Math ECG 2. 2023-2024

Mathématiques Appliquées - F. Gaunard
http://frederic.gaunard.com
ENC Bessières, Paris 17e.

Chapitre 15. Estimation

1 Introduction

Dans tous les chapitres précédents concernant les probabilités, les lois des processus aléatoires étaient connues, et c'est à partir de ces lois que nous avons fait les calculs.

Adoptons alors une approche différente d'ordre pratique; on **ne connait pas** la loi de X_i. Par exemple la durée de vie d'un composant électronique, le nombre de clients se présentant au guichet d'une banque un jour donné, la proportion d'un certain caractère dans une population, le résultat d'un scrutin à partir des premiers dépouillements...

Il peut être raisonnable de penser qu'elle appartient à une famille de lois usuelles (ce qui définit le **modèle** de loi) qui dépend d'un paramètre (inconnu) qu'on aimerait donc **estimer** à partir du résultat $(x_1, x_2, ..., x_n)$ de n expériences.

☞ Il s'agit bien d'une estimation car on ne peut pas faire une infinité d'expérience de même qu'on ne peut pas interroger l'ensemble de la population.

Ainsi, le $n-$uplet $(x_1, x_2, ..., x_n)$ correspond alors à l'**observation** d'un **échantillon**, c'est à dire la réalisation pour une certaine issue $\omega \in \Omega$ et ce sont des données dont on dispose.
On peut alors, à partir des observations, vouloir produire une valeur approchée du paramètre (par exemple de l'espérance), ce qui s'appelle une *estimation ponctuelle* (à l'aide d'une *formule* bien choisie appliquée aux données) ou vouloir fournir un intervalle qui contient le paramètre avec une très grande probabilité, ce qu'on appelle *estimation par intervalle de confiance*.

On ne résiste pas, en conclusion de ce cours de *Mathématiques appliquées*, de commencer ce chapitre par une application, qui devrait (un peu) motiver à l'introduction des notions qui suivront.

Un exemple historique. Le *German Tank Problem*

☞ À partir des numéros de série observés sur des tanks ennemis, peut-on *estimer* le nombre total d'engins des forces adverses ?

Le contexte. Été 1943, les Alliés essaient de percer le bloc de l'Axe en créant un nouveau front via l'Italie. Ils rencontrent un nouveau type de char allemand, le bien nommé *Sonderkraftfahrzeug 171* plus connu des afficionados de machines de combat sous le nom de *Panther*.

Ce dernier est mieux équipé et plus performant que ceux rencontrés jusqu'alors. Il a été conçu en réponse à l'excellent *T-34* utilisé par les soviétiques sur le front de l'Est.

Sans rentrer dans les détails d'armement de cette subtile et sympathique machine, il peut percer les défenses et détruire la majorité des tank alliés.

Néanmoins, malgré sa puissance théorique, celui-ci ne peut avoir un réel impact sur l'issue de la guerre que si le nombre d'unités produites est suffisant. Il apparait crucial pour les Alliés de déterminer ou plutôt d'*estimer* combien de *Panther* étaient produits. La tâche fut confiée à [la] *Economic Warfare Division of the American Embassy in London*[1].

La modélisation. On suppose que l'ennemi produit une série de chars immatriculés par des entiers en commençant par 1. En plus de cela, quelle que soit la date de production du char, ses années de service, ou encore son numéro de série, la distribution des numéros d'immatriculation est considérée comme étant uniforme dès l'instant où on mène l'analyse.

Dans notre modélisation, les allemands disposent de N tanks numérotés de 1 à N. Les force alliées observent aléatoirement, uniformément et "avec remise" n numéros de séries $(X_1, ..., X_n)$ et cherchent à estimer le paramètre N.

On considère dans tout le problème un $n-$échantillon $(X_1, ..., X_n)$ de la loi $\mathcal{U}(\llbracket 1, N \rrbracket)$ et une première idée serait de considérer la *moyenne* des valeurs observées, on commence donc par poser

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

(1) Que vaut $E(\overline{X}_n)$. Il serait *pratique* qu'en moyenne, la variable aléatoire choisie pour estimer N renvoie N. Expliciter alors une variable aléatoire T_n, fonction du $n-$échantillon $(X_1, ..., X_n)$ telle que $E(T_n) = N$.

(2) Calculer $V(T_n)$ et montrer, à l'aide de l'inégalité de Bienaymé-Tchebychev, que

$$\forall \varepsilon > 0, \qquad \lim_{n\to+\infty} P\left(|T_n - N| > \varepsilon\right) = 0.$$

(3) Ce résultat semble affirmer que l'*estimateur* T_n *converge* (dans un certain sens) vers N, c'est à dire que si n est assez grand (si on dispose de suffisamment de données) la valeur approchée de N obtenue avec la définition de T_n appliquée à l'observation est proche de N. En revanche, imaginons qu'on ait 5 données $X = [8, 322, 15, 135, 69]$, que vaut l'estimation obtenue avec T_5 correspondant à cette observation? Que cela motive-t-il ?

On introduit alors le nouvel *estimateur*, c'est à dire une nouvelle fonction du $n-$échantillon

$$M_n = \max(X_1, ..., X_n)$$

(4) Calculer, pour tout $k \in \mathbb{N}$, $P(M_n \leq k)$.

(5) Soit Y un v.a à valeurs dans $\llbracket 1; N \rrbracket$. Montrer que $E(Y) = \sum_{k=0}^{N-1} P(Y > k)$.

(6) Montrer alors que

$$E(M_n) = N - \sum_{k=0}^{N-1} \left(\frac{k}{N}\right)^n.$$

(7) Vérifier que, pour tout $k \in \llbracket 0; N-1 \rrbracket$,

$$0 \leq \left(\frac{k}{N}\right)^n \leq N \int_{k/N}^{(k+1)/N} t^n \mathrm{d}t.$$

(8) En déduire que

$$N - \frac{N}{n+1} \leq E(M_n) \leq N$$

puis que

$$\lim_{n\to+\infty} E(M_n) = N.$$

(On dit que l'estimateur M_n est *asymptotiquement sans biais.*)

[1]Comme le raconte l'article *An Empirical Approach to Economic Intelligence in World War II*, R. Ruggles & H. Brodie, Journal of the American Statistical Association **42**-237 (1947), 72-91

Si l'estimateur M_n parait naturel, il a clairement un défaut; il sous-estime nécessairement N (puisqu'il renverra toujours une valeurs inférieure (ou égale) à N). On va donc essayer d'y apporter une légère *correction*.

Commençons par introduire le numéro du plus petit tank observé $m_n = \min(X_1, ..., X_n)$.

Comme N est inconnu, on ne connait pas l'écart entre N et M_n, mais il parait raisonnable de penser qu'il y a (en moyenne) autant de tanks *non observés* entre M_n et N qu'entre 1 et m_n. Entre le plus petit numéro observé et le tank avec le numéro de série 1, il y a $m_n - 1$ numéros de tank.
On pense alors à ajouter la correction

$$\tilde{M}_n = M_n + (m_n - 1).$$

(9) En s'inspirant des calculs précédents pour M_n, déterminer $E(m_n)$ sous forme d'une somme qu'on ne cherchera pas à simplifier.

(10) Montrer que $\tilde{M}_n$ vérifie maintenant $E(\tilde{M}_n) = N$.

(11) **Comparaison des estimateurs**. On propose la fonction Python ci-dessous. Que fait-elle?

```
import numpy as np
import numpy.random as rd

def mystere(N, n) :
    T=[]
    M=[]
    for j in range(1000):
        X=[rd.randint(1, N+1) for k in range(n)]
        T.append(2*np.mean(X)-1)
        M.append(np.max(X)+np.min(X)-1)
    t=np.mean(T)
    m=np.mean(M)
    return [t,m]
```

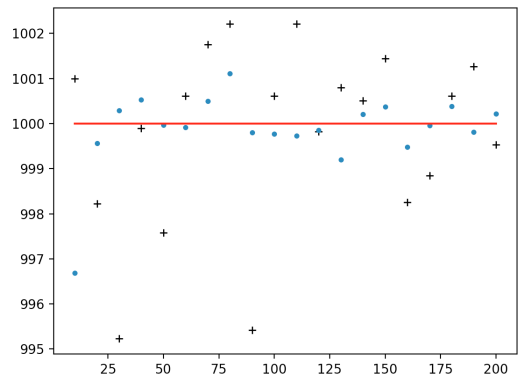
On ajoute les instructions ci-dessous dont l'exécution permet d'obtenir la figure ci-contre. Interpréter. Quel estimateur semble le plus performant?

