CHAPITRE III
ESTIMATION STATISTIQUE

I . GÉNÉRALITÉS

I.1. OBJET DE L'ESTIMATION STATISTIQUE

Le problème de l'estimation statistique consiste à donner des valeurs approchées aux paramètres inconnus d'une population (par exemple la moyenne m, et la variance σ^2 si le caractère est quantitatif ou la proportion p, si le caractère est qualitatif à 2 modalités) à l'aide d'un n-échantillon d'observations issu de la population.

Exemple
On se demande qu'elle est la proportion exacte de fumeurs chez les jeunes de 14 à 18 ans. Il est évident que l'examen de tous les individus concernés aboutirait à la valeur exacte cherchée, mais cette méthode est trop coûteuse à mettre en oeuvre. Par contre nous disposons de l'information collectée au niveau d'un échantillon : sur 100 jeunes observés 35 sont fumeurs ; cette valeur observée nous informe sur la "vraie valeur" de p, proportion de fumeurs dans la population étudiée. En effet, on peut en déduire que p est approximativement égal à 0,35.

Cependant on peut souhaiter disposer d'une information plus précise sur p en donnant un encadrement pour p ou fourchette d'estimation pour p ; ainsi, des observations on peut conclure plus précisément après des calculs ad-hoc, qu'avec un risque d'erreur de 5% on a : $0{,}26 \leq p \leq 0{,}44$.

La démarche d'inférence statistique consiste à étendre à la population toute entière des propriétés observées sur un/des échantillon(s) issu(s) de cette population par tirage au hasard. L'estimation des paramètres caractérisant la répartition d'un caractère statistique dans une population donnée, s'inscrit dans cette démarche .

Cette démarche est inverse de celle adoptée au chapitre précédent en théorie de l'échantillonnage. En effet, en théorie de l'échantillonnage, le statisticien étudie les relations liant la répartition d'un caractère statistique dans un échantillon d'observations et la répartition de ce même caractère dans la population toute entière dont il est issu.

I.2. CADRE DE L'ESTIMATION STATISTIQUE ET NOTATIONS

Comme dans le chapitre consacré à l'échantillonnage, on considère une population $\mathcal{P}$ de taille N, dans laquelle on s'intéresse à la distribution d'un certain caractère $\mathcal{X}$.

Dans cette population, un échantillon de taille n est prélevé au hasard ; la valeur du caractère sur le $i^{ème}$ individu de l'échantillon est une variable aléatoire notée X_i. A l'échantillon de taille n prélevé, on associe un n-uplet de variables $(X_1, X_2, ..., X_n)$.

Le prélèvement au hasard implique que :
(1) Au caractère $\mathcal{X}$ considéré, on associe une variable aléatoire X, dont la loi de probabilité est donnée par la répartition du caractère $\mathcal{X}$ dans la population.
(2) Les v.a. $X_1, ..., X_n$ sont de même loi que X.

CamScanner

(3) Les v.a. $X_1, ..., X_n$ sont indépendantes en probabilité :
- exactement, si les tirages ont lieu avec remise,
- de manière approchée s'ils ont lieu sans remise mais que la condition $n/N < 0,1$ est satisfaite.

Ceci nous a conduit à définir le n-uplet $(X_1, ..., X_n)$ de v.a. indépendantes et de même loi comme **n-échantillon** de la variable X.

Le problème de l'estimation statistique se présente alors de la façon suivante : on suppose que la loi de probabilité de X est d'un type connu mais que l'on ignore un ou plusieurs paramètres de cette loi.

On notera θ le (ou les) paramètre(s) inconnus et l'on utilisera, dans la suite du cours la ***notation*** suivante :

pour toute v.a. X et pour tout nombre réel x, on posera :

$$f(x, \theta) = \begin{cases} f_\theta(x) \text{ densité de X} & \text{si X est une v.a. continue} \\ P_\theta(X=x) & \text{si X est une v.a. discrète} \end{cases}$$

Exemples illustrant cette notation :

1) Si X suit la loi de Poisson de paramètre θ, alors :

$$f(x, \theta) = e^{-\theta} \frac{\theta^x}{x!} \qquad \text{où } x \in \mathbb{N}$$

2) Si X suit une loi normale $\mathcal{N}(\theta, \sigma)$, où l'écart-type σ est connu et où l'espérance mathématique θ est inconnue alors :

$$f(x,\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{1}{2\sigma^2} (x - \theta)^2 \right) \qquad \text{avec } x \in \mathbb{R}$$

3) Si X suit une loi normale $\mathcal{N}(\theta_1,\theta_2)$, où $\theta = (\theta_1,\theta_2)$ est inconnu, on pose :

$$f(x,\theta) = \frac{1}{\theta_2\sqrt{2\pi}} \exp\left(- \frac{1}{2\theta_2^2} (x - \theta_1)^2 \right) \qquad \text{avec } x \in \mathbb{R}.$$

L'objet de l'estimation statistique est de déterminer une valeur approchée du (ou des) paramètre(s) inconnu(s) θ, à partir de valeurs observées $x_1, ..., x_n$ de X.

