

Chapitre 14

Estimations

Sommaire

14.1 Estimation ponctuelle	336
14.1.1 Echantillonnage	336
14.1.2 Estimateur	337
14.1.3 Biais	339
14.1.4 Suite d'estimateurs	339
14.1.5 Risque quadratique	340
14.1.6 Estimateur convergent	340
14.2 Estimation par intervalle de confiance	341
14.2.1 Intervalle de confiance au niveau de confiance	341
14.2.2 Utilisation de l'inégalité de Bienaymé-Tchebychev	342
14.2.3 Intervalle de confiance asymptotique	343
14.2.4 Utilisation de la loi normale	344
14.2.5 Estimation de l'espérance d'une loi normale d'écart-type connu	344
14.2.6 Estimation du paramètre d'une variable de Bernoulli	345

Les statisticiens connaissent en général, le type de loi qui décrit tel ou tel phénomène (de par l'observation), mais souvent ils ne connaissent pas tous les paramètres de la dite loi. Ils doivent donc les estimer : c'est l'objectif de la statistique inférentielle.

L'objectif de ce dernier chapitre est d'introduire le vocabulaire et la démarche de la statistique inférentielle en abordant, sur quelques cas simples, le problème de l'estimation, ponctuelle ou par intervalle de confiance.

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.

Marquis Pierre Simon de Laplace (23 mars 1749 [Beaumont-en-Auge] - 5 mars 1827 [Paris])

Né à Beaumont-en-Auge, fils de cultivateur, Laplace s'initia aux mathématiques à l'École militaire de cette petite ville. Il y commença son enseignement. Il doit cette éducation à ses voisins aisés qui avait détecté son intelligence exceptionnelle.

18 ans, il arrive à Paris avec une lettre de recommandation pour rencontrer le mathématicien d'Alembert, mais ce dernier refuse de rencontrer l'inconnu. Mais Laplace insiste : il envoie à d'Alembert un article qu'il a écrit sur la mécanique classique. D'Alembert en est si impressionné qu'il est tout

heureux de patronner Laplace. Il lui obtient un poste d'enseignement en mathématique. En 1783, il devint examinateur du corps de l'artillerie et fut élu, en 1785, à l'Académie des Sciences. A la Révolution, il participa à l'organisation de l'École Normale et de l'École Polytechnique, et fut membre de l'Institut, dès sa création. Bonaparte lui confia le ministère de l'Intérieur, mais seulement pour 6 mois.

L'oeuvre la plus importante de Laplace concerne le calcul des probabilités et la mécanique céleste. Il établit aussi, grâce à ses travaux avec Lavoisier entre 1782 et 1784 la formule des transformations adiabatiques d'un gaz, ainsi que deux lois fondamentales de l'électromagnétisme. En mécanique, c'est avec le mathématicien Joseph-Louis de Lagrange, Laplace résume ses travaux et réunit ceux de Newton, Halley, Clairaut, d'Alembert et Euler, concernant la gravitation universelle, dans les cinq volumes de sa mécanique céleste (1798-1825).

On rapporte que, feuilletant la Mécanique céleste, Napoléon fit remarquer à Laplace qu'il n'y était nulle part fait mention de Dieu. "Je n'ai pas eu besoin de cette hypothèse", rétorqua le savant.

On considèrera Θ : un sous-ensemble de $\mathbb{R}$ (éventuellement de $\mathbb{R}^2$).

14.1 Estimation ponctuelle

14.1.1 Echantillonnage

Définition 14.1 (n -échantillon).

On appelle n -échantillon de la loi X tout n -uplet $(X_1, \dots, X_n)$ de var définies sur le même espace probabilisé $(\Omega, \mathcal{A}, P)$, **mutuellement indépendantes** et suivant toutes la même loi que X .

On dit que X est la loi parente de l'échantillon.

Exemples 14.1.

-Si on cherche à calculer la taille moyenne d'un homme de 40 ans en France, il est impossible de déterminer exactement la taille de tous les hommes français et d'en faire une moyenne. Pour donner une valeur approchée de cette moyenne on prend un échantillon d'hommes, par exemple 1000 hommes, on détermine leur taille puis on fait la moyenne. Avec un échantillon assez grand on considère que l'on a obtenu une valeur approchée, 'une estimation', de la taille moyenne d'un homme de 40 ans.

-On dispose d'une urne contenant des boules rouges et des boules blanches mais on ne connaît pas la composition de l'urne. En 100 tirages avec remise on obtient 30 boules rouges et 70 boules blanches. A combien peut-on estimer la proportion de boules rouges dans l'urne?

Prenons ici X la variable qui vaut 1 si on tire une boule rouge et 0 sinon. Alors on sait que X suit une loi de Bernoulli de paramètre p où p est la probabilité de tirer une boule rouge, c'est-à-dire que p est la proportion de boules rouges.

Dans cette expérience on cherche à connaître la valeur de p . On a donc ici $\theta = p$ et $\Theta =]0;1[$. L'énoncé de l'exemple 2 nous dit que lors de 100 réalisations de la variable X , X a pris 30 fois la valeur 1 et 70 fois la valeur 0.

Définition 14.2.

Si $(X_1, \dots, X_n)$ est un échantillon de la loi de X , on appelle réalisation de cet échantillon (ou aussi échantillon observé), tout n -uplet $(x_1, \dots, x_n)$, où, pour tout k de $[1, n]$, x_k est la valeur prise par la var X_k .

Remarque.

Dans la pratique, pour réaliser un échantillon, on effectue n épreuves, identiques et indépendantes, pour lesquelles, après la k ème épreuve, la var X_k (associée à cette k ème épreuve) a pris la valeur x_k .

Exemple 14.2.

Le résultat d'un lancer de dé, supposé équilibré, est régi par la loi uniforme sur $[[1, 6]]$; en lançant le dé n fois, on réalise un n -échantillon $(X_1, \dots, X_n)$.

14.1.2 Estimateur

On se restreindra à une famille de lois de probabilités indexées par un paramètre scalaire (ou vectoriel) dont la valeur (scalaire ou vectorielle) caractérise la loi. On cherche alors à estimer la valeur du paramètre (ou une fonction simple de ce paramètre) à partir des données disponibles.

Dans ce contexte, on considère un phénomène aléatoire et on s'intéresse à une variable aléatoire réelle X qui lui est liée, dont on suppose que la loi de probabilité n'est pas complètement spécifiée et appartient à une famille de lois dépendant d'un paramètre décrivant un sous-ensemble Θ de $\mathbb{R}$ (éventuellement de $\mathbb{R}^2$).

Le paramètre θ est une quantité inconnue, fixée dans toute l'étude, que l'on cherche à déterminer ou pour laquelle on cherche une information partielle.

Le problème de l'estimation consiste alors à estimer la vraie valeur du paramètre θ ou de $g(\theta)$ (fonction à valeurs réelles du paramètre θ), à partir d'un échantillon de données $x_1, \dots, x_n$ obtenues en observant n fois le phénomène.

Cette fonction du paramètre représentera en général une valeur caractéristique de la loi inconnue comme son espérance, sa variance, son étendue... On supposera que cet échantillon est la réalisation de n variables aléatoires $X_1, \dots, X_n$ définies sur un même espace probabilisable $(\Omega, \mathcal{A})$ muni d'une famille de probabilités $(P_\theta)_{\theta \in \Theta}$. Les $X_1, \dots, X_n$ seront supposées P_θ -indépendantes et de même loi que X pour tout θ .

Nous allons nous intéresser au cas où la loi de probabilité a des paramètres inconnus.

La situation de référence est celle où l'on se trouve en présence d'un phénomène dont les manifestations sont des mesures que l'on est incapable de prévoir.

Le travail du statisticien est rechercher la loi théorique qui " colle " le mieux à l'échantillon.

Exemples 14.3.

-A l'occasion d'un référendum, une certaine proportion p de personnes vote oui et une proportion $1 - p$ de personnes vote non.