(Les programmes Python de cet exemple sont disponibles en cliquant *ici*.)

```
T=[ ]
M=[ ]
N=1000
x=[10*n for n in range(1, 21)]
for n in x:
    [t,m]=mystere(N,n)
    T.append(t)
    M.append(m)

plt.plot(x, T, 'k+')
plt.plot(x, M, '.')
plt.plot(x, [N for k in x], 'red')
plt.show()
```

Affichage Python



(12) Choisir alors l'estimateur le plus performant pour proposer une estimation du nombre de tanks ennemis à partir des données top secrètes transmises par les service de renseignement, au péril de leur vie.

```
X=[14,   44,   50,   101,   117,   127,   134,   139,   165,   188,
192,  201,   204,  215,   234,   243,   244,   253,   269,   269,
282,  287,   288,  322,   345]
```

2 Estimation ponctuelle

2.1 Modèle statistique

Définition

Soit X une variable aléatoire définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. On appelle $n-$**échantillon** de X un $n-$uplet $(X_1, ..., X_n)$ de v.a.i.i.d, de même loi que X.

Soit $(X_1, ..., X_n)$ un $n-$échantillon de X. Pour tout $\omega \in \Omega$, on appelle réalisation du $n-$échantillon $(X_1, ..., X_n)$ le $n-$uplet $(X_1(\omega), ..., X_n(\omega))$.

☞ Par convention, on note souvent une réalisation $(X_1(\omega), ..., X_n(\omega))$ par des minuscules $(x_1, ..., x_n)$.

À retenir!

☞ Il faut bien distinguer le $n-$échantillon $(X_1, ..., X_n)$, qui est une variable aléatoire (et donc une fonction), de la réalisation $(X_1(\omega), ..., X_n(\omega))$ qui est un élément de $\mathbb{R}^n$.

Définition

Soit Θ un sous-ensemble de $\mathbb{R}$. Un **modèle statistique** est un ensemble de lois $\mathcal{M}_\Theta$. Étant donnée une observation $(x_1, x_2, ..., x_n)$ de X, le but (du statisticien) est d'identifier la loi, parmi celles de $\mathcal{M}_\Theta$, ayant permis de la générer.

Exemple

On dispose d'une pièce dont on ne connait pas la probabilité de tomber sur *Pile*, que l'on note p, et que l'on aimerait estimer. On considère alors une v.a. X (qui vaut 1 en cas de *Pile* et 0 sinon) et qu'on va bien entendu identifier parmi un ensemble de lois de Bernoulli, ici $\mathcal{M}_\Theta = \{\mathcal{B}(p); p \in [0; 1]\}$.

On lance cette pièce n fois et on note X_i la v.a. qui vaut 1 si la pièce tombe sur *Pile* au $i-$ème lancer et 0 sinon. le $n-$uplet $(X_1, X_2, ..., X_n)$ est un $n-$échantillon de X.

On décide de prendre $n = 10$. Le résultat des 10 lancers donne $(1, 1, 0, 0, 1, 1, 0, 0, 0, 1)$ qui représente donc une réalisation du $10-$échantillon $(X_1, X_2, ..., X_{10})$.

☞ Il va donc falloir introduire un *estimateur*, c'est à dire une *fonction* de l'échantillon qui, appliquée à l'observation donnera une estimation du paramètre cherché.

2.2 Estimateurs

Définition

Soit $(X_1, X_2, ..., X_n)$ un $n-$échantillon d'une v.a. X dont la loi dépend d'un paramètre θ, θ appartenant à une partie $\Theta \subset \mathbb{R}$. On appelle **estimateur** de θ toute variable aléatoire T_n de la forme

$$T_n = \varphi(X_1, X_2, ..., X_n),$$

où φ est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$,) valeurs dans Θ, éventuellement dépendante de n **mais indépendante de** θ. Toute réalisation d'un estimateur est appelée **estimation ponctuelle** de θ.

Exemple

Reprenons l'exemple précédent. La variable aléatoire

$$\overline{X}_{10} = \frac{X_1 + X_2 + ... + X_{10}}{10}$$

est un estimateur de p. À partir de l'observation $(1, 1, 0, 0, 1, 1, 0, 0, 0, 1)$, $\overline{X}_{10}$ donne une estimation (ponctuelle) de $1/2$ pour p.

☞ Certains estimateurs visent à estimer non pas le paramètre mais une fonction $g(\theta)$ du paramètre.

Par exemple, en reprenant le contexte ci-dessus, on peut vouloir estimer la variance de X, à savoir la quantité $g(p) = p(1-p)$, qui est bien une fonction du paramètre p. Les deux variables aléatoires ci-dessous alors sont des estimateurs de $g(p)$

$$\frac{1}{10}\sum_{i=1}^{10}(X_i - \overline{X}_{10})^2, \qquad \overline{X}_{10}(1 - \overline{X}_{10}).$$

Remarque

☞ La notion d'estimateur parait alors très floue (ou très générale), dans le sens "destinée à approcher la valeur de θ" puisqu'en réalité, toute fonction des données est donc un estimateur... Le choix de l'estimateur est une vraie question. La méthode du maximum de vraisemblance ci-après propose une réponse.

☞ La notion suivante a quitté le (nouveau) programme de Mathématiques appliquée. On peut quand même raisonnablement estimer que les exercice demanderont de calculer l'espérance d'un estimateur et que de toute façon, celle-ci reste pertinente pour la compréhension des problématiques du chapitre. Elle répond notamment à la remarque ci-avant.

Définition hors programme

Soit T_n un estimateur de $g(\theta)$ admettant une espérance $E_\theta(T_n)$. (pour tout $\theta \in \Theta$). On appelle **biais** de T_n le nombre

$$b_\theta(T_n) = E_\theta(T_n) - g(\theta)$$

c'est à dire l'écart "moyen" entre l'estimateur et le paramètre à estimer. Lorsque le biais est nul, on dit que T_n est **sans biais** ou **non biaisé**.

Une variable aléatoire T_n est donc un estimateur non biaisé de $g(\theta)$ si et seulement si c'est une fonction du $n-$échantillon indépendante de $g(\theta)$ vérifiant $E_\theta(T_n) = g(\theta)$.

Propriété hors programme

Soient X une v.a. d'espérance m et $(X_1, ..., X_n)$ un $n-$échantillon de X. Alors, la *moyenne empirique*

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + ... + X_n)$$

est un estimateur sans biais de m.

Exercice 1. Soit $(X_1, X_2, ..., X_n)$ un $n-$échantillon d'une variable $X \in \mathcal{M}_\Theta = \{\mathcal{U}([0;\theta]); \theta \in \mathbb{R}_+\}$.

(1) Montrer que $\overline{X}_n$ est un estimateur biaisé et θ et préciser $b_\theta(\overline{X}_n)$.
(2) Proposer alors un estimateur V_n de θ sans biais, obtenu comme transformation simple de $\overline{X}_n$.
(3) On considère maintenant l'estimateur $M_n = \max(X_1, X_2, ..., X_n)$.
   (a) Déterminer la fonction de répartition F de M_n et en déduire une densité f de M_n.

(b) Montrer alors que

$$E_\theta(M_n) = \frac{n\theta}{n+1}.$$

(c) En déduire un estimateur sans biais Z_n à partir de M_n.

Exercice 2. (**Estimation de la variance**) Soit $(X_1, ..., X_n)$ un $n-$échantillon d'une v.a. X admettant pour espérance m et pour variance σ^2.