Deux approches sont possibles :
(1) ou bien on souhaite disposer d'une valeur numérique ayant une forte probabilité d'être voisine du paramètre inconnu θ ; on est alors dans le cas de l'estimation ponctuelle et l'on dit que cette valeur numérique est une **estimation** du paramètre θ .
(2) ou bien on veut construire un intervalle qui a une forte probabilité de contenir le paramètre θ ; cette approche est celle de l'**estimation par intervalle de confiance**.
Cette seconde approche est plus complète que la première car elle permet d'apprécier à la fois la **précision** et la **crédibilité** de l'estimation obtenue.

I.3. NOTION D'ESTIMATEUR

Dans le cas de l'estimation ponctuelle, autrement dit lorsque l'on souhaite obtenir une valeur unique, appelée **estimation**, plus que la valeur numérique elle même, c'est la manière de l'obtenir qui intéresse le statisticien. En effet, il est courant que l'on ait à évaluer θ à partir de différents échantillons. On s'attachera donc à mettre en évidence une **règle d'obtention** de la valeur approchée de θ à partir des valeurs observées $x_1, ..., x_n$.

Cette règle sera caractérisée par une statistique, c'est-à-dire par une fonction algébrique des variables de l'échantillon, soit $T = g(X_1, ..., X_n)$.

CamScanner

Compte tenu de son emploi, cette statistique sera appelée **estimateur du paramère θ.**

Si l'échantillon observé est $(x_1, ..., x_n)$, réalisation de l'échantillon aléatoire $(X_1, ..., X_n)$, alors l'**estimateur** $T = g(X_1, ..., X_n)$, qui est une variable aléatoire, prend la valeur $t = g(x_1, ..., x_n)$ qui est une **estimation** de θ (un nombre certain) ; cette estimation est parfois notée $\hat{\theta}$.

Deux questions se posent désormais et feront l'objet des prochains paragraphes :
(1) Quelles qualités demander à un estimateur ?
(2) Comment obtenir des estimateurs pour un certain paramètre inconnu θ ?

II. L'ESTIMATION PONCTUELLE

II.1. PROPRIETES DES ESTIMATEURS

Un **estimateur** a été défini comme étant une statistique, c'est-à-dire une fonction des variables de l'échantillon, dont les valeurs ont une forte probabilité d'être proches de la vraie valeur du paramètre inconnu θ. Quelles qualités exiger de cette statistique pour que ses valeurs soient proches de θ ?

II.1.1. Biais d'un estimateur

Soit X une v.a. dont la loi, donnée par $f(x, \theta)$, est d'un type connu mais dont le paramètre θ est inconnu. Soient $(X_1, ..., X_n)$ un échantillon de taille n de la v.a. X et $T_n = g(X_1, ..., X_n)$ un estimateur de θ.

Définitions
L'estimateur T_n est dit **sans biais** si $E(T_n) = \theta$, pour tout n.
Sinon on appelle **biais de l'estimateur** T_n la quantité $b_n = E(T_n) - \theta$.
Dans le cas où l'estimateur est biaisé on peut avoir la propriété un peu moins forte suivante :
l'estimateur T_n est dit **asymptotiquement sans biais** si $\lim_{n\to\infty} b_n = 0$, i.e. si $\lim_{n\to\infty} E(T_n) = \theta$.

Remarque :
Si un estimateur T_n du paramètre θ est sans biais, ses valeurs seront "en moyenne" égales à θ cependant elles pourront s'en écarter plus ou moins selon les fluctuations d'échantillonnage.

exemples d'estimateurs sans biais :

a) la fréquence empirique
Considérons le cas d'un caractère qualitatif à deux modalités : autrement dit chaque individu possède une certaine propriété A ou non, alors $\theta = p$ est la proportion des individus qui possèdent la propriété A dans $\mathcal{P}$.

On tire au hasard n individus et l'on définit les variables aléatoires $X_1, ..., X_n$ par

$$X_i = \begin{cases} 1 \text{ si le } i^{ème} \text{ individu tiré possède la propriété A} \\ 0 \text{ sinon.} \end{cases}$$

On a vu dans le cours d'échantillonnage que la fréquence empirique

$$F_n = \frac{X_1 + ... + X_n}{n}$$

CamScanner

vérifie $E(F_n) = p$, autrement dit, F_n est un estimateur sans biais de p.

b) La moyenne empirique :
Si m est la moyenne théorique d'une certaine v.a. X, autrement dit $m = E(X)$ et si, à l'échantillon $(X_1, ..., X_n)$ de la v.a. X, on associe la statistique :

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

alors, on sait que : $E(T_n) = E(X) = m$ (cf. cours d'échantillonnage).

Ainsi, $T_n = \bar{X}_n$ est un estimateur sans biais du paramètre $\underset{\sim}{m}$.

c) La variance empirique corrigée :
Si l'on souhaite estimer la variance $\theta = \sigma^2$ de X (on suppose que σ^2 est finie), on a vu que la variance empirique de l'échantillon, c'est-à-dire la statistique :

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2, \text{ vérifie l'égalité : } E(S_n^2) = \frac{(n-1)}{n} \sigma^2.$$

S_n^2 est donc un estimateur asymptotiquement sans biais de σ^2.

Si au lieu de S_n^2, on prend la variance empirique corrigée de l'échantillon, c.a.d. la statistique :

$$S'^2_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n}{n-1} S_n^2 \text{ on a : } E(S'^2_n) = \sigma^2.$$

de sorte que S'^2_n est un estimateur sans biais de la variance théorique σ^2.

