Cours 4 – Estimation Ponctuelle.

Eya ZOUGAR *

Institut National des Sciences appliqués-INSA

Génie mathématiques GM3
Thursday 09^{th} Febrary, 2023

*Basé sur le cours de Bruno PORTIER

Contents

Rappel:
1.Notion de convergence.

Soit T_n un estimateur du paramètre réel θ.

En statistique, on souhaitera que cet estimateur soit convergent.

La qualité de l'estimation doit s'améliorer avec l'augmentation du nombre de données, c'est à dire:

$$T_n \underset{n\to\infty}{\longrightarrow} \theta.$$

Cependant, comme T_n est une variable aléatoire, il faudra préciser le type de convergence :

- en probabilité ;
- en moyenne quadratique ;
- presque sûre ;
- ou bien en loi.

1.2. Une mesure d'efficacité: l'erreur quadratique moyenne.

Pour mesurer l'efficacité d'un estimateur par rapport à un autre, on peut utiliser l'erreur quadratique moyenne.

Soit T_n un estimateur du paramètre θ. Son erreur quadratique moyenne est définie par :

$$EQM = \mathbb{E}\left[(T_n - \theta)^2\right]$$

Il faut bien évidemment que $\mathbb{E}\left[T_n^2\right] < \infty$.

On dira que T_n converge en moyenne quadratique vers θ lorsque son erreur quadratique moyenne convergera vers 0.

1.3. Décomposition Biais-Variance.

Soit T_n un estimateur du paramètre réel θ.

On a alors la décomposition Biais-Variance de l'erreur quadratique moyenne, suivante :

$$\mathbb{E}\left[(T_n-\theta)^2\right] = \underbrace{\mathbb{E}\left[(T_n-\mathbb{E}(T_n))^2\right]}_{\text{Variance}} + \underbrace{(\mathbb{E}[T_n]-\theta)^2}_{\text{Biais}}.$$

Lorsque l'estimateur est sans biais, cette égalité se réduit à :

$$\mathbb{E}\left[(T_n-\theta)^2\right] = \mathbb{V}\text{ar}(T_n).$$

Remarque. Un estimateur sans biais sera convergent en moyenne quadratique (et donc en probabilité), dès que sa variance tendra vers 0.

1.4. Rappels: Théorèmes LGN et TLC.

Nous utiliserons deux résultats importants de convergence pour les sommes de variables aléatoires indépendantes et de même loi.

Théorème de Loi forte des grands nombres:
Soient (Y_n) une suite de variables aléatoires indépendantes et de même loi. Alors si $\mathbb{E}[|Y_1|] < \infty$,

$$\frac{1}{n}\sum_{j=1}^{n} Y_j \xrightarrow[n\to\infty]{p.s.} \mathbb{E}[Y_1]$$

Théorème de limite centrale: Si de plus $\mathbb{E}\left[Y_1^2\right] < \infty$, alors

$$\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n} Y_j - \mathbb{E}[Y_1]\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathbb{V}\text{ar}(Y_1))$$

1.5. La delta-méthode.

Théorème: Soit θ un paramètre réel inconnu et soit T_n un estimateur du paramètre θ satisfaisant :

$$\sqrt{n}\,(T_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

Soit $g : \mathbb{R} \longrightarrow \mathbb{R}$ une application dérivable en θ.
Alors, si $g'(\theta) \neq 0$, on a le théorème de limite centrale suivant :

$$\sqrt{n}\,(g(T_n) - g(\theta)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, (g'(\theta))^2\sigma^2\right)$$

2. La méthode des moments.

Soit une variable aléatoire dont la loi dépend d'un paramètre θ .
Nous supposons qu'il existe une fonctions h entièrement connue, telle que :

$$\mathbb{E}[X] = h(\theta), \quad \forall \theta \in \mathbb{R}^d$$

Soit $\{X_2, \ldots, X_n\}$ un échantillon de X.

Définition

On appelle estimateur de $h(\theta)$ obtenu par **la méthode des moments** la statistique:

$$T_n(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

2.2. L'estimateur de la moyenne.

Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires réelles, indépendantes et de même loi, d'espérance μ et de variance σ^2. On souhaite estimer l'espérance μ, supposée inconnue.