(1) On suppose que m est connu. Montrer que T_n est un estimateur non biaisé de σ^2, où

$$T_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2.$$

(2) On suppose que m n'est pas connu. On note $\overline{X}_n$ la moyenne empirique de l'échantillon et

$$U_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

(a) Montrer que, pour tout $i \in [\![1; n]\!]$, $E((X_i - \overline{X}_n)^2) = V(X_i - \overline{X}_n)$.

(b) Montrer que, pour tout $i \in [\![1; n]\!]$,

$$V(X_i - \overline{X}_n) = \left(1 - \frac{1}{n}\right)^2 V(X_i) + \frac{1}{n^2}\sum_{k\neq i} V(X_k).$$

(c) En déduire que

$$V(X_i - \overline{X}_n) = \frac{n-1}{n}\sigma^2.$$

(d) Montrer alors que U_n est un estimateur biaisé de σ^2. En déduire un estimateur sans biais de σ^2.

☞ Disposer d'un estimateur sans biais semble présenter un réel intérêt puisque son espérance est égale au paramètre cherché, mais ce seul fait ne garantit pas que l'estimateur fournit de bonnes approximations. Pour évaluer le défaut de moyenne, et juger de la qualité d'un estimateur, on calcule la moyenne des carrés des écarts au paramètre.

Définition hors programme

Si T_n est un estimateur de $g(\theta)$ et admet un moment d'ordre 2 pour toute valeur de $\theta \in \Theta$, on appelle **risque quadratique** de T_n le réel

$$r_\theta(T_n) = E_\theta((T_n - g(\theta))^2).$$

Si T_n et U_n sont deux estimateurs de $g(\theta)$, on dit que T_n est meilleur que U_n si et seulement si $r_\theta(T_n) \leq r_\theta(U_n)$ pour tout $\theta \in \Theta$.

☞ Le risque quadratique dépend (potentiellement) de θ et de n.

☞ $r_\theta(T_n)$ est toujours positif (ou nul). Dans le cas où $r_\theta(T_n) = 0$, alors $T_n = \theta$ presque sûrement. C'est donc l'estimateur parfait!

Propriété hors programme

Soit T_n un estimateur de $g(\theta)$ admettant espérance et variance (pour tout $\theta \in \Theta$). Alors,

$$r_\theta(T_n) = (b_\theta(T_n))^2 + V_\theta(T_n).$$

☞ Si T_n est un estimateur sans biais de $g(\theta)$, son risque quadratique est alors égal à sa variance et il suit que, parmi deux estimateurs sans biais, celui avec la plus petite variance est le meilleur.

Exercice 3. On reprend les notations de l'Exercice 1. C'est à dire que $X \in \mathcal{M}_\Theta = \{\mathcal{U}([0;\theta]); \theta \in \mathbb{R}_+\}$ et on a les deux estimateurs sans biais de θ

$$V_n = \frac{2}{n}(X_1 + X_2 + ... + X_n), \qquad Z_n = \frac{n+1}{n}\max(X_1, X_2, ..., X_n).$$

(1) Établir que

$$r_\theta(V_n) = V_\theta(V_n) = \frac{\theta^2}{3n}.$$

(2) Montrer que

$$E_\theta(Z_n^2) = \frac{(n+1)^2}{n^2}\int_0^\theta \frac{nt^{n+1}}{\theta^n}\mathrm{d}t.$$

En déduire que

$$r_\theta(Z_n) = V_\theta(Z_n) = \frac{\theta^2}{n(n+2)}$$

(3) Quel estimateur aura-t-on tendance à préférer en pratique?

Définition

Un estimateur T_n de $g(\theta)$ est dit convergent si, pour tout $\theta \in \Theta$,

$$\forall \varepsilon > 0, \qquad \lim_{n\to+\infty} P_\theta(|T_n - g(\theta)| > \varepsilon) = 0.$$

☞ Un estimateur convergent s'écarte donc du paramètre à estimer avec très faible probabilité si la taille de l'échantillon est assez grande.

À retenir!

La loi faible des grands nombres implique notamment que la *moyenne empirique* $\overline{X}_n$ d'un $n-$échantillon est un estimateur (non biaisé et) convergent de l'espérance.

Propriété hors programme

Soit T_n un estimateur tel quel, pour tout $\theta \in \Theta$, on ait

$$\lim_{n\to+\infty} r_\theta(T_n) = 0.$$

Alors, T_n est un estimateur convergent.

Preuve. Soient $\theta \in \Theta$ et $n \in \mathbb{N}^*$. D'après l'inégalité de Markov, pour tout $\varepsilon > 0$, on a

$$P\left(|T_n - g(\theta)| > \varepsilon\right) = P\left(|\left(T_n - g(\theta)\right)^2| > \varepsilon^2\right) \leq \frac{r_\theta(T_n)}{\varepsilon^2} \underset{n\to+\infty}{\longrightarrow} 0.$$

□

2.3 Méthode du maximum de vraisemblance

Nous avons en vu plus haut notamment que la loi (faible) des grands nombres fournit *naturellement* un estimateur de l'espérance d'une loi, mais si l'on recherche une méthode un peu générale pour *deviner* un estimateur, la méthode du *maximum de vraisemblance* est une stratégie souvent efficace. En voici le principe :

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

$$[X_1 = x_1] \cap [X_2 = x_2] \cap ... \cap [X_n = x_n].$$

Par hypothèse d'indépendance sur les variables du $n-$échantillon, la probabilité de l'évènement ci-dessus 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 $(X_1, ..., X_n)$ un $n-$échantillon d'une loi discrète $\mathcal{L}(\theta)$ où $\theta \in \Theta$ est un paramètre qu'on cherche à estimer et $(x_1, ..., x_n) \in X_1(\Omega)^n$ fixé. La fonction L_n définie sur Θ par

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

s'appelle la **vraisemblance** de la loi $\mathcal{L}$.
En notant $\theta^* = \varphi(x_1, ..., x_n)$ la valeur où L_n est maximale (c'est à dire telle que, pour tout $\theta \in \Theta$, ${}_n(\theta) \leq L_n(\theta^*)$, l'**estimateur du maximum de vraisemblance** est l'estimateur défini par

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

Exemple

Estimateur du maximum de vraisemblance pour la loi $\mathcal{B}(p)$.

Soient $(X_1, ..., X_n)$ un $n-$échantillon de la loi $\mathcal{B}(p)$ et $(x_1, ...x_n) \in \{0; 1\}^n$. Ici le paramètre θ à estimer est p. En observant que

$$\text{Card}\left(\{i \in [\![1, n]\!] : x_i = 1\}\right) = \sum_{i=1}^{n} x_i = s_n,$$

on voit que

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

Les extrémités de $[0; 1]$ ne peuvent être des *extrema*, sauf si $s_n = 0$ ou $s_n = n$. On peut donc supposer que $\theta \in]0; 1[$. On pose alors

$$h_n(\theta) = \ln(L_n(\theta)) = s_n \ln(\theta) - (s_n - n)\ln(1-\theta).$$

Il est facile de dériver et d'étudier les variations de h_n. Pour $\theta \in]0; 1[$, $h_n'(\theta) = \dfrac{s_n - n\theta}{\theta(1-\theta)}$.

Ainsi, h_n est maximale en $\theta^* = \dfrac{s_n}{n} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$.

Par croissance de l'exponentielle, comme h_n est maximale en $\theta^* = s_n/n$, il en est de même pour L_n. L'estimateur du maximum de vraisemblance est alors, dans le cas de lois de Bernoulli

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

c'est à dire la moyenne empirique.

Exercice 4. (Maximum de vraisemblance pour la loi de Poisson).

On considère un $n-$échantillon $(X_1, .., X_n)$ d'une loi de Poisson de paramètre $\lambda > 0$ inconnu que l'on cherche à estimer, ainsi que $(x_1, .., x_n)$ un $n-$uplet de $\mathbb{N}^n$ fixé.

(1) En notant

$$s_n = \sum_{i=1}^{n} x_i, \qquad p_n = \prod_{i=1}^{n}(x_i!),$$

exprimer la fonction de vraisemblance L_n de la loi de Poisson, définie sur $\mathbb{R}_+^*$.