II.1.2. Les estimateurs convergents

Les définitions suivantes doivent nous permettre de préciser l'idée intuitive selon laquelle, plus n est grand, plus l'estimateur T_n a une forte probabilité de prendre des valeurs proches du paramètre inconnu θ.

Définition 1
Un estimateur T_n du paramètre θ est dit **convergent en probabilité** si, pour tout nombre réel positif ε, on a :
$\lim_{n\to\infty} P(|T_n - \theta| < \varepsilon) = 1$ ou, ce qui revient au même, si $\lim_{n\to\infty} P(|T_n - \theta| \geq \varepsilon) = 0$.

Définition 2
Un estimateur T_n du paramètre θ est dit **convergent en moyenne quadratique** (en **m.q.**) s'il vérifie : $\lim_{n\to\infty} E((T_n - \theta)^2) = 0$.

Remarque
On rappelle que la convergence en m.q. entraîne la convergence en probabilité, ceci justifie l'intérêt de la proposition suivante qui fournit un critère de convergence en moyenne quadratique.

Proposition 1 :
Un estimateur T_n du paramètre θ est convergent en moyenne quadratique si, et seulement si, il est sans biais ou asymptotiquement sans biais et si sa variance tend vers 0 lorsque n tend vers l'infini.

CamScanner

exemple d'application de la proposition 1 :

Pour estimer la moyenne théorique m d'une v.a. X on peut prendre $T_n = \bar{X}_n$. On a déjà remarqué que cet estimateur est sans biais. Si de plus, la variance V(X) est finie, la relation : $V(\bar{X}_n) = \frac{V(X)}{n}$ montre que $V(\bar{X}_n)$ tend vers 0 lorsque n tend vers l'infini. La proposition 1 permet alors d'affirmer que $\bar{X}_n$ est un estimateur convergent en m.q. de $m = E(X)$.

Remarquons que le critère de convergence en m.q. donné à la proposition 1 suppose le calcul de l'espérance mathématique et de la variance de l'estimateur T_n. En pratique, ce calcul n'est pas toujours aisé ; la méthode suivante permet, alors dans certains cas, de démontrer la convergence en probabilité de T_n vers θ.

Proposition 2 - Théorème de Slutsky :
Soient (Y_n) est une suite infinie de v.a. qui converge en probabilité vers la v.a. Y et g une fonction réelle, continue définie sur l'ensemble des valeurs possibles des v. a. Y et Y_n, alors la suite des v.a. $(g(Y_n))$ converge en probabilité vers g(Y), lorsque n tend vers l'infini. On note : $\lim_{n\to\infty} g(Y_n) = g(Y)$ (Pr).
Dans les *exemples* suivants l'application de ce résultat, permet d'obtenir la convergence en probabilité des estimateurs concernés :

Exemple 1 :

On a vu que l'estimateur S_n^2 (et aussi S'^2_n) converge en moyenne quadratique vers σ^2. Il converge également en probabilité. Par ailleurs, la fonction g définie par $g(u) = \sqrt{u}$ pour tout u positif, est continue. La proposition 2 permet alors d'affirmer que $S_n = \sqrt{S_n^2}$ tend en probabilité vers $\sigma = \sqrt{\sigma^2}$, lorsque n tend vers l'infini.
S_n est donc un estimateur convergent en probabilité du paramètre σ.

Exemple 2 :
Soit X une v.a. de loi exponentielle admettant pour densité :

$$f(x) = \begin{cases} a\, e^{-ax} & \text{si} \quad x \geq 0 \\ 0 & \text{si} \quad x < 0 \end{cases}$$

où a est un paramètre positif que l'on veut estimer. On sait que $E(X) = \frac{1}{a}$.

Montrons $T_n = \frac{1}{\bar{X}_n}$ est un estimateur convergent pour le paramètre a.

$T_n \approx \bar{X}_n$ ($E(X) \approx \bar{X}_n$)

Daprès la loi des grands nombres on a: $\lim_{n\to\infty} \bar{X}n = E(X)$ (Pr) où $E(X) = \frac{1}{a}$.

La fonction $g(u) = \frac{1}{u}$ est continue sur $\mathbb{R}^*$; en appliquant la proposition 2, on voit que la suite des v.a. $T_n = \frac{1}{\bar{X}_n}$ converge en probabilité vers $\frac{1}{E(X)} = a$ lorsque n tend vers l'infini. On a ainsi établi la convergence en probabilité de l'estimateur T_n vers la valeur inconnue a .

CamScanner

II.2 COMPARAISON D'ESTIMATEURS

Il est souvent possible d'obtenir deux ou plusieurs estimateurs pour un même paramètre θ. Par exemple, pour une distribution symétrique la moyenne et la médiane empirique sont des estimateurs de $m = E(X)$.
La définition suivante nous permettra d'élaborer un critère de choix entre deux estimateurs.

Définition
On appelle **écart quadratique moyen (e.q.m)** la quantité $E[(T_n - \theta)^2]$.
On peut l'interpréter comme le carré d'une distance entre la vraie valeur du paramètre inconnu θ et son estimateur T_n.