On peut assimiler chaque vote à la réalisation d'une var qui soit une loi de Bernoulli de paramètre p .

-Si on reprend l'exemple des boules (exemple 2) :

Un échantillon de donnée est une liste de 0 et de 1 où on a 30 fois le nombre 1 et 70 fois le nombre 0. Donc après nos 100 tirages, la moyenne empirique de X vaut : 30%. Une estimation de p est donc 0,3.

S'il est vrai que dans certains cas, on ne connaît rien, à priori, de la loi parente à l'échantillon observé, des observations permettent souvent de préciser qu'elle appartient à un certain ensemble de lois dépendant d'un paramètre réel θ .

Définition 14.3 (Estimateur).

Si $(X_1, \dots, X_n)$ est un échantillon de la loi de X , on appelle estimateur de θ , toute var fonction de $X_1, \dots, X_n$ et indépendante de θ .

Un estimateur se définit donc dans l'intention de fournir une estimation.

Remarque (Estimation).

Si $\phi_n(X_1, \dots, X_n)$ est un estimateur de θ , alors avec les notations de la définition 14.3, $\phi_n(x_1, \dots, x_n)$ est une estimation de θ , ie la valeur que le statisticien accordera à θ .

Exemple 14.4.

On considère un dé dont on ne sait s'il est pipé ou non. On le lance n fois et on désigne par X_k la var de Bernoulli, égale à 1, si le résultat obtenu est 6. Soit p son paramètre inconnu que l'on cherche à estimer.

Soit $(X_1, \dots, X_n)$ est un n -échantillon de $\mathcal{B}(p)$. On a ici $\Theta = [0, 1], \theta = p$.

On considère les var : $T_n = \frac{1}{n} \sum_{i=1}^n X_i$ et $U_n = \frac{2}{n(n+1)} \sum_{i=1}^n i X_i$.

Elles prennent toutes deux leurs valeurs dans $[0, 1]$. Ce sont deux estimateurs de p .

Exemple 14.5 (Le sondage). Dans une population, un certain caractère est présent dans une proportion p , que l'on veut déterminer.

On pose l'expérience aléatoire suivante : on choisit une personne au hasard, et on vérifie si elle a le paramètre ou non.

On pose alors $X = \begin{cases} 0 & \text{si elle n'a pas le caractère} \\ 1 & \text{si elle l'a.} \end{cases}$ Alors $X \hookrightarrow \mathcal{B}(p)$; le sondage est une estimation du paramètre p d'une loi de Bernoulli.

Remarque.

-Pour construire un estimateur de p , il faut construire une var fonction de $(X_1, \dots, X_n)$, qui ne s'éloigne pas trop de p .

-Si T_n est un estimateur, on notera, lorsque ces valeurs existent, $E_\theta(T_n)$ l'espérance de T_n et $V_\theta(T_n)$ la variance de T_n , pour la probabilité P_θ .

Exemple important d'estimateur :

Définition 14.4 (Moyenne empirique).

Si $(X_1, \dots, X_n)$ est un échantillon de la loi de X , on appelle moyenne empirique associée à $(X_1, \dots, X_n)$, la var $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

Propriété 14.1 (Espérance et variance).

On considère une var X d'espérance m et de variance σ^2 et un n -échantillon $(X_1, \dots, X_n)$ de la loi de X . En notant $\bar{X}_n$ la moyenne empirique associé à cet échantillon, on a : $E(\bar{X}_n) = m, V(\bar{X}_n) = \frac{\sigma^2}{n}$.

Preuve.



14.1.3 Biais

[HP]

Définition 14.5 (Biais).

Si T_n est un estimateur de θ , si T_n a une espérance pour tout $\theta \in \Theta$, on appelle biais de T_n le réel :

$$b_\theta(T_n) = E(T_n) - \theta.$$

Définition 14.6 (Estimateur sans biais).

On dit que T_n est un estimateur sans biais de θ , si :

$$E(T_n) = \theta.$$

(ce qui est équivalent à dire si $b_\theta(T_n) = 0$.)

Remarques.

- Le biais mesure l'écart moyen entre les valeurs prises par l'estimateur et le réel que l'on cherche à estimer.
- Lorsqu'on dit que l'estimateur est sans biais cela signifie que en moyenne les valeurs de l'estimateur sont très proches de θ .
- Attention rien n'empêche un estimateur sans biais de prendre des valeurs très éloignés de θ car en moyenne les écarts peuvent se compenser.

Exemple 14.6.

Soit un n -échantillon $(X_1, \dots, X_n)$ d'une loi d'espérance m .

Comme $E(\bar{X}_n) = m$, la moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur sans biais de m .

14.1.4 Suite d'estimateurs

[HP] Un estimateur ne dépend pas seulement des valeurs possibles du paramètre à estimer mais aussi de la taille de l'échantillon. C'est pourquoi on ne considère pas un estimateur seul mais, le plus souvent une suite d'estimateurs.

Une suite $(T_n)_{n \in \mathbb{N}^*}$ d'estimateurs de θ : chaque T_n est de la forme $\phi(X_1, \dots, X_n)$.

Définition 14.7 (Estimateur asymptotiquement sans biais).

Une suite $(T_n)_{n \in \mathbb{N}^*}$ d'estimateurs de θ est asymptotiquement sans biais si pour tout $\theta \in \Theta$, $\lim_{n \rightarrow +\infty} E_\theta(T_n) = \theta$.

Remarque.

Par abus de langage on dit aussi que l'estimateur est asymptotiquement sans biais.

On a besoin d'une mesure supplémentaire pour différencier deux estimateurs sans biais et dire lequel est le " meilleur ".

14.1.5 Risque quadratique

[HP]

Définition 14.8 (Risque quadratique).

Si pour tout $\theta \in \Theta$, T_n admet un moment d'ordre 2, on appelle risque quadratique (ou erreur quadratique moyenne) de T_n le réel :

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

Remarque.

Le risque quadratique mesure la moyenne de l'écart de T_n à θ au carré. Comme un carré est toujours positif, les écarts à θ en plus ou en moins ne peuvent plus se compenser mais se cumulent. On a donc bien ici une façon de mesurer si T_n est un ' bon ' estimateur de θ .

Exemple 14.7.

Soit un n -échantillon $(X_1, \dots, X_n)$ d'une loi de Bernoulli de paramètre p .

Alors le risque quadratique de $\bar{X}_n$ est aussi sa variance : $r_p(\bar{X}_n) = \frac{p(1-p)}{n}$.

Théorème 14.2 (Décomposition biais-variance du risque quadratique).

Si un estimateur T_n a un moment d'ordre 2, alors :

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

EXERCICE 14.1.

On considère une var X d'espérance m et de variance σ^2 et un n -échantillon $(X_1, \dots, X_n)$ de la loi de X . En notant $\bar{X}_n$ la moyenne empirique associé à cet échantillon, calculer le risque quadratique de $\bar{X}_n$.

Remarque.

Si l'estimateur est sans biais et s'il possède une variance, on a bien sûr :

$$r_\theta(T_n) = V(T_n).$$

14.1.6 Estimateur convergent

Définition 14.9 (Estimateur convergent).

Une suite $(T_n)_{n \in \mathbb{N}^*}$ d'estimateurs de $g(\theta)$ est convergent (ou consistant) si :
pour tout $\theta \in \Theta$, $\forall \epsilon > 0$, $\lim_{n \rightarrow +\infty} P_\theta(|T_n - \theta| > \epsilon) = 0$.

Remarque.

Par abus de langage on dit aussi que l'estimateur est convergent.

On établit la convergence de l'estimateur T_n grâce à l'inégalité de Bienaymé-Tchebychev appliquée à T_n lorsque $E(T_n) = \theta$, ou grâce à l'inégalité de Markov appliquée à $(T_n - \theta)^2$ si $E(T_n) \neq \theta$.