(2) Calculer $L_n'(\theta)$ pour tout $\theta > 0$ et vérifier que L_n est maximale en

$$\theta^* = \frac{s_n}{n}.$$

(3) Quelle est donc l'expression de l'estimateur de maximum de vraisemblance pour la loi de Poisson?

Remarque

☞ Dans les deux exemples précédents, c'est la *moyenne empirique* qui apparait être l'estimateur du maximum de vraisemblance, ce qui justifie aussi que ce soit un estimateur *naturel* à introduire (notamment dans notre problème des tanks). Cependant, ce n'est pas toujours le cas: l'estimateur du maximum de vraisemblance est parfois différent.

Exercice 5. Montrer que l'estimateur du maximum de vraisemblance de la loi $\mathcal{G}(p)$ est donné pour un $n-$échantillon $(X_1, ..., X_n)$ par la formule

$$\hat{\theta}_n = \frac{n}{X_1 + ... + X_n}.$$

Remarque

On peut aussi définit la vraisemblance d'une loi continue. En notant f_θ la densité d'une variable X échantillonnée de loi inconnue $\mathcal{L}(\theta)$, et $(x_1, ..., x_n)$ un $n-$uplet de valeurs de $X(\Omega)$, il s'agit de la fonction

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

L'estimateur du maximum de vraisemblance est alors défini de la même manière que précédemment.

L'exercice 16, extrait de **EDHEC 2012** en propose un exemple.

Exercice 6. On considère un $n-$échantillon de la loi $\mathcal{U}([0;\theta])$ et on cherche à estimer $\theta > 0$.
Soit $(x_1, ..;, x_n) \in \left(\mathbb{R}_+^*\right)^n$ fixé. On note f_θ la densité de notre loi uniforme. 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 $L_n(\theta) = \begin{cases} \theta^{-n}, & \text{si } \theta \geq \max(x_1, ..., x_n) \\ 0, & \text{sinon} \end{cases}$

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

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

3 Estimation par intervalles de confiance

3.1 Motivation

À chaque estimation (observation d'un estimateur), correspond une valeur approchée, de précision non spécifiée, du paramètre θ. On peut vouloir préciser l'erreur commise (avec grande probabilité), c'est à dire déterminer un intervalle, construit sur l'observation, contenant θ avec une probabilité très élevée. C'est l'estimation par intervalle de confiance.

Par exemple, considérons un $n-$échantillon d'une loi de Bernoulli $X \hookrightarrow \mathcal{B}(p)$ dont on veut estimer p. On a pu voir précédemment que l'estimateur $T_n = \overline{X}_n$ est un estimateur non biaisé et convergent (son risque quadratique vaut $r_\theta(T_n) = p(1-p)/n$) de p. À n fixé, l'inégalité de Bienaymé-Tchebychev donne alors

$$\begin{aligned} P\left(|T_n - p| > \varepsilon\right) \leq \frac{p(1-p)}{n\varepsilon^2} &\iff P\left(|T_n - p| \leq \varepsilon\right) \geq 1 - \frac{p(1-p)}{n\varepsilon^2} \\ &\iff P\left(T_n \in [p-\varepsilon; p+\varepsilon]\right) \geq 1 - \frac{p(1-p)}{n\varepsilon^2} \\ &\iff P\left(p \in [T_n-\varepsilon; T_n+\varepsilon]\right) \geq 1 - \frac{p(1-p)}{n\varepsilon^2} \end{aligned}$$

Considérons donné le problème suivant : étant donné un réel $\alpha \in]0;1[$ (appelé *niveau de confiance* ou seuil), déterminer un intervalle I_α (appelé *intervalle de confiance*) ne dépendant pas de p, contenant la vraie valeur de p avec probabilité supérieure à $1-\alpha$.

Par l'observation précédente, il suffit de déterminer ε tel que

$$1 - \frac{p(1-p)}{n\varepsilon^2} \geq 1 - \alpha \iff \varepsilon \geq \sqrt{\frac{p(1-p)}{\alpha n}}.$$

On a alors

$$P\left(p \in \left[T_n - \sqrt{\frac{p(1-p)}{\alpha n}}; T_n + \sqrt{\frac{p(1-p)}{\alpha n}}\right]\right) \geq 1 - \alpha.$$

Il apparaît alors un problème; l'intervalle censé encadrer p dépend lui-même de p. Mais qu'à cela ne tienne, on voit que $p \mapsto p(1-p)$ est majorée par $1/4$ sur $[0;1]$ ou encore que

$$\left[T_n - \sqrt{\frac{p(1-p)}{\alpha n}}; T_n + \sqrt{\frac{p(1-p)}{\alpha n}}\right] \subset \left[T_n - \sqrt{\frac{1}{4\alpha n}}; T_n + \sqrt{\frac{1}{4\alpha n}}\right]$$

et on peut donc proposer un encadrement

$$P\left(p \in \left[T_n - \sqrt{\frac{1}{4\alpha n}}; T_n + \sqrt{\frac{1}{4\alpha n}}\right]\right) \geq 1 - \alpha.$$

3.2 Intervalles de confiance

Définition

Soient $(X_1, ..., X_n)$ un $n-$échantillon de la loi X dépendant d'un paramètre θ, U_n et V_n deux estimateurs de $g(\theta)$ tels que, pour tous $n \in \mathbb{N}$ et $\theta \in \Theta$, $P(U_n \leq V_n) = 1$. Soit $\alpha \in [0;1]$.
On dit que l'intervalle $[U_n, V_n]$ est un **intervalle de confiance** au niveau $1-\alpha$ (ou au risque α) pour $g(\theta)$ si, pour tout $\theta \in \Theta$,

$$P\left(g(\theta) \in [U_n, V_n]\right) \geq 1 - \alpha$$

ou de manière équivalente

$$P\left(g(\theta) \notin [U_n; V_n]\right) \leq \alpha.$$

À retenir!

☞ En pratique, le risque α étant donné, on utilise un estimateur non biaisé T_n de $g(\theta)$ et on cherche $\varepsilon > 0$ tel que $P(|T_n - g(\theta)| \leq \varepsilon) \geq 1 - \alpha$ à l'aide de l'inégalité de Bienaymé-Tchébychev ou du théorème central limite (approximation par la loi normale).

☞ Les bornes de l'intervalle de confiance **ne doivent jamais dépendre** du paramètre à estimer.

Exercice 7. Dernier retour sur l'Exercice 1. On considère à nouveau l'estimateur V_n du paramètre θ de la loi $x \in \mathcal{M}_\Theta = \{\mathcal{U}([0;\theta]),\ \theta \geq 0\}$.

(1) Montrer, à l'aide de l'inégalité de Bienaymé-Tchébychev, que

$$P(|V_n - \theta| > \varepsilon) \leq \frac{\theta^2}{3n\varepsilon^2}.$$

(2) Montrer que

$$\theta \in \left[V_n - \sqrt{\frac{\theta^2}{3n\alpha}}; V_n + \sqrt{\frac{\theta^2}{3n\alpha}}\right] \iff \frac{V_n}{1 + \frac{1}{\sqrt{3n\alpha}}} \leq \theta \leq \frac{V_n}{1 - \frac{1}{\sqrt{3n\alpha}}}.$$

(3) En déduire un intervalle de confiance au risque α pour θ.

3.3 Intervalles de confiance asymptotiques

Définition

Soient $(X_1, ..., X_n)$ un $n-$échantillon de la loi X dépendant d'un paramètre θ, U_n et V_n deux estimateurs de $g(\theta)$ tels que, pour tous $n \in \mathbb{N}$ et $\theta \in \Theta$, $P(U_n \leq V_n) = 1$. Soit $\alpha \in [0;1]$.

On dit que l'intervalle $[U_n, V_n]$ est un **intervalle de confiance asymptotique** pour $g(\theta)$ au niveau $1-\alpha$ (ou au risque α) si il existe une suite (α_n) de réels de $[0;1]$ vérifiant $\alpha_n \longrightarrow \alpha$, $n \to +\infty$, et telle que, pour tout $\theta \in \Theta$,

$$P\left(g(\theta) \in [U_n, V_n]\right) \geq 1 - \alpha_n$$

ou de manière équivalente

$$\lim_{n\to+\infty} P\left(U_n \leq g(\theta) \leq V_n\right) \geq 1 - \alpha.$$

À retenir!

