Cours 3 – Estimations : Généralités.

Eya ZOUGAR *

Institut National des Sciences appliqués-INSA

Génie mathématiques GM3
Thursday 02^{th} Febrary, 2023

*Basé sur le cours de Bruno PORTIER

Contents

1. Introduction.
1.1. Le cadre et le modèle probabiliste.

On dispose de n données réelles $x_1, x_2, \ldots, x_n$ qui sont les mesures d'une variable quantitative.

On fait l'hypothèse fondamentale que ces données sont en fait les réalisations de n variables aléatoires $X_1, X_2, \ldots, X_n$ que l'on suppose indépendantes et de même loi F.

Remarque: On fait parfois l'hypothèse que ces données sont n réalisations indépendantes d'une variable aléatoire X de loi F.

On s'intéresse à une caractéristique de cette loi F (espérance, variance, etc ...) ou bien à un paramètre de cette loi, en supposant que cette loi soit paramétrée (loi de Bernouill $\mathcal{B}(p)$, loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, etc ...)

Notons θ cette caractéristique ou ce paramètre, supposé inconnu.

1.2. Les principaux objectifs.

On souhaite:

- construire une estimation de ce paramètre θ à partir des observations $x_1, \ldots, x_n$.
- Cette estimation n'estime pas autre chose que ce que l'on souhaite estimer (notion de biais), soit proche de la vraie valeur du paramètre et que la qualité de l'estimation s'améliore avec le nombre de données (notion de convergence).
- On peut vouloir aussi avoir des informations sur la confiance que l'on peut accorder à cette estimation (notion d'intervalle de confiance).
- On peut aussi avoir une idée a priori de la valeur du paramètre et vouloir vérifier si cette hypothèse est vraie (notion de test statistique).
- Nous allons dans ce cours introduire les notions qui nous permettrons de répondre à ces différentes questions.

2. Définitions des principales notions.
2.1. C'est quoi un estimateur ?

Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires indépendantes et de même loi, et soient $x_1, x_2, \ldots, x_n$ une réalisation de cet échantillon.

Un estimateur T_n du paramètre θ [†] est une fonction mesurable des variables $X_1, \ldots, X_n$, c'est à dire de la forme

$$T_n = T(X_1, X_2, \ldots, X_n)$$

On appelle Estimation la réalisation sur les données de cet estimateur, c'est à dire la valeur $t_n = T(x_1, x_2, \ldots, x_n)$.

[†] On note θ cette caractéristique ou ce paramètre, que l'on suppose inconnu et que l'on souhaite estimer.

2.2. Notion de biais.

Soit T_n un estimateur du paramètre réel θ.

- En statistique, on souhaitera que cet estimateur soit:
  - sans biais, c'est à dire que

$$\mathbb{E}[T_n] = \theta;$$

  - ou asymptotiquement sans biais, c'est à dire $(\mathbb{E}[T_n] \underset{n\to\infty}{\longrightarrow} \theta)$;
- On appelle Biais de l'estimateur la quantité :

$$b(T_n) = \mathbb{E}[T_n] - \theta$$

2.3. 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.

2.4. 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.

2.5. 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.

2.6. 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

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

2.7. 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 :

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

3. Estimation de la moyenne.
3.1. Introduction : la méthode des moments.

En statistique, la construction de nombreux estimateurs repose sur le résultat probabiliste suivant (loi des grands nombres).

Si le paramètre inconnu θ auquel on s'intéresse s'écrit sous la forme d'une espérance, il est facile de proposer un estimateur sans biais et convergent.

En effet, si $\theta = \mathbb{E}[Y_1]$, alors $T_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$ est un estimateur sans biais et convergent de θ.

3.2. Construction de 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 le résultat précédent (LGN) avec $Y_n = X_n$.

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

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

3.3 Propriétés.

1. L'estimateur $\overline{X}_n$ est un estimateur sans biais du paramètre μ.
   En effet, on montre facilement

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}(X_j) = \frac{1}{n}\sum_{j=1}^{n}\mu = \mu$$

2. Sa variance tend vers 0 lorsque n tend vers l'infini. En effet:

$$\mathbb{V}\text{ar}(\overline{X}_n) = \mathbb{V}\text{ar}\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{V}\text{ar}(X_j) = \frac{\sigma^2}{n}$$

3. Cet estimateur est donc convergent.
   De plus, puisque les hypothèses du TLC son tsatisfaites, on a aussi le résultat de convergence en loi suivant :

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

   Ce résultat de convergence en loi peut être utilisé pour construire un intervalle de confiance pour le paramètre μ.

