Chapitre 12 - Estimation et intervalles de confiance

I. Introduction

Les statisticiens connaissent en général, le type de loi qui décrit tel ou tel phénomène (de par l'observation), mais souvent, il ne connaissent pas tous les paramètres de cette loi. Ils cherchent alors à estimer ces paramètres non connus : c'est tout l'objectif de la ***statistique inférentielle*** (inférer = déduire).

On considère une variable aléatoire X dont le type de loi est connu et dépend d'un paramètre θ, qui peut être un réel, comme par exemple le paramètre λ d'une loi de Poisson, ou un élément de $\mathbb{R}^2$ comme le couple (m, σ^2) d'une loi normale. On suppose alors que θ appartient à une partie de $\mathbb{R}$ ou de $\mathbb{R}^2$, et on note Θ cette partie.

L'objectif est alors de donner une *estimation* de la vraie valeur de θ.
On cherchera parfois à estimer un réel $g(\theta)$ où g est une fonction simple. Dans ce cas $g(\theta)$ sera en général une valeur caractéristique de la loi inconnue, comme son espérance, sa variance, son étendue...

L'estimation de θ se fera à partir d'un échantillon de données x_1, ..., x_n obtenu en observant n fois le phénomène.

On suppose que cet échantillon est la réalisation de n variables aléatoires X_1, X_2, ..., X_n, définies sur un même espace probabilisé $(\Omega, \mathcal{A})$ muni d'une famille de probabilités $(P_\theta)_{\theta\in\Theta}$. Les variables aléatoires X_1, ..., X_n sont supposées mutuellement indépendantes (pour la probabilité P_θ) et de même loi que X.

L'idée est de prendre n grand pour bien rendre compte du phénomène.

Exemple 1 : On étudie les truites d'un lac de montagne et on s'intéresse à la probabilité p qu'une truite soit atteinte d'une maladie donnée. On considère $X \hookrightarrow \mathscr{B}(p)$ et on cherche à estimer p (X vaut 1 si la truite est malade et 0 sinon).
Pour cela, on prélève n truites et on note X_1, ..., X_n les variables de Bernoulli associées à la truite numéro 1, ... , numéro n.

Exemple 2 : on sait qu'il existe deux réels a et b tels que $a < b$ et $X \hookrightarrow \mathscr{U}([a, b])$ mais on ne connait pas a et b. Pour se simplifier la vie on cherchera à estimer un seul paramètre inconnu, ici la moyenne $m = \dfrac{a+b}{2}$ ou l'étendue $b - a$.

Exemple 3 : on souhaite modéliser le nombre N de voitures se présentant à un péage en une heure. Ce type de variable aléatoire suit une loi de Poisson; on cherchera à estimer le paramètre de cette loi.

Il y a deux types d'estimation : **l'estimation ponctuelle** et l'**estimation par intervalle de confiance**.

Dans tout ce cours, n désignera un entier naturel supérieur ou égal à 2.

II. Echantillonage

Définition II.1
Echantillon de taille n

Soit X une variable aléatoire réelle.
On appelle **n-échantillon (ou échantillon de taille n)** issue de la variable aléatoire X, tout vecteur aléatoire $(X_1, \cdots, X_n)$ vérifiant les propriétés suivantes :

- les variables aléatoires $X_1, X_2, \ldots, X_n$ ont même loi que la variable aléatoire X.
- les variables aléatoires $(X_i)_{1\leq i\leq n}$ sont indépendantes

On dit alors que $(X_1, \cdots, X_n)$ est un **n-échantillon i.i.d.** (indépendant, identiquement distribué)

On appelle **réalisation de l'échantillon** $(X_1, \cdots, X_n)$ toute n-liste de réels $(x_1, x_2, \cdots, x_n)$ où x_i est une valeur prise par la variable aléatoire X_i pour tout $i \in [\![1, n]\!]$.

Dans la pratique, pour constituer un échantillon, on effectue n épreuves, identiques et indépendantes, et X_k est la VAR égale au résultat de la k-ième épreuve.

Remarque
Soit X une variable aléatoire réelle, dont la loi dépend d'un paramètre $\theta \in \Theta$. On peut montrer que pour tout $\theta \in \Theta$, il existe des variables aléatoires X_1, ..., X_n définies sur un même espace probabilisable $(\Omega, \mathcal{A})$ et une probabilité P_θ (dépendant de θ) telles que les variables X_1, ..., X_n sont de même loi que X, et sont indépendantes pour la probabilité P_θ.