☞ Les intervalles de confiance asymptotiques sont toujours déduits d'une convergence en loi. En particulier, le théorème central limite est très utile lorsque l'estimateur T_n est la moyenne empirique.

Exercice 8. Afin d'avoir une idée du nombre de poissons N présents dans un étang, on en pêche une certaine quantité K, que l'on marque, puis que l'on remet à l'eau. On revient un jour plus tard, et l'on pêche n poissons (avec remise pour simplifier). Pour tout $i \in\in [\![1;n]\!]$, on note X_i la variable qui vaut 1 si le $i-$ième poisson pêché est marqué ou non. On suppose que ces variables sont indépendantes, et on introduit la moyenne empirique $\overline{X}_n$.

(1) Déterminer la loi de X_i puis calculer l'espérance et la variance de $\overline{X}_n$.
(2) Montrer que, si on note $p = K/N$, alors

$$\sqrt{\frac{n}{p(1-p)}}\left(\overline{X}_n - p\right) \xrightarrow{\mathcal{L}} X, \qquad X \hookrightarrow \mathcal{N}(0;1).$$

(3) Déterminer alors un intervalle de confiance asymptotique de risque α pour $p = K/N$ construit sur $\overline{X}_n$. En déduire un intervalle de confiance asymptotique de même risque pour N construit sur $\overline{X}_n$.
(4) *Application numérique.* $K = 50$, $n = 300$, $\overline{X}_n = 0.6, \alpha = 0.05$.

3.4 Paramètre d'une Bernoulli. Comparaison des intervalles de confiance

On considère un $n-$échantillon $(X_1, ..., X_n)$ d'une loi de Bernoulli $X \hookrightarrow \mathcal{B}(p)$. On cherche un intervalle de confiance pour p au risque α. On sait que $T_n = \overline{X}_n$ est un estimateur non biaisé de p.

- Comme vu précédemment, l'inégalité de Bienaymé-Tchébychev fournit l'intervalle de confiance cherché

$$P\left(p \in \left[T_n - \sqrt{\frac{1}{4\alpha n}}; T_n + \sqrt{\frac{1}{4\alpha n}}\right]\right) \geq 1 - \alpha.$$

- On peut aussi procéder avec le théorème central limite, ce qu'on a davantage tendance à faire. En effet, ce dernier permet d'affirmer que

$$T_n^* = \sqrt{\frac{n}{p(1-p)}}\left(T_n - p\right) \xrightarrow{\mathcal{L}} Z, \quad Z \hookrightarrow \mathcal{N}(0;1).$$

  On va encore utiliser le fait que $p(1-p) \leq 1/4$ (et donc que $\sqrt{n/p(1-p)} \geq 2\sqrt{n}$). On note Φ la fonction de répartition de la loi normale centrée réduite et, pour $\alpha \in]0;1[$, on utilise la notation standard

$$t_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

On a donc

$$\begin{aligned} 1-\alpha &= P(|Z| \leq t_\alpha) \\ &\underset{n\to+\infty}{\longleftarrow} P\left(\sqrt{\frac{n}{p(1-p)}}\,|T_n - p| \leq t_\alpha\right) = P\left(|T_n - p| \leq t_\alpha\sqrt{\frac{p(1-p)}{n}}\right) \\ &\leq P\left(|T_n - p| \leq t_\alpha\sqrt{\frac{1}{4n}}\right) \end{aligned}$$

Et on obtient alors un autre intervalle de confiance asymptotique au risque α

$$\lim_{n\to+\infty} P\left(p \in \left[T_n - \frac{1}{2\sqrt{n}}t_\alpha; T_n + \frac{1}{2\sqrt{n}}t_\alpha\right]\right) \geq 1-\alpha.$$

Examinons en détails les avantages et inconvénients de ces deux intervalle: ☞ Clairement, l'avantage du premier intervalle par rapport au deuxième est son caractère exact, c'est-à-dire non asymptotique.

☞ Cependant, l'intervalle asymptotique possède une étendue plus restreinte, et donne donc une estimation plus précise de p.

Lequel préférer? On voit que, pour de petites valeurs de n, l'intervalle asymptotique fournira des bornes trop restreintes, et que l'inégalité

$$P(g(\theta) \in [U_n; V_n]) \geq 1-\alpha$$

ne sera pas du tout respectée. La question se reformule donc mieux ainsi : à partir de quelle valeur de n est-il raisonnable d'échanger le caractère exact de l'intervalle de confiance contre un intervalle plus réduit? On considère qu'en pratique, on doit au moins avoir $n > 20$.

À connaître sur le bout des doigts

✎ La majoration de la quantité $p(1-p)$ par $1/4$ (sur l'intervalle $[0;1]$) est très pratique et très souvent utilisée, parfois sans rappel ni indication, dans les problèmes de concours.

Exercice 9. Dans une population d'un pays totalement fictif, des électeurs doivent choisir parmi 2 candidats, Jean-Michel Peste et Jean-Pierre Choléra, le futur président.
On note p la proportion d'électeurs désirant voter pour Jean-Michel. On choisit un échantillon $(X_1, ..., X_{100})$ où l'on a noté $X_i = 1$ si la personne a voté pour M. Peste, et 0 sinon.
Parmi ces personnes, 55% déclarent vouloir voter pour M. Peste.

Peut-on déclarer au risque $\alpha = 5\%$ que M. Peste sera élu président? On utilisera le théorème central limite.

Intervalle de confiance asymptotique avec variance inconnue

On a introduit dans l'Exercice 2 la *variance empirique* obtenue à partir de la moyenne empirique

$$\overline{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

On sait que c'est un estimateur biaisé de la variance σ^2 (on sait aussi comment corriger cet estimateur en un estimateur non biaisé). On peut montrer (mais pas avec les outils à disposition ici) que c'est aussi un estimateur convergent de σ^2.
Précédemment, on a *élargi* l'intervalle de confiance pour se débarasser de la présence du paramètre dans les bornes de l'intervalle. Ici, le problème est un peu différent, mais on peut utiliser l'estimateur ci-dessus pour former un intervalle de confiance pour l'espérance lorsque la variance est aussi inconnue.

Théorème

Soit X une variable aléatoire d'espérance m et de variance non nulle inconnue et $(X_1, X_2, ..., X_n)$ un $n-$échantillon de X. Alors, l'intervalle

$$\left[\overline{X}_n - t_\alpha \frac{\overline{S}_n}{\sqrt{n}};\ \overline{X}_n + t_\alpha \frac{\overline{S}_n}{\sqrt{n}}\right],$$

où $\Phi(t_\alpha) = 1 - \alpha/2$, est un intervalle de confiance asymptotique pour m au risque α.

☞ On peut remplacer $\overline{S}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$ par n'importe quel estimateur convergent de l'écart-type.

4 Sélection d'exercices - Travaux dirigés

4.1 Estimation ponctuelle

Exercice 10. On considère la variable aléatoire X dont la loi est donnée par

$$P(X = -1) = p, \quad P(X = 0) = 1 - 2p, \quad P(X = 1) = p,$$

pour un certain paramètre $p \in \left]0; \frac{1}{2}\right[$. On dispose d'un $n-$échantillon $(X_1, ..., X_n)$ de X, et on cherche à déterminer le paramètre p.

(1) Calculer $E(\overline{X}_n)$. En déduire que l'estimateur $\overline{X}_n$ est biaisé.
(2) Peut-on trouver des réels a et b tels que $aX_n + b$ soit un estimateur sans biais de p?

(3) On note $T_n = \frac{1}{2n}\sum_{i=1}^{n} X_i^2$.
  (a) Montrer que T_n est un estimateur sans biais de p.
  (b) Calculer la variance $V_\theta(T_n)$.
  (c) Montrer que l'estimateur est convergent.