Théorème 14.3 (Condition suffisante de convergence d'un estimateur). [HP]

Si pour tout $\theta \in \Theta$, l'estimateur T_n possède un moment d'ordre 2, alors la suite $(T_n)_{n \in \mathbb{N}^*}$ est un estimateur convergent de θ si

pour tout $\theta \in \Theta$, $\lim_{n \rightarrow +\infty} r_\theta(T_n) = 0$.

Preuve.

◇

Remarques (HP).

1. Si T_n est sans biais, alors la condition du théorème 14.3 s'écrit : $\lim_{n \rightarrow +\infty} V(T_n) = 0$.
2. Si T_n est asymptotiquement sans biais, alors grâce à la décomposition biais-variance du risque quadratique, la condition du théorème 14.3 s'écrit aussi : $\lim_{n \rightarrow +\infty} V(T_n) = 0$.
3. De deux estimateurs de θ , on choisit celui qui a le plus petit risque quadratique (ou mieux, celui qui a le risque quadratique qui tend le plus vite vers 0) car ainsi, avec l'inégalité de Markov ou celle de Bienaymé-Tchebychev, on minimise plus vite la probabilité que l'estimateur s'écarte de θ de plus de ϵ .

Remarque.

De deux estimateurs de θ , ayant tous deux une espérance égale à θ , on choisit celui qui a la plus petite variance ou mieux, qui a la variance qui tend le plus vite vers 0.

EXERCICE 14.2.

Montrer que la moyenne empirique est un estimateur convergent de l'espérance m de X .

EXERCICE 14.3.

Une var X suit une loi uniforme sur $[0, \theta]$, θ étant un réel strictement positif inconnu. On note F la fonction de répartition de X et f une densité de X .

On dispose d'un échantillon $(X_1, \dots, X_n)$ de la loi de X et on pose $Z_n = \max(X_1, \dots, X_n)$. Déterminer la fonction de répartition F_n ainsi qu'une densité de Z_n .

Montrer que Z_n est un estimateur asymptotiquement sans biais de θ , puis que c'est un estimateur convergent de θ .

14.2 Estimation par intervalle de confiance

14.2.1 Intervalle de confiance au niveau de confiance

S'il existe des critères pour juger des qualités d'un estimateur ponctuel T_n de $g(\theta)$ (biais, risque, convergence), aucune certitude ne peut jamais être apportée quant au fait que l'estimation donne la vraie valeur à estimer.

La démarche de l'estimation par intervalle de confiance consiste à trouver un intervalle aléatoire qui contienne θ avec une probabilité minimale donnée. L'utilisation dans certains cas du théorème limite central impose d'introduire la notion d'intervalle de confiance asymptotique.

Ce paragraphe a uniquement pour but de préciser le vocabulaire employé. Les situations seront

étudiées sous forme d'exercices, aucune connaissance autre que ce vocabulaire n'est exigible sur les intervalles de confiance.

Dans tout ce paragraphe $(U_n)_{n \geq 1}$ et $(V_n)_{n \geq 1}$ désigneront des suites d'estimateurs de θ tels que pour tout $n \geq 1$ et pour tout $\theta \in \Theta$, $P_\Theta([U_n \leq V_n]) = 1$.

Définition 14.10 (🔗 **Intervalle de confiance au niveau de confiance $1 - \alpha$ (ou risque α)).**

On dit que $[U_n, V_n]$ est un intervalle de confiance de θ au niveau de confiance $1 - \alpha$ ($\alpha \in [0, 1]$), si pour tout $\theta \in \Theta$, $P_\Theta([U_n \leq \theta \leq V_n]) \geq 1 - \alpha$.

Remarques.

- Si $U_n = \phi_n(X_1, \dots, X_n)$ et $V_n = \psi(X_1, \dots, X_n)$ alors, en posant $u_n = \phi_n(x_1, \dots, x_n)$ et $v_n = \psi(x_1, \dots, x_n)$, l'intervalle $[u_n, v_n]$ est l'intervalle de confiance observé nommé souvent fourchette.

- Les variables U_n et V_n sont en fait obtenues à partir d'un estimateur T_n , en posant $U_n = T_n - a$ et $V_n = T_n + a$. Ainsi il faudra, à l'aide d'une approximation ou d'un calcul exact, déterminer la valeur de a à poser pour obtenir le niveau de confiance ou de risque souhaité.

14.2.2 🔗 Utilisation de l'inégalité de Bienaymé-Tchebychev

D'après l'inégalité de Bienaymé-Tchebychev (B-T), si T_n est une var, d'espérance $E(T_n)$ et d'écart-type σ :

$$P(|T_n - E(T_n)| \leq \epsilon) \geq 1 - \frac{\sigma^2}{\epsilon^2}.$$

Si l'un au moins des paramètres est inconnu et si $g(\theta)$ est un estimateur sans biais de T_n , on utilisera une ou des majorations pour déterminer un intervalle de confiance de θ .

Exemple 14.8. [Estimation du paramètre d'une variable de Bernoulli avec B-T]

On veut déterminer un intervalle de confiance d'au moins 0,95 de la probabilité p d'obtenir 6 lors d'un lancer d'un dé.

D'après la loi faible des grands nombres, cet intervalle sera d'autant plus petit (au sens de l'inclusion) que le nombre de simulations effectuées sera grand.

Désignons pour un entier naturel non nul k , par T_k la var de Bernoulli égale à 1, si l'on obtient 6 au k -ième lancer.

La suite $(T_n)_{n \in \mathbb{N}^*}$ est une suite de var de Bernoulli indépendantes, de paramètre p .

On a :

$$E(T_n) = p \text{ et } V(T_n) = p(1 - p).$$

Soit $(\bar{T}_n)_{n \in \mathbb{N}^*}$ la suite des moyennes empiriques :

$$\bar{T}_n = \frac{1}{n} \sum_{k=1}^n T_k, E(\bar{T}_n) = p, V(\bar{T}_n) = \frac{p(1 - p)}{n}.$$

D'après l'inégalité de B-T :

$$P(|\bar{T}_n - p| \leq \epsilon) \geq 1 - \frac{p(1 - p)}{n\epsilon^2}.$$

Pour donner un encadrement de p , on va majorer $p(1 - p)$ par $\frac{1}{4}$ (en effet vérifier que la fonction $p \mapsto p(1 - p)$ sur $]0, 1[$, a un $\max = 1/4$ en $p = 1/2$)

On obtient alors :

$$P(|\bar{T}_n - p| \leq \epsilon) \geq 1 - \alpha \text{ et } 1 - \alpha = 1 - \frac{1}{4n\epsilon^2} \Leftrightarrow \epsilon = \frac{1}{2\sqrt{n\alpha}}.$$

On a donc l'intervalle de confiance de p à un niveau de confiance $1 - \alpha$:

$$P\left(p \in \left[\bar{T}_n - \frac{1}{2\sqrt{n\alpha}}, \bar{T}_n + \frac{1}{2\sqrt{n\alpha}}\right]\right) \geq 1 - \alpha.$$

Pour $1 - \alpha = 0,95$, on a donc :

$$P\left(p \in \left[\bar{T}_n - \frac{1}{2\sqrt{0,05n}}, \bar{T}_n + \frac{1}{2\sqrt{0,05n}}\right]\right) \geq 0,95.$$

Pour $n = 30$, on a comme intervalle de confiance :

$$[\bar{T}_n - 0,41, \bar{T}_n + 0,41].$$

Bref, compte tenu de la longueur de l'intervalle de confiance, il est presque impossible que p ne soit pas dedans.

Remarque (MÉTHODE).

