Chapitre 17

Estimation

Compétences attendues.

- ✓ Comparer deux estimateurs à l'aide de `Python`.
- ✓ Déterminer un intervalle de confiance par l'inégalité de Bienaymé-Tchebychev.
- ✓ Déterminer un intervalle de confiance asymptotique à partir du théorème limite central.

Anthony Mansuy
Professeur de Mathématiques en deuxième année de CPGE filière ECG au Lycée Clemenceau (Reims)
Page personnelle : `http://anthony-mansuy.fr`
E-mail : `mansuy.anthony@hotmail.fr`

1 Estimation ponctuelle

1.1 Modélisation mathématique

Introduction du problème

On étudie un phénomène aléatoire qui est reproductible dans des conditions identiques et indépendantes. On connait pour des raisons théoriques ou empiriques le type de loi le décrivant. Mais les paramètres de la dite loi sont souvent inconnus. On doit donc les estimer : c'est l'objectif de ce qu'on appelle la statistique inférentielle.

Exemples.

- À l'approche du second tour d'une élection présidentielle, on interroge une personne au hasard et on note $X = 1$ si elle se prononce pour le candidat A et $X = 0$ si c'est pour le candidat B. X suit une loi de Bernoulli de paramètre $\theta \in [0, 1]$ inconnu qui correspond à la proportion de français qui votent pour A.
- On souhaite modéliser le nombre N de voitures se présentant à un péage en une heure. Il s'agit du nombre de réalisations d'un évènement rare sur un grand nombre d'observations. On sait donc que ce nombre N suit une loi de Poisson. On cherche le paramètre de cette loi.
- On souhaite modéliser la durée de vie T d'un appareil électrique. Ce phénomène étant sans mémoire, il se modélise à l'aide d'une loi exponentielle. Il s'agit de trouver son paramètre.

Notons X la variable égale au résultat de notre expérience aléatoire. On suppose donc que l'on ne connaît qu'imparfaitement la loi de X : on sait de quel type elle est (elle appartient à une famille de lois $(\mathcal{L}(\theta))_{\theta\in\Theta}$ connue) mais elle dépend d'un paramètre θ inconnu appartenant à Θ l'espace des paramètres. Le paramètre θ peut être réel (Θ est une partie $\mathbb{R}$) ou vectoriel (Θ est une partie de $\mathbb{R}^k$, $k \geq 2$).

Exemple. Dans les trois cas précédents, on a :

- $X \hookrightarrow \mathcal{B}(\theta)$, $\Theta =]0, 1[$;
- $N \hookrightarrow \mathcal{P}(\theta)$, $\Theta =]0, +\infty[$;
- $T \hookrightarrow \mathcal{E}(\theta)$, $\Theta =]0, +\infty[$.

On peut aussi imaginer le cas de $X \hookrightarrow \mathcal{N}(m, \sigma^2)$ où $\theta = (m, \sigma^2)$ est inconnu (paramètre vectoriel, $\Theta = \mathbb{R}\times]0, +\infty[$).

L'objectif de la statistique inférentielle est d'estimer la vraie valeur du paramètre θ à partir de réalisations de la variable X.

Échantillonnage

Bien entendu, une seule réalisation d'une variable aléatoire de loi $\mathcal{L}(\theta)$ ne permettra pas d'obtenir beaucoup d'informations sur θ. On est donc amené à introduire la notion d'échantillon :

Définition.

- On appelle **n-échantillon de loi mère** $\mathcal{L}(\theta)$ sur un espace probabilisé $(\Omega, \mathscr{A}, P)$ un n-uplet $(X_1, \dots, X_n)$ de variables aléatoires i.i.d. suivant toutes la loi $\mathcal{L}(\theta)$.
- Un n-uplet $(x_1, \dots, x_n) \in \mathbb{R}^n$ est une **réalisation de l'échantillon** (ou **échantillon observé**) si c'est une réalisation du vecteur aléatoire $(X_1, \dots, X_n)$, c'est-à-dire si :

$$\exists \omega \in \Omega, \quad (x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega)).$$

Remarques.

1. Un échantillon est un n-uplet de variables aléatoires alors qu'un échantillon observé est un n-uplet de réels. Si par exemple $\mathcal{E}(\theta)$ est la loi mère du n-échantillon, alors `rd.exponential(1/theta, n)` fournit une réalisation de cet échantillon.
2. Soit $(x_1, \dots, x_n) \in \mathbb{R}^n$ un échantillon de données obtenu en observant n fois le phénomène aléatoire. On admettra l'existence de n variables aléatoires $X_1, \dots, X_n$, toutes définies sur un même espace probabilisable $(\Omega, \mathscr{A})$, telles que $(x_1, \dots, x_n)$ soit une réalisation de $(X_1, \dots, X_n)$, c'est-à-dire :

$$\exists \omega \in \Omega, \quad (X_1(\omega), \dots, X_n(\omega)) = (x_1, \dots, x_n).$$

Pour tout $\theta \in \Theta$, on admettra de plus l'existence d'une probabilité P_θ sur $(\Omega, \mathscr{A})$, dépendant de θ, telle que $(X_1, \dots, X_n)$ soit un n-échantillon P_θ-indépendant de loi mère $\mathcal{L}(\theta)$.

3. Si X est une variable aléatoire sur $(\Omega, \mathscr{A})$, on notera, en cas d'existence, $E_\theta(X)$ et $V_\theta(X)$ son espérance et sa variance pour la probabilité P_θ, qui dépendent donc eux aussi de θ.

Exemple. Reprenons l'exemple de l'élection présidentielle. On questionne 5 individus sur leurs intentions de vote et on obtient les résultats suivants (en notant 1 ou 0 selon que le choix se porte sur le candidat A ou B) :

$$1, \quad 0, \quad 0, \quad 1, \quad 0.$$

Ces résultats observés correspondent, pour tout $\theta \in [0,1]$, à la réalisation d'un 5-échantillon $(X_1, \dots, X_5)$ P_θ-indépendant de loi mère $\mathcal{B}(\theta)$. Pour tout $1 \leq i \leq 5$, on a $E_\theta(X_i) = \theta$ et $V_\theta(X_i) = \theta(1-\theta)$.

