# Colles

semaine 23 : 14 mars - 19 mars

# I. Estimation ponctuelle

## I.1. Notion de $n$-échantillon

Soit $X$ une v.a.r.

Soit $n \in \mathbb{N}^*$.

- On appelle $n$-échantillon de la v.a.r. $X$ tout $n$-uplet $(X_1, \dots, X_n)$ de v.a.r. :
  - × indépendantes,
  - × de même loi, la loi de $X$.

## I.2. Estimateur et estimation

### I.2.a) Définition

Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$.

Soit $n \in \mathbb{N}^*$ et soit $(X_1, \dots, X_n)$ un $n$-échantillon de la v.a.r. $X$.

- On appelle **estimateur de** $\theta$ toute v.a.r. $T_n$, qui s'exprime en fonction des v.a.r. $(X_1, \dots, X_n)$ et dont l'expression ne fait pas apparaître le paramèrtre $\theta$.
- Autrement dit, $T_n$ est un estimateur de $\theta$ s'il existe une fonction $\varphi$ de $n$ variables telle que :
$$T_n = \varphi(X_1, \dots, X_n)$$
- Si $\varphi$ permet de définir un estimateur de $\theta$ alors, on appelle **estimation de** $\theta$ tout réel de la forme :
$$\varphi(x_1, \dots, x_n)$$
où $(x_1, \dots, x_n) \in X_1(\Omega) \times \dots \times X_n(\Omega)$.
- Lorsque l'estimateur $T_n$ possède une espérance (resp. une variance), on la note $\mathbb{E}_\theta(T_n)$ (resp. $\mathbb{V}_\theta(T_n)$) *(lire espérance / variance sous $\theta$).*

**Remarque**
Le fait de noter $\mathbb{E}_\theta(T_n)$ l'espérance de $T_n$ rappelle que celle-ci dépend a priori du paramètre $\theta$. Il en va évidemment de même de la loi de probabilité de la variable $T_n$ : un événement (par exemple le succès dans une épreuve de Bernoulli) prend des probabilités différentes selon la valeur de $\theta$. En toute rigueur, il convient donc de l'indiquer en notant $\mathbb{P}_\theta$ l'application probabilité en jeu, car elle dépend à son tour de la valeur du paramètre estimé. Un estimateur $T_n$ a donc une dépendance en $\theta$ (issu de la dépendance en $\theta$ de la loi de $X$). Pour autant, cela n'autorise pas à faire apparaître $\theta$ dans l'expression de $T_n$ (c'est une contrainte différente).

**Exemple** *(d'estimateurs)*

Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$.

Soit $n \in \mathbb{N}^*$ et soit $(X_1, \ldots, X_n)$ un $n$-échantillon de la v.a.r. $X$.

- Les v.a.r. $S_n = X_1 + \ldots + X_n$ et $\overline{X_n}$ sont des estimateurs de $\theta$.
  *(attention, l'expression de la v.a.r. $\overline{X}_n^*$ peut dépendre de $\theta$ !)*
- La v.a.r. $X_1 \times \ldots \times X_n$ est un estimateur de $\theta$.
- Si $(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$, la v.a.r. $\sum_{i=1}^{n} \lambda_i \cdot X_i$ est un estimateur de $\theta$.
- Les v.a.r. $X_1$ et $X_2 - \sqrt{X_1 + X_n}$ sont des estimateurs de $\theta$.

Par contre :

$\times$ la v.a.r. constante égale à $\theta$ n'est pas un estimateur de $\theta$.

$\times$ la v.a.r. $X_1 + \ldots + X_n - \theta$ n'est pas un estimateur de $\theta$.

En effet, ces v.a.r. font apparaître $\theta$ dans leur expression ce qui est interdit dans la définition.

### I.2.b) Un exemple classique d'estimateur

Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$.

Soit $(X_1, \ldots, X_n)$ un $n$-échantillon de la v.a.r. $X$.

- Alors la v.a.r. $\overline{X}_n = \dfrac{1}{n} \sum_{i=1}^{n} X_i$ est un estimateur de $\theta$.
- Cet estimateur est appelée **moyenne empirique** de l'échantillon $(X_1, \ldots, X_n)$.

## I.3. Qualité de l'estimateur

### I.3.a) Biais d'un estimateur

Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$.

Soit $n \in \mathbb{N}^*$ et soit $T_n$ un estimateur de $\theta$.

On suppose que $T_n$ admet une espérance.

- Si $T_n$ admet une espérance, on appelle **biais de l'estimateur** $T_n$ le réel :

$$\boxed{b_\theta(T_n) = \mathbb{E}_\theta\big(T_n - \theta\big) = \mathbb{E}_\theta(T_n) - \theta}$$

- Si le paramètre estimé est $g(\theta)$, l'expression du biais devient :

$$b_\theta(T_n) = \mathbb{E}_\theta\big(T_n - g(\theta)\big) = \mathbb{E}_\theta(T_n) - g(\theta)$$

- On dit que $T_n$ est un **estimateur sans biais de** $\theta$ lorsque $b_\theta(T_n) = 0$, ou encore : $\mathbb{E}_\theta(T_n) = \theta$. Dans le cas contraire, on parlera d'estimateur biaisé.

### I.3.b) Estimateur asymptotiquement sans biais

Soit $X$ unve v.a.r. dont la loi dépend d'un paramètre $\theta$.

Soit $(T_n)_{n\in\mathbb{N}^*}$ une d'estimateurs de $\theta$ (resp. $g(\theta)$).

On suppose que pour tout $n \in \mathbb{N}^*$, $T_n$ admet une espérance.