Exercice 11. Soit a un réel strictement positif. On note f la fonction définie sur $\mathbb{R}$ par

$$f(x) = \begin{cases} 0, & \text{si } x < a \\ \dfrac{3a^3}{x^4}, & \text{si } x \geq a \end{cases}$$

(1) Montrer que f est une densité de probabilité.

Un capteur mesure **en permanence** le taux de gaz carbonique émis par un moteur. On suppose que le temps écoulé entre le démarrage du moteur et l'instant précis (en heures) où son taux de gaz carbonique devient non réglementaire est une variable aléatoire T de densité f.

(2) Montrer que T admet une espérance et une variance de valeurs:

$$E(T) = \frac{3a}{2}, \quad V(t) = \frac{3a^2}{4}.$$

(3) (a) Déterminer la fonction de répartition de T.
  (b) Calculer les probabilités $P(T > 2a)$ et $P_{(T>2a)}(T > 6a)$.

(4) On met en route n moteurs de modèle identique au précédent, et indépendants. On note $T_1, T_2, ..., T_n$ les temps respectifs pendant lesquels ces moteurs ont un taux de gaz carbonique réglementaire ($T_1, T_2, ..., T_n$ suivent donc la même loi que T et sont indépendantes).

  (a) Montrer que la variable

$$Z_n = \frac{2}{3n}\sum_{k=1}^{n} T_k$$

  est un estimateur sans biais du paramètre a.
  (b) Calculer son risque quadratique $r(Z_n)$. En déduire que T_n est un estimateur convergent.

Exercice 12. Soit (X_k) une suite de v.a.i.i.d. suivant une loi de Poisson $\mathcal{P}(\lambda)$, où $\lambda > 0$. On pose, pour tout $n \in \mathbb{N}^*$, $S_n = X_1 + X_2 + ... + X_n$.

(1) (a) Montrer que $S_2 \hookrightarrow \mathcal{P}(2\lambda)$.
(b) En déduire, par récurrence sur $n \in \mathbb{N}^*$, que $S_n \hookrightarrow \mathcal{P}(n\lambda)$.

(2) Pour $n \in \mathbb{N}$, $n \geq 2$, on pose, $Y_n = \left(1 - \dfrac{1}{n}\right)^{S_n}$.

(a) Montrer que Y_n est une v.a. discrète à valeurs dans $]0;1]$.
(b) Déterminer la loi de Y_n, son espérance et sa variance.
(c) En déduire un estimateur sans biais de $e^{-\lambda}$.

Exercice 13. La sécurité routière fait une enquête sur le nombre d'accidents survenus par semaine sur un tronçon d'autoroute. Soit X la v.a. égale au nombre d'accidents en une semaine. On suppose que $X \in \mathcal{M}_\Theta = \{\mathcal{P}(\lambda),\ \lambda > 0\}$. On se propose d'estimer le paramètre $e^{-\lambda} = P(X = 0)$. On note $(X_1, ..., X_n)$ un $n-$échantillon de X.

(1) Soit Y_n le nombre de fois où l'on a pas observé d'accident pendant la semaine, *i.e.*

$$Y_n = \text{Card}\left(\{i \in [\![1;n]\!];\ X_i = 0\}\right).$$

(a) Montrer que Y_n/n est un estimateur non biaisé de $e^{-\lambda}$.
(b) Déterminer son risque quadratique, noté ici $r(Y_n/n)$.

(2) Vérifier que $\overline{X}_n$ est un estimateur sans biais de λ.
(3) On pose $S_n = X_1 + ... + X_n$. Quelle est la loi de S_n? À l'aide du théorème de transfert, déterminer l'espérance de $e^{-\overline{X}_n}$. En déduire que $e^{-\overline{X}_n}$ est un estimateur biaisé de $e^{-\lambda}$.

Exercice 14. (D'après **EDHEC 2014**)

Dans cet exercice, θ désigne un réel strictement positif et n un entier naturel supérieur ou égal à 2. Pour tout k de $\mathbb{N}$, on pose

$$u_k = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^k.$$

(1) Montrer que la suite (u_k) définit bien une loi de probabilité.

On considère maintenant une variable aléatoire X prenant ses valeurs dans $\mathbb{N}$ et dont la loi est donnée par

$$\forall k \in \mathbb{N}, \qquad P(X = k) = u_k.$$

(2) (a) On pose $Y = X + 1$. Reconnaître la loi de Y et en déduire l'espérance et la variance de X.
(b) Compléter la fonction Python suivante pour qu'elle simule la loi d'une variable aléatoire X:

```
def simul_X(theta) :
    Y=1;
    while ..... :
        Y=Y+1;
    return .....
```

(3) Dans cette question, on souhaite estimer le paramètre θ par la méthode du *maximum de vraisemblance*. Pour ce faire, on considère un échantillon $(X_1, X_2, \ldots, X_n)$ composé de variables aléatoires indépendantes ayant toutes la même loi que X et on introduit la fonction L, de $\mathbb{R}_+^*$ dans $\mathbb{R}$, définie par

$$\forall \theta \in \mathbb{R}_+^*, \quad L(\theta) = \prod_{k=1}^{n} P(X_k = x_k),$$

où $x_1, x_2, \ldots, x_n$ désignent des entiers naturels éléments de $X(\Omega)$. L'objectif est de choisir la valeur de θ qui rend $L(\theta)$ maximale.

(a) Écrire $\ln(L(\theta))$ en fonction de θ et de $S_n = \sum_{k=1}^{n} x_k$.

(b) On considère la fonction φ, définie par

$$\forall \theta \in]0;+\infty[, \qquad \varphi(\theta) = S_n \ln\theta - (S_n + n)\ln(1+\theta).$$

Montrer que la fonction φ admet un maximum, atteint en un seul réel que l'on notera $\hat{\theta_n}$ et que l'on exprimera en fonction de S_n. Que représente $\hat{\theta_n}$ pour la fonction L?

On pose dorénavant

$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

La variable T_n est appelée *estimateur du maximum de vraisemblance* pour θ.

(c) Vérifier que T_n est un estimateur sans biais de θ.
(d) Calculer le risque quadratique $r_{T_n}(\theta)$ de T_n et vérifier que $\lim\limits_{n\to+\infty} r_{T_n}(\theta) = 0$.

Exercice 15. (D'après **EDHEC 2020**)

On considère une variable aléatoire X suivant la loi normale $\mathcal{N}\left(0, \sigma^2\right)$, où σ est strictement positif dont on note f_X une densité et F_X la fonction de répartition.

(1) Montrer que : $\forall x \in \mathbb{R}, F_X(-x) = 1 - F_X(x)$.

(2) On pose $Y = |X|$ et on admet que Y est une variable aléatoire.

(a) Montrer que la fonction de répartition de Y est la fonction, notée F_Y, définie par:

$$F_Y(x) = \begin{cases} 2F_X(x) - 1, \text{ si } x \geq 0 \\ 0, \text{ si } x < 0 \end{cases}$$

(b) En déduire que Y est une variable à densité et donner une densité f_Y de Y.
(c) Montrer que Y possède une espérance et que l'on a $E(Y) = \sigma\sqrt{\frac{2}{\pi}}$.

(3) On suppose, dans cette question seulement, que σ est inconnu et on se propose de l'estimer.
Soit n un entier naturel supérieur ou égal à 1. On considère un échantillon $(Y_1, Y_2, \ldots, Y_n)$ composé de variables aléatoires, mutuellement indépendantes, et ayant toutes la même loi que Y
On note S_n la variable aléatoire définie par $S_n = \frac{1}{n}\sum_{k=1}^{n} Y_k$.

(a) Montrer que S_n est un estimateur de σ, donner la valeur de son biais, puis proposer un estimateur sans biais de σ, que l'on notera T_n, construit de façon affine à partir de S_n.
(b) Rappeler la valeur du moment d'ordre 2 de X, puis déterminer $E\left(Y^2\right), V(Y)$ et $V\left(S_n\right)$.
(c) Déterminer le risque quadratique de T_n en tant qu'estimateur de σ. En déduire que T_n est un estimateur convergent de σ.

