



N° Réf :.....

Centre Universitaire
Abd elhafid boussouf Mila

Institut des sciences et de la technologie

Département de Mathématiques et Informatique

Mémoire préparé En vue de l'obtention du diplôme de Licence

En: - Filière mathématiques

Estimation par réduction de la variance

Préparé par :

- Bouchouk Hadda
- Laibi Imane
- Meziti Meriem

Encadrer par : Mehazzem Allal

Grade.M.A.A

Année universitaire : 2014/2015

Remerciements

Avant tous nous remercions Allah qui nous à donne la capacité d'écrire et de réfléchir, la patience d'aller jusqu'au bout du rêve et le bonheur de lever nos mains vers le ciel et de dire " Ya Kayoum ".

Nous tiens tout d'abord à remercier notre encadreur Monsieur Mehazzem Allal, qui nous 'a offert cette chance formidable de travailler avec-lui. Grâce à ses suggestions et à son suivi continu. Son énergie et sa confiance, son aide pour la réalisation de ce travail. Il est important pour nous de remercions nos familles nos très chers Parents, frères, sœurs, collègues et amis, qui ont toujours été une Source inépuisable d'encouragements.

Nous souhaitons la réussite à tous les étudiants des sections Mathématiques et Informatique.

Table des matières

Introduction Générale	2
1 Estimation par intervalle de confiance	3
1.1 Introduction	3
1.2 Principe	3
1.3 Intervalles de confiance	4
1.4 Intervalles de confiance pour une espérance	4
1.4.1 Espérance d'une loi normal, écart-type σ connu	4
1.4.2 Espérance d'une loi normal, écart-type σ inconnu	5
1.5 Intervalles de confiance pour une variance	5
1.5.1 Variance d'une loi normale, espérance μ connu	5
1.5.2 Variance d'une loi normale, espérance μ inconnu	6
1.6 Intervalle de confiance pour une proportion	6
1.6.1 Principe de construction	6
1.6.2 Les différences méthodes de construction	7
1.7 Intervalles de confiance pour un coefficient de corrélation linéaire	7
1.7.1 Coefficient de corrélation linéaire simple	7
1.7.2 Estimation et intervalle de confiance	8
2 Estimation par réduction de variance	9
2.1 Description de la méthode de monte-carlo	9
2.1.1 La définition de la méthode de monte-carlo	9
2.1.2 Premier exemple de la méthode de Monte-Carlo	9
2.1.3 Description de la méthode	10
2.1.4 Convergence de la méthode	10
2.1.5 Estimation de la variance d'un calcul	11
2.2 Méthodes de réduction de variance	12
2.2.1 Echantillonnage préférentiel	12
2.2.2 Méthode des variables de controles	13

2.2.3	Variables antithétiques	13
2.2.4	Méthode de l'échantillonnage stratifié	13
Bibliographie		13

Introduction Générale

La notion de base en statistique est celle de population : ensemble d'individus (ou objets ou unités statistiques) pouvant être décrits par un ensemble des variables (ou propriétés ou caractéristiques) communes. La variabilité d'une population signifie que les variables décrivant les individus peuvent prendre des valeurs différentes d'un individu à l'autre. L'analyse statistique est l'étude de cette variabilité. Souvent, il est matériellement impossible d'étudier tous les individus d'une population. Si l'on se limite à une partie de la population, on fait un sondage, la partie étudiée s'appelle un échantillon. Afin d'assurer la représentativité de l'échantillon, on va voir estimation

L'estimation statistique consiste à donner une valeur approchée à une caractéristique d'une population, à partir d'un échantillon d'observations issues de cette population. Nous intéressons dans un premier temps à l'estimation par l'intervalle de confiance. Dans un second temps, nous chercherons à décrire des méthodes pour l'estimation comme la méthode de monte-carlo Cette méthode est basée sur des simulations.

Ce mémoire est partagé en deux chapitres.

Dans le premier chapitre, on s'intéresse l'estimation par intervalle de confiance, plus précisément on rappelle la définition de l'intervalle de confiance, intervalle de confiance pour une espérance, intervalle de confiance pour une variance.

Enfin, dans le deuxième chapitre nous allons donner l'estimation par réduction de la variance, la méthode de monte-carlo, les méthodes de réduction de variance.

Chapitre 1

Estimation par intervalle de confiance

1.1 Introduction

Ce chapitre est consacré à la construction d'intervalles contenant le paramètre inconnu supposé réel, avec un niveau de confiance fixé. Il contient essentiellement des méthodes, sans rentrer dans un formalisme excessif.