On notera ainsi parfois P_θ au lieu de P. Si X est une variable aléatoire, on notera parfois $E_\theta(X)$ et $V_\theta(X)$ l'espérance et la variance de X pour la probabilité P_θ.

III. Estimateur

III.1) Définition

Définition III.1
Soit X une variable aléatoire qui suit une loi $\mathcal{L}$. On suppose que la loi de X dépend d'un paramètre inconnu θ.
Soit $(X_1, \cdots, X_n)$ un n-échantillon i.i.d. issu de la variable aléatoire X.

On appelle **estimateur de θ (ou de $g(\theta)$)** une variable aléatoire de la forme

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

où φ est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$. La réalisation $\varphi(x_1, \cdots, x_n)$ de l'estimateur T_n est l'estimation de θ (ou de $g(\theta)$). Cette estimation ne dépend que de l'échantillon $(x_1, \cdots, x_n)$ observé.

Exemple
Soit X une variable aléatoire qui suit une loi $\mathcal{L}$ dépendant d'un paramètre θ. On cherche à connaître θ (ou son image $g(\theta)$ par une fonction g).
Soit $(X_1, X_2, \ldots, X_n)$ un échantillon de la loi de X.

- La variable aléatoire $\frac{1}{n}(X_1 + \cdots + X_n)$ est un estimateur de θ associé à l'échantillon $(X_1, \cdots, X_n)$.
- La variable aléatoire $X_1 \times \cdots \times X_n$ est un estimateur de θ associé à l'échantillon $(X_1, \cdots, X_n)$.
- La variable aléatoire $min(X_1, \cdots, X_n)$ est un estimateur de θ associé à l'échantillon $(X_1, \cdots, X_n)$.

Ces estimateurs peuvent être pertinents ou non !!!

III.2) Moyenne empirique

Définition III.2
Un estimateur classique : la moyenne empirique
Soit X une variable aléatoire qui suit une loi $\mathcal{L}$. On suppose que la loi de X dépend d'un paramètre inconnu θ.
Soit $(X_1, X_2, \ldots, X_n)$ un échantillon de la loi de X.
On suppose que X admet une espérance notée m et une variance $\sigma^2 > 0$.

On appelle **moyenne empirique de** X sur l'échantillon $(X_1, \cdots, X_n)$, la variable aléatoire

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Proposition III.1

- La moyenne empirique $\overline{X_n}$ admet un moment d'ordre 2.
- $E(\overline{X_n}) = E(X) = m$
- $V(\overline{X_n}) = \dfrac{V(X)}{n} = \dfrac{\sigma^2}{n}$
- La suite $\left(\overline{X_n}\right)_{n\in\mathbb{N}^*}$ converge en probabilité vers la variable aléatoire certaine égale à $m = E(X)$.
- La variable aléatoire centrée réduite associée à $\overline{X_n}$ i.e. $\overline{X_n}^* = \dfrac{\overline{X_n} - m}{\frac{\sigma}{\sqrt{n}}}$ converge en loi vers une variable aléatoire de loi normale centrée réduite.

Remarque
Il s'agit de reformulations de la loi faible des grands nombres et du Théorème Central Limite.

IV. Estimation ponctuelle

IV.1) Estimateur sans biais

Définition IV.1
Soit X une variable aléatoire qui suit une loi $\mathcal{L}$ dépendant d'un paramètre θ. On cherche à estimer $g(\theta)$.
Soit $(X_1, X_2, \ldots, X_n)$ un échantillon i.i.d. de la loi de X.
On considère un estimateur de $g(\theta)$ noté $T_n = \varphi_n(X_1, X_2, \ldots, X_n)$.
On suppose que T_n admet une espérance.
On dit que l'estimateur T_n de $g(\theta)$ est **sans biais** lorsque $E(T_n) = g(\theta)$.

IV.2) Estimateur convergent

Définition IV.2
Soit X une variable aléatoire qui suit une loi $\mathcal{L}$ dépendant de θ, on cherche à connaître $g(\theta)$.
Pour tout $n \in \mathbb{N}$, $n \geq 1$, on considère T_n un estimateur de $g(\theta)$.

On dit que la suite d'estimateurs $(T_n)_{n\geq 1}$ est **convergente** lorsque la suite de variables $(T_n)_{n\in\mathbb{N}^*}$ converge en probabilité vers la variable aléatoire certaine égale à $g(\theta)$. c'est à dire lorsque :

$$T_n \xrightarrow{P} g(\theta)$$

Remarque
Par abus de langage, on dit que l'estimateur T_n est convergent.