Estimateur

À partir d'un échantillon observé, on souhaite estimer une valeur caractéristique de la loi $\mathcal{L}(\theta)$ telle que son espérance, sa variance, son étendue… On notera $g(\theta)$ cette valeur, où $g : \Theta \to \mathbb{R}$.

Définition.

- On appelle **estimateur** (d'ordre n) de $g(\theta)$ toute variable aléatoire T_n de la forme $\varphi_n(X_1, \dots, X_n)$ indépendante de θ, où $(X_1, \dots, X_n)$ est un n-échantillon.
- On appelle **estimation** de $g(\theta)$ une réalisation $t_n = \varphi_n(x_1, \dots, x_n)$ de $T_n = \varphi_n(X_1, \dots, X_n)$.

Remarques.

- Un estimateur de $g(\theta)$ est une variable aléatoire alors qu'une estimation de $g(\theta)$ est un réel.
- L'estimateur T_n ne dépend que du n-échantillon $(X_1, \dots, X_n)$. En particulier, φ_n ne doit pas dépendre du paramètre θ qui est inconnu et qu'on cherche justement à estimer.

Définition.

On appelle **moyenne empirique** l'estimateur :

$$\overline{X_n} = \frac{X_1 + \cdots + X_n}{n}.$$

Exemple. Reprenons l'exemple de l'élection présidentielle. La moyenne empirique $\overline{X_n} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$ est l'estimateur le plus naturel de θ. On obtient avec l'échantillon observé que :

$$\overline{X_5}(\omega) = \frac{1}{5}(1+0+0+1+0) = \frac{2}{5} \text{ est une estimation de } \theta.$$

On peut envisager bien d'autres estimateurs pour θ, comme par exemple

$$A_n = \frac{2}{n(n+1)} \sum_{k=1}^{n} k X_k$$

qui fournit une autre estimation de θ à partir de notre échantillon observé :

$$A_5(\omega) = \frac{2}{5 \times 6}(1+0+0+4+0) = \frac{1}{3}.$$

On verra dans la suite comment décider :

- s'il est pertinent de prendre ces nombres comme estimation de θ ou non ;
- si l'un de ces estimateurs est plus pertinent que les autres, c'est-à-dire s'il donne en moyenne des estimations « plus proche » de la véritable valeur de θ.

Exemple. Si $(X_1, \dots, X_n)$ est un n-échantillon de loi mère $\mathcal{U}([a,b])$ de paramètre vectoriel inconnu $\theta = (a,b) \in \mathbb{R}^2$, alors :

- $T_n = \max(X_1, \dots, X_n) - \min(X_1, \dots, X_n)$ est un estimateur de $g(\theta) = b - a$.
- $U_n = 2\overline{X_n}$ est un estimateur de $g(\theta) = a + b$.

Et beaucoup d'autres estimateurs sont possibles...

Remarque. Très souvent, on prendra $g = Id_{\mathbb{R}}$ et on estimera θ directement. Mais on peut aussi chercher à estimer une fonction $g(\theta)$ des paramètres, par exemple pour l'étendue d'une loi uniforme comme dans l'exemple précédent. On peut aussi vouloir estimer $g(\theta)$, et non θ directement, par facilité de calculs : pour la loi exponentielle de paramètre θ par exemple, il est plus facile d'estimer $g(\theta) = \dfrac{1}{\theta}$ en utilisant la moyenne empirique.

1.2 Étude qualitative d'un estimateur

Comme on l'a déjà vu, la définition très générale d'un estimateur ne présume en rien de sa pertinence. L'objectif à présent est de comparer avec `Python` différents estimateurs et d'essayer d'isoler les plus intéressants pour l'estimation de $g(\theta)$.

Il existe deux façons de les comparer :

- Numériquement : le plus efficace est celui qui s'approche le plus de la valeur de $g(\theta)$.
- Graphiquement : le plus efficace est celui dont la distribution est « concentrée » autour de $g(\theta)$.

Comparaison numérique d'estimateurs

Méthode.

Pour comparer numériquement deux estimateurs X_n et Y_n d'un même paramètre $g(\theta)$:

1. *On choisit une valeur de $g(\theta)$.*
2. *On commande sur `Python` plusieurs réalisations de X_n et Y_n.*
3. *Si les valeurs de l'un sont notablement plus proches de $g(\theta)$ que celles de l'autre, alors celui-ci est un meilleur estimateur.*

Exemple. Reprenons le cas d'une élection présidentielle, et donc d'un échantillon $(X_1, \dots, X_n)$ de loi mère $\mathcal{B}(\theta)$. Comparons les estimateurs $\overline{X}_n$ et A_n. Pour cela, on va choisir un paramètre θ au hasard dans $[0,1]$, correspondant à la proportion réelle de votants pour le candidat A.

```
1 theta = rd.random()
```

`Python` connait déjà le résultat du scrutin, et on pourrait lui demander avec l'instruction `print(theta)`, mais gardons pour le moment le mystère... On va estimer ce paramètre à l'aide des estimateurs $\overline{X_n}$ et A_n. Il nous faut pour cela un échantillon observé, qu'on prendra de taille $n = 60$ par exemple.

```
2 E = rd.binomial(1, theta, 60)
```

Grâce à cet échantillon, on peut obtenir deux estimations de θ à l'aide respectivement de $\overline{X_{60}}$ et A_{60}. Demandons à `Python` de les calculer :

```
3 X = np.mean(E)
4 A = (2/(60*61))*np.sum(np.arange(1, 61, 1)*E)
5 print(X, A)
```

On obtient des estimations de $\overline{X_{60}}$ et A_{60} égales respectivement à `0.3166667` et `0.2737705`.
On peut enfin comparer ces estimations ponctuelles avec la valeur réelle de θ : `print(theta)` renvoie `0.3760119`. On remarque que l'estimation associée à $\overline{X_{60}}$ est la plus proche de θ. Donc $\overline{X_n}$ serait un meilleur estimateur de θ que A_n.

