{B}_s\},\ E_k = (y_k - \bar{y}_U) - B(x_k - \bar{x}_U).$$

Exemple 4: Estimation de la variance

- Lors de l'étude de la technique de linéarisation, de l'étude du ratio ou des estimateurs par régression, on a proposé des estimateurs de variance de la variance approximative.
- L'estimateur de la variance approximative (calculée sur s) devrait être sans biais pour la vraie variance de l'estimateur (calculée sur U).
- L'estimateur de la variance approximative (calculée sur s) devrait être sans biais pour la variance approximative de l'estimateur (calculée sur U).
- L'estimateur de variance de l'estimateur par régression, pour le tirage SI, proposé était:

$$\hat{V}(\hat{t}_{y,reg}) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{es},$$

$$\hat{t}_{y,reg} = N\{\bar{y}_s + (\bar{x}_U - \bar{x}_s)\hat{B}_s\},\ e_k = (y_k - \bar{y}_s) - \hat{B}(x_k - \bar{x}_s).$$

Exemple: test de l'hypothèse $H_0 : \mu = \mu_0$

- Soient $X_1, \ldots, X_n$ iid de loi $\mathcal{L}(\mu; \sigma^2)$, avec $E(X) = \mu$, $\text{var}(X) = \sigma^2$.
- On veut tester:

$$\begin{aligned} H_0 &: \mu = \mu_0, \\ H_1 &: \mu > \mu_0. \end{aligned}$$

- Pour faire le test, on utilise le t-test, qui est:

$$t = n^{1/2} \left(\frac{\bar{x}_n - \mu_0}{s} \right),$$

$\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$, $s^2 = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$.

Exemple: test de l'hypothèse $H_0 : \mu = \mu_0$ (suite)

- La règle de décision du t-test dans le cas unilatéral est de rejeter H_0 si $t > t_{n-1;1-\alpha}$.
- On sait que si $\mathcal{L}(\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2)$, alors le niveau du test est α.
- Si $\mathcal{L}(\mu, \sigma^2) \neq \mathcal{N}(\mu, \sigma^2)$, mais que n est assez grand, alors le niveau du test est approximativement α.

Étude du niveau du test

- On pourrait considérer l'étude du niveau du test de $H_0 : \mu = \mu_0 = 1$, dans deux cas.
- Situation 1: $X_1, \ldots, X_n$ iid $\mathcal{N}(1, 1)$.
- Situation 2: $X_1, \ldots, X_n$ iid $\mathcal{U}(1 - \sqrt{3}, 1 + \sqrt{3})$, avec $\mathcal{U}(a, b)$ la loi uniforme sur $[a, b]$.
- Dans les deux cas H_0 est vraie.

Étude du niveau du test (suite)

On procède comme suit. On considère la situation 1, ou 2. On fixe donc $\mathcal{L}$. On fixe également la taille de l'échantillon n. On choisit un niveau α.

- On choisit un nombre de répliques, B, typiquement grand. Par exemple, $B = 10^6$.
- Pour chaque réplique k, on génère un échantillon de taille n selon $\mathcal{L}$.
- À la réplique k, on fait le t-test. On regarde si on rejette H_0 ou si on l'accepte (faute de mieux).
- On calcule le pourcentage de rejets.

Étude du niveau du test (suite)

- Plus précisément, pour le tirage k, on peut associer une variable aléatoire de Bernoulli:

$$I_k = \begin{cases} 1 & \text{rejette le test de } H_0 : \mu = 1, \\ 0 & \text{ne rejette pas le test de } H_0 : \mu = 1. \end{cases}$$

- La variable I_k est une variable indicatrice.
- On peut définir:

$$\hat{\alpha} = \frac{1}{B} \sum_{k=1}^{B} I_k.$$

- La variable $\hat{\alpha}$ mesure le pourcentage de rejets.
- Les variables aléatoires I_k sont Bernoulli(α) sous l'hypothèse:

$$H_0^{MC} : \text{ le niveau du test est } \alpha$$

Tester le niveau du test

- On peut tester le niveau du test et vérifier qu'il est bien égal à α.
- On peut tester on niveau $\alpha^{(MC)} = 5\%$, disons, pour les différents tests (de niveaux α).
- Sous H_0^{MC}:

$$\text{var}(\hat{\alpha}) = \frac{1}{B^2} \sum_{k=1}^{B} \text{var}(I_k) = \frac{\alpha(1-\alpha)}{B}.$$

- On rejette H_0^{MC} lorsque:

$$\left| \frac{\hat{\alpha} - \alpha}{\sqrt{(\alpha(1-\alpha))/B}} \right| > 1.96.$$

Complémentaire de la zone de rejet

- Il est souvent commode de considérer des limites de significations (attention ce ne sont pas des intervalles de confiance):

$$\left[\alpha - 1.96\sqrt{\frac{\alpha(1-\alpha)}{B}},\ \alpha + 1.96\sqrt{\frac{\alpha(1-\alpha)}{B}}\right]$$

- On a le tableau suivant, si $B = 10000$, $\alpha^{MC} = 5\%$:

$$\begin{aligned} \alpha = 1\%: &\quad [0.00805, 0.01195], \\ \alpha = 5\%: &\quad [0.04573, 0.05427], \\ \alpha = 10\%: &\quad [0.09412, 0.10588], \end{aligned}$$

Exemple: Comparaison empirique entre deux estimateurs

- Supposons que l'on désire comparer empiriquement:

$$\begin{aligned} \bar{y}_s &= n^{-1} \sum_s y_k, \\ \hat{\bar{y}}_{U,reg} &= \bar{y}_s + \hat{B}_s(\bar{x}_U - \bar{x}_s) \end{aligned}$$

- On pourrait vouloir calculer le biais exact de ces estimateurs pour n donné, dans une population U déterminée, sous le plan SI.