Pour estimer le paramètre $\mu = \mathbb{E}[X_1]$, on utilise la méthode des moments avec $h = Id$.

Estimation de la moyenne

Pour estimer l'espérance μ, on prend la moyenne empirique des (X_j), notée $\overline{X}_n$ et définie par :

$$\overline{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j$$

2.3 Propriétés.

1. L'estimateur $\overline{X}_n$ est un estimateur sans biais du paramètre μ.

2. Cet estimateur est convergent:
   - en moyenne quadratique.
   - en loi suivant:

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

   - en probabilité.

3.1. Estimation de la variance d'un échantillon.

Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires réelles, indépendantes et de même loi, d'espérance μ et de variance σ^2.

On suppose que μ et σ^2 sont inconnus.

On souhaite estimer la variance σ^2.

3.2. Construction de l'estimateur.

On a $\sigma^2 = \mathbb{V}\text{ar}(X_1) = \mathbb{E}\left((X_1 - \mu)^2\right)$.

- Supposons ds un premier temps μ connu. Estimer σ^2, c'est en fait estimer une espérance, et donc un estimateur naturel est:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2$$

- Cependant, en génèral μ est inconnu. On l'estime par $\overline{X}_n$ et un estimateur de la variance σ^2 est alors donné par :

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 = \frac{1}{n}\sum_{j=1}^{n}X_j^2 \; - \; \overline{X}_n^2.$$

Cette méthode, qui consiste à remplacer un paramètre par son estimateur, est dite "du plug-in".

Remarque. En écrivant la variance σ^2 sous la forme $\sigma^2 = \mathbb{E}(X_1^2) - (\mathbb{E}(X_1))^2$, nous serions directement tombés sur cet estimateur.

3.3. Propriétés.

- Cet estimateur est biaisé, mais asymptotiquement sans biais et convergent.

  En effet, à partir de la décomposition suivante,

$$\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 = \sum_{j=1}^{n}(X_j - \mu)^2 - n(\overline{X}_n - \mu)^2 \qquad (1)$$

  on montre facilement que: $\mathbb{E}\left(\widehat{\sigma}^2\right) = \dfrac{(n-1)}{n}\,\sigma^2$

- Si les variables $(X_n)_{n\geq 1}$ possèdent un moment d'ordre 4 fini, c'est à dire si $\mathbb{E}(X_n^4) < \infty$ pour tout $n \geq 1$, alors, on a le théorème de limite centrale suivant :

$$\sqrt{n}\left(\widehat{\sigma}^2 - \sigma^2\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \tau^4 - \sigma^4) \qquad (2)$$

  où $\tau^4 = \mathbb{E}\left[(X_1 - \mu)^4\right]$.

3.4. Eléments de preuve.

On établit la convergence presque sûre de $\widehat{\sigma}^2$ vers σ^2 en utilisant la décomposition en somme de carrés précédente et le fait que $\overline{X}_n \xrightarrow[n\to\infty]{p.s.} \mu$ et que grâce à la loi forte des grands nombres, pour la suite de variables aléatoires $\left((X_n-\mu)^2\right)_{n\geq 1}$, on a

$$\frac{1}{n}\sum_{j=1}^{n}(X_j-\mu)^2 \xrightarrow[n\to\infty]{p.s.} \mathbb{E}\left[(X_1-\mu)^2\right]=\sigma^2$$

On démontre le TLC à partir de la décomposition en somme de carrés précédente, et en utilisant le fait que, pour la suite de variables aléatoires $\left((X_n-\mu)^2\right)_{n\geq 1}$, on a le TLC suivant :

$$\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}(X_j-\mu)^2-\sigma^2\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,\tau^4-\sigma^4), \qquad (3)$$

et le fait que grâce à l'inégalité de Markov, on a $\sqrt{n}(\overline{X}_n-\mu)^2 \xrightarrow[n\to\infty]{P} 0$.