Lorsque nous procédons à l'estimation de l'une des caractéristiques, notée θ , de la population, il est plus pertinent de fournir un intervalle $\theta_1 < \theta < \theta_2$ plutôt que de donner simplement une estimation ponctuelle, c'est-à-dire d'écrire uniquement $\hat{\theta} = c$. Un tel intervalle $] \theta_1 ; \theta_2 [$ s'appelle une estimation par intervalle de confiance du paramètre ou estimation ensembliste de la caractéristique.

1.2 Principe

La méthode sur laquelle repose la construction des intervalles de confiance est la suivante :

Soit $\hat{\theta}_n$ un estimateur de θ dont nous connaissons la loi de probabilité pour chaque valeur de θ .

Étant donné une valeur θ_0 du paramètre θ , nous déterminons un intervalle de probabilité bilatéral de niveau $(1-\alpha)$ pour l'estimateur $\hat{\theta}_n$, c'est-à-dire deux bornes θ_1^n et θ_2^n telle que :

$$P(\theta_1^n < \hat{\theta}_n < \theta_2^n \mid \theta = \theta_0) \geq 1 - \alpha$$

Remarque 1.2.1 1. Ces deux bornes dépendent de la valeur θ_0 .

2. Nous pourrions également construire des intervalles unilatéraux pour lesquels $\theta_1^n = -\infty$ ou $\theta_2^n = +\infty$.

3. Nous choisissons dans la plupart des cas un intervalle de probabilité à risques symétriques $\frac{\alpha}{2}$ et $\frac{\alpha}{2}$.

1.3 Intervalles de confiance

Définition 1.3.1 : Nous appelons intervalle de confiance de niveau de confiance $1 - \alpha$ du paramètre θ tout intervalle $] \theta_1 ; \theta_2 [$ tel que : $P(] \theta_1 ; \theta_2 [\ni \theta) = 1 - \alpha$ pour $\alpha \in [0; 1]$ fixé.

Propriétés :

1. La probabilité α pour que l'intervalle de confiance ne contienne pas la vraie valeur θ peut être répartie différemment de part et d'autre des bornes de l'intervalle de confiance. Posons $\alpha = \alpha_1 + \alpha_2$, où α_1 mesure le risque que la vraie valeur inconnue de θ soit plus petite que la borne inférieure θ_1 de l'intervalle de confiance et α_2 mesure le risque que la vraie valeur inconnue de θ soit plus grande que la borne supérieure θ_2 de l'intervalle de confiance.

2. L'intervalle de confiance est dit bilatéral lorsque $\alpha_1 \neq 0$ et $\alpha_2 \neq 0$. Si $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, l'intervalle est dit symétrique, sinon il est dit dissymétrique.

3. L'intervalle de confiance est dit unilatéral lorsque $\alpha_1 = 0$ ou $\alpha_2 = 0$. Il est toujours dissymétrique. Si nous sommes uniquement intéressés par le calcul d'une valeur minimale pour la quantité estimée, nous fixons $\alpha_1 = \alpha$ et $\alpha_2 = 0$. L'intervalle de confiance est alors de la forme $] \theta_1 ; \theta_2 [=] -\infty ; b [$.

4. $] \theta_1 ; \theta_2 [$ est un intervalle aléatoire car il dépend de l'estimateur $\hat{\theta}_n$. $] \theta_1 ; \theta_2 [$ s'obtient par :

$$\begin{cases} \theta_1 = (\theta_2^n)^{-1} \left(\hat{\theta}_n \right) \\ \theta_2 = (\theta_1^n)^{-1} \left(\hat{\theta}_n \right) \end{cases}$$

5. Si nous augmentons le niveau $1 - \alpha$ ou encore parfois appelé également le coefficient de confiance, nous augmentons la longueur de l'intervalle de probabilité

6. Nous associons souvent aux intervalles de confiance, un test basé sur la règle de décision suivante : soit $\hat{\theta}(obs)$ la valeur observée de l'estimateur $\hat{\theta}_n$:

.Si la valeur observée $\hat{\theta}(obs) \in] \theta_1^n ; \theta_2^n [$ nous conservons la valeur θ_0 comme valeur possible du paramètre θ .

.Si la valeur observée $\hat{\theta}(obs) \notin] \theta_1^n ; \theta_2^n [$ nous éliminons la valeur θ_0 .

1.4 Intervalles de confiance pour une espérance

1.4.1 Espérance d'une loi normale, écart-type σ connu

Soit $(X_1, \dots, X_n)$ un échantillon aléatoire de loi parente la loi normale d'espérance μ et d'écart-type σ .