Remarque. Notons qu'on n'est jamais à l'abri d'un « mauvais » échantillon, qui donnerait exceptionnellement une meilleure valeur pour un des estimateurs alors qu'il est en moyenne moins bon. Pour éviter ce problème, on peut répéter ce procédé pour plusieurs échantillons de taille 60, et donc pour plusieurs estimations de θ par $\overline{X_{60}}$ et A_{60}, et comparer les écarts des estimations.

```
n = 60
for k in range(10):
    theta = rd.random()
    E = rd.binomial(1, theta, n) #échantillon observé
    X = np.mean(E) #estimateur moyenne empirique
    A = 2/(n*(n+1))*np.sum(np.arange(1, n+1)*E) #estimateur A_n
    print(X, A, theta)
```

On obtient les résultats suivants :

$\overline{X_{60}}$	A_{60}	θ
0.95	0.957377	0.914636
0.233333	0.179781	0.329055
0.033333	0.024043	0.044351
0.033333	0.014754	0.019852
0.3	0.271038	0.257691
0.95	0.965573	0.942797
0.666666	0.716393	0.645562
0.483333	0.510928	0.387532
0.316666	0.269945	0.147319
0.733333	0.703278	0.702468

Il est difficile ici de déterminer si les estimations associées à $\overline{X_n}$ sont plus proches de θ que celles associées à A_n. Les résultats obtenus ne permettent donc pas de conclure si $\overline{X_n}$ est un meilleur estimateur θ que A_n. Une comparaison graphique de ces estimateurs semble nécessaire.

Exemple. On suppose que la loi mère de l'échantillon est une loi uniforme $\mathcal{U}([0,\theta])$. On considère les estimateurs de θ suivants :

$$T_n = \max(X_1,\dots,X_n), \quad U_n = 2\overline{X_n} \quad \text{et} \quad V_n = 2X_n.$$

On cherche à comparer numériquement ces estimateurs avec `Python`. On prend par exemple $n = 100$ et on compare T_{100}, U_{100} et V_{100}.

```
n = 100
for k in range(10):
    theta = rd.random()
    E = rd.uniform(0, theta, n) #échantillon observé
    T = np.max(E) #estimateur T_n
    U = 2*np.mean(E) #estimateur U_n
    V = 2*E[n-1] #estimateur V_n
    print(T, U, V, theta)
```

On obtient les résultats suivants :

T_{100}	U_{100}	V_{100}	θ
0.022780	0.022874	0.024675	0.022915
0.166384	0.179372	0.222005	0.167108
0.029078	0.028053	0.009866	0.029368
0.101533	0.107619	0.008192	0.102894
0.301755	0.305395	0.309860	0.302198
0.441948	0.403083	0.076368	0.444671
0.957815	0.937929	1.465579	0.967351
0.367195	0.380379	0.542416	0.368645
0.807158	0.741833	1.028595	0.809040
0.320507	0.352034	0.435721	0.328980

On remarque que les valeurs obtenues pour T_{100} sont les plus proches de θ alors que celles obtenues pour V_{100} sont les plus éloignées. Donc T_n semble être le meilleur estimateur de θ et V_n le moins bon. Une comparaison graphique de ces estimateurs va confirmer ce résultat.

Comparaison graphique d'estimateurs

Méthode.

Pour comparer graphiquement deux estimateurs X_n et Y_n d'un même paramètre $g(\theta)$:

1. *On choisit une valeur de $g(\theta)$.*
2. *On commande sur* **Python** *plusieurs réalisations de X_n et Y_n et on trace l'histogramme des résultats obtenus.*
3. *Plus la distribution est « ressérée » autour de $g(\theta)$, meilleure elle est du point de vue de l'estimation de cette valeur (car on a plus de chance de la trouver à proximité de cette valeur).*

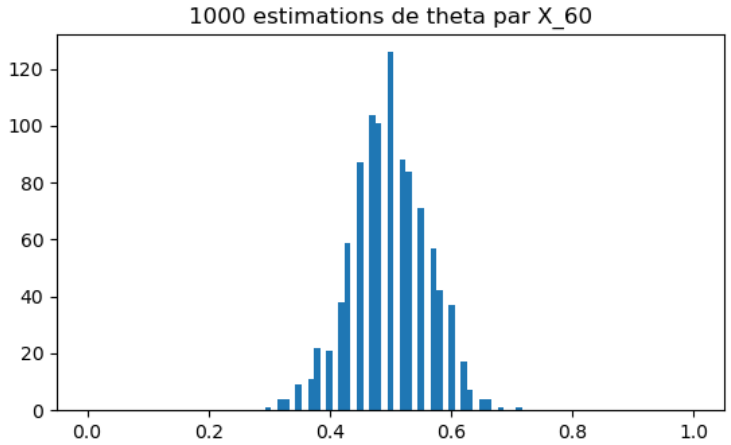
Exemple. On suppose que la loi mère de l'échantillon est $\mathcal{B}(\theta)$. On considère les estimateurs de θ suivants :

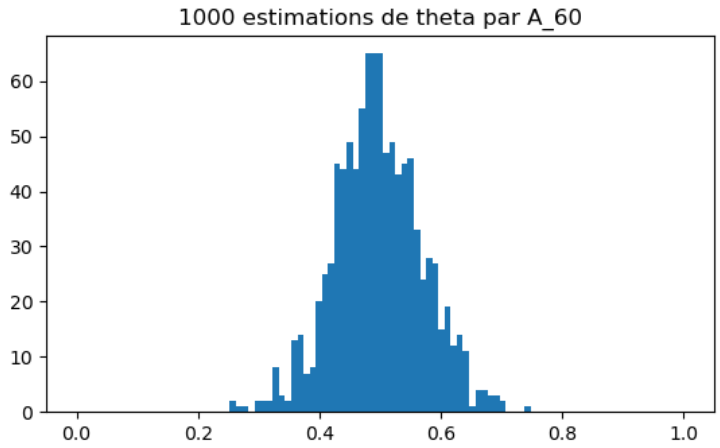
$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \quad \text{et} \quad A_n = \frac{2}{n(n+1)}\sum_{k=1}^{n} kX_k.$$