Calcul de la variance

A partir de la décomposition

$$\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 = \sum_{j=1}^{n}(X_j - \mu)^2 - n(\overline{X}_n - \mu)^2 \qquad (4)$$

on déduit que

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\widehat{\sigma}^2) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2\right) + \mathbb{V}\mathrm{ar}\left((\overline{X}_n - \mu)^2\right) \\
&- \frac{2}{n}\mathbb{C}\mathrm{ov}\left(\sum_{j=1}^{n}(X_j - \mu)^2 ; (\overline{X}_n - \mu)^2\right) \\
&= A + B - 2C
\end{aligned}$$

Puisque les variables (X_j) sont indépendandtes et de même loi, on a

$$A = \frac{1}{n}\mathbb{V}\mathrm{ar}\left((X_1 - \mu)^2\right) = \frac{\tau^4 - \sigma^4}{n} \qquad (5)$$

Calcul de la variance

On a

$$C = \frac{1}{n}\sum_{j=1}^{n}\mathbb{C}\mathrm{ov}\left((X_j-\mu)^2;(\overline{X}_n-\mu)^2\right) \quad (6)$$

$$= \mathbb{C}\mathrm{ov}\left((X_1-\mu)^2;(\overline{X}_n-\mu)^2\right) \quad (7)$$

$$= \frac{1}{n^2}\mathbb{C}\mathrm{ov}\left((X_1-\mu)^2;\left(\sum_{j=1}^{n}(X_j-\mu)\right)^2\right) \quad (8)$$

Or

$$\left(\sum_{j=1}^{n}(X_j-\mu)\right)^2 = \sum_{j=1}^{n}(X_j-\mu)^2 + 2\sum_{j<k}(X_j-\mu)(X_k-\mu)$$

Comme les variables (X_j) sont indépendantes, on a

$$\mathbb{C}\mathrm{ov}\left((X_1-\mu)^2;(X_j-\mu)^2\right)=0$$

Calcul de la variance

et pour tout $1 \leq j < k \leq n$,

$$\mathbb{C}\text{ov}\left((X_1-\mu)^2\,;\,(X_j-\mu)(X_k-\mu)\right)=0$$

Ainsi, en rassemblant les différents résultats, on obtient que

$$C=\frac{\mathbb{V}\text{ar}((X_1-\mu)^2)}{n^2}=\frac{\tau^4-\sigma^4}{n^2}$$

Il reste maintenant à calculer $B=\mathbb{V}\text{ar}\left((\overline{X}_n-\mu)^2\right)$. On a

$$\begin{aligned} B &= \mathbb{C}\text{ov}\left((\overline{X}_n-\mu)^2\,;\,(\overline{X}_n-\mu)^2\right) && (9)\\ &= \frac{1}{n^4}\mathbb{C}\text{ov}\left(\left(\sum_{j=1}^{n}(X_j-\mu)\right)^2\,;\,\left(\sum_{j=1}^{n}(X_j-\mu)\right)^2\right) && (10)\end{aligned}$$

en développant et en utilisant le fait que les (X_j) sont indépendantes, on montre facilement que

$$B=\frac{\tau^4+2\sigma^4}{n^3}$$

3.5. L'estimateur sans biais de la variance (Corrigé).

Cependant, lorsque la taille de l'échantillon est petite, on préfère prendre l'estimateur sans biais,
Généralement noté S^2 et défini par :