Pour trouver deux variables U_n et V_n telles que $P_\Theta([U_n \leq \theta \leq V_n]) \geq 1 - \alpha$, en utilisant l'inégalité de Bienaymé-Tchebychev on peut procéder ainsi :

1. On montre que l'estimateur T_n dont on dispose est un estimateur sans biais de θ .
2. On écrit l'inégalité de Bienaymé-Tchebychev appliquée à T_n , pour une valeur quelconque de ϵ . On obtient une majoration de la forme $P_\theta(|T_n - \theta| \geq \epsilon) \leq \dots$
3. On transforme la probabilité écrite (événement contraire) pour obtenir qqch du style $P_\theta(|T_n - \theta| \leq \epsilon) \geq \dots$
4. On choisit ϵ (dépendant toujours de $V(T_n)$) de sorte que le membre de droite de l'inégalité soit égal à $1 - \alpha$
5. on transforme à l'intérieur de la probabilité pour présenter le résultat sous la forme $P_\theta([U_n \leq \theta \leq V_n]) \geq 1 - \alpha$.

EXERCICE 14.4 (Intervalle de confiance grâce à l'inégalité de Bienaymé-Tchebychev).

Soit X une var de Bernoulli de paramètre p , $0 < p < 1$ et soit un n -échantillon $(X_1, \dots, X_n)$ de la loi de X . En notant $\bar{X}_n$ la moyenne empirique associé à cet échantillon et $\alpha \in]0, 1]$, calculer l'espérance et la variance de $\bar{X}_n$.

Majorer $p(1-p)$ pour $0 < p < 1$ et en déduire un intervalle de confiance de p au niveau de confiance de $1 - \alpha$.

14.2.3 Intervalle de confiance asymptotique

Définition 14.11 (Intervalle de confiance asymptotique).

On dit que la suite $([U_n, V_n])_{n \in \mathbb{N}^*}$ est un intervalle de confiance asymptotique de θ au niveau de confiance $1 - \alpha$ ($\alpha \in [0, 1]$), si pour tout $\theta \in \Theta$, il existe une suite (α_n) à valeurs dans $[0, 1]$, de limite α telle que pour tout $n \in \mathbb{N}^*$, $P_\theta([U_n \leq \theta \leq V_n]) \geq 1 - \alpha_n$.

Remarques.

-Par abus de langage on dit aussi que $[U_n, V_n]$ est un intervalle de confiance asymptotique.

-Il est équivalent de dire : la suite $([U_n, V_n])_{n \in \mathbb{N}^*}$ est un intervalle de confiance asymptotique de θ au niveau de confiance $1 - \alpha$ ($\alpha \in [0, 1]$), si l'on a : pour tout $\theta \in \Theta$, $\lim_{n \rightarrow +\infty} P_\theta([U_n \leq \theta \leq V_n]) \geq 1 - \alpha$.

-Dans la pratique la suite α_n devra obligatoirement être fournie par l'énoncé.

Nous allons maintenant donner deux exemples d'intervalles de confiance utilisant le théorème central limite.

14.2.4 Utilisation de la loi normale

14.2.5 Estimation de l'espérance d'une loi normale d'écart-type connu

Exemple 14.9. Soit $(X_1, \dots, X_n)$ un n -échantillon de loi parente, une loi normale d'espérance m à estimer et d'écart-type σ connu.

On considère $\bar{X}_n$ la moyenne empirique de l'échantillon.

C'est un estimateur sans biais de m . De plus, $\bar{X}_n$ suit une loi normale, comme combinaison linéaire de lois normales indépendantes.

Remarque.

En effet, on a le résultat suivant :

Si $X \hookrightarrow \mathcal{N}(m, \sigma^2)$ alors $aX + b \hookrightarrow \mathcal{N}(am + b, (a\sigma)^2)$.

D'après la proposition sur la transformation affine d'une variable aléatoire à densité, $Y = aX + b$ est une variable aléatoire à densité avec pour densité $f_Y(x) = \frac{1}{|a|} f_X\left(\frac{x-b}{a}\right) = \frac{1}{(|a|\sigma)\sqrt{2\pi}} e^{-\frac{(x-(am+b))^2}{2(a\sigma)^2}}$.

Donc, $\bar{X}_n \hookrightarrow \mathcal{N}(m, \frac{\sigma^2}{n})$ et donc d'après le théorème central limite $Y_n = \sqrt{n} \frac{\bar{X}_n - m}{\sigma}$ converge en loi vers la loi normale centrée réduite. Donc pour n assez grand ($n \geq 30$), on peut donc approximer la var Y_n par la loi normale centrée réduite et écrire que $Y_n \hookrightarrow \mathcal{N}(0, 1)$.

Montrons qu'il existe un réel t_α positif tel que $P(-t_\alpha \leq Y_n \leq t_\alpha) = 1 - \alpha$.

On a : $P(-t_\alpha \leq Y_n \leq t_\alpha) = \Phi(t_\alpha) - \Phi(-t_\alpha) = 2\Phi(t_\alpha) - 1$ donc

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

Donc un intervalle de confiance de m avec un niveau de confiance égal à $1 - \alpha$ est :

$$\left[\bar{X}_n - t_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X}_n + t_\alpha \frac{\sigma}{\sqrt{n}} \right].$$

En général la valeur de t_α est donnée dans l'énoncé, mais on peut aussi la calculer avec Python.

On utilisera la fonction `sp.ndtri(1 -  $\alpha/2$ )` qui renvoie $\Phi^{-1}(1 - \alpha/2)$ (librairie : `numpy.scipy as sp`).

Remarque.

1. Pour un risque de 10%, $\alpha = 0,1$, $1 - \alpha/2 = 0,95$, $\Phi^{-1}(0,95) = 1,65$.
2. Pour un risque de 5%, $\alpha = 0,05$, $1 - \alpha/2 = 0,975$, $\Phi^{-1}(0,975) = 1,96$.
3. Pour un risque de 1%, $\alpha = 0,01$, $1 - \alpha/2 = 0,995$, $\Phi^{-1}(0,995) = 2,58$.

De plus l'intervalle de confiance obtenu par cette méthode sera bien meilleur que celui obtenu avec l'inégalité de B-T.

EXERCICE 14.5.

On admet que la durée de vie, exprimée en jours, d'un composant électronique suit une loi normale d'écart-type 70. Les durées de vie de 250 composants ont donné une moyenne de 450 jours. Donner un intervalle de confiance à 99% de durée de vie d'un composant électronique.

Si on note, X_i la variable égale à la durée de vie du i ème composant et m la durée de vie moyenne d'un composant, on a :

$$n = 250, X_i \hookrightarrow \mathcal{N}(m, 70^2), \bar{X}_n = 450, t_\alpha = 2,58.$$

$$\text{On a alors } P\left(450 - 2,58 \times \frac{70}{\sqrt{250}} \leq m \leq 450 + 2,58 \times \frac{70}{\sqrt{250}}\right) = 0,99$$

D'où $m \in [438; 462]$ avec un risque inférieur à 1%.

14.2.6 Estimation du paramètre d'une variable de Bernoulli

Exemple 14.10.

Soit $(X_1, \dots, X_n)$ un n -échantillon de loi parente, une loi de Bernoulli de paramètre p à estimer. On considère $\bar{X}_n$ la moyenne empirique de l'échantillon.

C'est un estimateur sans biais de p . De plus, $\sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0, 1)$.

Si l'on choisit $t_\alpha = \Phi^{-1}(1 - \frac{\alpha}{2})$ alors en suivant le même raisonnement que dans l'exemple 14.9 précédent :

$$p \in \left[\bar{X}_n - t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}, \bar{X}_n + t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$$

avec un risque de α .

Seulement cet intervalle est donné en fonction de p qui est le paramètre à estimer, on utilise donc de nouveau la majoration $p(1-p) \leq 1/4$ sur $[0, 1]$.

Par suite :

$$p \in \left[\bar{X}_n - \frac{t_\alpha}{2\sqrt{n}}, \bar{X}_n + \frac{t_\alpha}{2\sqrt{n}} \right].$$

Si nous comparons les deux types d'intervalles de confiance obtenus (exemples 8 et 10) par l'inégalité de $B - T$ et celui obtenu par l'approximation de la loi binomiale par la loi normale, on peut voir que celui obtenu dans l'exemple 10 est plus précis.

En effet cela revient à comparer t_α à $\frac{1}{\sqrt{\alpha}}$.