Remarque
La moyenne empirique est un estimateur sans biais et convergent de $m = E(X)$.
Ce résultat n'est pas explicitement dans le programme : il doit être redémontré à chaque fois via la loi faible des grands nombres.

Théorème IV.1
Composition par une fonction continue

Soit X une variable aléatoire qui suit une loi $\mathcal{L}$ dépendant de θ
Soit T_n un estimateur de $g(\theta)$. Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction **continue** de $\mathbb{R}$ dans $\mathbb{R}$.
Si T_n est un estimateur convergent alors $(f(T_n))_{n\geqslant 1}$ est une suite convergente d'estimateurs de $f(g(\theta))$.

Remarque
Conséquence immédiate du cours sur la convergence en probabilités.

Théorème IV.2
Condition suffisante de convergence

Soit (T_n) une suite d'estimateurs de $g(\theta)$.
Si $\lim_{n\to+\infty} E(T_n) = g(\theta)$ et $\lim_{n\to+\infty} V(T_n) = 0$ alors (T_n) est convergente

Preuve
Preuve de cours à connaître
Soit $\epsilon > 0$. Par stricte croissance de la fonction carré sur $\mathbb{R}_+$, puis d'après l'inégalité de Markov,

$$P(|T_n - g(\theta)| > \epsilon) = P\big((T_n - g(\theta))^2 > \epsilon^2\big) \leq \frac{E((T_n - g(\theta))^2)}{\epsilon^2}$$

De plus,

$$\begin{aligned} E\big((T_n - g(\theta))^2\big) &= E\big(T_n^2 - 2.g(\theta).T_n + g(\theta)^2\big) \\ &= E(T_n^2) - 2g(\theta)E(T_n) + g(\theta)^2 \quad \text{(par linéarité de l'espérance)} \\ &= E(T_n^2) - E(T_n)^2 + E(T_n)^2 - 2g(\theta)E(T_n) + g(\theta)^2 \\ &= V(T_n) + (E(T_n) - g(\theta))^2 \end{aligned}$$

Ainsi

$$P(|T_n - g(\theta)| > \epsilon) \leq \frac{V(T_n)}{\epsilon^2} + \frac{(E(T_n) - g(\theta))^2}{\epsilon^2}$$

et comme $\lim_{n\to+\infty} E(T_n) = g(\theta)$ et $\lim_{n\to+\infty} V(T_n) = 0$ on a bien $\lim_{n\to+\infty} P(|T_n - g(\theta)| > \epsilon) = 0$, donc $T_n \xrightarrow{P} g(\theta)$.

Remarque
Comme

$$[|T_n - g(\theta)| \leq \epsilon] = [T_n - \epsilon \leq g(\theta) \leq T_n + \epsilon],$$

on en déduit que

$$\lim_{n\to+\infty} P(g(\theta) \in [T_n - \epsilon, T_n + \epsilon]) = 1$$

On dit que $[T_n - \epsilon, T_n + \epsilon]$ est un intervalle de confiance asymptotique de $g(\theta)$.

IV.3) Comparaison de deux estimateurs

On souhaite comparer deux estimateurs de $g(\theta)$, afin de retenir le plus pertinent, c'est-à-dire retenir celui qui en moyenne fournit le résultat le plus proche de $g(\theta)$.
Une première idée pourrait être d'éliminer les estimateurs biaisés. Cependant, un estimateur non biaisé ayant une variance importante prendra souvent des valeurs éloignées de $g(\theta)$, alors qu'un estimateur faiblement biaisé mais ayant une variance faible pourra fournir des résultats plus proches.

Si T_n est un estimateur de $g(\theta)$, il s'agit d'obtenir $E(T_n) - g(\theta)$ et $V(T_n)$ les plus faibles possibles.

Dans la preuve de cours ci-dessus, nous avons montré que pour tout $\epsilon > 0$,

$$P(|T_n - g(\theta)| > \epsilon) \leq \frac{V(T_n) + (E(T_n) - g(\theta))^2}{\epsilon^2}$$

La quantité

$$V(T_n) + (E(T_n) - g(\theta))^2$$

parait être un bon outil de comparaison : si cette quantité est faible alors à la fois $|E(T_n) - g(\theta)|$ et $V(T_n)$ sont faibles.