$$S^2 = \frac{1}{(n-1)} \sum_{j=1}^{n} (X_j - \overline{X}_n)^2 \tag{11}$$

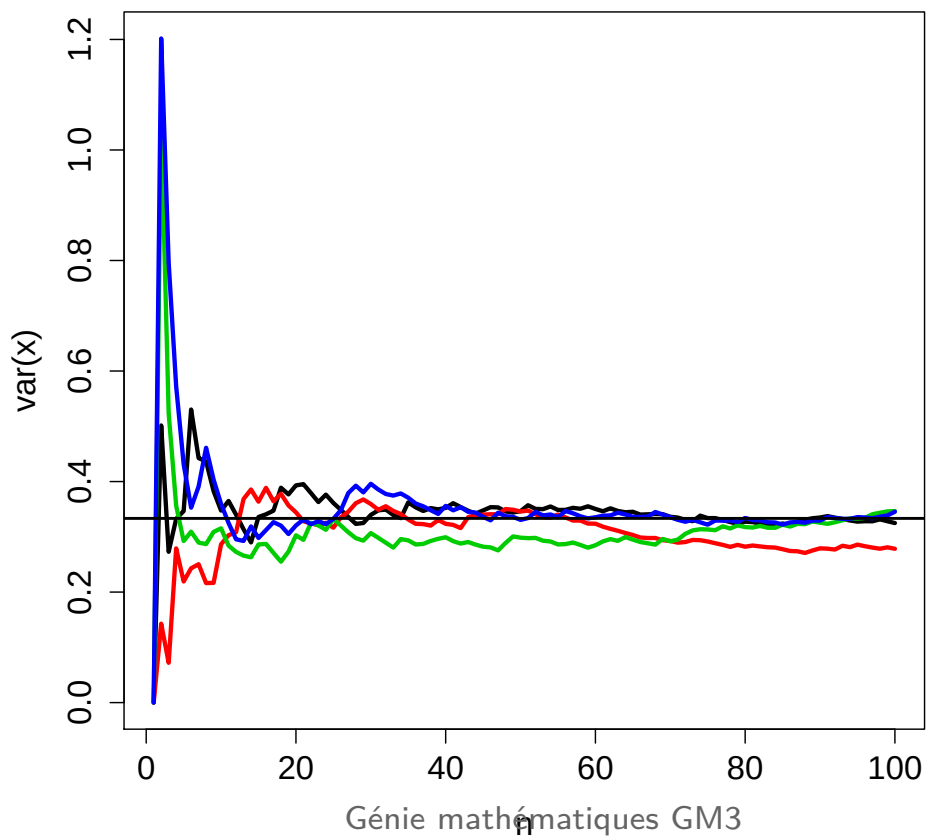
On peut montrer que, si les variables $X_1, \ldots, X_n$ possèdent un moment d'ordre 4 fini (ie. $\mathbb{E}(X_1^4) < \infty$), alors

$$\mathbb{V}\text{ar}(S^2) = \frac{1}{n} \left(\tau^4 - \sigma^4 + \frac{2}{(n-1)} \sigma^4 \right) \underset{n \to \infty}{\longrightarrow} 0$$

où l'on a noté $\tau^4 = \mathbb{E}\left((X_1 - \mu)^4\right)$.

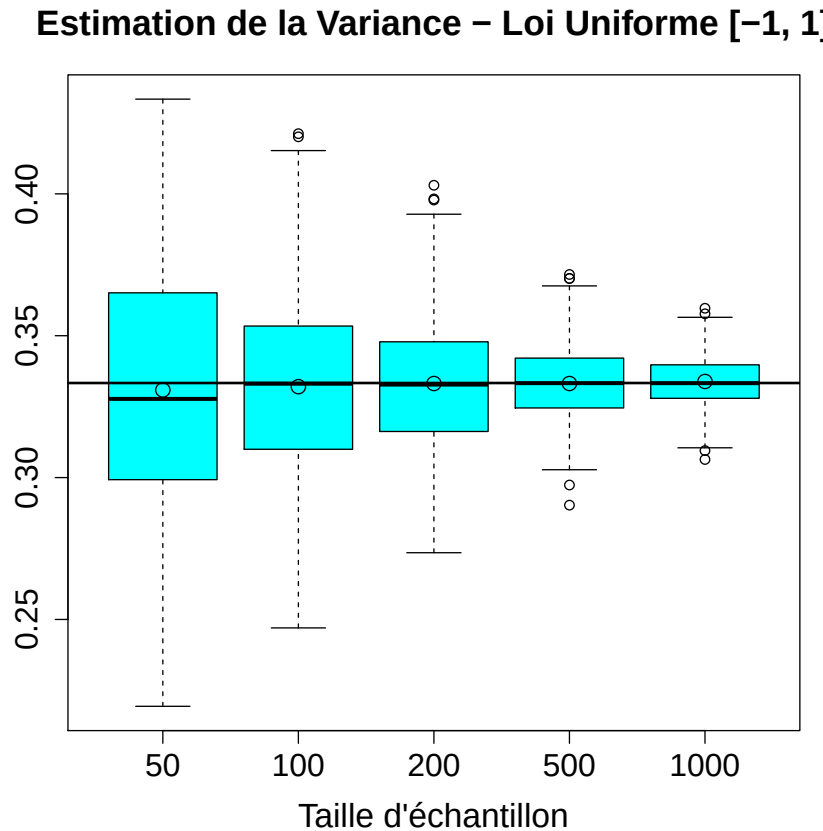
3.6. Etude par simulation dans le cas de la loi uniforme.