3.4. Etude par simulations dans le cas de la loi uniforme.

3.4.1 Introduction: pk faire des simulations.

L'objectif du travail de simulation consistera à illustrer certains résultats théoriques.
On pourra visualiser sur quelques trajectoires :

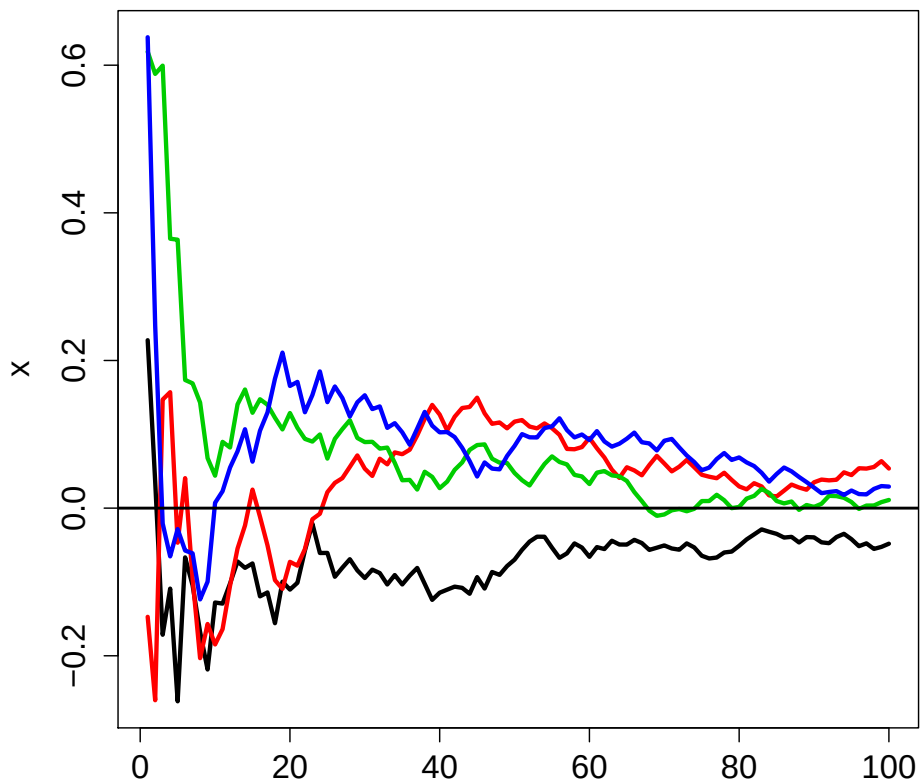
- La convergence de l'estimateur.
- L'effet des fluctuations d'échantillonnage.

Avec un plus grand nombre d'estimations, et plusieurs tailles d'échantillon, on pourra vérifier de manière empirique que :

- l'estimateur est sans biais ou non.
  On calculera la moyenne des estimations qui correspondra à une estimation de $\mathbb{E}[T_n]$, et on la comparera à la valeur du paramètre.
- que la variance de l'estimateur tend bien vers 0.