Exercice 16. (Maximum de vraisemblance pour variables continues, d'après **EDHEC 2012**).

On désigne par λ, un réel strictement positif et on considère la fonction f, définie sur $\mathbb{R}$, par

$$f(x) = \lambda |x| e^{-\lambda x^2}.$$

(1) (a) Montrer que f est paire.
(b) Établir la convergence et calculer la valeur de l'intégrale $\displaystyle\int_0^{+\infty} f(x)\,\mathrm{d}x$.
(c) Montrer que la fonction f peut être considérée comme densité d'une variable aléatoire X que l'on suppose, dans la suite, définie sur un certain espace probabilisé $(\Omega; \mathcal{A}; P)$.

(2) (a) Justifier la convergence de l'intégrale $\displaystyle\int_0^{+\infty} xf(x)\,\mathrm{d}x$.
(b) En déduire que la variable aléatoire X possède une espérance, notée $E(X)$, et donner sa valeur.

(3) (a) Montrer, grâce à une IPP, la convergence et donner la valeur de l'intégrale $\int_0^{+\infty} x^2 f(x)\,\mathrm{d}x$.

(b) En déduire que la variable aléatoire X possède une variance, notée $V(X)$, et donner sa valeur.

(4) On pose $Y = X^2$ et on admet que Y est une variable aléatoire à densité, elle aussi définie sur le même espace probabilisé.

(a) Donner l'expression de la fonction de répartition F_Y de la variable aléatoire Y à l'aide de la fonction de répartition F_X de la variable aléatoire X .

(b) Déterminer une densité f_Y de Y, puis vérifier que Y suit la loi exponentielle de paramètre λ .

(c) Retrouver alors sans calcul la valeur de $V(X)$.

(5) Soit U une variable aléatoire suivant la loi uniforme sur $[0;1[$. On pose $W = -\dfrac{1}{\lambda}\ln(1-U)$ et on admet que W est une variable aléatoire.

(a) Déterminer la fonction de répartition de W et en déduire la loi suivie par la variable aléatoire W.

(b) Vérifier que la probabilité que X prenne des valeurs positives est égale à la probabilité que X prenne des valeurs négatives.

On suppose, dans la suite, que le paramètre λ est inconnu et on souhaite l'estimer en utilisant la loi de Y.

On désigne par n un entier naturel supérieur ou égal à 2 et on considère un *échantillon* $Y_1, Y_2, .., Y_n$ de la loi de Y, c'est à dire des variables $Y_1, Y_2, .., Y_n$ définies sur $(\Omega; \mathcal{A}; P)$, indépendantes et de même loi que Y.

(6) On considère des réels $x_1, x_2, .., x_n$ strictement positifs, ainsi que la fonction L , à valeurs dans $\mathbb{R}$, définie sur $]0;+\infty[$ par

$$L(\lambda) = \prod_{k=1}^{n} f_Y(x_k), \qquad \lambda > 0.$$

(a) Exprimer $L(\lambda)$, puis $\ln(L(\lambda))$ en fonction de λ, $x_1, ..., x_n$.

(b) On considère la fonction φ, définie pour tout réel $\lambda > 0$ par

$$\varphi(\lambda) = n\ln(\lambda) - \lambda\sum_{=1}^{n} x_k.$$

Montrer que la fonction φ admet un maximum, atteint en un seul réel que l'on notera z et que l'on exprimera en fonction de $x_1, x_2, ..., x_n$. Que peut-on dire de z pour la fonction L?

(7) On pose dorénavant, toujours avec $n \geq 2$,

$$Z_n = \frac{n}{Y_1 + Y_2 + ... + Y_n}.$$

On admet que Z_n est une variable aléatoire définie, elle aussi, sur l'espace probabilisé $(\Omega; \mathcal{A}; P)$. La suite $(Z_n)_{n\geqslant 2}$ est appelée *estimateur du maximum de vraisemblance* pour λ.

On admet que la variable aléatoire $\overline{Y}_n = Y_1 + Y_2 + ... + Y_n$ admet pour densité la fonction f_n définie par

$$f_n(t) = \begin{cases} 0, & \text{si } t < 0 \\ \dfrac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t}, & \text{si } t \geqslant 0 \end{cases}.$$

(a) En remarquant que $\int_0^{+\infty} f_{n-1}(t)\,\mathrm{d}t = 1$, montrer que Z_n possède une espérance et que

$$E(Z_n) = \frac{n}{n-1}\lambda.$$

(b) Déterminer un estimateur non biaisé de λ, noté Z'_n, construit à partir de Z_n.

4.2 Estimation par intervalles de confiance

Exercice 17. (D'après **ESSEC II 2016**) Soit X d'espérance m et de variance σ^2. On considère alors un $n-$échantillon $(X_1, ..., X_n)$ de X et on pose $T_n = X_1 + X_2 + ... + X_n$. On suppose que σ^2 est connue, mais pas m.

(1) Déterminer $E(T_n)$ et $V(T_n)$.
(2) Montrer que, pour tout $\varepsilon > 0$

$$P\left(|T_n - nm| > n\varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

(3) Quel est le risque de l'intervalle de confiance

$$\left[\frac{T_n}{n} - \varepsilon;\ \frac{T_n}{n} + \varepsilon\right]$$

pour m?

Exercice 18. Soit $n \in \mathbb{N}^*$. On considère un échantillon $(X_1, \ldots, X_n)$ de la loi normale $\mathcal{N}(m, m/5)$, avec $m > 0$, le paramètre m étant inconnu.

Soit $\alpha \in]0;1[$ et t_α le réel positif tel que $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$. On suppose que $n > \frac{t_\alpha^2}{25}$.

On note

$$\overline{X_n} = \frac{X_1 + \ldots + X_n}{n}, \qquad \text{et} \qquad Y_n = 5\sqrt{n}\frac{\overline{X_n} - m}{m}.$$

(1) Justifier que la variable aléatoire Y_n converge en loi vers une loi normale centrée réduite.
(2) Justifier alors que pour n assez grand, on peut écrire

$$P\left(|Y_n| \leq t_\alpha\right) = 1 - \alpha.$$

(3) Montrer que l'intervalle

$$\left[5\sqrt{n}\frac{\overline{X_n}}{5\sqrt{n} + t_\alpha}; 5\sqrt{n}\frac{\overline{X_n}}{5\sqrt{n} - t_\alpha}\right]$$

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

(4) Pour $n = 100$, une réalisation de ce n-échantillon nous donne une moyenne empirique de 12.
Déterminer une estimation d'un intervalle de confiance de m à 95%. *On donne* $\Phi(1,96) = 0,975$.

Exercice 19. Soient $n \in \mathbb{N}^*$ et $a > 0$. On considère une suite de v.a. (X_n) dont la fonction de répartition est donnée par

$$F_{X_n}(t) = \begin{cases} 0, & \text{si } x < 0 \\ 1 - \exp\left(-ax - \frac{x^2}{2n}\right), & \text{si } 0 \leq x < 2n \\ 1, & \text{si } x \leq 2n \end{cases}$$

(1) Montrer que (X_n) converge en loi vers une v.a dont on précisera la loi.

(2) Soient $Z \hookrightarrow \mathcal{E}(1)$ et $\alpha \in]0;1[$.

(a) Déterminer deux réels c et d strictement positifs tels que

$$P(c \leq Z \leq d) = 1 - \alpha, \quad \text{et} \quad P(Z \leq c) = \frac{\alpha}{2}.$$

(b) Quelle est la loi de aZ?
(c) Montrer que

$$\lim_{n \to +\infty} P\left(a \in \left[\frac{c}{X_n}; \frac{d}{X_n}\right]\right) = 1 - \alpha.$$

(d) Que peut-on dire de l'intervalle

$$\left[\frac{c}{X_n}; \frac{d}{X_n}\right]?$$

Exercice 20. Soit $p \in]0\,;\,1[$, on considère une variable aléatoire X qui suit la loi géométrique de paramètre p. On pose $q = 1-p$. Pour tout $n \in \mathbb{N}^*$, on considère un $n-$échantillon $(X_1, \dots, X_n)$ de la loi de X. On pose

$$S_n = \sum_{k=1}^{n} X_k, \quad Y_n = \frac{n}{S_n} \quad \text{et} \quad \overline{X}_n = \frac{1}{Y_n} = \frac{1}{n} S_n.$$

(1) Montrer que $\overline{X}_n$ est un estimateur sans biais de $\frac{1}{p}$. Quel est son risque quadratique ?