$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n X_i$ est le meilleur estimateur de μ et $\hat{\mu}_n$ suit la loi $N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$

a) Intervalle de la probabilité : L'intervalle de probabilité pour $\hat{\mu}_n$ à $1 - \alpha$ est égal à :

$$\mu - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \hat{\mu}_n < \mu + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},$$

où $u_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ pour la loi normale centrée-réduite.

b) Intervalle de confiance : L'intervalle de confiance pour μ à $1 - \alpha$ est égal à :

$$\hat{\mu}_n(obs) - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu}_n(obs) + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},$$

où $u_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ pour la loi normale centrée-réduite.

Remarque 1.4.1 Souvent $1 - \alpha$ sera fixé à 0.95 et par conséquent $u_{1-\frac{\alpha}{2}}$ sera égale à 1.96.

1.4.2 Espérance d'une loi normal, écart-type σ inconnu

a) Théorème de Fisher : Soit $x_1, \dots, x_n$ n variables aléatoires indépendantes de la loi normale d'espérance μ et d'écart-type σ .

$\hat{\mu}_n$ et $\frac{ns_n^2}{\sigma^2}$ sont deux variables aléatoires indépendantes et $\frac{\hat{\mu}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ suit la loi normale

$N(0; 1)$ et $\frac{ns_n^2}{\sigma^2}$ suit la loi du Khi-deux $x^2(n-1)$. $\frac{\hat{\mu}_n - \mu}{\frac{s_n}{\sqrt{n-1}}}$ suit de la loi de student $t(n-1)$.

Remarque 1.4.2 s_n^2 est défini $s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_n)^2$.

b) Intervalle de la probabilité : L'intervalle de probabilité pour $\frac{\hat{\mu}_n - \mu}{\frac{s_n}{\sqrt{n-1}}}$ à $1 - \alpha$ est

$$\text{égale à : } -t_{n-1; 1-\frac{\alpha}{2}} < \frac{\hat{\mu}_n - \mu}{\frac{s_n}{\sqrt{n-1}}} < t_{n-1; 1-\frac{\alpha}{2}},$$

où $t_{n-1; 1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ pour la loi student $t(n-1)$.

c) Intervalle de confiance : L'intervalle de confiance pour μ à $1 - \alpha$ est égale à :

$$\hat{\mu}_n(obs) - t_{n-1; 1-\frac{\alpha}{2}} \frac{s_n(obs)}{\sqrt{n-1}} < \mu < \hat{\mu}_n(obs) + t_{n-1; 1-\frac{\alpha}{2}} \frac{s_n(obs)}{\sqrt{n-1}},$$

où $t_{n-1; 1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ pour la loi student $t(n-1)$.

1.5 Intervalles de confiance pour une variance

1.5.1 Variance d'une loi normale, espérance μ connu

$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$ est un estimateur de la variance σ^2 et $\frac{n\hat{\sigma}_n^2}{\sigma^2}$ suit la loi du khi-deux à n degrés de liberté, notée $X^2(n)$.

a) Intervalle de probabilité : k_1 et k_2 sont les bornes de l'intervalle de probabilité pour $\frac{n\hat{\sigma}_n^2}{\sigma^2}$ si $P\left(k_1 < \frac{n\hat{\sigma}_n^2}{\sigma^2} < k_2\right) = 1 - \alpha$.

Par exemple, nous pouvons prendre la borne k_1 égale au quantile d'ordre $\frac{\alpha}{2}$ pour la loi du Khi-deux à n degrés de liberté $X^2(n)$ et la borne k_2 égale au quantile d'ordre $1 - \frac{\alpha}{2}$ pour la loi du Khi-deux à n degrés de liberté $X^2(n)$.

b) Intervalle de confiance : L'intervalle de confiance pour σ^2 à $1 - \alpha$ est égal

$$\text{à : } \frac{n\hat{\sigma}_n^2(obs)}{k_2} < \sigma^2 < \frac{n\hat{\sigma}_n^2(obs)}{k_1}.$$

Par exemple, nous pouvons prendre la borne k_1 égale au quantile d'ordre $\frac{\alpha}{2}$ pour la loi de Khi-deux à n degrés de liberté $X^2(n)$ et la borne k_2 égale au quantile d'ordre $1 - \frac{\alpha}{2}$ pour la loi du Khi-deux à n degrés de liberté $X^2(n)$

1.5.2 Variance d'une loi normale, espérance μ inconnu

$s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_n)^2$ est un estimateur de σ^2 et $\frac{ns_n^2}{\sigma^2}$ suit la loi du Khi-deux à $(n-1)$ degrés de liberté $X(n-1)^2$.