Méthode : pour comparer deux estimateurs T_n et U_n de $g(\theta)$, on calcule leurs espérances et leurs variances. Puis, si

$$V(T_n) + (E(T_n) - g(\theta))^2 < V(U_n) + (E(U_n) - g(\theta))^2$$

alors l'estimateur T_n sera plus intéressant que l'estimateur U_n.

IV.4) Exercices

Exercice 1
Soit X une variable aléatoire de loi $\mathcal{N}(0, \sigma^2)$ de variance σ^2 inconnue où $\sigma > 0$.
Pour tout $n \geq 2$, on considère un n-échantillon i.i.d $(X_1, \ldots, X_n)$ de même loi que X.
On note $T_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k^2$

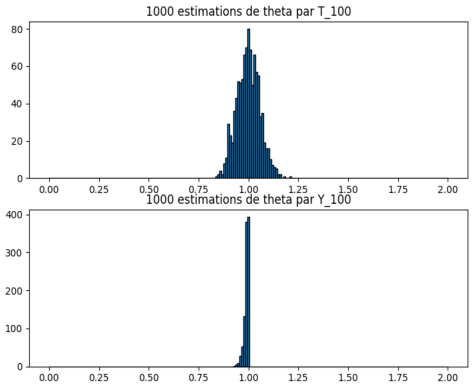
1. Justifier que la variable X admet un moment d'ordre 4.
2. Montrer que T_n est un estimateur sans biais et convergent de σ^2.

Exercice 2
Une variable aléatoire X suit la loi uniforme sur le segment $[0, \theta]$, θ étant un paramètre strictement positif inconnu.

1. Rappeler une densité f_X, la fonction de répartition F_X, l'esperance $E(X)$ et la variance $V(X)$ de la variable X.

   Soit X_1, X_2, …, X_n un échantillon de n observations indépendantes de la variable aléatoire X $(n \geq 2)$. On note $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$.
2. Justifier que $T_n = 2\overline{X_n}$ est un estimateur sans biais et convergent de θ que l'on notera T_n.
3. On considère la variable aléatoire suivante $Y_n = Max(X_1, ..., X_n)$.
   (on note aussi $Y_n = Sup(X_1, ..., X_n)$).
   (a) Justifier que Y_n est une variable à densité et déterminer une densité de probabilité f_{Y_n} de Y_n. Calculer $E(Y_n)$ et $V(Y_n)$.
   (b) Montrer que Y_n est un estimateur convergent de θ. Cet estimateur est-il sans biais ?
   (c) Comparer les estimateurs T_n et Y_n. Lequel est le plus pertinent ?
   (d) Définir à partir de Y_n un estimateur sans biais de θ, que l'on notera β_n.
4. Simulation informatique : le programme suivant permet d'effectuer 1000 simulations de T_{100} et Y_{100}, puis affiche les histogrammes des valeurs obtenues. On se place dans le cas où $\theta = 1$.

```
n=100
E =rd.random([1000,n]) # 1000 échantillons observés
T=np.zeros(1000) # estimateur T_n
Y=np.zeros(1000) # estimateur Y_n
for k in range(1000):
    Y[k]=np.max(E[k,:])
    T[k]=2*np.mean(E[k,:])
c = np.linspace(0,2,200)
plt.subplot(2,1,1)
plt.hist(T,c,edgecolor='k')
plt.title('1000 estimations de theta par T_100')
plt.subplot(2,1,2)
plt.hist(Y,c,edgecolor='k')
plt.title('1000 estimations de theta par Y_100')
plt.show()
```



Ceci confirme-t-il les résultats obtenus précédemment ?

V. Intervalles de confiance

S'il existe des critères pour juger des qualités d'un estimateur ponctuel T_n de $g(\theta)$, aucune certitude ne peut jamais être apportée quand au fait que l'estimation est proche de la vraie valeur à estimer.
La démarche de l'estimation par intervalle de confiance consiste à trouver un intervalle aléatoire qui contienne $g(\theta)$ avec une probabilité minimale donnée.

V.1) Définition

Définition V.1
Soit X une variable aléatoire qui suit une loi $\mathcal{L}$ dépendant de θ; on cherche à connaître $g(\theta)$.
Soit $(X_1, \cdots, X_n)$ un n-échantillon i.i.d. issu de X. Soit $\alpha \in]0,1[$ un réel **fixé**.
Soient U_n et V_n deux variables aléatoires fonctions de $(X_1, \cdots, X_n)$.