On compare 1000 estimations de θ par $\overline{X_n}$ et A_n à l'aide du programme `Python` suivant (avec $n = 60$ et $\mathcal{B}(1/2)$ pour loi mère, c'est-à-dire avec $\theta = 1/2$) :

```
n = 60
E = rd.binomial(1, 1/2, [1000, n]) #1000 échantillons observés
X = [] #estimateur moyenne empirique
A = [] #estimateur A_n
for k in range(1000):
    X.append(np.mean(E[k, :]))
    A.append(2/(n*(n+1))*np.sum(np.arange(1, n+1)*E[k, :]))
c = np.linspace(0, 1, 100)
plt.subplot(2, 1, 1)
plt.hist(X, c)
plt.title("1000 estimations de theta par X_60")
plt.subplot(2, 1, 2)
plt.hist(A, c)
plt.title("1000 estimations de theta par A_60")
plt.show()
```

On obtient le résultat graphique suivant :





On peut remarquer que :

- Les estimations obtenues par $\overline{X_{60}}$ et par A_{60} sont en moyenne égales à θ (les histogrammes sont "centrés" en θ).
- Les estimations de θ obtenues par A_{60} sont (très légèrement) plus dispersées que celles obtenues par $\overline{X_{60}}$.

On peut donc conclure que $\overline{X_n}$ semble être un meilleur estimateur de θ que A_n.

Remarques.

1. Pour déterminer si un estimateur T_n est en moyenne égal au paramètre à estimer $g(\theta)$, on peut calculer son **biais** :
$$b_\theta(T_n) = E_\theta(T_n - g(\theta)) = E_\theta(T_n) - g(\theta).$$
Un estimateur T_n est **sans biais** si $b_\theta(T_n) = 0$ et donc si l'erreur commise en moyenne est nulle.
2. Pour évaluer la « dispersion moyenne » d'un estimateur T_n par rapport au paramètre à estimer $g(\theta)$, on peut calculer son **risque quadratique** :
$$r_\theta(T_n) = E_\theta\left((T_n - g(\theta))^2\right).$$
Plus $r_\theta(T_n)$ est petit, plus les valeurs de T_n sont concentrées autour du paramètre à estimer $g(\theta)$.
3. On a la décomposition suivante du risque quadratique en fonction du biais et de la variance :
$$r_\theta(T_n) = (b_\theta(T_n))^2 + V_\theta(T_n).$$
En effet :

Exercice.

1. Déterminer les biais des estimateurs $\overline{X_n}$ et A_n.

2. Comparer les risques quadratiques de $\overline{X_n}$ et A_n et en déduire le meilleur des deux estimateurs.

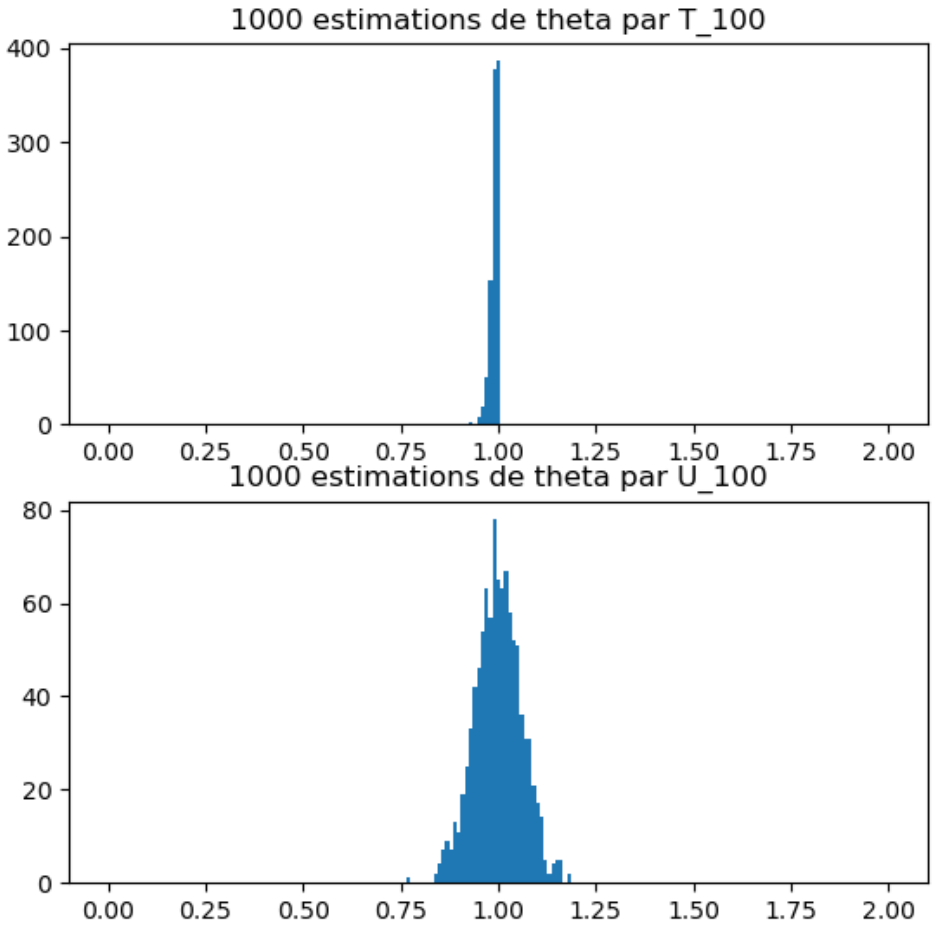
Exemple. On suppose que la loi mère de l'échantillon est $\mathcal{U}([0,\theta])$. On considère les estimateurs de θ suivants :