Avec les valeurs données dans la remarque ci-dessus, on peut voir que $t_\alpha < \frac{1}{\sqrt{\alpha}}$.

EXERCICE 14.6.

On effectue un contrôle de fabrication sur des pièces usinées dont une proportion p est défectueuse. On en contrôle 200 et on en trouve 20 de défectueuses.

Donner des intervalles de confiance de p avec un niveau de confiance de 95% puis de 99%.

On note X_i la var de Bernoulli égale à 1 si la i ème pièce est défectueuse, alors $X_i \hookrightarrow \mathcal{B}(p)$.

$n = 200$, $\bar{X}_n = 20/200 = 1/10$, $t_\alpha = 1,96$ ou $2,58$.

On a alors $P(1/10 - t_\alpha \times \frac{1}{2\sqrt{200}} \leq p \leq 1/10 + t_\alpha \times \frac{1}{2\sqrt{200}}) \geq 1 - \alpha$

D'où $p \in [0,0307; 0,1693]$ avec un risque inférieur à 5% et $p \in [0,00880, 0,1913]$ avec un risque inférieur à 1%.

Feuille d'exercices n°14

ESTIMATIONS

Exercice 1. Etudier l'espérance d'un estimateur

1. Soit X une var définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. On suppose que X suit une loi de Bernoulli de paramètre p inconnu. Pour $n \in \mathbb{N}^*$, on désigne par $(X_1, X_2, \dots, X_n)$ un n -échantillon de var indépendantes, définies sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, de même loi que X . On note $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$.
Montrer que l'estimateur $\bar{X}_n$ a une espérance égale à p .
2. Soit X une var définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. On suppose que X suit une loi de Bernoulli de paramètre p inconnu. Pour $n \in \mathbb{N}^*$, on désigne par $(X_1, X_2, \dots, X_n)$ un n -échantillon de var indépendantes, définies sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, de même loi que X . On note $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$.
Montrer que, pour tout $n \in \mathbb{N}^*$, l'estimateur $\chi_n = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2$ est un estimateur biaisé de $\nu(X) = p(1-p)$.
Démontrer que la suite $(\chi_n)_{n \in \mathbb{N}^*}$ est une suite d'estimateurs dont l'espérance tend vers $\nu(X)$.
Construire à partir de χ_n un estimateur dont l'espérance est égale à $\nu(X)$.
3. Soit Y une var définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. On suppose que Y suit une loi de Poisson de paramètre $\lambda > 0$ inconnu.
Pour $n \in \mathbb{N}^*$, on désigne par $(Y_1, Y_2, \dots, Y_n)$ un n -échantillon de var indépendantes, définies sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, de même loi que Y .
On note $\bar{Y}_n = \frac{Y_1 + Y_2 + \dots + Y_n}{n}$.
Montrer que $e^{-\bar{Y}_n}$ a une espérance qui tend vers $e^{-\lambda}$.

Exercice 2. Calculer le risque quadratique d'un estimateur [HP]

1. Soit X une var définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. On suppose que X suit une loi de Bernoulli de paramètre p inconnu. Pour $n \in \mathbb{N}^*$, on désigne par $(X_1, X_2, \dots, X_n)$ un n -échantillon de var indépendantes, définies sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, de même loi que X . On note $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$.
Déterminer le risque quadratique de l'estimateur $\bar{X}_n$ de p .

2. Soit Y une var définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. On suppose que Y suit une loi de Poisson de paramètre $\lambda > 0$ inconnu.

Pour $n \in \mathbb{N}^*$, on désigne par $(Y_1, Y_2, \dots, Y_n)$ un n -échantillon de var indépendantes, définies sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, de même loi que Y .

On note $\bar{Y}_n = \frac{Y_1 + Y_2 + \dots + Y_n}{n}$.

Déterminer le risque quadratique de l'estimateur $e^{-\bar{Y}_n}$ en tant qu'estimateur du paramètre $e^{-\lambda}$.

Exercice 3. Prouver qu'un estimateur est convergent

1. Soit X une var définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. On suppose que X suit une loi de Bernoulli de paramètre p inconnu. Pour $n \in \mathbb{N}^*$, on désigne par $(X_1, X_2, \dots, X_n)$ un n -échantillon de var indépendantes, définies sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, de même loi que X . On note $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$.

Montrer que la suite $(\bar{X}_n)_{n \in \mathbb{N}^*}$ est une suite convergente d'estimateurs du paramètre p .

2. Soit Y une var définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. On suppose que Y suit une loi de Poisson de paramètre $\lambda > 0$ inconnu.

Pour $n \in \mathbb{N}^*$, on désigne par $(Y_1, Y_2, \dots, Y_n)$ un n -échantillon de var indépendantes, définies sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, de même loi que Y .

On note $\bar{Y}_n = \frac{Y_1 + Y_2 + \dots + Y_n}{n}$.

Montrer que la suite $(e^{-\bar{Y}_n})_{n \in \mathbb{N}^*}$ est une suite convergente d'estimateurs du paramètre $e^{-\lambda}$.

3. Soit Y une var de loi $\mathcal{N}(m, 1)$. Pour $n \in \mathbb{N}^*$, on désigne par $(Y_1, Y_2, \dots, Y_n)$ un n -échantillon de var indépendantes, définies sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, de même loi que Y .

On note $\bar{Y}_n = \frac{Y_1 + Y_2 + \dots + Y_n}{n}$.

Montrer que $\bar{Y}_n$ est un estimateur convergent du paramètre m .

Exercice 4. Construire un intervalle de confiance

1. Soit $\alpha \in]0, 1[$. Soit X une var définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. On suppose que X suit une loi de Bernoulli de paramètre $p \in]0, 1[$ inconnu. Pour $n \in \mathbb{N}^*$, on désigne par $(X_1, X_2, \dots, X_n)$ un n -échantillon de var indépendantes, définies sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, de même loi que X . On note $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$.

En majorant pour tout $\epsilon > 0$, $P(|\bar{X}_n - p| \geq \epsilon)$, construire un intervalle de confiance du paramètre p au niveau de risque α .

Quelle est l'amplitude de l'intervalle de confiance obtenu pour $n = 1000$ et $\alpha = 0.05$?

2. Soit $\alpha \in]0, 1[$. Soit X une var définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. On suppose que X suit une loi de Bernoulli de paramètre $p \in]0, 1[$ inconnu. Pour $n \in \mathbb{N}^*$, on désigne par $(X_1, X_2, \dots, X_n)$ un n -échantillon de var indépendantes, définies sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, de même loi que X . On note $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ et $\bar{X}_n^*$, la var centrée réduite associée à la var X_n .

Quelle convergence en loi peut-on établir pour la suite $(\bar{X}_n^*)_{n \in \mathbb{N}^*}$?

En déduire, un intervalle de confiance du paramètre p au niveau de risque α .

Quelle est l'amplitude de l'intervalle de confiance obtenu pour $n = 1000$ et $\alpha = 0.05$?

Exercice 5. EML 2012

Soit $a \in \mathbb{R}_+^*$.

1. Montrer que, pour tout entier n tel que $n \geq 0$, l'intégrale $I_n = \int_0^{+\infty} x^n e^{-\frac{x^2}{2a^2}} dx$ est convergente.
2. (a) Rappeler une densité d'une variable aléatoire. qui suit la loi normale d'espérance nulle et de variance a^2 .

En déduire : $I_0 = a\sqrt{\frac{\pi}{2}}$.

- (b) Calculer la dérivée de l'application $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ définie, pour tout $x \in \mathbb{R}$, par : $\varphi(x) = e^{-\frac{x^2}{2a^2}}$
En déduire : $I_1 = a^2$.

3. (a) Montrer, pour tout entier n . tel que $n \geq 2$ et pour tout $t \in [0; +\infty[$:

$$\int_0^t x^n e^{-\frac{x^2}{2a^2}} dx = -a^2 t^{n-1} e^{-\frac{t^2}{2a^2}} + (n-1) a^2 \int_0^t x^{n-2} e^{-\frac{x^2}{2a^2}} dx$$

- (b) En déduire, pour tout entier n tel que $n \geq 2$: $I_n = (n-1) a^2 I_{n-2}$.