On appelle **intervalle de confiance de** $g(\theta)$ **au niveau de confiance** $1-\alpha$
(ou au **niveau de risque** α) tout intervalle $[U_n, V_n]$ vérifiant :

- $U_n \leq V_n$.
- U_n et V_n sont des estimateurs de $g(\theta)$ dont l'expression ne dépend pas de θ.
- $P\Big(U_n \leqslant g(\theta) \leqslant V_n\Big) \geqslant 1-\alpha$.

Toute réalisation de $[U_n, V_n]$ est une estimation de cet intervalle de confiance.

Remarque
- Il n'y a pas unicité de U_n et V_n !!!
  Par exemple si $[U_n, V_n]$ est un intervalle de confiance au niveau de confiance $1-\alpha$, il en est de même pour $[U_n - 1, V_n + 1]$, même si ce deuxième intervalle est moins intéressant.
- On parle d'intervalle de confiance de niveau de confiance $1-\alpha$. En pratique, le risque α sera petit et le niveau de confiance $1-\alpha$ sera "grand" c'est à dire proche de 1.
  Par exemple : $\alpha_1 = 0.05 = 5\%$ et $1-\alpha_1 = 0.95 = 95\%$
  Ou $\alpha_2 = 0.01 = 1\%$ et $1-\alpha_2 = 0.99 = 99\%$.
- Souvent U_n et V_n sont construits à l'aide du même estimateur T_n de θ.
- Le réel $V_n - U_n$ est appelé amplitude ou étendue de l'intervalle de confiance (on cherche à le minimiser pour avoir un intervalle de confiance le plus précis possible).

Remarque
Supposons que $U_n = \phi_n(X_1, X_2, \ldots, X_n)$ et $V_n = \psi_n(X_1, X_2, \ldots, X_n)$.
La réalisation observée x_1,..., x_n fournit l'**intervalle de confiance observé** $[u_n, v_n]$ où $u_n = \phi_n(x_1, x_2, \ldots, x_n)$
et $v_n = \psi_n(x_1, x_2, \ldots, x_n)$.
L'intervalle $[u_n, v_n]$ est parfois appelé "fourchette".

V.2) Intervalle de confiance du paramètre d'une loi de Bernoulli

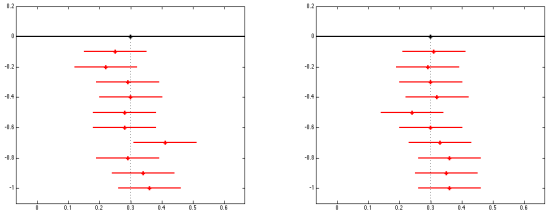
Un bureau d'études teste une machine qui fabrique un certain type d'objets. En fonctionnement normal, elle produit des objets défectueux en proportion $p \in]0,1[$ inconnue. Un étude cherche à estimer la valeur de p.
On prélève à la production un échantillon de n objets et on considère les variables de Bernoulli X_1, ..., X_n où $X_k = 1$ ssi le k-ième objet est défectueux.
Ces variables sont supposées indépendantes : il s'agit d'un n-échantillon i.i.d. de la loi $\mathcal{B}(p)$.

On note : $\overline{X_n} = \dfrac{1}{n}\sum_{k=1}^{n} X_k$. Soit α un réel de $]0,1]$.

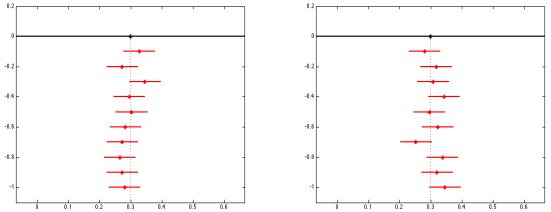
1. Montrer que pour tout $x \in [0,1]$, $x(1-x) \leq \frac{1}{4}$ (**à retenir et à savoir redémontrer rapidement**).
2. Calculer l'espérance et la variance de $\overline{X_n}$.
3. Estimation ponctuelle
   Justifier que $\overline{X_n}$ est un estimateur sans biais et convergent de p.
4. Estimation par intervalle de confiance
   (a) A l'aide de l'inégalité de Bienaymé-Tchebychev, montrer pour tout réel ϵ strictement positif,
   $$P(\overline{X_n} - \varepsilon \leqslant p \leqslant \overline{X_n} + \varepsilon) \geqslant 1 - \frac{p(1-p)}{n\varepsilon^2}$$
   (b) En déduire que : $P\left(\overline{X_n} - \sqrt{\dfrac{p(1-p)}{n\alpha}} \leq p \leq \overline{X_n} + \sqrt{\dfrac{p(1-p)}{n\alpha}}\right) \geq 1-\alpha$
   (c) En déduire un intervalle de confiance de p à un niveau de confiance au moins égal à $1-\alpha$
5. Application numérique :
   On a prélevé un échantillon de $n = 1000$ objets et constaté que 10 d'entre eux étaient défectueux.
   (a) Quelle estimation ponctuelle de p peut-on donner avec l'estimateur $\overline{X_n}$?
   (b) Déterminer une estimation de l'intervalle de confiance ci-dessus au niveau de confiance 0.95 ("fourchette").