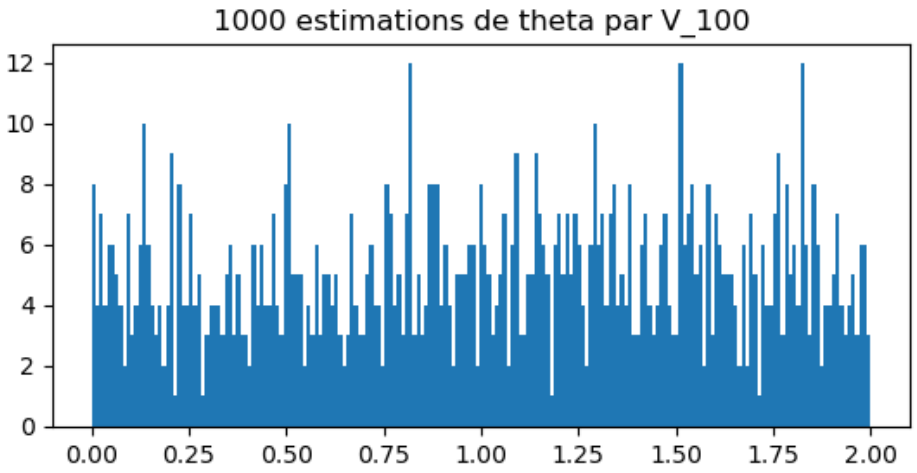
$$T_n = \max(X_1,\dots,X_n), \quad U_n = 2\overline{X_n} \quad \text{et} \quad V_n = 2X_n.$$

On compare 1000 estimations de θ par T_n, U_n et V_n à l'aide du programme `Python` suivant (avec $n = 100$ et $\mathcal{U}([0,1])$ pour loi mère, c'est-à-dire avec $\theta = 1$) :

```
n = 100
E = rd.uniform(0, 1, [1000, 100]) #1000 échantillons observés
T = [] #estimateur T_n
U = [] #estimateur U_n
V = [] #estimateur V_n
for k in range(1000):
    T.append(np.max(E[k, :]))
    U.append(2*np.mean(E[k, :]))
    V.append(2*E[k, n-1])
c = np.linspace(0, 2, 200)
plt.subplot(3, 1, 1)
plt.hist(T, c)
plt.title("1000 estimations de theta par T_100")
plt.subplot(3, 1, 2)
plt.hist(U, c)
plt.title("1000 estimations de theta par U_100")
plt.subplot(3,1,3)
plt.hist(V, c)
plt.title("1000 estimations de theta par V_100")
plt.show()
```

On obtient le résultat graphique suivant :





L'estimateur T_n semble être celui qui fournit les meilleurs estimations parmi les trois estimateurs de θ considérés. En effet, c'est celui qui à la distribution la plus concentrée sur la valeur à estimer ($\theta = 1$). A l'opposé, V_n semble être le moins bon estimateur de θ.

On pourrait calculer et comparer les risques quadratiques de ces trois estimateurs T_n, U_n et V_n pour confirmer ces observations.

1.3 Méthode du maximum de vraisemblance

Nous avons vu comment comparer des estimateurs pour choisir le meilleur. Nous présentons ici une méthode générale pour « deviner » un « bon » estimateur, la méthode du **maximum de vraisemblance**.

Cas discret

Considérons qu'on dispose d'une observation $(x_1, \dots, x_n)$ d'un n-échantillon $(X_1, \dots, X_n)$ d'une loi mère discrète $\mathcal{L}(\theta)$ et on cherche à estimer θ. L'idée est alors de choisir comme estimateur $\hat{\theta}_n = \varphi_n(X_1, \dots, X_n)$ une fonction du n-échantillon $(X_1, \dots, X_n)$ où l'expression de la fonction φ_n est choisie de sorte que $\theta^* = \varphi_n(x_1, \dots, x_n)$ soit la valeur rendant maximale la probabilité de l'événement

$$(X_1 = x_1) \cap (X_2 = x_2) \cap \dots \cap (X_n = x_n).$$

Par hypothèse d'indépendance sur les variables du n-échantillon, la probabilité de l'événement ci-dessous vaut

$$P_\theta\left(\bigcap_{i=1}^{n}(X_i = x_i)\right) = \prod_{i=1}^{n} P_\theta(X_i = x_i)$$

ce qui justifie les définitions ci-dessous.

Définition.

- Soient $(X_1, \dots, X_n)$ un n-échantillon d'une loi discrète $\mathcal{L}(\theta)$ où $\theta \in \Theta$ est un paramètre qu'on cherche à estimer et $(x_1, \dots, x_n) \in X_1(\Omega)^n$ fixé. La fonction L_n définie sur Θ par

$$L_n : \theta \mapsto \prod_{i=1}^{n} P_\theta(X_i = x_i)$$

  s'appelle la **vraisemblance** de la loi $\mathcal{L}$.

- En notant $\theta^* = \varphi_n(x_1, \dots, x_n)$ la valeur où L_n est maximale (c'est-à-dire telle que, pour tout $\theta \in \Theta$, $L_n(\theta) \leq L_n(\theta^*)$, l'**estimateur du maximum de vraisemblance** est l'estimateur défini par

$$\hat{\theta}_n = \varphi_n(X_1, \dots, X_n).$$

Exercice. Soient $(X_1, \dots, X_n)$ un n-échantillon de la loi $\mathcal{B}(p)$ et $(x_1, \dots, x_n) \in \{0,1\}^n$. Le paramètre θ à estimer est p. On pose :

$$L_n(\theta) = \prod_{i=1}^{n} P_\theta(X_i = x_i) = \theta^{s_n}(1-\theta)^{n-s_n}$$

où $s_n = \sum_{i=1}^{n} x_i$.

1. On pose $h_n(\theta) = \ln(L_n(\theta))$. Étudier les variations de h_n.

2. En déduire la valeur θ^* où L_n est maximale, puis donner l'expression de l'estimateur du maximum de vraisemblance.

Cas continu

Il existe également une version continue de la méthode du maximum de vraisemblance.

Définition.

- Soient $(X_1, \dots, X_n)$ un n-échantillon d'une loi continue $\mathcal{L}(\theta)$ où $\theta \in \Theta$ est un paramètre qu'on cherche à estimer, f_θ une densité de X_1 et $(x_1, \dots, x_n) \in X_1(\Omega)^n$ fixé. La fonction L_n définie sur Θ par

$$L_n : \theta \mapsto \prod_{i=1}^{n} f_\theta(x_i)$$

  s'appelle la **vraisemblance** de la loi $\mathcal{L}$.