Remarque 1.5.1 $\hat{\mu}_n$ est défini $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n X_i$

a) Intervalle de probabilité : l_1 et l_2 sont les bornes de l'intervalle de probabilité pour $\frac{ns_n^2}{\sigma^2}$ si $P\left(l_1 < \frac{ns_n^2}{\sigma^2} = \frac{(n-1)s_{n,c}^2}{\sigma^2} < l_2\right) = 1 - \alpha$.

Par exemple, nous pouvons prendre la borne l_1 égale au quantile d'ordre $\frac{\alpha}{2}$ pour la loi du Khi-deux à $(n-1)$ degrés de liberté $X^2(n-1)$ et la borne l_2 égale au quantile d'ordre $1 - \frac{\alpha}{2}$ pour la loi du Khi-deux à $(n-1)$ degrés de liberté $X^2(n-1)$.

b) Intervalle de confiance : Intervalle de confiance pour σ^2 à $1 - \alpha$ est égal à :

$$\frac{(n-1)s_{n,c}^2(obs)}{l_2} = \frac{ns_n^2(obs)}{l_2} < \sigma^2 < \frac{ns_n^2(obs)}{l_1} = \frac{(n-1)s_{n,c}^2(obs)}{l_1}$$

1.6 Intervalle de confiance pour une proportion

1.6.1 Principe de construction

Nous souhaitons construire un intervalle de confiance pour proportion π_A d'individus de la population qui possèdent un certain caractère A . Pour estimer π_A , nous allons nous servir de l'estimateur $\hat{\pi}_{n,A}$, qui a été défini auparavant à $\hat{\pi}_{n,A} = \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n X_i$ à partir du moment où nous faisons l'hypothèse que le tirage est aléatoire avec remise, se qui correspond à une population infinie. D'autre part, nous pouvons montrer que $n\hat{\pi}_{n,A}$ suit la loi binomiale de paramètres n et π_A . À partir de ce résultat, nous pouvons construire un intervalle de confiance pour π_A .

1.6.2 Les différences méthodes de construction

Les trois méthodes pour construire un intervalle de confiance pour une proportion π_A que nous rencontrerons le plus souvent sont :

a. La méthode exacte ou encore appelée la méthode de clopper-pearson qui maintenant est réalisable avec la plupart des logiciels de statistique.

b. La méthode des score ou encore appelée la méthode de wilson ou encore la méthode de l'ellipse.

c. La méthode asymptotique ou encore appelée la méthode de wald.

Des études statistiques ont montré que, parmi ces trois méthodes, la méthode à privilégier est la méthode des score. Néanmoins, la méthode de wald reste très utilisée et présentée dans de nombreux manuels alors qu'elle nous permet généralement pas d'obtenir des intervalles de confiance de qualité convenable.

a) Intervalle de confiance par la méthode du score ou la méthode wilson

L'intervalle de confiance pour la proportion π_A au niveau de confiance $(1 - \alpha)$

est égale à :

$$\frac{\hat{\pi}_{n,A}(obs) + \frac{1}{2n}u_{1-\frac{\alpha}{2}}^2 - u_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\hat{\pi}_{n,A}(obs) \times (1 - \hat{\pi}_{n,A}(obs))}{n} + \frac{u_{1-\frac{\alpha}{2}}^2}{4n^2}}}{1 + \frac{1}{n}u_{1-\frac{\alpha}{2}}^2} < \pi_A$$

$$< \frac{\hat{\pi}_{n,A}(obs) + \frac{1}{2n}u_{1-\frac{\alpha}{2}}^2 + u_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\hat{\pi}_{n,A}(obs) \times (1 - \hat{\pi}_{n,A}(obs))}{n} + \frac{u_{1-\frac{\alpha}{2}}^2}{4n^2}}}{1 + \frac{1}{n}u_{1-\frac{\alpha}{2}}^2},$$

où $u_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ pour la loi normale centrée réduite.

b) Intervalle de confiance par la méthode asymptotique ou la méthode de wald

L'intervalle de confiance pour la proportion π_A au niveau de confiance $(1 - \alpha)$

est égale à :

$$\hat{\pi}_{n,A}(obs) - u_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\hat{\pi}_{n,A}(obs) \times (1 - \hat{\pi}_{n,A}(obs))}{n}} < \pi_A$$

$$< \hat{\pi}_{n,A}(obs) + u_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\hat{\pi}_{n,A}(obs) \times (1 - \hat{\pi}_{n,A}(obs))}{n}},$$

où $u_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ pour la loi normale centrée réduite.

1.7 Intervalles de confiance pour un coefficient de corrélation linéaire

1.7.1 Coefficient de corrélation linéaire simple

Le coefficient de corrélation linéaire $\rho(x, y)$ mesure le degré d'association

linéaire entre deux variable aléatoires X et Y : $\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)}$

l'encadrement suivant est toujours vrai $-1 \leq \rho(x, y) \leq 1$.