- (c) Calculer I_2 et I_3 .

On considère l'application $g_a : \mathbb{R} \rightarrow \mathbb{R}$ définie, pour tout $x \in \mathbb{R}$, par :

$$g_a(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \frac{x}{a^2} e^{-\frac{x^2}{2a^2}} & \text{si } x > 0 \end{cases}$$

4. Montrer que g_a est une densité.

On considère une variable aléatoire X admettant g_a comme densité.

5. Déterminer la fonction de répartition de la variable aléatoire X .

6. Montrer que la variable aléatoire X admet une espérance $E(X)$ et que $E(X) = a\sqrt{\frac{\pi}{2}}$

7. Montrer que la variable aléatoire X admet une variance $V(X)$ et calculer $V(X)$.

8. (a) On considère une variable aléatoire U suivant la loi uniforme sur l'intervalle, $]0; 1]$. Montrer que la variable aléatoire $Z = a\sqrt{-2\ln(U)}$ suit la même loi que la variable aléatoire X

- (b) En déduire un programme en Python, simulant la variable aléatoire X , le réel a strictement positif étant entré par l'utilisateur.

Soit un entier n tel que $n \geq 2$.

On dit. que les. variables aléatoires à densité $X_1, X_2, \dots, X_n$ sont indépendantes si, pour tout, n -uplet $(x_1, x_2, \dots, x_n)$ de réels, les événements $(X_1 \leq x_1), (X_2 \leq x_2), \dots, (X_n \leq x_n)$ sont mutuellement indépendants.

On admet que si n variables aléatoires à densité $X_1, X_2, \dots, X_n$ admettent une espérance, alors la variable aléatoire $X_1 + \dots + X_n$ admet une espérance qui est égale à la somme des espérances.

On admet que si n variables aléatoires à densité $X_1, X_2, \dots, X_n$ sont indépendantes et admettent variance alors la variable aléatoire $X_1 + \dots + X_n$ admet une variance qui est égale à la. somme des variances.

On considère n variables aléatoires indépendantes $X_1, X_2, \dots, X_n$ suivant toutes la même loi que la variable aléatoire X .

9. On considère la variable aléatoire $A_n = \frac{\sqrt{2}}{n\sqrt{\pi}} (X_1 + X_2 + \dots + X_n)$.

- (a) Montrer que la variable -aléatoire A_n , est un estimateur sans biais de a .
- (b) Déterminer le risque quadratique de l'estimateur A_n .

On définit la variable aléatoire $M_n = \min(X_1, X_2, \dots, X_n)$.

Ainsi : $\forall t \in \mathbb{R}, (M_n > t) = (X_1 > t) \cap (X_2 > t) \cap \dots \cap (X_n > t)$.

- 10. (a) Montrer, pour tout $t \in]0, +\infty[$: $P(M_n > t) = e^{-\frac{nt^2}{2a^2}}$.
- (b) En déduire la fonction de répartition de M_n .
- (c) Montrer que M_n est une variable aléatoire à densité, admettant g_b comme densité avec $b = \frac{a}{\sqrt{n}}$.
- (d) Montrez- que la variable aléatoire M_n , admet une espérance $E(M_n)$ et une variance $V(M_n)$
Calculer $E(M_n)$ et $V(M_n)$.
- 11. (a) En déduire un estimateur B_n , sans biais de a , de la forme $\lambda_n M_n$ avec $\lambda_n \in \mathbb{R}$.
- (b) Déterminer le risque quadratique de l'estimateur B_n .

Exercice 6. HEC 2012

Sous réserve d'existence, on note $E(U)$ et $V(U)$ respectivement, l'espérance mathématique et la variance d'une variable aléatoire U définie sur l'espace probabilisé $(\Omega; \mathcal{A}; P)$.

Pour p entier supérieur ou égal à 2, on dit que les variables aléatoires à densité $U_1; \dots; U_p$ sont indépendantes si pour tout p -uplet $(u_1; \dots; u_p)$ de réels, les événements $[U_1 \leq u_1]; \dots; [U_p \leq u_p]$ sont indépendants.

L'objet du problème est l'étude de quelques propriétés d'une loi de probabilité utilisée notamment en fiabilité.

Les parties I et II sont largement indépendantes. La partie III est indépendante des parties I et II.

Partie I. Loi à 1 paramètre.

On note λ un paramètre réel strictement positif. On considère la fonction f_λ de $\mathbb{R}$ dans $\mathbb{R}$ définie par :

$$f_\lambda(x) = \begin{cases} \frac{\lambda}{2\sqrt{x}} e^{-\lambda\sqrt{x}} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}.$$

- 1. (a) Montrer que la fonction f_λ est de classe C^2 sur $\mathbb{R}_+^*$.
- (b) Dresser le tableau de variation de f_λ sur $\mathbb{R}_+^*$ et préciser les limites suivantes :
 $\lim_{x \rightarrow 0^+} f_\lambda(x)$, $\lim_{x \rightarrow +\infty} f_\lambda(x)$.
- (c) Établir la convexité de la fonction f_λ sur $\mathbb{R}_+^*$.

- (d) Tracer l'allure de la courbe représentative de f_λ dans le plan rapporté à un repère orthogonal.
2. (a) Vérifier que la fonction $x \mapsto -e^{-\lambda\sqrt{x}}$ est une primitive de f_λ sur $\mathbb{R}_+^*$.
- (b) Établir la convergence de l'intégrale $\int_0^{+\infty} f_\lambda(x) dx$ et calculer sa valeur.
- (c) En déduire que la fonction f_λ est une densité de probabilité sur $\mathbb{R}_+^*$.
3. Soit X une variable aléatoire définie sur un espace probabilisé $(\Omega; \mathcal{A}; P)$, à valeurs strictement positives, ayant f_λ pour densité. On note F_λ la fonction de répartition de X et on pose : $Y = \lambda\sqrt{X}$.
- (a) Calculer pour tout x réel, $F_\lambda(x)$.
- (b) Montrer que Y suit la loi exponentielle de paramètre 1.
- (c) Établir pour tout r de $\mathbb{N}^*$, l'existence de $E(Y^r)$.
- (d) Montrer que pour tout r de $\mathbb{N}^*$, on a : $E(Y^{r+1}) = (r+1)E(Y^r)$.
- (e) En déduire pour tout r de $\mathbb{N}^*$, $E(Y^r)$ et $E(X^r)$. En particulier, calculer $E(X)$ et $V(X)$.
4. Soit $(X_n)_{n \in \mathbb{N}^*}$ une suite de variables aléatoires définies sur $(\Omega; \mathcal{A}; P)$, indépendantes et de même loi que X .
Soit $(a_n)_{n \in \mathbb{N}^*}$ et $(b_n)_{n \in \mathbb{N}^*}$ deux suites de réels strictement positifs vérifiant $\lim_{n \rightarrow +\infty} n^2 a_n = 1$ et $\lim_{n \rightarrow +\infty} n^2 b_n = 0$.
On pose pour tout n de $\mathbb{N}^*$, $M_n = \min_{1 \leq k \leq n} (X_k)$ et $J_n = \frac{M_n - b_n}{a_n}$.
On admet que M_n et J_n sont des variables aléatoires à densité définies sur $(\Omega; \mathcal{A}; P)$.
Montrer que la suite $(J_n)_{n \in \mathbb{N}^*}$ converge en loi vers une variable aléatoire dont on précisera la loi.

Partie II. Estimation ponctuelle de λ .

Pour n entier de $\mathbb{N}^*$, on note $(X_1; \dots; X_n)$ un n -échantillon de variables aléatoires à valeurs strictement positives, indépendantes et de même loi que la variable aléatoire X définie dans la question 3. On rappelle que $Y = \lambda\sqrt{X}$, et on pose pour tout k de $[[1; n]]$, $Y_k = \lambda\sqrt{X_k}$, $S_k = \sum_{j=1}^k Y_j$ et g_k une densité de S_k .