- On dit que la suite d'estimateurs $(T_n)_{n\in\mathbb{N}^*}$ est **asymptotiquement sans biais** lorsque :

$$\lim_{n\to+\infty} b_\theta(T_n) = 0 \quad \text{ou encore} \quad \lim_{n\to+\infty} \mathbb{E}_\theta(T_n) = \theta$$

$$(\text{resp. } \lim_{n\to+\infty} \mathbb{E}_\theta(T_n) = g(\theta))$$

- On dit aussi, par abus de langage, que l'estimateur $T_n$ est asymptotiquement sans biais.

### I.3.c) Risque quadratique

Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$ (resp. $g(\theta)$).

Soit $n \in \mathbb{N}^*$ et soit $T_n$ un estimateur de $\theta$.

On suppose que $T_n$ admet un moment d'ordre 2.

- On appelle **risque quadratique de l'estimateur** $T_n$ le réel :

$$\boxed{r_\theta(T_n) \;=\; \mathbb{E}_\theta\left((T_n-\theta)^2\right)} \quad (\text{resp. } \mathbb{E}_\theta\left((T_n-g(\theta))^2\right))$$

`À RETENIR`

Entre deux estimateurs de $\theta$, on préférera toujours celui dont le risque quadratique est le plus faible.

Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$ (resp. $g(\theta)$).

Soit $n \in \mathbb{N}^*$ et soit $T_n$ un estimateur de $\theta$.

On suppose que $T_n$ admet un moment d'ordre 2.

***1)*** On a alors :

$$\boxed{r_\theta(T_n) \;=\; \mathbb{V}_\theta(T_n) + \big(b_\theta(T_n)\big)^2}$$

***2)*** Si on suppose de plus que $T_n$ est sans biais alors :

$$\boxed{r_\theta(T_n) \;=\; \mathbb{V}_\theta(T_n)}$$

*Démonstration.*

- Comme la v.a.r. $T_n$ admet un moment d'ordre 2, alors elle admet un risque quadratique.
- De plus :

$$\begin{aligned}
& \mathbb{V}_\theta(T_n) + \big(b_\theta(T_n)\big)^2 \\
= \ & \mathbb{E}_\theta\big((T_n)^2\big) - \big(\mathbb{E}_\theta(T_n)\big)^2 + \big(\mathbb{E}_\theta(T_n) - \theta\big)^2 && \textit{(par la formule de Kœnïg-Huygens)} \\
= \ & \mathbb{E}_\theta\big((T_n)^2\big) - \cancel{\big(\mathbb{E}_\theta(T_n)\big)^2} + \cancel{\big(\mathbb{E}_\theta(T_n)\big)^2} - 2\,\theta\,\mathbb{E}_\theta(T_n) + \theta^2 \\
= \ & \mathbb{E}_\theta\big(T_n^2 - 2\,\theta\,T_n + \theta^2\big) && \textit{(par linéarité de l'espérance)} \\
= \ & \mathbb{E}_\theta\big((T_n - \theta)^2\big) \\
= \ & r_\theta(T_n)
\end{aligned}$$

□

## I.4. Estimateur convergent

### I.4.a) Définition

Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$.

Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de $\theta$ (resp. $g(\theta)$).

- On dit que la suite d'estimateurs $(T_n)_{n\in\mathbb{N}^*}$ de $\theta$ (resp. $g(\theta)$) est convergente si :

$$\boxed{\forall \varepsilon > 0,\ \lim_{n\to+\infty} \mathbb{P}\big(\,[\ |T_n - \theta\ | > \varepsilon\ ]\,\big) = 0}$$

$$\Big(\text{resp. } \forall \varepsilon > 0,\ \lim_{n\to+\infty} \mathbb{P}\big(\,[\ |T_n - g(\theta)\ | > \varepsilon\ ]\big) = 0\Big)$$

On dit aussi, par abus de langage, que l'estimateur $T_n$ est convergent.

- On peut réécrire la propriété précédente :

$$\boxed{\text{L'estimateur } T_n \text{ est convergent} \quad \Leftrightarrow \quad T_n \xrightarrow[n\to+\infty]{\mathbb{P}} \theta}$$

Ainsi, un estimateur convergent est un estimateur qui converge vers le paramètre à estimer.

### I.4.b) Démontrer en pratique qu'un estimateur est convergent

Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$ (resp. $g(\theta)$).

Soit $n \in \mathbb{N}^*$ et soit $T_n$ un estimateur de $\theta$.

On suppose que $T_n$ admet un moment d'ordre 2.

$$\boxed{\lim_{n\to+\infty} r_\theta(T_n) = 0 \quad \Rightarrow \quad T_n \text{ est un estimateur convergent}}$$

**Remarque**
La propriété annoncée par ce théorème est très naturelle. On suppose que la limite du risque quadratique est nulle. Autrement dit, asymptotiquement parlant (en $+\infty$), l'estimateur $T_n$ a un écart quadratique nul avec le paramètre $\theta$ à estimer. On en déduit que $T_n$ est asymptotiquement égal à $\theta$. Cela s'énonce rigoureusement par le fait que $T_n$ converge (en probabilité) vers $\theta$.

*Démonstration.*
Soit $\varepsilon > 0$. La v.a.r. $(T_n - \theta)^2$ :

× admet une espérance,

× est à valeurs positives.