Lorsque $\rho(X, Y) = 0$, les deux variables aléatoires X et Y sont dites non corrélées. L'indépendance implique toujours l'absence de corrélation. Pour deux variables aléatoires ne pas être corrélées ne signifie pas généralement être l'indépendantes. Deux exceptions où l'absence de corrélation et l'indépendance sont équivalentes : les couples (X, Y) qui suivent une loi normale à deux dimensions et les couples de variables de bernoulli.

1.7.2 Estimation et intervalle de confiance

Nous allons utiliser l'estimateur $\widehat{\rho(X, Y)}$ du coefficient de corrélation linéaire

$\rho(X, Y)$ défini par :

$$\widehat{\rho(X, Y)} = \frac{\frac{1}{n-1} \sum_{i=1}^n [(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)]}{\sqrt{\hat{\sigma}_x^2 \times \hat{\sigma}_y^2}} = \frac{\sum_{i=1}^n (X_i \times Y_i) - \bar{X}_n \times \bar{Y}_n}{\sqrt{(\sum_{i=1}^n X_i^2) - \bar{X}_n^2} \sqrt{(\sum_{i=1}^n Y_i^2) - \bar{Y}_n^2}}$$

où $\bar{X}_n$ est la moyenne des X_i et $\bar{Y}_n$ est la moyenne des Y_i

L'étude des propriétés de cet estimateur est plus complexe puisqu'il s'agit d'un rapport de variables aléatoires qui ne sont indépendantes. Il s'agit d'un estimateur biaisé et convergent de $\rho(X, Y)$:

$$E\left(\widehat{\rho(X, Y)}\right) = \rho(X, Y) - \frac{\rho(X, Y)(1 - \rho(X, Y)^2)}{2n}$$

$$\text{var}\left(\widehat{\rho(X, Y)}\right) = \frac{(1 - \rho(X, Y)^2)^2}{n} + o\left(\frac{1}{n}\right),$$

où $o\left(\frac{1}{n}\right)$ est une fonction qui tend vers 0 lorsque n tend vers $+\infty$ plus vite que $\frac{1}{n}$.

Propriété :

La distribution asymptotique de la variable aléatoire $\sqrt{n}\left(\widehat{\rho(X, Y)} - \rho(X, Y)\right)$ est la loi normale $N\left(0; (1 - \rho(X, Y)^2)^2\right)$.

Remarque 1.7.1 : Ce resultat n'est pas assez précis et nous ne l'utiliserons que si ne pouvons pas faire autrement et si la taille de l'échantillon vérifie la condition suivante :

$n \geq 30$ voici un intervalle de confiance, basé sur la transformation de fisher, pour ρ de niveau approximatif $(1 - \alpha) \%$ plus précis : $\left[\tanh\left(\widehat{z}(\text{obs}) - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right), \tanh\left(\widehat{z}(\text{obs}) + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right) \right]$

où $\widehat{z}(\text{obs}) = \tanh^{-1}(\widehat{\rho}(\text{obs}))$, $\tanh^{-1}(x) = \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right)$, $\tanh(x) = \frac{\exp(2x)-1}{\exp(2x)+1}$ et $u_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ pour la loi normale centrée -réduite.

Chapitre 2

Estimation par réduction de variance

2.1 Description de la méthode de monte-carlo

2.1.1 La définition de la méthode de monte-carlo

La méthode de Monte-Carlo est une méthode d'approximation, au sens statistique du terme. Il n'y a pas de définition précise de ce qu'est une technique de type Monte Carlo, mais la description la plus habituelle consiste à dire que les méthodes de ce type se caractérisent par l'utilisation du hasard pour résoudre des problèmes centrés sur le calcul d'une valeur numérique. La réponse fournie sera une réponse statistique de type "la valeur cherchée I se trouve très probablement (par exemple avec une probabilité au moins égale à 0,95) dans l'intervalle de confiance $[I_1, I_2]$ ". La précision est mesurée par la taille $I_2 - I_1$ de l'intervalle de confiance. Si on utilise n points échantillonnés de manière indépendante, la méthode converge en $O(\frac{1}{\sqrt{n}})$, quelque soit la dimension du problème, d'où son intérêt en grande dimension (parfois pour des fonctions à plusieurs dizaines de variable, la méthode de Monte-Carlo est le seul outil capable de donner une réponse en un temps raisonnable).

Applications de la méthode : en finance, en sismologie, télécommunication, ingénierie physique, biologie, sciences sociales.