- En notant $\theta^* = \varphi_n(x_1, \dots, x_n)$ la valeur où L_n est maximale (c'est-à-dire telle que, pour tout $\theta \in \Theta$, $L_n(\theta) \leq L_n(\theta^*)$, l'**estimateur du maximum de vraisemblance** est l'estimateur défini par

$$\hat{\theta}_n = \varphi_n(X_1, \dots, X_n).$$

Exercice. On considère un n-échantillon de la loi $\mathcal{U}([0,\theta])$ et on cherche à estimer $\theta > 0$.
Soit $(x_1,\dots,x_n) \in \left(\mathbb{R}_+^*\right)^n$ fixé. On note f_θ une densité associée à la loi $\mathcal{U}([0,\theta])$. On introduit la fonction de vraisemblance, définie sur $\mathbb{R}_+^*$ par :

$$L_n(\theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

1. Montrer que, pour tout $\theta \geq 0$, on a :

$$L_n(\theta) = \begin{cases} \theta^{-n} & \text{si } \theta \geq \max(x_1,\dots,x_n), \\ 0 & \text{sinon.} \end{cases}$$

2. En déduire que l'estimateur du maximum de vraisemblance pour la loi $\mathcal{U}([0,\theta])$ est donné par :

$$\hat{\theta}_n = \max(X_1,\dots,X_n).$$

2 Estimation par intervalles de confiance

S'il existe des critères pour juger des qualités d'un estimateur ponctuel T_n de $g(\theta)$ (biais, risque quadratique...), aucun de ces critères ne permet de garantir que la valeur prise par un estimateur T_n à partir d'un échantillon observé $(x_1, \dots, x_n)$ sera « proche » de la valeur $g(\theta)$ du paramètre à estimer. Ainsi, même si T_n est un « bon » estimateur (risque quadratique faible), on n'est jamais à l'abri de tomber sur un « mauvais » échantillon qui nous donnerait une mauvaise estimation de $g(\theta)$.

La démarche de l'estimation par intervalle de confiance est de contrôler cette incertitude. Elle consiste à construire, à partir de l'échantillon, un intervalle (le plus petit possible) dans lequel se trouve la valeur exacte de $g(\theta)$ avec une grande probabilité, fixée à l'avance.

2.1 Intervalles de confiance

Dans tout ce paragraphe :

- $(X_1, \dots, X_n)$ est un échantillon i.i.d. de même loi mère de paramètre inconnu $\theta \in \Theta$,
- pour tout $n \in \mathbb{N}^*$, $U_n = \varphi_n(X_1, \dots, X_n)$ et $V_n = \psi_n(X_1, \dots, X_n)$ sont des estimateurs de $g(\theta)$ tels que $P_\theta(U_n \leq V_n) = 1$ pour tout $\theta \in \Theta$ (U_n est inférieur à V_n P_θ-presque sûrement).

Définition.

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

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

$$\forall \theta \in \Theta, \quad P_\theta(U_n \leq g(\theta) \leq V_n) \geq 1 - \alpha.$$

  Le réel α est appelé le **risque**.

- Soit $\omega \in \Omega$. L'intervalle $[U_n(\omega), V_n(\omega)]$ est une **réalisation** de l'intervalle de confiance $[U_n, V_n]$, aussi appelé **intervalle de confiance observé**.

Remarques.

1. Très souvent, on recherche un intervalle de confiance de $g(\theta)$ sous la forme d'un intervalle centré en une estimation ponctuelle de $g(\theta)$.
2. C'est à celui qui réalise l'étude de fixer le niveau de confiance $1 - \alpha$ qu'il souhaite, et donc le risque α de commettre une erreur qu'il accepte. Par exemple pour $\alpha = 0.05$, et si $[u_n, v_n]$ est une réalisation de $[U_n, V_n]$, alors on a
$$u_n \leq g(\theta) \leq v_n$$
avec une probabilité de 95%. Il y a cependant 5% de (mal)chance de tomber sur un « mauvais échantillon » qui nous donnera un intervalle de confiance ne contenant pas $g(\theta)$.

   La plupart du temps, c'est ce niveau de risque de 0.05 qui est utilisé, et qui est communément accepté par exemple en sciences humaines. Mais dans des domaines plus sensibles où l'on n'a pas vraiment de droit à l'erreur (aérospatiale, physique nucléaire, etc), on travaille avec des niveaux de risque de 0.01, voir moins.

Propriété 1 (Intervalle de confiance par l'inégalité de Bienaymé-Tchebychev)

Si $(X_1, \dots, X_n)$ est un n-échantillon i.i.d. de loi de Bernoulli de paramètre p inconnu, alors

$$\left[\overline{X_n} - \frac{1}{2\sqrt{n\alpha}}, \overline{X_n} + \frac{1}{2\sqrt{n\alpha}}\right]$$

est un intervalle de confiance de p au niveau de confiance $1 - \alpha$.

Preuve.

□

Remarque. Nous pouvons observer sur cet intervalle deux résultats intuitifs :

- plus le niveau de risque souhaité est petit et plus l'intervalle de confiance est grand ;
- plus l'échantillon est de taille n importante et plus l'intervalle de confiance est petit.

Applications numériques.

- Prenons pour risque $\alpha = 0.05$ et pour taille de notre échantillon $n = 100$. Alors un intervalle de confiance de p au niveau de confiance 0.95 est

$$\left[\overline{X_n} - 0.22, \overline{X_n} + 0.22\right].$$

  Notons que l'amplitude de cet intervalle est énorme : 0.44 alors que $p \in [0, 1]$.

- Pour $\alpha = 0.05$ et $n = 1000$ (taille de l'échantillon généralement utilisé par les instituts de sondage), l'intervalle de confiance de p au niveau de confiance 0.95 est

$$\left[\overline{X_n} - 0.07, \overline{X_n} + 0.07\right].$$

- Prenons la démarche inverse : on souhaite estimer p avec une erreur d'au plus 0.01 et à un niveau de risque $\alpha = 0.05$. Alors la taille n de notre échantillon doit satisfaire

$$\frac{1}{2\sqrt{n\alpha}} \leq 0.005 \quad \Rightarrow \quad n \geq 50000.$$

  Il faut donc un échantillon de taille 50000 (difficile en pratique...).