Ainsi, par inégalité de Markov (avec $a = \varepsilon^2$) :

$$\mathbb{P}_\theta\left(\left[(T_n-\theta)^2 > \varepsilon^2\right]\right) \leqslant \frac{\mathbb{E}_\theta((T_n-\theta)^2)}{\varepsilon^2}$$

Or, par stricte croissance de $x \mapsto \sqrt{x}$ sur $[0, +\infty[$ :

$$\mathbb{P}_\theta\left(\left[(T_n-\theta)^2 > \varepsilon^2\right]\right) = \mathbb{P}_\theta\big([|T_n-\theta| > \varepsilon]\big)$$

On en déduit :

$$0 \leqslant \mathbb{P}_\theta\big([|T_n-\theta| > \varepsilon]\big) \leqslant \frac{\mathbb{E}_\theta((T_n-\theta)^2)}{\varepsilon^2} = \frac{r_\theta(T_n)}{\varepsilon^2}$$

Or :

× $\lim\limits_{n\to+\infty} 0 = 0$,

× $\lim\limits_{n\to+\infty} \dfrac{r_\theta(T_n)}{\varepsilon^2} = 0$.

Ainsi, par théorème d'encadrement : $\lim\limits_{n\to+\infty} \mathbb{P}_\theta([|T_n-\theta| > \varepsilon]) = 0$. □

### I.4.c) Estimation de l'espérance

Soit $X$ une v.a.r. admettant une espérance $m$ (paramètre inconnu à estimer) et une variance (notée $\sigma^2$).

Soit $n \in \mathbb{N}^*$ et soit $(X_1, \ldots, X_n)$ un $n$-échantillon de la v.a.r. $X$.

Alors la moyenne empirique $\overline{X_n} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i$ est un estimateur sans biais et convergent du paramètre $m$.

# II. Intervalle de confiance

## II.1. Intervalle de confiance (exact)

Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$ (à estimer).

Soient $(U_n)_{n\in\mathbb{N}^*}$ et $(V_n)_{n\in\mathbb{N}^*}$ deux suites d'estimateurs de $\theta$ (resp. $g(\theta)$).

Soit $\alpha \in [0,1]$.

On suppose de plus que : $\mathbb{P}_\theta\big([U_n \leqslant V_n]\big) = 1$.

- On dit que $[U_n, V_n]$ est un intervalle de confiance de $\theta$ (resp. $g(\theta)$) au niveau de confiance $1-\alpha$ si :

$$\mathbb{P}_\theta\Big([U_n \leqslant \theta \leqslant V_n]\Big) \geqslant 1-\alpha$$

$$\text{II}$$

$$\mathbb{P}_\theta\big([\theta \in [U_n, V_n]\,]\big)$$

$$\Big(\text{resp. } \mathbb{P}_\theta\Big([U_n \leqslant g(\theta) \leqslant V_n]\Big) \geqslant 1-\alpha\Big)$$

- Le réel $\alpha \in [0,1]$ est appelé le niveau de risque de l'intervalle.

MÉTHODO **Utilisation de l'inégalité de Bienaymé-Tchebychev pour obtenir un intervalle de confiance d'un paramètre $\theta$ à l'aide d'un estimateur sans biais de $\theta$**

Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$ (à estimer).

Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de $\theta$.

On suppose que pour tout $n \in \mathbb{N}^*$ : $\mathbb{E}_\theta(T_n) = \theta$.
*($T_n$ est un estimateur sans biais de $\theta$)*

*1)* Soit $\varepsilon > 0$. D'après l'inégalité de Bienaymé-Tchebychev :

$$\mathbb{P}_\theta\big([\,|T_n - \mathbb{E}_\theta(T_n)| > \varepsilon\,]\big) \leqslant \frac{\mathbb{V}_\theta(T_n)}{\varepsilon^2}$$

$$\text{II}$$

$$\mathbb{P}_\theta\big([\,|T_n - \theta| > \varepsilon\,]\big)$$

*2)* Supposons qu'on arrive à majorer $\mathbb{V}_\theta(T_n)$ indépendamment de $\theta$.
Autrement dit, supposons qu'il existe une suite réelle $(v_n)$ telle que : $\forall n \in \mathbb{N}^*$, $\mathbb{V}_\theta(T_n) \leqslant v_n$. Alors on obtient :

$$\mathbb{P}_\theta\big([\,|T_n - \theta| > \varepsilon\,]\big) \leqslant \frac{v_n}{\varepsilon^2}$$

*3)* On en déduit alors :

$$1 - \mathbb{P}_\theta\big([\,|T_n - \theta| > \varepsilon\,]\big) \geqslant 1 - \frac{v_n}{\varepsilon^2}$$

$$\text{II}$$

$$\mathbb{P}_\theta\big([\,|T_n - \theta| \leqslant \varepsilon\,]\big)$$

$$\text{II}$$

$$\mathbb{P}_\theta\big([\,T_n - \varepsilon \leqslant \theta \leqslant T_n + \varepsilon\,]\big)$$

On en conclut que $[T_n - \varepsilon, T_n + \varepsilon]$ est un intervalle de confiance de $\theta$ au niveau de confiance $1 - \dfrac{v_n}{\varepsilon^2}$.

**Illustration classique** : estimation de l'espérance pour une loi de Bernoulli

Soit $X$ une v.a.r. de loi de Bernoulli $\mathcal{B}(p)$ où $p$ est un paramètre à estimer.
Soit $n \in \mathbb{N}^*$ et soit $(X_1, \dots, X_n)$ un $n$-échantillon de la v.a.r. $X$.
Notons $\overline{X_n} = \dfrac{1}{n} \sum_{i=1}^{n} X_i$.