2.1.2 Premier exemple de la méthode de Monte-Carlo

Un premier exemple de méthode de Monte-Carlo, est la détermination du nombre π , via l'estimation de la surface du cercle de rayon 1. On considère un grand nombre (N) de paires de nombres (r_1, r_2) tirés avec une distribution uniforme. L'ensemble de ces points sont donc dans le carré de côté $[0, 1] * [0, 1]$. La proportion p de points qui sont dans le cercle unité va tendre vers quand le nombre de points N tend vers

$+\infty$ (Le point (r_1, r_2) est dans le cercle unité si $r_1^2 + r_2^2 \leq 1$)

2.1.3 Description de la méthode

Pour utiliser une méthode de Monte-Carlo, on doit tout d'abord mettre sous la forme d'une espérance la quantité que l'on cherche à calculer. C'est souvent simple (calcul d'intégrale par exemple) mais peut être plus compliqué (équations aux dérivées partielles par exemple)

A l'issue de cette étape, il reste à calculer une quantité de la forme $E(X)$, c'est à dire l'espérance de la variable aléatoire X . Pour calculer $E(X)$, il convient de savoir simuler une variable aléatoire selon la loi de X .

On dispose alors d'une suite $(X_i)_{1 \leq i \leq N}$ de N réalisations de la variable aléatoire X .

On approxime alors $E(X)$ par : $E(X) \approx \frac{1}{N}(X_1 + \dots + X_N)$.

Application au cas du calcul d'une intégrale :

on cherche à calculer
$$I = \int_{[0,1]^d} f(u_1, \dots, u_d) du_1 \dots du_d.$$

1. Mise sous forme d'espérance : on pose $X = f(U_1, \dots, U_d)$ ou $U_1, \dots, U_d$ sont des réalisations de la loi uniforme sur l'intervalle $[0, 1]$

alors $E(X) = E(f(U_1), \dots, U_d) = I$

2. Simulation de la variable aléatoire : on suppose que l'on dispose d'une suite $(U_i)_{i \geq 1}$ de réalisations de la loi uniforme sur $[0, 1]$

On pose alors $X_1 = f(U_1, \dots, U_d)$, $X_2 = f(U_{d+1}, \dots, U_{2d})$, etc. Alors les (X_i) sont des réalisations de la variable X et $I \approx \frac{1}{N}(X_1 + \dots + X_N)$

Remarque 2.1.1 Cette méthode est facilement programmable, et ne dépend pas de la régularité de f (on lui demande juste d'être mesurable). Souvent on cherche à évaluer une

intégrale plus générale :
$$I \approx \int_{\mathbb{R}^d} g(x) f(x) dx = \int_{\mathbb{R}^d} g(x_1, \dots, x_d) f(x_1, \dots, x_d) dx_1, \dots, dx_d$$
 avec f

positive, $\int_{\mathbb{R}^d} f(x) dx = 1$. Alors $I = E(g(X))$ ou X est une variable aléatoire à valeur dans

$\mathbb{R}^d$ de loi f . Ainsi, on approche I par $I \approx \frac{1}{N} \sum_{i=1}^N g(X_i)$ si (X_i) est un échantillon tiré selon la loi f

2.1.4 Convergence de la méthode

C'est la loi forte des grands nombres qui permet de justifier la convergence de la méthode, et le théorème de la limite centrale qui précise la vitesse de convergence.

Théorème 2.1.2 Loi forte des grands nombres Soit $(X_i, i \geq 1)$ une suite de réalisations, de la variable aléatoire X . On suppose que $E(|X|) < +\infty$. Alors, pour presque tout ω (i.e. $\exists N \in \Omega$ avec $p(N) = 0$ et $\omega \notin N$) $E(X) = \lim_{n \rightarrow +\infty} \frac{1}{n}(X_1(\omega) + \dots + X_n(\omega))$
Première remarque : la méthode de Monte-Carlo ne peut donc s'utiliser que pour des variables aléatoires intégrables. Pour avoir une idée de l'intérêt de la méthode, il faut pouvoir évaluer l'erreur commise.

Définition 2.1.3 L'erreur est définie par $\epsilon_n = E(X) - \frac{1}{n}(X_1 + \dots + X_n)$ Le théorème de la limite centrale donne un asymptotique de l'erreur n , mais de nature aléatoire. Il dit que la loi de l'erreur finit par ressembler à une loi gaussienne centrée.