On reprend les 4 échantillons simulés dans la partie précédente et on calcule pour chaque échantillon, la suite des variances empiriques successives $s_2^2, s_3^2, \ldots, s_{100}^2$. (avec $s_n^2 = (1/(n-1)) \sum_{j=1}^{n} (x_j - \bar{x}_n)^2$) que l'on représente dans le graphique ci-dessous.



3.7. Etude par simulations du biais et de la variance.

On simule 400 échantillons de 1000 réalisations indépendantes d'une loi uniforme sur $[-1, 1]$. On construit pour chaque échantillon la suite des variances empiriques successives. On trouvera dans le graphique ci-dessous les boites à moustaches des 400 estimations pour les tailles d'échantillon $n = 50, 100, 200, 500$ et 1000.



4.8. Commentaires.

On constate que la taille des boites à moustaches se réduit avec l'augmentation de la taille de l'échantillon. La moyenne des 400 estimations est proche de 1/3, ce qui illustre le caractère sans biais de l'estimateur.

5. Estimation de l'écart-type d'un échantillon.
5.1. Le cadre et le problème.

Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires réelles, indépendantes et de même loi, d'espérance μ et de variance σ^2.

On suppose que μ et σ^2 sont inconnus.

On souhaite estimer l'écart-type σ.

5.2. Construction de l'estimateur.

A priori, la construction d'un estimateur de σ est simple puisque l'on dispose d'un estimateur sans biais de σ^2.

On peut en effet prendre, pour estimer σ,

$$S = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2} \tag{12}$$

puisque S^2 est un estimateur sans biais de σ^2. On rappelle que S^2 est défini par :

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2. \tag{13}$$

5.3. Propriétés.

Cet estimateur est biaisé, mais asymptotiquement sans biais et convergent.

On peut de plus montrer en utilisant la delta-méthode (en prenant $g(x) = \sqrt{x}$) que :

$$\sqrt{n}(S - \sigma) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau^4 - \sigma^4}{4\sigma^2}\right)$$

5.4. Quelques éléments de preuve.

Comme $S = \sqrt{\frac{n}{(n-1)}\widehat{\sigma}_n^2}$ et que $\widehat{\sigma}_n^2 \xrightarrow[n\to\infty]{p.s.} \sigma^2$, on déduit facilement que $S \xrightarrow[n\to\infty]{p.s.} \sigma$.

Par ailleurs, on a le TLC

$$\sqrt{n}(S^2 - \sigma^2) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \tau^4 - \sigma^4)$$

Soit g la fonction définie pour tout $x \in \mathbb{R}^+$ par $g(x) = \sqrt{x}$.

La fonction g est dérivable sur $\mathbb{R}^{*+}$ et on a $g'(x) = \frac{1}{2\sqrt{x}}$.

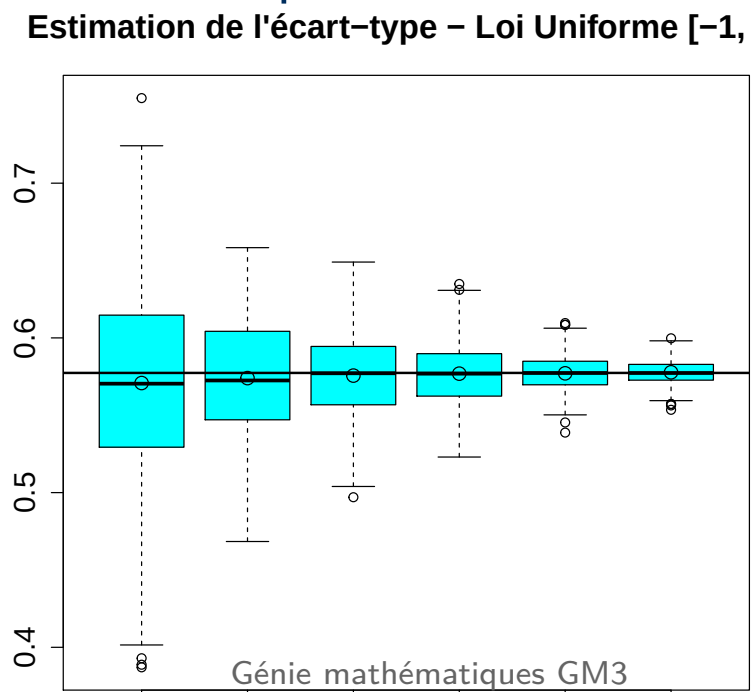
Puisque $g'(\sigma^2) \neq 0$, la delta-méthode s'applique et on a :