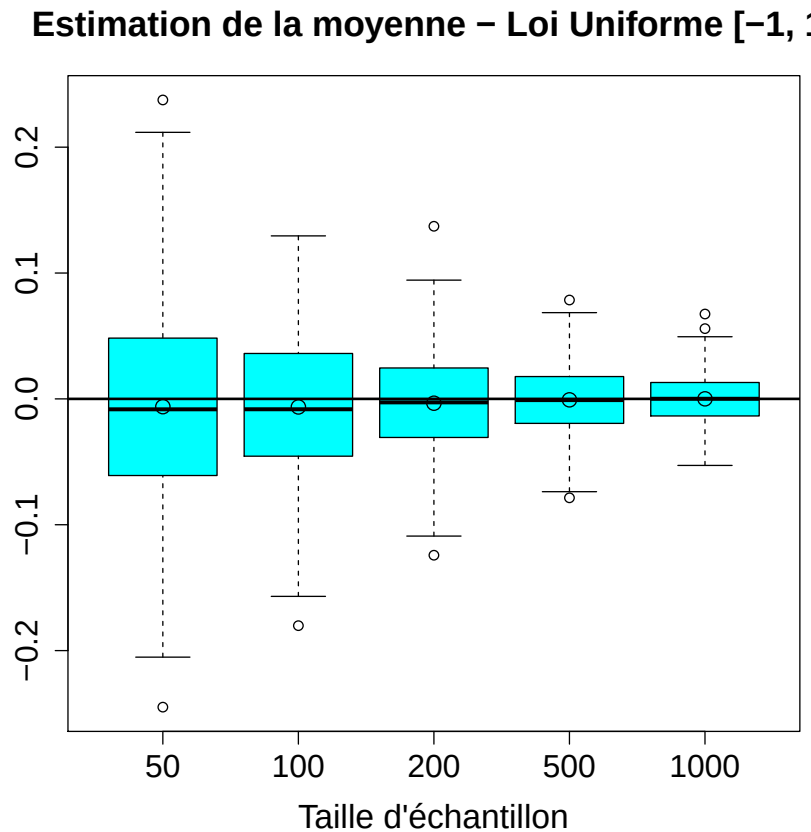
3.4.2. Protocole et étude des fluctuations d'échantillonnage.

On souhaite étudier le comportement de $\overline{X}_n$ dans le cas de la loi uniforme sur $[-1, 1]$, qui est d'espérance nulle. On simule 4 échantillons de 100 réalisations indépendantes. On construit pour chaque échantillon la suite des moyennes empiriques successives que l'on représente dans le graphique ci-dessous.



3.4.3. 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 moyennes 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.



3.4.4. Commentaires.

1. On constate que la taille des boites à moustaches se réduit avec l'augmentation de la taille de l'échantillon, ce qui illustre le fait que la variance de l'estimateur diminue avec l'augmentation de la taille n de l'échantillon.
2. La moyenne des 400 estimations est proche de 0, ce qui illustre le caractère sans biais de l'estimateur.

4. Estimation de la variance d'un échantillon.
4.1. 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 la variance σ^2.

4.2. 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:

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

- 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 :

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 = \frac{1}{n}\sum_{j=1}^{n}X_j^2 \; - \; \overline{X}_n^2. \tag{3}$$

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.

4.3. Propriétés.

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

  En effet, à partir de la décomposition suivante,

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

  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 :

$$\sqrt{n}\left(\widehat{\sigma}^2 - \sigma^2\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \tau^4 - \sigma^4) \qquad (5)$$

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

4.4. 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 :

$$\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}(X_j-\mu)^2-\sigma^2\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,\tau^4-\sigma^4), \qquad (6)$$