Simulation. Prenons le cas du deuxième tour d'une élection présidentielle avec deux candidats A et B. Soit p la proportion (inconnue) de personnes interrogées se prononçant pour le candidat A.

```
1 p = rd.random()
```

On cherche un intervalle de confiance de p au niveau de confiance $1 - \alpha = 0.95\%$. On sonde pour cela $n = 1000$ personnes.

```
2 E = rd.binomial(1, p, 1000) #1000-echantillon observé
```

On calcule la moyenne empirique sur cet échantillon observé.

```
3 Xbar = np.mean(E)
4 print(Xbar)
```

On obtient que `Xbar` est égal à `0.204` .

On en déduit que $\left[\overline{X_n} - 0.07, \overline{X_n} + 0.07\right] = [0.134, 0.274]$ est une réalisation de l'intervalle de confiance de p au niveau de confiance 0.95.

Vérifions pour finir si p appartient bien à notre intervalle de confiance (il y a théoriquement 95% de chance que ce soit bien le cas) :

```
4 print(p)
```

On obtient que `p` est égal à `0.2113249` .

On retiendra la méthode suivante pour déterminer un intervalle de confiance :

Méthode.

Supposons qu'on dispose d'un estimateur T_n ***sans biais*** *(c'est-à-dire tel que $E(T_n) = g(\theta)$) et dont on connait (un majorant de) la variance qui* ***ne fait pas intervenir*** *$g(\theta)$. Pour déterminer un intervalle de confiance de $g(\theta)$ à l'aide de l'inégalité de Bienaymé-Tchebychev, on procèdera comme suit :*

Étape 1 : Appliquer l'inégalité de Bienaymé-Tchebychev.

On calcule $E(T_n)$ (qui doit valoir $g(\theta)$) et $V(T_n)$ et on applique l'inégalité de Bienaymé-Tchebychev à T_n :

$$\forall \varepsilon > 0, \qquad \underbrace{P(|T_n - g(\theta)| \geq \varepsilon)}_{\text{Proba d'être hors de l'I. de C.}} \leq \underbrace{\frac{V(T_n)}{\varepsilon^2}}_{\text{Niveau de risque}}.$$

Étape 2 : Fixer le niveau de confiance.

On majore si nécessaire $\dfrac{V(T_n)}{\varepsilon^2}$ par une quantité qui ***ne fait pas intervenir*** *$g(\theta)$. On fixe ensuite ε afin que ce majorant soit égal à α.*

Étape 3 : Expliciter l'intervalle de confiance.

On résout l'inéquation $|T_n - g(\theta)| \leq \varepsilon$ afin d'isoler $g(\theta)$ et d'obtenir ainsi l'intervalle de confiance de $g(\theta)$ au niveau de confiance $1 - \alpha$.

2.2 Intervalles de confiance asymptotiques

Outre l'inégalité de Bienaymé-Tchebychev, le théorème limite central permet aussi d'obtenir des estimations par intervalle de confiance. Mais celui-ci donne seulement un résultat asymptotique, d'où la notion suivante :

Définition.

Soit $\alpha \in [0, 1]$. On appelle **intervalle de confiance asymptotique** de $g(\theta)$ au **niveau de confiance** $1-\alpha$ toute suite $([U_n, V_n])_{n\in\mathbb{N}^*}$ vérifiant : pour tout $\theta \in \Theta$, il existe une suite de réels $(\alpha_n)_{n\in\mathbb{N}^*}$ à valeurs dans $[0, 1]$ et de limite α, telle que

$$\forall n \in \mathbb{N}^*, \quad P_\theta(U_n \leq g(\theta) \leq V_n) \geq 1 - \alpha_n.$$

Remarque. Notons la différence avec un intervalle de confiance de niveau $1 - \alpha$: un intervalle de confiance asymptotique sera à un niveau de confiance "acceptable" pour n grand (α_n proche de α), sans plus d'information sur le n à considérer. D'où une perte de précision ici.

Propriété 2 (Intervalle de confiance asymptotique par le théorème limite central)

Si (X_n) est une suite de variables aléatoires i.i.d. suivant la même loi de Bernoulli de paramètre p inconnu, alors

$$\left[\overline{X_n} - \frac{t_\alpha}{2\sqrt{n}}, \overline{X_n} + \frac{t_\alpha}{2\sqrt{n}}\right].$$

est un intervalle de confiance asymptotique de p au niveau de confiance $1 - \alpha$ (où $t_\alpha = \Phi^{-1}(1 - \frac{\alpha}{2})$).

Preuve.

□

Remarque. Voici deux valeurs de t_α qu'on rencontrera souvent (et qu'on peut retrouver dans la table de la loi normale si besoin) :

$$t_{0.05} = \Phi^{-1}(0.975) \approx 1.96 \text{ pour } \alpha = 0.05 \quad \text{et} \quad t_{0.01} = \Phi^{-1}(0.995) \approx 2.57 \text{ pour } \alpha = 0.01.$$

Applications numériques.

- Prenons pour risque $\alpha = 0.05$ et une taille raisonnable pour l'échantillon $n = 1000$. On obtient l'intervalle de confiance asymptotique
$$\left[\overline{X_n} - 0.031, \overline{X_n} + 0.031\right].$$
Il est d'amplitude 0.062, à comparer au 0.14 obtenu pour celui avec l'inégalité de Bienaymé-Tchebychev.
- Si on souhaite estimer p avec une erreur d'au plus 0.01 et un risque $\alpha = 0.05$, alors la taille n de notre échantillon doit satisfaire
$$\frac{1.96}{2\sqrt{n}} \leq 0.01 \quad \Leftrightarrow \quad n \geq \left(\frac{0.98}{0.01}\right)^2 = 9604.$$
On a donc $n = 9604$. Là aussi, c'est bien meilleur que le $n = 50000$ obtenu à l'aide de l'inégalité de Bienaymé-Tchebychev.