Rappelons que la v.a.r. $\overline{X_n}$ admet une espérance et une variance. De plus :

× $\mathbb{E}(\overline{X_n}) \;=\; \mathbb{E}\left(\dfrac{1}{n} \sum_{i=1}^{n} X_i\right) \;=\; \dfrac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i) \;=\; \dfrac{1}{n} \sum_{i=1}^{n} p = p.$

× $\mathbb{V}(\overline{X_n}) \;=\; \mathbb{V}\left(\dfrac{1}{n} \sum_{i=1}^{n} X_i\right) \;=\; \dfrac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}(X_i) = \dfrac{1}{n^2} \sum_{i=1}^{n} p(1-p) \;=\; \dfrac{p(1-p)}{n}.$

*1)* Soit $\varepsilon > 0$. D'après l'inégalité de Bienaymé-Tchebychev :

$$\mathbb{P}\big(\,[\,|\overline{X_n} - p| \geqslant \varepsilon\,]\,\big) \;\leqslant\; \frac{p(1-p)}{n\,\varepsilon^2}$$

*2)* À l'aide de la majoration classique $p\,(1-p) \leqslant \dfrac{1}{4}$, on obtient :

$$\mathbb{P}\big(\,[\,|\overline{X_n} - p| > \varepsilon]\,\big) \;\leqslant\; \frac{p(1-p)}{n\,\varepsilon^2} \;\leqslant\; \frac{1}{4\,n\,\varepsilon^2}$$

*3)* On en déduit :

$$\begin{array}{ccc} 1 - \mathbb{P}\big(\,[\,|\overline{X_n} - p\,| > \varepsilon\,]\,\big) & \geqslant & 1 - \dfrac{1}{4\,n\,\varepsilon^2} \\ \shortparallel & & \\ \mathbb{P}\big(\,[\,|\overline{X_n} - p\,| \leqslant \varepsilon\,]\,\big) & & \\ \shortparallel & & \\ \mathbb{P}\big(\,[\,\overline{X_n} - \varepsilon \leqslant p \leqslant \overline{X_n} + \varepsilon\,]\,\big) & & \end{array}$$

En posant : $U_n = \overline{X_n} - \varepsilon$, $V_n = \overline{X_n} + \varepsilon$ et $\alpha = \dfrac{1}{4\,n\,\varepsilon^2}$ on en conclut :

$$\boxed{\mathbb{P}\big(\,[\,U_n \leqslant p \leqslant V_n\,]\,\big) \;\geqslant\; 1 - \alpha}$$

Ainsi, $[\overline{X_n} - \varepsilon, \overline{X_n} + \varepsilon]$ est un intervalle de confiance du paramètre $p$ au niveau de confiance $1 - \dfrac{1}{4\,n\,\varepsilon^2}$.

**Remarque**
Il y a plusieurs remarques à faire sur l'intervalle obtenu :

× cet intervalle est centré en $\overline{X_n}$.

× l'amplitude de cet intervalle est : $V_n - U_n = (\overline{X_n} + \varepsilon) - (\overline{X_n} - \varepsilon) = 2\,\varepsilon$.
Il est à noter que le réel $\varepsilon$ a été choisi en début de démonstration avec pour seule contrainte : $\varepsilon > 0$.
Ce réel $\varepsilon$ est appelé **marge d'erreur** de l'intervalle. Il est possible, avec cette méthode de produire des intervalles avec marge d'erreur fixée à l'avance.
A priori, une marge d'erreur faible est préférable.

× la marge d'erreur de l'intervalle influe directement sur le niveau de risque car : $\alpha = \frac{1}{4\,n\,\varepsilon^2}$. Ainsi, une marge d'erreur faible produit un niveau de risque élevé et un niveau de confiance faible.

Rappelons que le but d'un estimateur est de fournir une estimation. Lorsque l'on procède au sondage, on détermine une valeur $\overline{x_n}$ prise par l'estimateur $\overline{X_n}$. On peut alors considérer, grâce à l'analyse précédente, que le paramètre $p$ à estimer se trouve dans l'intervalle $[\overline{x_n} - \varepsilon, \overline{x_n} + \varepsilon]$ avec niveau de confiance $1 - \frac{1}{4\,n\,\varepsilon^2}$. Il s'agit alors de trouver un bon équilibre entre la précision de l'intervalle (mesure de l'amplitude ou de la marge d'erreur) et le niveau de confiance qu'on peut lui accorder (mesure du risque ou du niveau de confiance).

### Équilibre marge d'erreur / niveau de confiance

Explicitons le lien entre marge d'erreur et niveau de confiance.
Dans la suite, on fixe $n = 2500$ ce qui correspond à un sondage auprès de 2500 personnes.

- Pour $n = 2500$ fixé et $\alpha$ donné, on obtient les valeurs suivantes pour la marge d'erreur $\varepsilon = \dfrac{1}{\sqrt{4\,n\,\alpha}}$.

| | Calcul de la marge d'erreur $\varepsilon$ (en %) en fonction du niveau de confiance $1-\alpha$ (en %) avec $n = 2500$ fixé | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $1-\alpha$ | 70 | 75 | 80 | 85 | 90 | 95 | 97,5 | 99 |
| $\varepsilon$ | 1,8 | 2 | 2,2 | 2,6 | 3,2 | 4,5 | 6,3 | 10 |

La précision se dégrade lorsque le niveau de confiance augmente.

- Pour $n = 2500$ fixé et $\varepsilon$ donné, on obtient les valeurs suivantes pour le niveau de confiance $1 - \alpha = 1 - \dfrac{1}{4\,n\,\varepsilon^2}$.

| | Calcul du niveau de confiance $1-\alpha$ (en %) en fonction de la marge d'erreur $\varepsilon$ (en %) avec $n = 2500$ fixé | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\varepsilon$ | 0,5 | 1 | 1,5 | 2 | 2,5 | 3 | 3,5 | 4 |
| $1-\alpha$ | | 0 | 56 | 75 | 84 | 89 | 92 | 94 |

Le niveau de confiance augmente lorsque l'on dégrade la précision.