On admet que pour tout entier n supérieur ou égal à 2, les variables aléatoires $Y_1; \dots; Y_n$ sont indépendantes et que pour tout k de $[[1; n]]$, les variables aléatoires S_k et Y_{k+1} sont indépendantes. On admet que si T et Z sont deux variables aléatoires à densité indépendantes définies sur le même espace probabilisé, de densités respectives f_T et f_Z telles que f_T et f_Z soit bornée, alors la variable aléatoire $T + Z$ admet une densité f_{T+Z} définie pour tout x réel par

$$f_{T+Z}(x) = \int_{-\infty}^{+\infty} f_T(y) f_Z(x-y) dy.$$

5. (a) En utilisant les propriétés admises, montrer que $g_2(x) = \begin{cases} xe^{-x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$.
- (b) À l'aide d'un raisonnement par récurrence, montrer que pour tout n de $\mathbb{N}^*$, on a $g_n(x) = \begin{cases} \frac{1}{(n-1)!} x^{n-1} e^{-x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$.

- (c) On admet que pour tout n de $\mathbb{N}^*$, $\frac{1}{S_n}$ est une variable aléatoire à densité. Pour quelles valeurs de n , l'espérance $E\left(\frac{1}{S_n}\right)$ et la variance $V\left(\frac{1}{S_n}\right)$ existent-elles? Calculer alors leurs valeurs respectives.
6. On note $(x_1; \dots; x_n)$ un n -uplet de $(\mathbb{R}_+^*)^n$ constituant une réalisation du n -échantillon $(X_1; \dots; X_n)$.
- On suppose que le paramètre λ est inconnu. Soit H la fonction de $\mathbb{R}_+^*$ dans $\mathbb{R}$ définie par :

$$H(\lambda) = \ln \left(\prod_{k=1}^n f_{\lambda}(x_k) \right).$$

Montrer que la fonction H admet un maximum atteint en un unique point λ_0 dont on donnera la valeur.

7. On pose pour tout entier n supérieur ou égal à 3 ou $\lambda_n^* = \frac{n}{\sum_{k=1}^n \sqrt{X_k}}$.

- (a) Que représente λ_0 pour λ_n^* ?
- (b) Construire à partir de λ_n^* un estimateur sans biais $\hat{\lambda}_n$ de λ et calculer le risque quadratique $\rho(\hat{\lambda}_n)$ de $\hat{\lambda}_n$.
- (c) Calculer $\lim_{n \rightarrow +\infty} \rho(\hat{\lambda}_n)$. Commenter.

Exercice 7. ESSEC 2009

Notations

- Tout au long du sujet $(\Omega, \mathcal{F}, P)$ désignera un espace probabilisé et les variables aléatoires utilisées seront toutes définies sur cet espace probabilisé. Sous réserve d'existence, l'espérance mathématique d'une variable aléatoire réelle X sera notée $E(X)$ et sa variance sera notée $V(X)$.
- Pour un événement A , on notera $P_B(A)$ la probabilité conditionnelle de A sachant B où B est un événement non négligeable.

Le sujet est composé de quatre parties. Les parties I, II, III et IVA sont **indépendantes**.

Il s'agit de variations autour de la notion de risque quadratique en théorie de l'estimation.

Premier problème d'estimation

Dans ce premier problème d'estimation, on dispose d'une seule observation notée X . On suppose que X admet pour densité f_{θ} définie sur $\mathbb{R}$ par

$$\begin{aligned} f_{\theta}(x) &= \frac{k+1}{\theta^{k+1}} x^k \text{ pour tout } x \in [0, \theta] \\ &= 0 \text{ sinon} \end{aligned}$$

où k est un entier naturel non nul et θ un paramètre réel inconnu strictement positif que l'on souhaite estimer.

1. Montrer que f_{θ} est bien une densité de probabilité.
2. Calculer $E(X)$.

3. Déterminer λ_0 réel dépendant uniquement de k tel que $\lambda_0 X$ estime θ sans biais.
4. Calculer $V(X)$.

On définit le risque quadratique de T estimateur de θ par

$$R(T, \theta) = E((T - \theta)^2)$$

Redémontrer le résultat du cours précisant que pour tout T estimateur de θ

$$R(T, \theta) = (E(T) - \theta)^2 + V(T)$$

6. Donner la valeur de $R(\lambda_0 X, \theta)$.

Le but de la fin de cette partie I est de déterminer un estimateur de θ ayant un plus petit risque quadratique que celui de $\lambda_0 X$.

7. En utilisant I.5 montrer que pour tout λ réel

$$R(\lambda X, \theta) = \theta^2 Q(\lambda)$$

où Q est un polynôme de degré 2 dont les coefficients ne dépendent que de k .

8. Montrer que la fonction $\lambda \mapsto Q(\lambda)$ atteint son minimum en un unique réel noté λ^* que l'on exprimera en fonction de k .
9. Conclure sur le but recherché.

Second problème d'estimation

Dans ce second problème d'estimation, on dispose de n observations indépendantes ($n \geq 2$) notées $X_1, \dots, X_n$ de même loi de Poisson de paramètre θ inconnu ($\theta \in]0, +\infty[$). On souhaite estimer le paramètre $\exp(-\theta)$.

On définit pour tout i élément de $\{1, \dots, n\}$ la variable aléatoire Y_i par

$$Y_i = 1 \text{ si } X_i = 0 \text{ et } Y_i = 0 \text{ sinon.}$$

Puis on note

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$$

1. Pour tout i élément de $\{1, \dots, n\}$, donner la loi de Y_i .
2. Donner la loi de $\sum_{i=1}^n Y_i$, puis montrer que $E(\overline{Y}_n) = \exp(-\theta)$. On dira dans ce cas que $\overline{Y}_n$ est un estimateur sans biais de $\exp(-\theta)$.
3. Calculer $V(\overline{Y}_n)$.

Pour tout k élément de $\{1, \dots, n\}$ on définit $S_k = \sum_{i=1}^k X_i$.

4. Rappeler sans démonstration la loi de S_k pour tout k élément de $\{1, \dots, n\}$.

On définit jusqu'à la fin de cette partie II pour tout j entier naturel

$$\varphi(j) = P_{\{S_n=j\}}(X_1 = 0)$$

5. Montrer que pour tout j entier naturel

$$\varphi(j) = \left(1 - \frac{1}{n}\right)^j$$

On a donc $\varphi(j)$ indépendant du paramètre θ inconnu.

D'après la question II.5, on peut définir l'estimateur

$$\varphi(S_n) = \left(1 - \frac{1}{n}\right)^{S_n}$$

6. Montrer que $\varphi(S_n)$ admet une espérance et que $E(\varphi(S_n)) = \exp(-\theta)$. On dira dans ce cas que $\varphi(S_n)$ est un estimateur sans biais de $\exp(-\theta)$.

7. Montrer que $\varphi(S_n)$ admet une variance vérifiant

$$V(\varphi(S_n)) = \exp(-2\theta) \left(\exp\left(\frac{\theta}{n}\right) - 1 \right)$$

8. On souhaite comparer les performances de $\overline{Y}_n$ et $\varphi(S_n)$ en tant qu'estimateurs de $\exp(-\theta)$.

(a) En utilisant le théorème des accroissements finis, démontrer que

$$1 \leq \frac{\exp(\theta) - 1}{\theta} \leq \exp(\theta)$$

(b) Soit la fonction $h : [0, 1] \rightarrow \mathbb{R}$ définie par

$$h(t) = t \exp(\theta) + (1-t) - \exp(t\theta)$$

pour tout $t \in [0, 1]$. Etudier les variations de h .