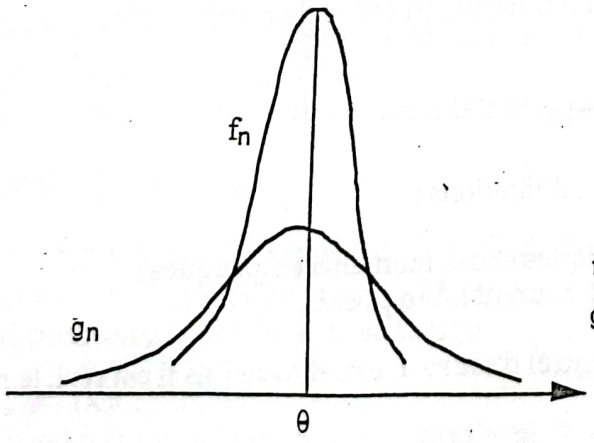
Remarque
La relation suivante exprime l'écart quadratique moyen en fonction de la variance et du biais de l'estimateur :

(*) $$E((T_n - \theta)^2) = Var(T_n) + b_n^2$$

Comparaison d'estimateurs :
Si T_n et U_n sont deux estimateurs d'un même paramètre θ, on préférera en général celui qui, à taille d'échantillon égale, a le plus faible e.q.m.
En effet l'e.q.m. d'un estimateur mesure sa dispersion autour de la vraie valeur du paramètre inconnu θ, par conséquent plus l'e.q.m. sera faible, meilleure sera la précision de l'estimation.

(1) *Cas de deux estimateurs sans biais :*
Si deux estimateurs U_n et T_n d'un même paramètre θ sont sans biais , on préférera en général celui qui, à taille d'échantillon égale, a la plus faible variance. En effet la variance d'une v.a. mesure sa dispersion autour de son espérance ; par conséquent, plus la dispersion sera faible, meilleure sera la précision de l'estimation.

Le graphique ci-dessous illustre cette situation :

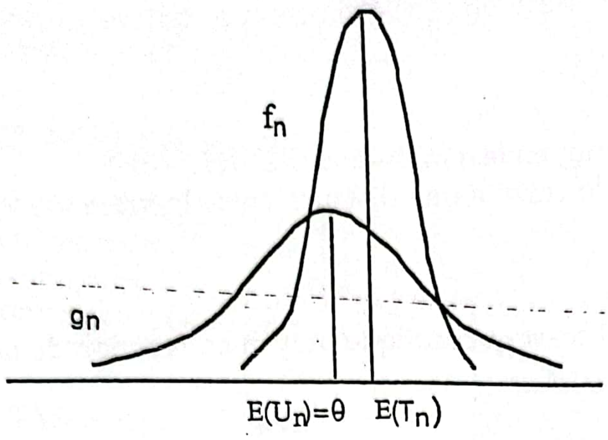


f_n densité de T_n
g_n densité de U_n

T_n et U_n sont sans biais et on a : $E(T_n) = E(U_n) = \theta$
de plus $V(T_n) < V(U_n)$
On est ainsi conduit à préférer T_n à U_n.

CamScanner

(2) *Cas de deux estimateurs dont l'un au moins est biaisé :*
Le graphique ci-dessous permet de comparer graphiquement deux estimateurs T_n et U_n admettant pour densité de probabilité, respectivement f_n et g_n.



Ici l'estimateur U_n est sans biais et on a $E(U_n) = \theta$; l'estimateur T_n est biaisé, et on a : $E(T_n) > \theta$.
Si on a : $E((T_n - \theta)^2) < E((U_n - \theta)^2) = V(U_n)$ autrement dit si e.q.m(T_n) < e.q.m(U_n) alors T_n sera préféré à U_n.
On est donc amené ici à préférer un estimateur biaisé à un estimateur non biaisé si l'erreur quadratique du premier est inférieure à la variance du second.

I. 3 CONSTRUCTION D'ESTIMATEURS

Certains estimateurs sont évidents. Par exemple la **moyenne empirique** $\overline{X}$ est un estimateur sans biais, convergent de la moyenne théorique m ; de même la **variance empirique corrigée** S'^2 est un estimateur sans biais et convergent de la variance théorique σ^2 .

Mais il n'en est pas toujours ainsi, nous allons donc indiquer deux méthodes de calcul d'estimateurs. Ces méthodes permettront, en outre de confirmer les estimateurs que nous connaissons déjà.

II.3.1. Méthode des moments

Au préalable donnons quelques définitions :

Définition des moments théoriques et des moments empiriques :
Soit X une variable aléatoire et r un nombre positif,

Le moment théorique (non centré) d'ordre r est, si toutefois il est réel, le nombre :

$$m_r(X) = E(X^r).$$

Le moment théorique centré d'ordre r est, si il existe, le nombre :

$$\mu_r(X) = E[(X - E(X))^r].$$

On a par exemple :
pour r = 1 $\quad m_1(X) = E(X)$ $\quad$ et pour r = 2 $\quad m_2(X) = E(X^2)$
$\quad \mu_1(X) = 0$ $\quad \mu_2(X) = V(X)$.

Si maintenant $(X_1, ..., X_n)$ est un n-échantillon de la v.a. X on peut définir les moments empiriques suivants, qui sont des statistiques d'échantillonnage, donc des variables aléatoires :

Le **moment empirique (non centré) d'ordre r** est la statistique :

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

Si $r = 1$ on obtient la moyenne empirique $\bar{X}_n$.