À RETENIR

- Améliorer la précision (diminuer la marge d'erreur $\varepsilon$) de l'intervalle, c'est augmenter le risque et ainsi diminuer le niveau de confiance.
- Dégrader la précision (augmenter la marge d'erreur $\varepsilon$) de l'intervalle, c'est diminuer le risque et ainsi augmenter le niveau de confiance.

**Le point de vue des instituts de sondage**

Lorsqu'un sondage est effectué, il faut systématiquement se poser la question des garanties aléatoires de précision sur lesquelles il se fonde. Pour les instituts de sondage, la question est donc de savoir combien de personnes il faut interroger pour obtenir un niveau de confiance $(1-\alpha)$ élevé et une précision importante (marge d'erreur $\varepsilon$ faible).

Dans le tableau suivant, on calcule $n = \dfrac{1}{4\,\alpha\,\varepsilon^2}$ pour différentes valeurs du couple $(\varepsilon, \alpha)$.

| | Nombre de sondés en fonction de la marge d'erreur $\varepsilon$ (en %) et du niveau de confiance $1-\alpha$ (en %) souhaités | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\varepsilon$ \ $1-\alpha$ | 70 | 75 | 80 | 85 | 90 | 95 | 97,5 | 99 |
| 0,5 | 33333 | 40000 | 50000 | 66667 | 100000 | 200000 | 400000 | 1000000 |
| 1 | 8333 | 10000 | 12500 | 16667 | 25000 | 50000 | 100000 | 250000 |
| 1,5 | 3704 | 4444 | 5556 | 7407 | 11111 | 22222 | 44444 | 111111 |
| 2 | 2083 | 2500 | 3125 | 4167 | 6250 | 12500 | 25000 | 62500 |
| 2,5 | 1333 | 1600 | 2000 | 2667 | 4000 | 8000 | 16000 | 40000 |
| 3 | 926 | 1111 | 1389 | 1852 | 2778 | 5556 | 11111 | 27778 |
| 3,5 | 680 | 816 | 1020 | 1361 | 2041 | 4082 | 8163 | 20408 |
| 4 | 521 | 625 | 781 | 1042 | 1563 | 3125 | 6250 | 15625 |

- Comme mentionné précédemment, le niveau de confiance $1-\alpha = 0,95$ est assez classique. Avec un tel niveau de confiance, on considère qu'il y a 95% de chances de tomber sur un panel standard. Lorsque c'est le cas, le paramètre réel se retrouve dans l'intervalle $[\overline{x_n} - \varepsilon, \overline{x_n} + \varepsilon]$.
  - × On peut souhaiter obtenir un résultat très précis. Pour un sondage concernant des élections, savoir qu'un candidat est évalué à 19% plus ou moins 0,5% serait idéal. Du point de vue du sondeur, cela voudrait dire interroger $n = 200000$ personnes. C'est inenvisageable pour des raisons évidentes de coût.
  - × L'institut de sondage doit alors revoir ses objectifs à la baisse. En interrogeant $n = 3125$ personnes, il assure avec une probabilité de 95% qu'un candidat est évalué à 19% plus ou moins 4%. Le coût est tout à fait envisageable mais le résultat semble alors un peu trop imprécis.
- Les résultats de ce dernier tableau démontrent que la méthode permettant d'obtenir un intervalle de confiance par inégalité de Bienaymé-Tchebychev est peu exploitable lorsque l'on cherche à obtenir des résultats relativement précis. Cela provient du fait que l'inégalité de Bienaymé-Tchebychev qui s'applique à toute v.a.r. (sans exploitation de la loi de celle-ci) est assez peu précise. Les instituts de sondage se basent sur une autre méthode consistant à obtenir un intervalle de confiance à l'aide du théorème central limite. Ce théorème énonce un résultat de convergence en loi. On sait que cette convergence se fait rapidement (des valeurs faibles de $n$ fournissent de très bonnes approximations du résultat). En conséquence, on peut espérer obtenir des garanties aléatoires de précision fortes avec un nombre de sondés plus faible. C'est l'objet du paragraphe suivant.

## II.2. Intervalle de confiance asymptotique

Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$ (à estimer).

Soient $(U_n)_{n\in\mathbb{N}^*}$ et $(V_n)_{n\in\mathbb{N}^*}$ deux suites d'estimateurs de $\theta$ (resp. $g(\theta)$).

On suppose de plus : $\forall n \in \mathbb{N}^*$, $\mathbb{P}_\theta\big([U_n \leqslant V_n]\big) = 1$.

Soit $\alpha \in [0,1]$.

- On dit que la suite d'intervalle d'estimateurs $\Big([U_n, V_n]\Big)_{n\in\mathbb{N}^*}$ est un intervalle de confiance asymptotique de $\theta$ (resp. $g(\theta)$) au niveau de confiance $1-\alpha$ s'il existe une suite $(\alpha_n)_{n\in\mathbb{N}^*}$

vérifiant $\begin{array}{l} \times\ \forall n \in \mathbb{N}^*,\ \alpha_n \in [0,1] \\ \times\ \lim\limits_{n\to+\infty} \alpha_n = \alpha \end{array}$ telle que :

$$\forall n \in \mathbb{N}^*,\quad \mathbb{P}_\theta\Big([U_n \leqslant \theta \leqslant V_n]\Big) \ \geqslant\ 1-\alpha_n \quad \Big(\text{resp. } \mathbb{P}_\theta\Big([U_n \leqslant g(\theta) \leqslant V_n]\Big) \ \geqslant\ 1-\alpha_n\Big)$$

$$\mathbb{P}_\theta\Big([\ \theta \in [U_n, V_n]\ ]\Big)$$

- Par abus de langage, on dira que $[U_n, V_n]$ est un intervalle de confiance asyptotique du paramètre $\theta$ au niveau de confiance $1-\alpha$.
- Le réel $\alpha \in [0,1]$ est appelé le niveau de risque de l'intervalle.