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

4.5. L'estimateur sans biais de la variance.

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

$$S^2 = \frac{1}{(n-1)} \sum_{j=1}^{n} (X_j - \overline{X}_n)^2 \tag{7}$$

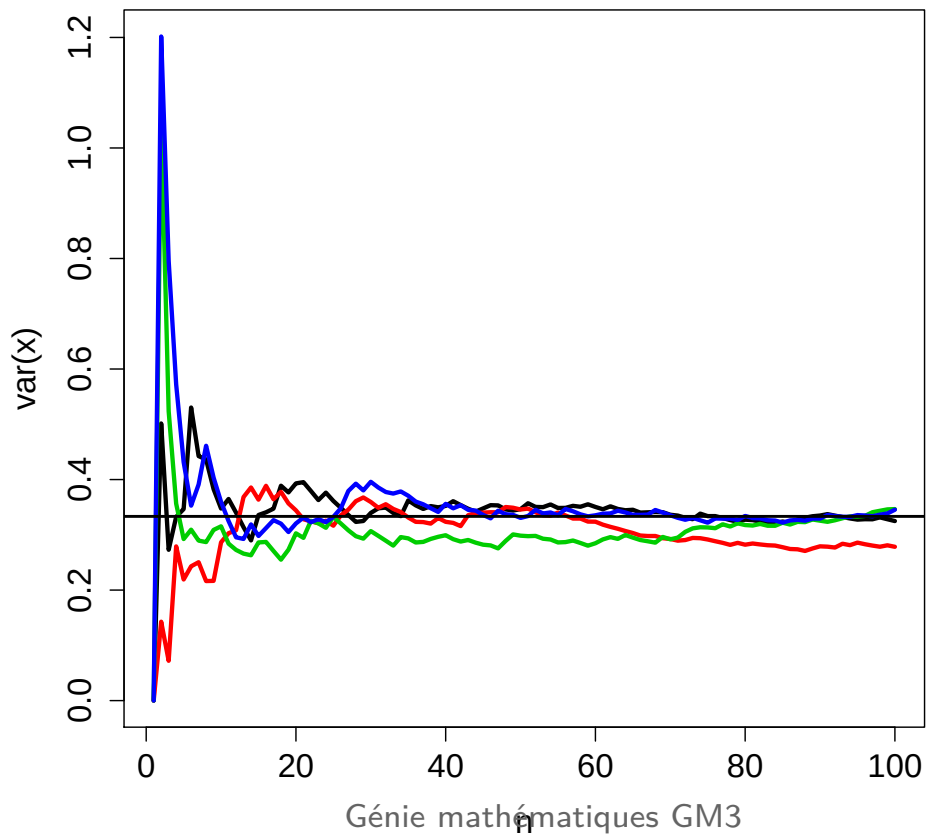
On peut montrer que, si les variables $X_1, \ldots, X_n$ possèdent un moment d'ordre 4 fini (ie. $\mathbb{E}(X_1^4) < \infty$), alors

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

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

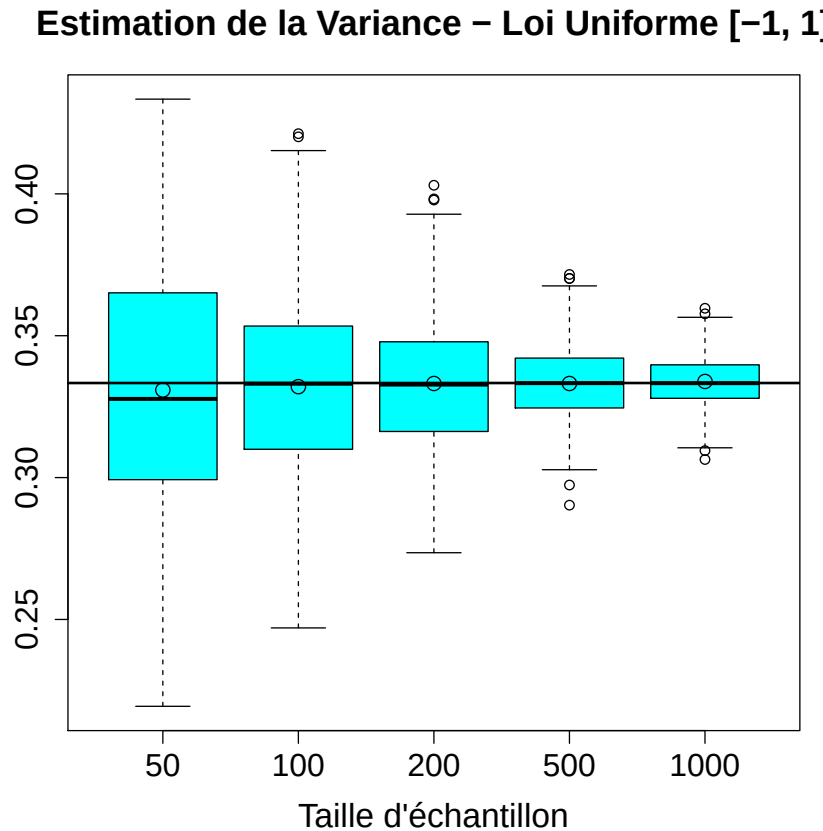
4.6. Etude par simulation dans le cas de la loi uniforme.

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.



4.7. 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.