$$\sqrt{n}\left(g(S^2) - g(\sigma^2)\right) = \sqrt{n}(S - \sigma) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, (g'(\sigma^2))^2(\tau^4 - \sigma^4) = \frac{\tau 4 - \sigma^4}{4\sigma^2}\right)$$

5.5. Etude par simulation dans le cas de la loi uniforme.

5.5.1. Etude du biais et de la variance.

On simule 400 échantillons de 1000 réalisations indépendantes d'une loi uniforme sur $[-1, 1]$. On construit pour chaque échantillon la suite des écart-types empiriques successifs. On trouvera dans le graphique ci-dessous les boites à moustaches des 400 estimations pour les tailles d'échantillon $n = 50, 100, 200, 500$ et 1000. Qu'observe-t-on et qu'illustre-t-on ?



5.5.2. Commentaires.

L'examen du graphique montre que :

- la taille des boites à moustaches se réduit avec l'augmentation de la taille de l'échantillon. Ceci illlustre le fait que la variance de l'estimateur décroit avec l'augmentation de la taille de l'échantillon (la variance tend vers 0 lorsque n tend vers l'infini).
- pour une petite taille déchantillon, la moyenne (représentée par un cercle) est légèrement éloignée de l'écart-type théorique, ce qui illustre le fait que l'estimateur est biaisé. Lorsque la taille de l'échantillon augmente, la moyenne se rapproche de la valeur théorique ce qui illustre le fait que l'estimateur est asymptotiquement sans biais.

6. Estimation ponctuelle d'une proportion.
6.1. Le cadre et le problème.

Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires à valeurs dans un ensemble E, indépendantes et de même loi.

Soit A un sous-ensemble de E tel que $\mathbb{P}[X_1 \in A] \neq 0$. On souhaite estimer cette probabilité inconnue et on la notera $p = \mathbb{P}[X_1 \in A]$. Comment estimer cette probabilité ?

6.2. Construction de l'estimateur.

Tout d'abord, remarquons que

$$p = \mathbb{P}[X_1 \in A] = \mathbb{E}[1_A(X_1)]$$

ainsi, estimer la proportion p, c'est estimer une espérance mathématique.

il est donc facile de proposer un estimateur de la proportion p. en utilisant le principe de la loi des grands nombres.

Ainsi pour estimer la proportion p, on prendra

$$\widehat{p}_{nn} = \frac{1}{n}\sum_{j=1}^{n} 1_A(X_j)$$

6.3. Propriétés.

L'estimateur $\widehat{p}_{nn}$ de la proportion est un estimateur sans biais de p et convergent en probabilité, presque sûrement et en moyenne quadratique. On a de plus

$$\mathbb{E}\left[(\widehat{p}_{nn} - p)^2\right] = \frac{p(1-p)}{n} \leq \frac{1}{4n}$$

et le théorème de limite centrale suivant :

$$\sqrt{n}\,(\widehat{p}_{nn} - p) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, p(1-p))$$

6.4. Eléments de preuve.

Pour tout $j = 1, 2, \ldots, n$, on pose $Y_j = 1_A(X_j)$.

Les variables aléatoires $Y_1, Y_2, \ldots, Y_n$ sont des variables aléatoires indépendantes et de même loi de Bernouilli de paramètre p.

L'estimateur $\widehat{p}_{nn}$ est donc une moyenne empirique de variables aléatoires indépendantes, de même loi, d'espérance p et de variance $p(1-p)$.

Les résultats de la section 3 s'appliquent alors directement.

Pour la majoration uniforme de l'erreur quadratique moyenne, il suffit de remarquer que pour tout $p \in]0, 1[$, on a $p(1-p) \leq 1/4$.