**Remarque**

Pour que $[U_n, V_n]$ est un intervalle de confiance asyptotique du paramètre $\theta$ au niveau de confiance $1-\alpha$, il suffit que :

$$\lim_{n\to+\infty} \mathbb{P}_\theta\Big([U_n \leqslant \theta \leqslant V_n]\Big) = 1-\alpha$$

En effet, en notant, pour tout $n \in \mathbb{N}^*$ : $\mathbb{P}_\theta\Big([U_n \leqslant \theta \leqslant V_n]\Big) = 1-\alpha_n$, on a bien :

$\times\ \forall n \in \mathbb{N}^*,\ \alpha_n = 1 - \mathbb{P}_\theta\Big([U_n \leqslant \theta \leqslant V_n]\Big) = \mathbb{P}_\theta\Big(\ \overline{[U_n \leqslant \theta \leqslant V_n]}\ \Big) \in [0,1]$

$\times\ \alpha_n = 1 - \mathbb{P}_\theta\Big([U_n \leqslant \theta \leqslant V_n]\Big) \underset{n\to+\infty}{\longrightarrow} 1-(1-\alpha) = \alpha$

et cette suite est telle que :

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}_\theta\Big([U_n \leqslant \theta \leqslant V_n]\Big) = 1-\alpha_n \ \geqslant\ 1-\alpha_n$$

MÉTHODO

**Utilisation du TCL pour obtenir un intervalle de confiance de l'espérance $m$ d'une v.a.r. $X$**

Soit $X$ une v.a.r. . On suppose que $X$ :

$\times$ admet une espérance $m$ **inconnue**, qu'on cherche à estimer.

$\times$ admet une variance $\sigma^2$ **connue** et non nulle.

Soit $(X_1, \ldots, X_n)$ un $n$-échantillon de la v.a.r. $X$.

Soit $\alpha \in [0,1]$ (on cherche un intervalle de confiance de $m$ au niveau de confiance $1-\alpha$).

Enfin, on note : $\overline{X_n} = \dfrac{X_1 + \cdot + X_n}{n}$.

Rappelons que la v.a.r. $\overline{X_n}$ admet une espérance et une variance. De plus : $\mathbb{E}(\overline{X_n}) = m$ et $\mathbb{V}(\overline{X_n}) = \dfrac{\sigma^2}{n}$.

*1)* Alors les v.a.r. $X_1, \ldots, X_n$ :

- × sont indépendantes,
- × ont même loi,
- × admettent la même espérance $m$.
- × admettent une variance non nulle.

Ainsi, par théorème central limite : $\overline{X_n}^* = \dfrac{\overline{X_n} - m}{\frac{\sigma}{\sqrt{n}}} \underset{n\to+\infty}{\mathscr{L}} Z$ où $Z \hookrightarrow \mathcal{N}(0,1)$.

*2)* Soit $x \in [0, +\infty[$. On a de plus :

$$\begin{aligned}
\mathbb{P}\Big(\Big[-x \leqslant \overline{X_n}^* \leqslant x\Big]\Big) &= \mathbb{P}\left(\left[-x \leqslant \sqrt{n}\,\frac{\overline{X_n} - m}{\sigma} \leqslant x\right]\right) \\
&= \mathbb{P}\left(\left[-\frac{\sigma\,x}{\sqrt{n}} \leqslant \overline{X_n} - m \leqslant \frac{\sigma\,x}{\sqrt{n}}\right]\right) \\
&= \mathbb{P}\left(\left[-\overline{X_n} - \frac{\sigma\,x}{\sqrt{n}} \leqslant -m \leqslant -\overline{X_n} + \frac{\sigma\,x}{\sqrt{n}}\right]\right) \\
&= \mathbb{P}\left(\left[\overline{X_n} + \frac{\sigma\,x}{\sqrt{n}} \geqslant m \geqslant \overline{X_n} - \frac{\sigma\,x}{\sqrt{n}}\right]\right)
\end{aligned}$$

On obtient ainsi :

$$\begin{aligned}
\mathbb{P}\left(\left[\overline{X_n} - \frac{\sigma\,x}{\sqrt{n}} \leqslant m \leqslant \overline{X_n} + \frac{\sigma\,x}{\sqrt{n}}\right]\right) &= \mathbb{P}\Big(\Big[-x \leqslant \overline{X_n}^* \leqslant x\Big]\Big) \\
&\underset{n\to+\infty}{\longrightarrow} \mathbb{P}([-x \leqslant Z \leqslant x]) \\
&= \Phi(x) - \Phi(-x) \\
&= \Phi(x) - (1 - \Phi(x)) \;=\; 2\,\Phi(x) - 1
\end{aligned}$$

*3)* On cherche alors une valeur $x$ telle que : $2\,\Phi(x) - 1 = 1 - \alpha$. Or :

$$\begin{aligned}
2\,\Phi(x) - 1 = 1 - \alpha &\Leftrightarrow 2\,\Phi(x) = 2 - \alpha \\
&\Leftrightarrow \Phi(x) = 1 - \tfrac{\alpha}{2} \\
&\Leftrightarrow x = \Phi^{-1}\big(1 - \tfrac{\alpha}{2}\big)
\end{aligned}$$

On note alors : $q_{1-\frac{\alpha}{2}} = \Phi^{-1}\big(1 - \frac{\alpha}{2}\big)$.