Que savons-nous sur $\bar{y}_s$?

- Il est *exactement* sans biais:

$$E(\bar{y}_s) = \bar{y}_U;$$

- La variance *théorique* est:

$$\text{var}(\bar{y}_s) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}^2;$$

- Un estimateur de variance est:

$$\hat{\text{var}}(\bar{y}_s) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{ys}^2;$$

- L'intervalle de confiance de niveau 95% est:

$$\bar{y}_s \pm 1.96\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_{ys}^2}.$$

Que savons-nous sur $\hat{\bar{y}}_{U,reg}$?

- Il est *approximativement* sans biais: $E(\hat{\bar{y}}_{U,reg}) \approx \bar{y}_U$;
- La variance *approximative* est:

$$AV(\hat{\bar{y}}_{U,reg}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{EU},$$

$E_k = (y_k - \bar{y}_U) = B(x_k - \bar{x}_U)$.

- Un estimateur de variance est:

$$\hat{\text{var}}(\hat{\bar{y}}_{U,reg}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{es},$$

$e_k = (y_k - \bar{y}_s) - \hat{B}_s(x_k - \bar{x}_s)$.

- L'intervalle de confiance de niveau 95% est

$$\hat{\bar{y}}_{U,reg} \pm 1.96\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S^2_{es}}.$$

Comparaison entre $\bar{y}_s$ et $\hat{\bar{y}}_{U,reg}$

- On se rappelle que si n est grand, il est attendu que:

$$AV(\hat{\bar{y}}_{U,reg}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}^2 \left(1 - r_{xyU}^2\right) \leq \left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}^2;$$

- C'est un résultat asymptotique, puisque le résultat repose sur la comparaison entre une variance approximative (asymptotique) et une variance théorique.

Retour vers les concepts fondamentaux

- Comme vu au début du cours, on dispose d'une population finie de taille N et il existe 2^N échantillons.
- Idéalement, si on peut lister tous les échantillons possibles $s \in \mathcal{S}$, alors pour $\hat{\theta}(s)$, estimateur de θ, on a
- L'espérance: $E\{\hat{\theta}(s)\} = \sum_{s\in\mathcal{S}} \hat{\theta}(s)p(s)$;
- La variance:

$$\text{var}\{\hat{\theta}(s)\} = \sum_{s\in\mathcal{S}} \left\{\hat{\theta}(s) - E\{\hat{\theta}(s)\}\right\}^2 p(s)$$

- Le biais exact:

$$B\{\hat{\theta}(s)\} = E\{\hat{\theta}(s)\} - \theta.$$

Retour vers les concepts fondamentaux (suite)

- Le niveau de confiance d'un intervalle de confiance $IC(s) = [L(s), U(s)]$:

$$\text{Niveau de confiance: } = \sum_{s \in \mathcal{S}} u(s)p(s),$$

avec:

$$u(s) = \begin{cases} 1, & \text{si } IC(s) \ni \theta, \\ 0, & \text{sinon.} \end{cases}$$

- Avec la méthode Monte Carlo, on se contente de sélectionner B échantillons indépendamment dans U selon le plan choisit p.

Retour vers les concepts fondamentaux (suite)

On disposera alors de $\hat{\theta}^{(i)}(s)$, $i = 1, \ldots, B$.

- Espérance Monte Carlo:

$$E_{MC}\left\{\hat{\theta}\right\} = B^{-1}\sum_{i=1}^{B}\hat{\theta}^{(i)}(s);$$

- Biais Monte Carlo:

$$B_{MC}\left\{\hat{\theta}\right\} = E_{MC}\left\{\hat{\theta}\right\} - \theta;$$

- Variance Monte Carlo:

$$V_{MC}\left\{\hat{\theta}\right\} = B^{-1}\sum_{i=1}^{B}\left[\hat{\theta}^{(i)}(s) - E_{MC}\left\{\hat{\theta}\right\}\right]^2;$$

Retour vers les concepts fondamentaux (suite et fin)

- Erreur Quadratique Moyenne Monte Carlo:

$$EQM_{MC}\left\{\hat{\theta}\right\} = B^{-1}\sum_{i=1}^{B}\left[\hat{\theta}^{(i)}(s) - \theta\right]^2;$$

- Niveau de confiance Monte Carlo:

$$\text{Niveau de confiance Monte Carlo: } = B^{-1}\sum_{i=1}^{B} u_i(s),$$

avec:

$$u_i(s) = \begin{cases} 1, & \text{si } IC^{(i)}(s) \ni \theta, \\ 0, & \text{sinon.} \end{cases}$$