Le **moment empirique centré d'ordre r** de la v.a. X est la statistique :

$$\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)^r$$

Pour $r = 1$ le moment empirique centré est nul ; pour $r = 2$ on retrouve la variance empirique non corrigée :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

Remarques :

(1) Les moments théoriques sont des **nombres certains** tandis que les moments empiriques sont des **variables aléatoires.**

(2) Seuls les moments empiriques sont **observables** ; en effet, à partir d'une réalisation $(x_1, ..., x_n)$ de l'échantillon on obtient une réalisation de l'un quelconque de ses moments empiriques. Au contraire, les moments théoriques, sont le résultat d'une définition ou d'un calcul mathématique mais pas d'une observation.

Méthode des moments

Cas où le paramètre θ est unidimensionnel.

Soit $(X_1, ..., X_n)$ un n-échantillon de la variable X. D'après la loi des grands nombres la statistique $\bar{X}_n$ tend en probabilité vers E(X), lorsque n tend vers l'infini. Dans un échantillon de taille suffisamment grande, les valeurs $\bar{x}_n$ de la v.a. $\bar{X}_n$ auront donc une forte probabilité d'être voisines de E(X).

Ainsi lorsque $\theta = E(X)$ la méthode des moments consiste à prendre pour estimateur de la moyenne, la statistique $T_n = \bar{X}_n$.

Si maintenant θ apparaît dans l'expression de E(X) c'est-à-dire si l'on a :

$E(X) = \varphi(\theta)$ où φ est une fonction bijective sur $\mathbb{R}$ et donc inversible,

alors on prend pour estimateur de θ la statistique

$$T_n = \varphi^{-1}(\bar{X})$$

Lorsque E(X) n'est pas fonction de θ (par exemple, si X est centrée), il faut rechercher un moment théorique d'ordre supérieur qui le soit. Si, par exemple, $E(X^2)$ est fonction de θ autrement dit si on a $E(X^2) = \varphi(\theta)$ où φ est une fonction bijective et donc inversible, la convergence du moment empirique d'ordre deux, soit $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ vers $E(X^2)$ le moment

théorique d'ordre 2, lorsque n tend vers l'infini nous amène à prendre pour estimat de θ la statistique :

$$T_n = \varphi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)$$

Exemples de calcul d'estimateurs par la méthode des moments :

(1) ***Paramètre d'une loi exponentielle :***

Si X est une v.a. de loi exponentielle de paramètre θ, on a alors $E(X) = \frac{1}{\theta}$ autrement dit φ est la fonction qui à u associe $\varphi(u) = \frac{1}{u}$ $(u > 0)$.

On a alors $\varphi^{-1} : v \to \varphi^{-1}(v) = \frac{1}{v}$ et l'estimateur donné par la méthode des moments est

$$T_n = \varphi^{-1}(\bar{X}) = \frac{1}{\bar{X}}.$$

(2) ***Variance d'une loi normale d'espérance connue :***

Soit X une v.a. de loi normale $\mathcal{N}(m, \sigma)$ où m est connu mais σ est inconnu. On cherche un estimateur du paramètre $\theta = \sigma^2$.

Le moment théorique d'ordre 2 s'écrit $E(X^2) = m^2 + \sigma^2 = m^2 + \theta = \varphi(\theta)$

autrement dit la fonction φ est définie par $\varphi(u) = m^2 + u$ et $\varphi^{-1}(v) = v - m^2$ ce qui donne l'estimateur :

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

Cas d'un paramètre θ multidimensionnel :

Plus généralement si θ n'est pas un nombre mais un vecteur $\theta = (\theta_1,...,\theta_k)$, il y a k paramètres à estimer. Il faudra alors écrire k équations faisant intervenir les moments théoriques.

Le choix des moments théoriques sera fait de manière que chacun des $\theta_1,...,\theta_k$ figure dans l'expression de l'un au moins de ces moments.

Si par exemple ces k moments sont les k premiers, on sera conduit au système d'équations suivant :

$$\varphi_r(\theta_1,...,\theta_k) = E(X^r) \quad (r = 1,...,k)$$

Il suffira alors de résoudre ce système par rapport aux inconnues $\theta_1, ...,\theta_k$.

Exemple : Estimation simultanée des paramètres de la loi normale $\mathcal{N}(m, \sigma)$

Si m et $v = \sigma^2$ sont tous deux inconnus, on a $\theta = (m,v)$.

Or on a : $m = E(X)$ et $v + m^2 = E(X^2)$.

Compte tenu de la relation :

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{1}{n}\sum_{i=1}^{n} (Xi - \bar{X})^2 + \bar{X}^2.$$

on obtient les estimateurs :

$$T_1 = \bar{X} \quad \text{et} \quad T_2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2 = S^2$$

L'estimateur du vecteur $\theta = (m, v)$ est donc $T = (T_1, T_2) = (X, S^2)$.

CamScanner

Propriétés des estimateurs obtenus par la méthode des moments

(1) D'après la loi des grands nombres, les moments empiriques tendent en probabilité vers les moments théoriques de même ordre, par suite les estimateurs obtenus par la méthode des moments sont convergents en probabilité.

(2) Si n est assez grand, les moments empiriques suivent approximativement une loi normale.

D'autres propriétés des estimateurs obtenus par la méthode des moments pourront être mises en évidence par une étude plus approfondie dans chaque cas considéré.

II.3.2. Méthode du maximum de vraisemblance

Exemple introductif
Une urne contient des boules blanches et noires en proportion 1/4 et 3/4 mais l'on ignore quelle couleur figure en proportion 1/4. On choisit au hasard avec remise, trois boules dans cette urne, et l'on observe le nombre de boules blanches obtenues.