Ce nombre est appelé **quantile d'ordre** $1 - \frac{\alpha}{2}$. On le notera par la suite $t_\alpha$ pour plus de lisibilité.

*4)* D'après ce qui précède, on a :

$$\lim_{n\to+\infty} \mathbb{P}\left(\left[\overline{X_n} - \frac{\sigma}{\sqrt{n}}\,t_\alpha \leqslant m \leqslant \overline{X_n} + \frac{\sigma}{\sqrt{n}}\,t_\alpha\right]\right) \;=\; 2\Phi(t_\alpha) - 1 \;=\; 1 - \alpha$$

Cela démontre que $\left[\overline{X_n} - \dfrac{\sigma}{\sqrt{n}}\,t_\alpha \;,\; \overline{X_n} + \dfrac{\sigma}{\sqrt{n}}\,t_\alpha\right]$ est un intervalle de confiance asymptotique de $m$ au niveau de confiance $1 - \alpha$.

**Illustration classique** : estimation de l'espérance pour une loi de Bernoulli

Reprenons l'exemple précédent où $X \hookrightarrow \mathcal{B}(p)$ où $p$ est le paramètre à estimer.

- En reprenant l'étude précédente, on obtient :

$$\lim_{n\to+\infty} \mathbb{P}\left(\left[\overline{X_n} - \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,t_\alpha \leqslant p \leqslant \overline{X_n} + \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,t_\alpha\right]\right) = 1-\alpha \quad (*)$$

Attention : les extrémités de l'encadrement dépendent du paramètre à estimer $p$. On utilise donc la majoration classique $p(1-p) \leqslant \dfrac{1}{4}$ pour obtenir des estimateurs de $p$.

- Comme $p(1-p) \leqslant \dfrac{1}{4}$, par croissance de $x \mapsto \sqrt{x}$ sur $[0,+\infty[$ :

$$\sqrt{p(1-p)} \leqslant \frac{1}{2} \quad \text{et} \quad -\sqrt{p(1-p)} \geqslant -\frac{1}{2}$$

On en déduit :

$$\left[\overline{X_n} - \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,t_\alpha \leqslant p \leqslant \overline{X_n} + \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,t_\alpha\right] \subset \left[\overline{X_n} - \frac{1}{2\sqrt{n}}\,t_\alpha \leqslant p \leqslant \overline{X_n} + \frac{1}{2\sqrt{n}}\,t_\alpha\right]$$

Ainsi, par croissance de $\mathbb{P}$ :

$$\mathbb{P}\left(\left[\overline{X_n} - \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,t_\alpha \leqslant p \leqslant \overline{X_n} + \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,t_\alpha\right]\right) \leqslant \mathbb{P}\left(\left[\overline{X_n} - \frac{1}{2\sqrt{n}}\,t_\alpha \leqslant p \leqslant \overline{X_n} + \frac{1}{2\sqrt{n}}\,t_\alpha\right]\right)$$

En posant, pour tout $n \in \mathbb{N}^*$, $\alpha_n = \mathbb{P}\left(\left[\overline{X_n} - \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\,t_\alpha \leqslant p \leqslant \overline{X_n} + \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\,t_\alpha\right]\right)$, on a bien

exhibé une suite $(\alpha_n)$ vérifiant : $\times$ $\forall n \in \mathbb{N}^*$, $\alpha_n \in [0,1]$ ; $\times$ $\lim\limits_{n\to+\infty} \alpha_n = \alpha$ d'après $(*)$ et telle que pour tout $n \in \mathbb{N}^*$ :

$$\mathbb{P}\left(\left[\overline{X_n} - \frac{1}{2\sqrt{n}}\,t_\alpha \leqslant p \leqslant \overline{X_n} + \frac{1}{2\sqrt{n}}\,t_\alpha\right]\right) \geqslant 1-\alpha_n$$

Ainsi, $\left[\overline{X_n} - \dfrac{1}{2\sqrt{n}}\,t_\alpha \ ,\ \overline{X_n} + \dfrac{1}{2\sqrt{n}}\,t_\alpha\right]$ est un intervalle de confiance asymptotique de $p$ au niveau de confiance $1-\alpha$.

**Remarque**

On peut faire le même genre de remarques sur l'intervalle obtenu :

$\times$ cet intervalle est centré en $\overline{X_n}$.

$\times$ l'amplitude de cet intervalle est : $\left(\overline{X_n} + \dfrac{1}{2\sqrt{n}}\ t_\alpha\right) - \left(\overline{X_n} - \dfrac{1}{2\sqrt{n}}\ t_\alpha\right) = \dfrac{t_\alpha}{\sqrt{n}}$.

**Équilibre marge d'erreur / niveau de confiance**

Rappelons les liens entre marge d'erreur $\varepsilon$ et niveau de confiance $1-\alpha$ :

$$\varepsilon = \frac{t_\alpha}{2\sqrt{n}} \quad \text{et} \quad 1-\alpha = 2\,\Phi(t_\alpha) - 1$$

Lorsque le niveau de confiance $1-\alpha$ augmente, $2\,\Phi(t_\alpha)-1$ augmente, ce qui démontre que $t_\alpha$ augmente. Dans ce cas, la marge d'erreur $\varepsilon$ augmente elle aussi. Cet intervalle de confiance asymptotique possède les mêmes propriétés que l'intervalle de confiance exact (obtenu par l'inégalité de BT).

À RETENIR

- Améliorer la précision (diminuer la marge d'erreur $\varepsilon$) de l'intervalle, c'est augmenter le risque et ainsi diminuer le niveau de confiance.
- Dégrader la précision (augmenter la marge d'erreur $\varepsilon$) de l'intervalle, c'est diminuer le risque et ainsi augmenter le niveau de confiance.