Proposition V.1
Soit X une variable aléatoire de Bernoulli de paramètre p.
On pose $U_n = \overline{X_n} - \dfrac{1}{2\sqrt{n\alpha}}$ et $V_n = \overline{X_n} + \dfrac{1}{2\sqrt{n\alpha}}$.
Alors $[U_n, V_n]$ est un intervalle de confiance de p de risque α.
A savoir redémontrer

Simulations pour $\alpha = 0,25$ et $n = 100$.



Simulations pour $\alpha = 0,25$ et $n = 400$.



V.3) Intervalle de confiance de l'espérance d'une loi normale dont la variance est connue

Un autre bureau d'études teste des cables de téléphérique. En fonctionnement normal, la charge maximale (en tonnes) X supportée par le cable suit une loi normale de paramètre (m, σ^2).
On suppose que la valeur de m est inconnue et que la valeur de σ^2 est connue.

Une étude cherche à estimer la valeur de m. Pour cela, on teste successivement un échantillon aléatoire de n cables. On considère les variables X_1, ..., X_n i.i.d. où X_k est la charge maximale supportée par le k-ième cable testé. Ces variables sont supposées indépendantes.

On note $\forall n \in \mathbb{N}^* \quad \overline{X}_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$.

1. Estimation ponctuelle
   Préciser $E(\overline{X_n})$ et $V(\overline{X_n})$. Prouver que $\overline{X_n}$ est un estimateur sans biais convergent de m.
2. Estimation par intervalle de confiance
   (a) Justifier que la variable aléatoire $\overline{X_n}^* = \dfrac{\overline{X_n} - m}{\frac{\sigma}{\sqrt{n}}}$ suit une loi normale centrée réduite.
   (b) Soit $t \in]0, +\infty[$.
      i. Justifier que : $P\left(|\overline{X_n}^*| \leqslant t\right) = 2\Phi(t) - 1$ où Φ est la fonction de répartition de la loi normale centrée réduite.
      ii. En déduire que : $P\left(\overline{X_n} - t\frac{\sigma}{\sqrt{n}} \leqslant m \leqslant \overline{X_n} + t\frac{\sigma}{\sqrt{n}}\right) = 2\Phi(t) - 1$.
   (c) Montrer qu'il existe un unique u_α de $\mathbb{R}$ tel que $2\Phi(u_\alpha) - 1 = 1 - \alpha$. Justifier que $u_\alpha > 0$
   (d) On note $U_n = \overline{X_n} - u_\alpha\frac{\sigma}{\sqrt{n}}$ et $V_n = \overline{X_n} + u_\alpha\frac{\sigma}{\sqrt{n}}$.
   Montrer que $[U_n, V_n]$ est un intervalle de confiance de m au niveau de confiance $1 - \alpha$.
   (e) En fait, σ est inconnu mais on sait que $\sigma \leq 0.5$. Donner un intervalle de confiance de m au niveau de confiance $1 - \alpha$.
3. Application numérique ($\Phi(2.58) = 0.995$)
   On sait que $\sigma \leq 0.5$. On a testé $n = 50$ cables et la moyenne des charges maximales observées sur cet échantillon est 12.2 tonnes. Donner une estimation de l'intervalle de confiance (fourchette) de m au risque 0.01.

V.4) Intervalle de confiance asymptotique

Définition V.2
Soit X une variable aléatoire qui suit une loi $\mathcal{L}$ dont on cherche à connaître $g(\theta)$.
Soit $(X_1, \cdots, X_n)$ un échantillon i.i.d. de taille n issu de X.
Soient U_n et V_n deux variables aléatoires fonctions de $(X_1, \cdots, X_n)$.