(2) Pour tout $n \in \mathbb{N}^*$, on pose $T_n = \sqrt{\frac{n}{p^2 q}}(Y_n - p)$. On **admet** que T_n converge en loi vers une variable aléatoire $T \hookrightarrow \mathcal{N}(0,1)$.

(a) Soit $\alpha \in]0\,;\,1[$ et a_α l'unique réel vérifiant $P([T > a_\alpha]) = \frac{\alpha}{2}$.
Montrer que $P(-a_\alpha \leq T \leq a_\alpha) = 1 - \alpha$.

(b) Pour n assez grand on considère alors que $P(-a_\alpha \leq T_n \leq a_\alpha) = 1 - \alpha$. En déduire que :

$$P\left(Y_n - a_\alpha p\sqrt{\frac{q}{n}} \leq p \leq Y_n + a_\alpha p\sqrt{\frac{q}{n}}\right) = 1 - \alpha.$$

(c) Étudier la fonction $f : x \mapsto x\sqrt{1-x}$ sur l'intervalle $[0;1]$.

(d) Déduire des deux questions précédentes que pour n assez grand :

$$P\left(Y_n - \frac{2a_\alpha}{3\sqrt{3n}} \leq p \leq Y_n + \frac{2a_\alpha}{3\sqrt{3n}}\right) \geq 1 - \alpha.$$

(e) On suppose que $n = 900$, une réalisation de l'échantillon $(X_1, \dots, X_{900})$ a donné la valeur 4 à $\overline{X}_{900}$. Donner alors la réalisation y_{900} de la variable Y_{900}.
On se donne un niveau de risque $\alpha = 0,05$, le nombre $a_{0,05}$ vaut à peu près 2.
Trouver une fourchette pour p avec un niveau de confiance d'au moins $0,95$. *On donne* $\frac{2}{45\sqrt{3}} \simeq 0,026$.

Exercice 21. (D'après **Oral ESCP**, voie S)

Soit $n \in \mathbb{N}^*$. On considère un $n-$échantillon $(X_1, X_2, ..., X_n)$ de la loi exponentielle $\mathcal{E}(\lambda)$ de paramètre $\lambda > 0$. On pose $L_n = \min(X_1, ..., X_n)$ et $M_n = \max(X_1, ..., X_n)$.

(1) Déterminer les fonctions de répartition des variables L_n puis M_n.

(2) Quelle est la loi de la variable $Y_n = n\lambda L_n$?

(3) On maintenant alors $Z_n = \lambda M_n - \ln(n)$.

(a) Déterminer la fonction de répartition F_n de Z_n.

(b) Pour $t \in \mathbb{R}$, déterminer la limite $F(t)$ de $F_n(t)$ lorsque $n \to +\infty$.

(c) En déduire soigneusement que Z_n converge en loi vers une variable aléatoire à densité, que l'on notera Z. (*On dit que Z suit la loi de Gumbel.*)

La durée de vie d'une ampoule est modélisée par une variable aléatoire $X \hookrightarrow \mathcal{E}(\lambda)$ où $\lambda > 0$ est inconnu. On cherche à estimer la durée de vide moyenne $\mu = E(X) = 1/\lambda$ et on dispose d'un échantillon de n ampoules (dont les durées de vie sont supposées indépendantes).

(4) Dans cette question, on suppose que la seule information dont on dispose est la durée de vie de l'ampoule qui a *grillé* le plus tôt.

(a) À l'aide de la Question (2), proposer un estimateur $\tilde{L}_n$ de μ, construit à partir de L_n, qui soit sans biais.

(b) Quel est son risque quadratique?

(c) Montrer que, pour tout $\varepsilon > 0$, $P\left(\left|\tilde{L}_n - \mu\right| > \varepsilon\right) = 1 - e^{-1+\lambda\varepsilon} + e^{-1-\lambda\varepsilon}$.

(d) L'estimateur $\tilde{L}_n$ est-il convergent?

(e) Soit $\alpha \in]0;1[$. Montrer que si $Y \hookrightarrow \mathcal{E}(1)$, alors

$$P\left(Y < -\ln\left(1-\frac{\alpha}{2}\right)\right) = P\left(Y > -\ln\left(\frac{\alpha}{2}\right)\right) = \frac{\alpha}{2}.$$

(f) Montrer alors que l'intervalle

$$I_{\alpha,n} = \left[\frac{\tilde{L}_n}{-\ln(\alpha/2)},\ \frac{\tilde{L}_n}{-\ln(1-\alpha/2)}\right]$$

est un intervalle de confiance au seuil $1-\alpha$ pour μ.

Exercice 22. (Question confidentielle, extrait **CB n°4**, Printemps 2020)

Certains sujets abordés dans les enquêtes d'opinion sont parfois assez intimes, et on court le risque que les personnes interrogées se refusent à répondre franchement à l'enquêteur, faussant ainsi le résultat.

On peut alors avoir recours à une astuce consistant à inverser aléatoirement les réponses.

Considérons une *question confidentielle* pour laquelle on veut estimer la probabilité p de réponses positives. L'enquêteur procède alors à une expérience aléatoire (dont il ne connait pas le résultat) avant de poser la question à chaque individu.

Si **l'enquêteur ignore le résultat de l'expérience**, il ne pourra pas savoir si la réponse est franche ou non, et on peut espérer que la personne sondée acceptera de jouer le jeu.

Plus précisément, soit $n \in \mathbb{N}^*$ le nombre de personnes interrogées. Pour tout $n \in \mathbb{N}^*$, on introduit une variable de Bernoulli X_n de paramètre $\alpha \in]0;1[$ et une variable Y_n de Bernoulli définie comme suit:

- Pour tout $n \in \mathbb{N}^*$, on pose une question (dont la réponse ne peut être que "oui" ou "non") à la $n-$ième personne;
- Si $X_n = 1$, la $n-$ième personne doit être franche, sinon la réponse est inversée.
- Si la $n-$ième personne répond "oui", alors $Y_n = 1$ et 0 sinon.

On a donc $P_{[X_n=1]}(Y_n = 1) = p$.

Le but de l'exercice est de donner une estimation de p.

(1) Montrer que : $P(Y_n = 1) = \alpha p + (1-\alpha)(1-p)$.
(2) Sachant qu'une personne a répondu "oui", quelle est la probabilité qu'elle ait été franche?

On introduit alors la variable aléatoire T_n définie par

$$F_n = \frac{Y_1 + Y_2 + ... + Y_n}{n}.$$

(3) Montrer que F_n est un estimateur sans biais et convergent de $P(Y_n = 1)$.
(4) Pour $\alpha \neq 1/2$, exprimer p en fonction de $P(Y_n = 1)$.
(5) En déduire que

$$T_n = \frac{F_n - 1 + \alpha}{2\alpha - 1}$$

est un estimateur sans biais et convergent de p.

(6) (a) Déterminer le risque quadratique de T_n.
(b) Pour $n \in \mathbb{N}^*$ fixé, quelle valeur attribuer à α pour que le risque quadratique de T_n soit minimal? Est-ce acceptable?

(7) Soit $\beta \in]0;1[$. Montrer que

$$I_n = \left[T_n - \frac{1}{2(2\alpha-1)\sqrt{n\beta}}; T_n + \frac{1}{2(2\alpha-1)\sqrt{n\beta}}\right]$$

est un intervalle de confiance pour p au risque β.

5 Appendice: Table de la loi normale centrée réduite $\mathcal{N}(0,1)$

$$\Phi(t)(x) = P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \mathrm{d}t \quad \text{et} \quad \Phi(-x) = 1 - \Phi(x).$$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000