Théorème 2.1.4 Soit $(X_i, i \geq 1)$ une suite de réalisation de la variable aléatoire X . On suppose que $E(X^2) < +\infty$. On note σ^2 la variance de X ($\sigma^2 = E(X^2) - (E(X))^2$), alors $\frac{\sqrt{n}}{\sigma}\epsilon_n$ converge en loi vers une gaussienne centrée réduite G . Cela signifie que si G est une variable aléatoire de loi $\frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}} dx$ et si f est une fonction continue bornée, $E(f(\frac{\sqrt{n}}{\sigma}\epsilon_n))$ converge vers $E(f(G)) = \int_{-\infty}^{+\infty} f(x) \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}} dx$

Remarque 2.1.5 :Le théorème de la limite centrale ne permet jamais de borner l'erreur (le support d'une gaussienne est égale à $\mathbb{R}$ en entier). On présente souvent l'erreur de la méthode de Monte-Carlo, soit en donnant l'écart-type de l'erreur ϵ_n , c'est à dire $\frac{\sigma}{\sqrt{n}}$, soit en donnant un intervalle de confiance à 95% pour le résultat. Comme $P(|N| < 1,96) = 0,95$, si N est une variable aléatoire de loi la gaussienne centrée réduite, cela signifie que $P(|\epsilon_n| \leq 1,96 \frac{\sigma}{\sqrt{n}}) \approx 0,95$.

Définition 2.1.6 L'intervalle de confiance à 95% est donc

$$\left[\frac{1}{n} \sum_{i=1}^n X_i - 1,96 \frac{\sigma}{\sqrt{n}}, \frac{1}{n} \sum_{i=1}^n X_i + 1,96 \frac{\sigma}{\sqrt{n}} \right]$$

Remarque 2.1.7 La vitesse de convergence est en $\frac{1}{\sqrt{n}}$, ce qui n'est pas très rapide, mais c'est parfois la seule méthode accessible, de plus cette vitesse ne change pas si on est en grande dimension, et elle ne dépend pas de la régularité de f

2.1.5 Estimation de la variance d'un calcul

On désire évaluer la variance d'un calcul par Monte-Carlo d'une intégrale de type

$$\int_{\mathbb{R}} x f(x) dx = E(X) \text{ avec } X \text{ une variable aléatoire de loi } f. \text{ Soit } X_i \text{ des réalisations de } X, \text{ on approche } I \text{ par } \bar{I} = \frac{1}{n} \sum_{i=1}^n X_i$$

Définition 2.1.8 *Variance empirique de X*

$\bar{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i^2 - \bar{I}^2)$. $\bar{\sigma}_n^2$ est un estimateur sans biais de σ^2 c'est à dire que $E(\bar{\sigma}_n^2) = \sigma^2$ et que la suite $\bar{\sigma}_n^2$ converge vers σ^2 presque sûrement quand $n \rightarrow +\infty$. Cela signifie qu'au cours d'un calcul Monte-Carlo, on peut simultanément calculer l'approximation de l'espérance et de sa variance, ce qui permet d'obtenir un intervalle de confiance à 95%.

2.2 Méthodes de réduction de variance

Nous venons de voir que la vitesse de convergence de la méthode Monte-Carlo est de l'ordre de $\frac{\sigma}{\sqrt{n}}$. Pour améliorer cette méthode, il existe de nombreuses techniques, dites de réduction de variance, qui cherchent à diminuer la valeur de σ^2 . L'idée générale de ces méthodes est de donner une autre représentation sous forme d'espérance de la quantité à calculer, c'est à dire : $E(X) = E(Y)$ avec $\text{var}(Y) < \text{var}(X)$.

2.2.1 Echantillonnage préférentiel

Soit X une variable aléatoire de loi f (sur $\mathbb{R}$ par exemple), et on cherche à calculer

$$E(g(X)) = \int_{\mathbb{R}} g(x)f(x)dx.$$

Soit h une autre loi de probabilité ($h > 0$ et $\int_{\mathbb{R}} h(x)dx = 1$).

$$\text{Alors on écrit : } E(g(X)) = \int_{\mathbb{R}} \frac{g(x)f(x)}{h(x)} h(x) dx = E \frac{g(y)f(y)}{h(y)}$$

avec Y une variable aléatoire de loi h .

En effectuant un tirage de n valeurs $Y_1, \dots, Y_n$, on approxime $E(g(X))$ par $\frac{1}{n} \sum_{i=1}^n \frac{g(y_i)f(y_i)}{h(y_i)}$

Si on pose $Z = \frac{g(y)f(y)}{h(y)}$, on a accéléré le calcul si $\text{var}(Z) < \text{var}(X)$.

Si h est choisie aussi proche que possible de $|g(x)f(x)|$ (et ensuite on la normalise), et qu'elle est facilement calculable, on a réduit la variance.