(c) En déduire que

$$\exp\left(\frac{\theta}{n}\right) \leq \frac{\exp(\theta)}{n} + \frac{n-1}{n}$$

puis l'inégalité

$$V(\varphi(S_n)) \leq V(\overline{Y}_n)$$

(d) On définit le risque quadratique de T_n estimateur de $\exp(-\theta)$ par

$$R(T_n, \exp(-\theta)) = E\left((T_n - \exp(-\theta))^2\right)$$

Comparer les risques quadratiques de $\overline{Y}_n$ et $\varphi(S_n)$.

On reprendra à la fin de la partie IV l'étude de $\varphi(S_n)$.

Information de Fisher

Cas discret

Dans cette section III.A, on considère I un intervalle de $\mathbb{R}$, θ un paramètre inconnu de I et X une variable aléatoire à valeurs dans $\mathbb{N}$ ($X(\Omega) \subseteq \mathbb{N}$). On suppose qu'il existe une fonction p définie sur $I \times X(\Omega)$ telle que pour tout k élément de $X(\Omega)$

$$P(X = k) = p(\theta, k)$$

et vérifiant pour tout k de $X(\Omega)$ $\theta \mapsto p(\theta, k)$ dérivable sur I .

On définit sous réserve d'existence l'**information de Fisher** de X par

$$I_X(\theta) = \sum_{k \in X(\Omega)} (\partial_1 \ln p(\theta, k))^2 p(\theta, k)$$

1. Dans cette question 1, on considère X variable aléatoire de loi de Bernoulli de paramètre θ ($\theta \in]0, 1[$).

On a alors $X(\Omega) = \{0, 1\}$, $P(X = 1) = p(\theta, 1) = \theta$, $P(X = 0) = p(\theta, 0) = 1 - \theta$ et

$$I_X(\theta) = (\partial_1 \ln p(\theta, 1))^2 p(\theta, 1) + (\partial_1 \ln p(\theta, 0))^2 p(\theta, 0)$$

Montrer que

$$I_X(\theta) = \frac{1}{\theta(1-\theta)}$$

2. Dans cette question 2, on considère X variable aléatoire de loi binomiale de paramètres N et θ ($N \in \mathbb{N}^*$, $\theta \in]0, 1[$)

(a) Montrer que

$$I_X(\theta) = \frac{1}{(\theta(1-\theta))^2} \sum_{k=0}^N (k - N\theta)^2 \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

(b) En déduire que

$$I_X(\theta) = \frac{V(X)}{(\theta(1-\theta))^2}$$

puis donner la valeur de $I_X(\theta)$.

3. Dans cette question 3, on considère X une variable aléatoire de Poisson de paramètre θ ($\theta \in]0, +\infty[$). Puisque $X(\Omega) = \mathbb{N}$, on a sous réserve de convergence

$$I_X(\theta) = \sum_{k=0}^{+\infty} (\partial_1 \ln p(\theta, k))^2 p(\theta, k)$$

(a) Montrer que la série de terme général $(\partial_1 \ln p(\theta, k))^2 p(\theta, k)$ converge et calculer sa somme $I_X(\theta)$.

(b) Justifier que

$$I_X(\theta) = E\left((\partial_1 \ln p(\theta, X))^2\right)$$

Cas d'une variable gaussienne

Soit X une variable aléatoire qui suit la loi normale de moyenne θ ($\theta \in \mathbb{R}$) et de variance 1 dont la densité est notée $x \mapsto f(\theta, x)$. On définit sous réserve de convergence l'**information de Fisher** de X par

$$I_X(\theta) = \int_{-\infty}^{+\infty} (\partial_1 \ln(f(\theta, x)))^2 f(\theta, x) dx$$

1. Montrer que sous réserve de convergence

$$I_X(\theta) = \int_{-\infty}^{+\infty} (x - \theta)^2 f(\theta, x) dx$$

2. En déduire l'existence et la valeur de $I_X(\theta)$.

3. Justifier que

$$I_X(\theta) = E\left((\partial_1 \ln(f(\theta, X)))^2\right)$$

Minoration du risque quadratique

Inégalité de Cramer-Rao

Dans cette section IV.A, on considère I un intervalle de $\mathbb{R}$, θ un paramètre inconnu de I et X une variable aléatoire telle que $X(\Omega) = \{0, \dots, N\}$ ($N \in \mathbb{N}$). On suppose qu'il existe une fonction p définie sur $I \times X(\Omega)$ telle que pour tout $k \in \{0, \dots, N\}$

$$P(X = k) = p(\theta, k)$$

et vérifiant :

- pour tout $k \in \{0, \dots, N\}$, $\theta \mapsto p(\theta, k)$ dérivable sur I ,
- l'information de Fisher de X notée $I_X(\theta)$ définie dans la partie III est non nulle pour tout $\theta \in I$.

Le but de la section IV.A est de démontrer l'inégalité suivante due à Cramer et Rao.

Théorème de Cramer-Rao

Soit $f(X)$ un estimateur sans biais de $g(\theta)$ à savoir tel que $E(f(X)) = g(\theta)$ où g est dérivable sur I . On a alors

$$V(f(X)) \geq \frac{(g'(\theta))^2}{I_X(\theta)}$$

1. Montrer que pour tout θ élément de I

$$\sum_{k=0}^N \partial_1(p(\theta, k)) = 0$$

2. En déduire que pour tout θ élément de I

$$E(\partial_1 \ln(p(\theta, X))) = 0 \quad (E)$$

3. En dérivant partiellement par rapport à θ les deux membres de l'égalité (E), montrer que pour tout θ élément de I

$$E(\partial_{1,1}^2 \ln(p(\theta, X))) = -E((\partial_1 \ln(p(\theta, X)))^2)$$

4. Montrer que pour tout θ élément de I

$$g'(\theta) = \sum_{k=0}^N f(k) (\partial_1 \ln(p(\theta, k))) p(\theta, k)$$

puis que

$$g'(\theta) = E((f(X) - g(\theta)) \partial_1 \ln(p(\theta, X)))$$

5. On pose pour tout t réel

$$L(t) = E\left(\left((f(X) - g(\theta)) + t \partial_1 \ln(p(\theta, X))\right)^2\right)$$

- (a) Développer le polynôme L suivant les puissances décroissantes de t .

- (b) Calculer le discriminant Δ de L et justifier que $\Delta \leq 0$.
- (c) En déduire l'inégalité de Cramer-Rao.

Extension du théorème de Cramer-Rao

On reprend dans cette section IV.B les notations et hypothèses de la partie II. On admet que, dans ce contexte, le théorème de Cramer-Rao peut se généraliser comme suit :

Théorème de Cramer-Rao

Soit $T_n = f(X_1, \dots, X_n)$ un estimateur sans biais de $g(\theta)$ à savoir tel que $E(f(X_1, \dots, X_n)) = g(\theta)$ où g est dérivable sur $]0, +\infty[$. On a alors

$$V(T_n) \geq \frac{(g'(\theta))^2}{nI_{X_1}(\theta)}$$

où $I_{X_1}(\theta)$ est l'information de Fisher d'une variable aléatoire de loi de Poisson de paramètre θ définie et calculée à la partie III.

Il s'agit dans cette section d'exploiter cette nouvelle inégalité de Cramer-Rao. On note

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

1. Calculer $E(\overline{X}_n)$ et $V(\overline{X}_n)$.
2. Déduire de la généralisation de Cramer-Rao, que $\overline{X}_n$ a le plus petit risque quadratique parmi les estimateurs sans biais de θ .
3. Montrer que pour $g(\theta) = \exp(-\theta)$ où $\theta \in]0, +\infty[$

$$V(\varphi(S_n)) \underset{n \rightarrow +\infty}{\sim} \frac{(g'(\theta))^2}{nI_{X_1}(\theta)}$$

4. Que prouve ce résultat en terme d'optimalité de $\varphi(S_n)$ dans l'estimation de $\exp(-\theta)$?
5. A la lumière de la partie I, peut-on conclure que lorsque n est grand $\varphi(S_n)$ est le meilleur estimateur de $\exp(-\theta)$ en terme de risque quadratique?