Ainsi, la population considérée est l'ensemble des boules de l'urne. Le caractère statistique est qualitatif et comporte deux modalités : "blanc" ou "noir", codées respectivement 1 et 0. La v.a. X associée à ce caractère suit la loi de Bernoulli de paramètre $\theta = p$ où p est la proportion de boules blanches qui vaut soit 1/4, soit 3/4.

On observe une réalisation de (X_1, X_2, X_3) qui est un échantillon de taille 3 de X et l'on pose : $Z = X_1 + X_2 + X_3$.

Alors la statistique Z représente le nombre de boules blanches tirées ; Z suit la loi $\mathcal{B}(3,p)$. On désigne par z l'une quelconque de ses réalisations et l'on pose :

$$h(z, p) = P(Z = z) = C_3^z\, p^z\, (1-p)^{3-z}$$

où z = nombre de boules blanches dans l'échantillon (z peut prendre les valeurs 0, 1, 2 ou 3).

On peut dresser le tableau suivant :

valeurs de Z	0	1	2	3
h(z, 3/4)	1/64	9/64	24/64	27/64
h(z,1/4)	27/64	27/64	9/64	1/64

On décide de choisir comme estimation la valeur de θ (1/4 ou 3/4) qui, à z fixé, maximise la fonction h(z, θ). Autrement dit, on choisit la valeur de θ qui donne à la probabilité de réalisation de z la valeur maximale. On est ainsi amené à définir l'estimateur W de θ par la formule

$$W = \begin{cases} 1/4 & \text{si } z = 0 \text{ ou } 1 \\ 3/4 & \text{si } z = 2 \text{ ou } 3 \end{cases}$$

La méthode du maximum de vraisemblance

Définition
Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon de la v.a. X dont la loi dépend du paramètre inconnu θ. On appelle **vraisemblance** de θ étant données les observations $(x_1, x_2, \ldots, x_n)$, le nombre :

$$L(x_1, \ldots, x_n ; \theta) = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta).$$

CamScanner

La fonction $\theta \to L(x_1, ..., x_n ; \theta)$ est appelée **fonction de vraisemblance** de θ.

Remarque :
On rappelle que $f(x, \theta)$ définit la loi de la v.a. X par la formule :

$$f(x, \theta) = \begin{cases} f_\theta(x) \text{ densité de X} & \text{si X est une v.a. absolument continue} \\ P_\theta(X=x) & \text{si X est une v.a. discrète} \end{cases}$$

alors $L(x_1, ..., x_n ; \theta)$ est la densité du n-uplet $(X_1, X_2, ..., X_n)$ sachant θ si X est une v.a. absolument continue et la probabilité $P_\theta(X_1 = x_1, ..., X_n = x_n)$ si X est une v.a. discrète.

La méthode du maximum de vraisemblance (en abrégé **méthode du M.V.**) consiste, comme dans l'exemple précédent, à choisir la (ou les) valeur(s) de θ qui rend(ent) maximum la vraisemblance du paramètre c.à.d. qui rend(ent) maximum la probabilité de réalisation des valeurs $x_1, ..., x_n$ observées lors du tirage de l'échantillon.

Il s'agit donc d'un **problème d'optimisation** : on cherche une valeur $\hat{\theta}$ de θ qui maximise la fonction :

$$\theta \to L(x_1, ..., x_n ; \theta)$$

Autrement dit, il faut que $\hat{\theta}$ vérifie :

$$L(x_1, ..., x_n ; \theta) \le L(x_1, ..., x_n ; \hat{\theta}) \text{ pour tout } \theta.$$

En général la fonction ci-dessus est dérivable et on peut utiliser les dérivées pour la recherche du maximum.

On remarque que la vraisemblance étant positive, on peut se contenter de chercher le maximum de la fonction $\theta \to L(x_1, ..., x_n ; \theta)$ sur l'ensemble des θ où cette fonction est **strictement** positive.

Sur cet ensemble, il est alors équivalent de chercher la valeur de θ qui rend maximum la fonction :

$$\theta \to \ell n\, L(x_1, ..., x_n ; \theta) = \sum_{i=1}^{n} \ell n\, f(x_i, \theta).$$

Une **condition nécessaire** est alors :

$$\frac{\partial}{\partial \theta} \ell n\, L(x_1, ..., x_n ; \theta) = 0$$

Si $\hat{\theta}$ satisfait cette condition, une **condition suffisante** pour que $\hat{\theta}$ soit un maximum s'écrit :

$$\frac{\partial^2}{\partial \theta^2} \ell n\, L(x_1, ..., x_n ; \theta) \le 0 \text{ au point critique } \theta = \hat{\theta}.$$

L'estimateur du maximum de vraisemblance est alors la variable aléatoire :

$$W_n = \hat{\Theta}(X_1, ..., X_n).$$

Propriétés des estimateurs obtenus par la méthode du M.V.

(1) L'estimateur W_n obtenu par la méthode du maximum de vraisemblance est asymptotiquement sans biais et convergent en probabilité.