Soit $\alpha \in]0, 1[$ un réel **fixé**.
On appelle **intervalle de confiance asymptotique de** $g(\theta)$ **au niveau de confiance** $1 - \alpha$ tout intervalle $[U_n, V_n]$ vérifiant

- $U_n \leq V_n$.
- U_n et V_n sont des estimateurs de $g(\theta)$ dont l'expression ne dépend pas de $g(\theta)$.
- il existe une suite $(\alpha_n)_{n\in\mathbb{N}^*}$ de réels vérifiant :

$$P\Big(U_n \leqslant g(\theta) \leqslant V_n\Big) \geqslant 1 - \alpha_n$$

où $\lim_{n\to+\infty}(\alpha_n) = \alpha$

Remarque
Si $[U_n, V_n]$ est un **intervalle de confiance asymptotique de** θ **au niveau de confiance** $1 - \alpha$ alors :

$$\lim_{n\to+\infty}\Big(P\Big(U_n \leqslant \theta \leqslant V_n\Big)\Big) \geqslant 1 - \alpha$$

Remarque
En pratique, on peut montrer que

$$\lim_{n\to+\infty}(P(U_n \leq \theta \leq V_n)) = 1 - \alpha$$

et alors $[U_n, V_n]$ est un intervalle de confiance asymptotique de θ au risque α.

En effet, il suffit de poser $\alpha_n = 1 - P(U_n \leq \theta \leq V_n)$ pour revenir à la définition.

V.5) Intervalle de confiance asymptotique du paramètre p d'une loi de Bernoulli

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes de même loi $\mathcal{B}(p)$ où $p \in]0, 1[$.
Soit $\alpha \in]0, 1[$.
On note $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$.

1. Rappel : $\overline{X_n}$ est un estimateur sans biais convergent de p et $E(\overline{X_n}) =$ et $V(\overline{X_n}) =$
2. Soit Φ la fonction de répartition d'une variable suivant $\mathcal{N}(0, 1)$.
   Justifier qu'il existe un unique réel $t_\alpha > 0$ tel que $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$.
3. (a) Justifier que
$$\lim_{n\to+\infty} P(\overline{X_n} - t_\alpha.\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq p \leq \overline{X_n} + t_\alpha.\frac{\sqrt{p(1-p)}}{\sqrt{n}}) = 1 - \alpha$$
   (b) En déduire que
$$\lim_{n\to+\infty} P(\overline{X_n} - t_\alpha.\frac{1}{2\sqrt{n}} \leq p \leq \overline{X_n} + t_\alpha.\frac{1}{2\sqrt{n}}) \geq 1 - \alpha$$
   En déduire un intervalle de confiance asymptotique de p au niveau de confiance $1 - \alpha$.
4. Application numérique :
   On reprend l'exemple final du V.3
   On donne $\Phi(1.96) = 0.975$. On a prélevé un échantillon de taille $n = 1000$ et observé 10 objets sans défaut.
   Donner une estimation de cet intervalle de confiance asymptotique au niveau de risque 0.05.

VI. Méthode de Monte-Carlo

La **méthode de Monte-Carlo** permet d'obtenir par simulation une estimation ponctuelle d'un réel M s'exprimant comme l'espérance $E(Y)$ d'une certaine variable aléatoire, ou comme une certaine probabilité.

Les outils sont :

- les lois usuelles (via `rd.random()` ou `rd. .... ` en Python),
- le théorème de transfert
- L'utilisation de la moyenne empirique $\overline{Y_n} = \frac{1}{n}\sum_{k=1}^{n} Y_k$ associée à une suite Y_1, Y_2, de variables i.i.d. de même loi que Y. La moyenne empirique est un estimateur sans biais et convergent (loi faible des grands nombres) de $E(Y)$.

Méthode VI.1
Méthode de Monte-Carlo (estimation d'une espérance)

1. On écrit à l'aide du théorème de transfert $M = E(Y)$ avec $Y = g(X)$, où X est une variable aléatoire que l'on sait simuler.
2. On simule, pour n assez grand, un n-échantillon i.i.d. Y_1, ..., Y_n de même loi que Y. On prend la valeur observée de $\overline{Y_n}$ comme estimation de M.
3. En Python, on peut déclarer la fonction g, et utiliser `rd. ...`, `np.mean`, `np.sum`, boucle `for`...

Remarque
On peut être amené par l'énoncé à affiner par la mise en place d'un intervalle de confiance.