Simulation. Reprenons notre simulation. On a obtenu sur notre échantillon de taille $n = 1000$ une moyenne empirique observée égale à `Xbar = 0.204`. On obtient donc l'intervalle de confiance asymptotique de `p` au niveau de risque $\alpha = 0.05$ suivant :

$$\left[\overline{X_n} - 0.031, \overline{X_n} + 0.031\right] = [0.173, 0.235],$$

`p` valant en réalité 0.2113249.

Remarque. Il s'agit d'un intervalle de confiance asymptotique, dont on ne contrôle donc pas le risque (il faut que n soit « grand » pour que $\alpha_n \approx \alpha$, mais « grand » comment ?). En pratique, on considère que c'est bien un intervalle de confiance de risque α dès que $n \geq 30$, $np \geq 5$, $n(1-p) \geq 5$, conditions d'approximation d'une loi binomiale par une loi normale (ce qu'on fait ici).

Exemple. Le premier tour de l'élection présidentielle de 2002.
Quelques jours avant les élections, des sondages réalisés auprès de 1000 personnes donnaient (estimations ponctuelles par la moyenne empirique) :

14.5% d'intentions de vote à Jean-Marie Le Pen et 17% à Lionel Jospin.

Pourtant les scores finaux ont été de

16.83% pour Le Pen et 16,18% pour Jospin.

Comment l'expliquer ?

À l'aide des intervalles de confiance que nous venons d'obtenir, on pouvait affirmer avec une certitude de 95% que

le score final de Le Pen serait entre 11.5% et 17.5%, et celui de Jospin entre 14% et 20%.

L'intersection de ces intervalles étant loin d'être vide, il était douteux de conclure uniquement sur la base d'une estimation ponctuelle.

Remarque. Soit (X_n) une suite de variables aléatoires i.i.d. de loi de Bernoulli de paramètre p inconnu. Soit $\alpha \in]0,1[$. On a obtenu deux intervalles de confiance au niveau de confiance $1-\alpha$:

- grâce à l'inégalité de Bienaymé-Tchebychev :
$$\left[\overline{X_n} - \frac{1}{2\sqrt{n\alpha}}, \overline{X_n} + \frac{1}{2\sqrt{n\alpha}}\right] \tag{BT}$$
- grâce au théorème limite central :
$$\left[\overline{X_n} - \frac{t_\alpha}{2\sqrt{n}}, \overline{X_n} + \frac{t_\alpha}{2\sqrt{n}}\right]. \tag{TLC}$$

Reprenons notre simulation `Python`. On voulait un intervalle de confiance de `p` à un niveau de risque $\alpha = 0.05$. Pour cela, on avait créé un échantillon de taille $n = 1000$. On avait obtenu `Xbar = 0.204`. On obtient ainsi les intervalles de confiance suivants :

$$BT \ : \ [0.134, 0.274] \qquad \text{et} \qquad TLC \ : \ [0.173, 0.235].$$

Et on a bien $p = 0.2113249$ qui appartient à ces deux intervalles de confiance (on avait en théorie 95% de chance que ce soit effectivement le cas).

Répétons cela pour $m = 10000$ échantillons de taille $n = 1000$ à l'aide du programme suivant, en testant si p est dans chacun des m intervalles de confiance obtenus.

```
p = 0.2113249
alpha = 0.05
t = 1.96
n = 1000
m = 10000
BT = 0
TLC = 0
for k in range(m):
    Xn = np.mean(rd.binomial(1, p, n))
    if np.abs(Xn-p)<t/(2*np.sqrt(alpha*n)):
        BT = BT+1
    if np.abs(Xn-p)<t/(2*np.sqrt(n)):
        TLC = TLC+1
print(100*BT/m)
print(100*TLC/m)
```

Ce programme renvoie la proportion d'intervalles de confiance contenant p. Il estime ainsi (à l'aide de la méthode de Monte Carlo) le niveau de confiance réel de chaque intervalle. On obtient :

$$BT = 100 \qquad \text{et} \qquad TLC = 98.42.$$

C'est donc l'intervalle TLC qui répond le mieux à notre problème : il nous donne une meilleur approximation de p, l'intervalle de confiance étant plus petit, et est bien d'un niveau de confiance estimé d'approximativement 0.95.

Voici d'autres résultats pour différentes valeurs de p (pour $n = 1000$ et $\alpha = 0.05$ avec $m = 10000$ répétitions) :

p **réel**	**BT**	**TLC**
0.5238291	100.0	95.04
0.7667777	100.0	97.94
0.1610254	100.0	99.26
0.0131476	100.0	100.0
0.9775233	100.0	100.0
0.2489265	100.0	97.73
0.3863217	100.0	95.57

On retiendra la méthode suivante pour déterminer un intervalle de confiance asymptotique :

Méthode.

Supposons qu'on dispose d'un estimateur T_n ***sans biais*** *(c'est-à-dire tel que $E(T_n) = g(\theta)$) dont on connait (un majorant de) la variance qui* ***ne fait pas intervenir*** *$g(\theta)$. Pour déterminer un intervalle de confiance asymptotique d'un paramètre $g(\theta)$ à l'aide du théorème limite central, on procèdera comme suit :*

Étape 1 : Appliquer le théorème limite central.

On calculer $E(T_n)$ (qui doit valoir $g(\theta)$) et $V(T_n)$ et on applique le théorème limite central.

On disposera ainsi d'une convergence en loi $T_n^* \xrightarrow{\mathcal{L}} T$ avec $T \hookrightarrow \mathcal{N}(0,1)$.

Étape 2 : Fixer le niveau de confiance.

Pour tout $a < b$, on a :

$$\lim_{n\to+\infty} P(a < T_n^* \leq b) = P(a < T \leq b) = \Phi(b) - \Phi(a).$$

On choisit a et b de telle sorte que $\Phi(b) - \Phi(a) = 1 - \alpha$.

Étape 3 : Expliciter l'intervalle de confiance.

On résout l'inéquation $a < T_n^* \leq b$ afin d'isoler $g(\theta)$ et d'obtenir ainsi l'intervalle de confiance asymptotique de $g(\theta)$ au niveau de confiance $1 - \alpha$.