Exemple 2.2.1 $\int_0^1 \cos\left(\frac{\pi x}{2}\right) dx$: on approche le cosinus par un polynôme du second degré

$h(x) = (1 - x^2)$. Par normalisation

$\lambda = \frac{1}{3}$. En calculant les variances, on voit qu'on a réduit la variance d'un facteur 100.

2.2.2 Méthode des variables de controles

Cette fois-ci, on prend h une fonction voisine de g facilement simulable, pour laquelle $E(h(X))$ est explicite et on écrit : $E(g(X)) = E(g(X) - h(X)) + E(h(X))$.

On estime par un calcul Monte-Carlo l'espérance $E(g(X) - h(X))$ et on a une formule directe pour $E(h(X))$. Si h est bien choisie, alors $Var(g(X) - h(X)) < Var(g(X))$

Exemple 2.2.2 $\int_0^1 \exp(x) dx = \int_0^1 (\exp(x) - 1 - x) dx + \frac{3}{2}$ en finance :

$$C - P = E(\exp(\sigma G) - K) = \exp\left(\frac{\sigma^2}{2}\right) - K$$

2.2.3 Variables antithétiques

Cette méthode consiste, quand on effectue des réalisations X_i de la variable aléatoire x , à associer à X_i , la réalisation $-X_i$.

En effet, $I = \int_0^1 f(x) dx = \frac{1}{2}(\int_0^1 (f(x) + f(1-x))dx)$.

Ainsi, $\tilde{I}_{2n} = \frac{1}{2n} \sum_i (1^{2N} \frac{1}{2} (f(U_i) + f(1-U_i)))$ au lieu de $I_{2n} = \frac{1}{2n} \sum_i (1^{2N} f(U_i))$.

Si f est continue, monotone, cette méthode est plus efficace que le Monte-Carlo initial

Cette idée est généralisable à toute dimension et à toute transformation préservant la loi de la variable aléatoire.

2.2.4 Méthode de l'échantillonnage stratifié

C'est la méthode qui est utilisée dans les sondages :

On cherche à calculer $I = E(g(X)) = \int_{\mathbb{R}^d} g(x)f(x)dx$, X est ici une variable aléatoire de loi f dans $\mathbb{R}^d$.

On suppose que l'on dispose d'une partition $(D_i, 1 \leq i \leq m)$ de $\mathbb{R}^d$.

Alors $I = \sum_{i=1}^m E(1_{X \in D_i} g(X)) = \sum_{i=1}^m E(g(X)|X \in D_i)P(X \in D_i)$. Si on connaît les $p_i = P(X \in D_i)$, il reste à calculer les $I_i = E(g(X)|X \in D_i)$ par des méthodes de Monte-Carlo (calcul approché : $\tilde{I}_i$) ainsi, I sera approché par $\tilde{I} = \sum_{i=1}^m p_i \tilde{I}_i$

On suppose que pour calculer $\tilde{I}_i$, on effectue n_i tirages, la variance de l'erreur d'approximation est donnée par $\frac{\sigma_i^2}{n_i}$, si on note $\sigma_i^2 = \text{var}(g(X)|X \in D_i)$. La variance du calcul de $\tilde{I}$ vaut alors $\sum_{i=1}^m p_i^2 \frac{\sigma_i^2}{n_i}$. Si on pose $n = \sum_{i=1}^m n_i$. On peut vérifier que les n_i qui minimise la variance de $\tilde{I}$ sont donnés par : $n_i = n \frac{p_i \sigma_i}{\sum_{i=1}^m p_i \sigma_i}$ et le minimum de la variance est alors $\frac{1}{n} (\sum_{i=1}^m p_i \sigma_i)^2$. On peut montrer que cette variance est inférieure à la variance qu'on obtiendrait en effectuant n tirages aléatoires par la méthode de Monte-Carlo : $\text{var}(g(X)) \geq \sum_{i=1}^m p_i \text{var}(g(X)|X \in D_i) \geq (\sum_{i=1}^m p_i \sigma_i)^2$

Bibliographie

- [1] F. Bertrand, M. Maumy-Bertrand, Statistique pour les scientifiques, Dunod, paris, 2011.
- [2] Josselin Garnier (Université Paris Diderot), Méthodes de réduction de variance pour Monte Carlo.
- [3] Céline Baranger - Julien Mathiaud, Méthode de Monte-Carlo, 2012/2013.
- [4] [Http ://www.odilejacob.fr/catalogue/sciences/mathematiques/estimation](http://www.odilejacob.fr/catalogue/sciences/mathematiques/estimation).
- [5] [Http ://fr.wikipedia.org/wiki/Statistique_inferentielle](http://fr.wikipedia.org/wiki/Statistique_inferentielle).