Méthode VI.2
Méthode de Monte-Carlo (estimation d'une probabilité)

1. On reconnait $M = P(A)$, où A est un événement souvent lié à une VAR X. On considère la VAR de Bernoulli Y indicatrice de l'événement A et on a $M = E(Y)$.
2. On simule, pour n assez grand, un n-échantillon i.i.d. Y_1, ..., Y_n de même loi que Y. On prend la valeur observée de $\overline{Y_n}$ comme estimation de M.
3. En Python, on utilise si possible `rd.random(n)`, ou `rd. ...(n)` (pour toute loi usuelle) pour générer directement les n valeurs de Y_1,..., Y_n (sinon boucle `for`). On compte le nombre de succès S_n, puis $\overline{Y_n} = \frac{S_n}{n}$.

Exemple
Un classique : méthode de Monte-Carlo pour trouver une valeur approchée de π
On considère une cible carrée ABCD de côtés les points $A(1,1)$, $B(-1,1)$, $C(-1,-1)$, $D(1-,-1)$ ainsi que le cercle unité $\mathcal{C}$ de centre $O(0,0)$ et de rayon 1.

1. Représenter le carré et le cercle dans le plan. Quelle est l'équation du cercle $\mathcal{C}$? A quelle condition un point $M(x,y)$ est-il à l'intérieur du cercle $\mathcal{C}$?
2. On effectue une succession de lancers d'une fléchette. On suppose que les lancers sont indépendants, que chaque lancer atteint la cible carrée, de façon aléatoire, c'est-à-dire que l'abscisse X et l'ordonnée Y d'une fléchette sont des variables aléatoires indépendantes suivant la loi uniforme $\mathcal{U}([-1,1])$.
Quelle est la probabilité qu'une fléchette soit à l'intérieur du cercle ?
3. En déduire un programme `MonteCarloPi(n):` qui simule n lancers de fléchettes, calcule la fréquence avec laquelle on atteint l'intérieur du cercle, et retourne une valeur approchée de π. Exécuter ce programme pour $n = 1000$, pour $n = 100000$, pour $n = 10000000$.
Que peut on dire quand à cette méthode ?
4. Une méthode déterministe : on sait que $\sum_{k=1}^{n} \frac{1}{k^2} = \frac{\pi^2}{6}$. En déduire une autre méthode de calcul d'une valeur approchée de π et taper le programme correspondant. Comparer les temps de calcul.

Exercice 3
Soit $M = \int_0^1 \frac{4}{1+t^2} dt$.

1. Ecrire M comme l'espérance d'une variable aléatoire Y. Justifier que $V(Y)$ existe et $V(Y) \leq 16$.
2. (a) Ecrire un programme Python, qui définit la fonction g où $g(t) = \frac{4}{1+t^2}$ puis qui affiche une estimation de M par la méthode de Monte-Carlo pour un échantillon de taille n.
   (b) Calculer la valeur exacte de M.
   Comparer avec les résultats obtenus précédemment.

Exercice 4
De quel réel le résultat affiché par le script Python ci-dessous devrait-il être proche pour n suffisamment grand ?

```
n=int(input('n='))
Y=rd.exponential(3,n)
M=np.mean(np.exp(-Y))
print(M)
```

Exercice 5
Dans chaque cas, justifier la convergence de l'intégrale puis écrire un programme utilisant la méthode de Monte-Carlo qui permet d'obtenir une estimation ponctuelle de M.

1. $M = \int_0^{+\infty} \sin(t).e^{-2t}\ dt$
   Adapter le programme pour estimer $M_1 = \int_1^{+\infty} \sin(t).e^{-2t}\ dt$
2. $M = \int_0^{+\infty} \frac{e^{-t^2}}{1+t^2}\ dt$.

Exercice 6
Extrait Edhec 2017
On peut montrer par un changement de variables que

$$I(x) = \int_x^{+\infty} t^k.e^{-t}\ dt = e^{-x}.\int_0^{+\infty} (u+x)^k.e^{-u} du$$

Compléter le script suivant pour qu'il calcule une estimation de $I(x)$ par la méthode de Monte-Carlo.

```
x=float(input('x='))
k=int(input('k='))
Z=rd.exp(1,100000)
s=np.exp(-x)*np.mean(.............................)
print(s)
```

Exercice 7
Dans chaque cas, écrire M comme l'espérance d'une certaine variable aléatoire, puis écrire un programme Python pour déterminer ,via la méthode de Monte-Carlo, une estimation ponctuelle de M.

1. $M = \sum_{k=1}^{+\infty} \frac{1}{k^2.2^k}$
2. $M = \sum_{k=0}^{+\infty} \frac{\sqrt{k}}{k!}$
3. $M = \sum_{k=0}^{10} \frac{\binom{10}{k}}{k^2+1}$