(2) Lorsque n est grand, la loi de la v.a. $W_n - \theta$, est voisine de la loi normale $\mathcal{N}(0, \sigma_n)$ où l'on a posé :

$$\sigma_n = \frac{1}{\sqrt{I_n(\theta)}} \quad \text{avec} \quad I_n(\theta) = n\, E\left(\left[\frac{\partial}{\partial \theta} \ell n\, f(x,\theta) \right]^2 \right)$$

Exemple : *estimateurs pour la moyenne et la variance d'une loi normale*

Soit X une v.a. de loi normale $\mathcal{N}(m, \sigma)$ où m et σ^2 sont inconnus. Le paramètre θ est alors un vecteur à deux coordonnées : $\theta = (m,v)$ et la densité de X s'écrit :

$$f(x, \theta) = f(x; m, v) = \frac{1}{\sqrt{2\pi v}} \exp\left(- \frac{(x-m)^2}{2v} \right).$$

La vraisemblance de l'échantillon s'écrit alors :

$$L(x_1, ..., x_n ; \theta) = \frac{1}{(2\pi v)^{n/2}} \exp\left(-\frac{1}{2v} \sum_{i=1}^{n} (x_i - m)^2 \right)$$

En prenant le logarithme des deux membres, il vient

$$\ell n\, L(x_1, ..., x_n ; \theta) = -\frac{n}{2} \ell n\, v - \frac{1}{2v} \sum_{i=1}^{n} (x_i - m)^2 + \frac{n}{2} \ell n\, (2\pi)$$

Ici, les conditions nécessaires consistent en l'annulation des dérivées partielles premières par rapport à m et à v. Elles s'écrivent respectivement :

$$\frac{\partial \ell n\, L}{\partial m} = \frac{1}{v} \sum_{i=1}^{n} (x_i - m) = 0$$

$$\frac{\partial \ell n\, L}{\partial v} = -\frac{n}{2v} + \frac{1}{2v^2} \sum_{i=1}^{n} (x_i - m)^2 = 0.$$

On tire de ces deux relations :

$$\hat{m} = \bar{x} \quad \text{et} \quad \hat{v} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Le lecteur vérifiera que ceci correspond bien à un maximum de la fonction de vraisemblance (la démonstration de la condition suffisante sera faite en TD).

Les estimateurs cherchés pour m et v sont donc la moyenne et la variance empiriques :

$$W_1 = \bar{X}_n \quad \text{et} \quad W_2 = S^2.$$

Conclusion

Dans ce paragraphe consacré à l'estimation ponctuelle, nous avons vu comment construire des estimateurs qui étant donné un échantillon d'observations fournissent une estimation, c'est-à-dire une valeur numérique approchée, du paramètre inconnu.

Il faut désormais s'interroger sur l'écart entre la vraie valeur du paramètre et son estimation, autrement dit, essayer de chiffer l'erreur que l'on risque de commettre en considérant que l'estimation obtenue est la "vraie valeur" du paramètre à estimer.

La connaissance de la loi de probabilité de l'estimateur, qui renseigne, à la fois sur les valeurs possibles de l'estimateur et leur probabilité, permettra d'évaluer simultanément la précision et la crédibilité de l'estimation.

III. ESTIMATION PAR INTERVALLE DE CONFIANCE

Le problème de l'estimation par intervalle de confiance est le même que celui de l'estimation ponctuelle, c'est la nature de la solution recherchée qui est différente : on souhaite construire un intervalle dont les bornes sont aléatoires et ayant une forte

CamScanner

probabilité de contenir la valeur du paramètre inconnu. Autrement dit à partir d'échantillon $(X_1, ..., X_n)$ de la v.a. X et s'étant donné un nombre α, compris entre 0 et 1, doit déterminer deux statistiques de $A(X_1, ..., X_n)$ et $B(X_1, ..., X_n)$ vérifiant :

(1) $A(X_1, ..., X_n) \leq B(X_1, ..., X_n)$ et

(2) $P(\theta \notin [A(X_1, ..., X_n), B(X_1, ..., X_n)]) = \alpha$ ou, ce qui revient au même

(2bis) $P(A(X_1, ..., X_n) \leq \theta \leq B(X_1, ..., X_n)) = 1 - \alpha$.

Définitions :
L'intervalle $[A(X_1, ..., X_n), B(X_1, ..., X_n)]$ est appelé **intervalle de confiance au seuil de risque α** pour le paramètre θ ou intervalle de confiance de niveau $1-\alpha$ pour le paramètre θ.

Si l'on dispose d'un échantillon d'observations $(x_1, x_2, ..., x_n)$, réalisation de $(X_1, ..., X_n)$, la statistique $A(X_1, ..., X_n)$ prendra la valeur $a = A(x_1, ..., x_n)$, et la statistique $B(X_1, ..., X_n)$ prendra la valeur $b = B(x_1, ..., x_n)$ alors l'intervalle $[a, b]$, qui est une réalisation de l'intervalle de confiance sera appelé **fourchette d'estimation au risque α** pour le paramètre θ.

Le nombre α est appelé **seuil de risque** et la valeur $1-\alpha$ est le **niveau de confiance.**

L'intérêt de cette approche tient au fait que l'on peut, évaluer à la fois la précision et la crédibilité de l'estimation $\hat{\theta}$ de θ.

La **précision** peut être mesurée par l'erreur absolue qui, si θ est dans l'intervalle $[a, b]$, vaut : $e = \max(b-\theta, \theta-a)$ et est majorée par $b-a$. $\theta \in [a,b]$; $e = \max(b-\theta, \theta-a)$
La précision peut aussi être mesurée par l'erreur relative qui, si θ est positif, vaut e/θ et est majorée par $(b-a)/a$, lorsque a est positif.