**Le point de vue des instituts de sondage**

| | Nombre de sondés en fonction de la marge d'erreur $\varepsilon$ (en %) et du niveau de confiance $1-\alpha$ (en %) souhaités | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\varepsilon$ \ $1-\alpha$ | 70 (1,04) | 75 (1,17) | 80 (1,28) | 85 (1,44) | 90 (1,64) | 95 (1,96) | 97,5 (2,26) | 99 (2,57) |
| 0,5 | 10816 | 13689 | 16384 | 20736 | 26896 | 38416 | 51076 | 66049 |
| 1 | 2704 | 3422 | 4096 | 5184 | 6724 | 9604 | 12769 | 16512 |
| 1,5 | 1202 | 1521 | 1820 | 2304 | 2988 | 4268 | 5674 | 7339 |
| 2 | 676 | 856 | 1024 | 1296 | 1681 | 2401 | 3192 | 4128 |
| 2,5 | 433 | 548 | 655 | 829 | 1076 | 1537 | 2043 | 2642 |
| 3 | 300 | 380 | 455 | 576 | 747 | 1067 | 1419 | 1835 |
| 3,5 | 221 | 279 | 334 | 423 | 549 | 784 | 1042 | 1348 |
| 4 | 169 | 214 | 256 | 324 | 420 | 600 | 798 | 1032 |

- Sur la première ligne, on a placé entre parenthèse la valeur de $t_\alpha$ correspondante au niveau de confiance $1-\alpha$ considéré.
  Par exemple, si $1-\alpha = 0,95$ alors $1-\dfrac{\alpha}{2} = 0,975$ et $t_\alpha \simeq 1,96$.
- Comme mentionné précédemment, le niveau de confiance $1-\alpha = 0,95$ est assez classique. Avec un tel niveau de confiance, on considère qu'il y a 95% de chances de tomber sur un panel standard. Lorsque c'est le cas, le paramètre réel se retrouve dans l'intervalle $[\overline{x_n}-\varepsilon, \overline{x_n}+\varepsilon]$.
  - × On peut souhaiter obtenir un résultat très précis. Pour un sondage concernant des élections, savoir qu'un candidat est évalué à 19% plus ou moins 0,5% serait idéal. Du point de vue du sondeur, cela voudrait dire interroger $n = 38416$ personnes. C'est presque 6 fois moins que pour l'intervalle de confiance obtenu par inégalité de Bienaymé-Tchebychev. Pour autant, c'est toujours inenvisageable pour des raisons de coût.
  - × L'institut de sondage doit alors revoir ses objectifs à la baisse. En interrogeant $n = 1537$ personnes, il assure avec une probabilité de 95% qu'un candidat est évalué à 19% plus ou moins 2,5%. C'est 5 fois moins que dans le cas de l'intervalle de confiance obtenu par inégalité de Bienaymé-Tchebychev. Le coût est tout à fait envisageable et le résultat offre une précision correcte.
- Notons que les intervalles de confiance obtenus par les deux méthodes ont été réalisés avec la majoration : $p(1-p) \leqslant \dfrac{1}{4}$ $(*)$.
  - × La valeur $\frac{1}{4}$ est atteinte dans le cas où $p = \frac{1}{2}$. La majoration $(*)$ est donc la meilleure que l'on puisse faire en l'absence d'information sur $p$.
  - × Le rôle d'un sondage est justement d'obtenir de l'information sur $p$ (une valeur approchée). Si un candidat est évalué à 20% (resp. 80%) alors on a $p(1-p) \simeq 0,16$.
    Avec ce calcul et pour $1-\alpha = 0,95$ et $n = 1500$, on obtient alors :
    $$\varepsilon \;=\; \frac{\sqrt{p(1-p)}}{\sqrt{n}}\, t_\alpha \;\simeq\; \frac{\sqrt{0,16}}{\sqrt{1500}}\, 1,96 \;\simeq\; 0,02$$
    Ainsi, les sondages font souvent valoir une marge d'erreur qui dépend de l'estimation du candidat.

# Informations concernant cette semaine de colles

Les questions de cours pour cette semaine se trouvent dans le document « programme_23_B.pdf ».

## Exercices types

Les compétences attendues sur le chapitre convergence et approximations sont les suivantes :

- savoir démontrer qu'une v.a.r. est un estimateur d'un paramètre $\theta$ inconnu.
- savoir détrerminer le biais d'un estimateur d'un paramètre $\theta$ inconnu.
- savoir, à partir d'un estimateur biaisé, obtenir un estimateur non biaisé d'un paramètre $\theta$ inconnu.
- savoir démontrer qu'une suite d'estimateurs d'un paramètre $\theta$ inconnu est asymptotiquement sans biais.
- savoir démontrer qu'une suite d'estimateurs d'un paramètre $\theta$ inconnu est asymptotiquement sans biais.
- connaître la définition du risque quadratique d'un estimateur d'un paramètre $\theta$ inconnu.
- savoir déterminer le risque quadratique d'un estimateur d'un paramètre $\theta$ inconnu à l'aide de la décomposition biais-variance.
- connaître la notion de convergence d'une suite d'estimateurs d'un paramètre $\theta$ inconnu.
- savoir démontrer qu'une suite d'estimateurs d'un paramètre $\theta$ inconnu est convergente à l'aide du risque quadratique.

La notion d'intervalle de confiance n'est pas au programme de colles de cette semaine (elle est présenté dans ce document pour préparation de la prochaine semaine de colles). Les colleurs devront donc amener cette notion auprès des élèves s'ils souhaitent les interroger dessus.