La crédibilité, est appréciée par le nombre $1-\alpha$, probabilité pour que θ appartienne à l'intervalle de confiance. En effet, du fait du tirage aléatoire de l'échantillon, il n'est jamais certain que la vraie valeur de θ soit contenue dans cet intervalle.

Remarques :

a) On peut noter que n et α étant donnés, il existe une infinité d'intervalles de confiance possibles. En effet, la relation :

(2) $P(\theta \notin [A(X_1, ..., X_n), B(X_1, ..., X_n)]) = \alpha$

peut se décomposer, d'une infinité de façons : si α_1 et α_2 sont deux nombres positifs tels que $\alpha_1 + \alpha_2 = \alpha$ on voit que les deux relations

(3) $P(\theta < A(X_1, ..., X_n) = \alpha_1$ et

(4) $P(\theta > B(X_1, ..., X_n) = \alpha_2$. impliquent la relation (2)

Souvent, on répartit le risque de façon symétrique c'est à dire avec $\alpha_1 = \alpha_2 = \alpha/2$ mais cela n'est, pas obligatoire, en outre celà ne constitue pas toujours la meilleure solution face à un problème donné.

b) Si la vraie valeur du paramètre θ appartient à l'intervalle $[a,b]$ et si l'on choisit, comme valeur approchée de θ, le point milieu $(a+b)/2$ de cet intervalle, alors l'erreur absolue est majorée par $(b-a)/2$ et l'erreur relative par $(a+b)/2a$ (lorsque a est positif).

c) Il faut souligner que l'on a considéré ci-dessus le cas d'un intervalle **bilatéral** c'est-à-dire un intervalle dont les deux bornes sont aléatoires et pour lequel l'erreur peut avoir lieu,

CamScanner

soit à cause de la réalisation de l'évènement de la relation (3), soit à cause de la réalisation de celui de la relation (4).

Dans un autre problème, il pourra se faire que la recherche d'un intervalle unilatéral, autrement dit, de la forme : $]-\infty,\ C(X_1, \ldots, X_n)]$ ou $[C(X_1, \ldots, X_n),\ +\infty[$, s'avère plus pertinente, compte-tenu de la nature du problème.

III.1 INTERVALLE DE CONFIANCE POUR LA MOYENNE D'UNE LOI $\mathcal{N}(m,\sigma)$

On considère une population dans laquelle la v.a. X, associée au caractère quantitatif étudié, suit la loi normale $\mathcal{N}(m, \sigma)$. Soient $(X_1, \ldots, X_n)$ un n-échantillon de X et $\bar{X}$ la moyenne empirique de cet échantillon. On suppose que le paramètre m est inconnu et l'on souhaite construire un intervalle de confiance pour m.

Deux situations peuvent se présenter : ou bien l'écart-type σ est connu, ou bien il ne l'est pas.

III.1.1. L'écart-type σ est connu :

On rappelle que la moyenne empirique $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ suit la loi normale $\mathcal{N}\left(m, \frac{\sigma}{\sqrt{n}}\right)$ et

donc que la v.a. centrée réduite associée : $U = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$ suit la loi normale $\mathcal{N}(0, 1)$.

Par ailleurs, si on désigne par F_0 la fonction de répartition de la loi $\mathcal{N}(0, 1)$, pour tout t positif, on a :

$$P(-t \le U \le t) = F_0(t) - F_0(-t) = 2F_0(t) - 1$$

$P(-t \le U \le t) = 1 - \alpha$

Alors si le seuil de risque choisi est α, et si t vérifie $F_0(t) = 1 - \alpha/2$, en observant que la double inégalité $-t \le U \le t$ est équivalente à :

$$m - \frac{t\sigma}{\sqrt{n}} \le \bar{X} \le m + \frac{t\sigma}{\sqrt{n}} \text{ soit encore à : } \bar{X} - \frac{t\sigma}{\sqrt{n}} \le m \le \bar{X} + \frac{t\sigma}{\sqrt{n}}$$

on a : $$P(-t \le U \le t) = P\left(\bar{X} - \frac{t\sigma}{\sqrt{n}} \le m \le \bar{X} + \frac{t\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Autrement dit, lorsque σ est connu, l'intervalle de confiance de niveau $1 - \alpha$ pour m est donné par :

$$\left[\bar{X} - t\frac{\sigma}{\sqrt{n}} \quad \bar{X} + t\frac{\sigma}{\sqrt{n}}\right]$$

où t vérifie $F_0(t) = 1 - \alpha/2$ (lu dans la table de la loi $\mathcal{N}(0;1)$)

Remarquons que le nombre t est fonction du seuil de risque α, choisi. Le tableau suivant résume, à titre d'exemple, les valeurs de t lues dans la table et correspondant à trois valeurs possibles de α les plus couramment utilisées.

α	0,01	0,05	0,10
$1-\alpha/2$	0,995	0,975	0,95
t	2,575	1,96	1,645

CamScanner

Si maintenant on dispose d'un échantillon d'observations $(x_1, x_2, ..., x_n)$ de la v.a. statistique $\bar{X}$ prend une valeur, notée $\bar{x}$ telle que l'on ait la **fourchettte d'estimation p m au risque α suivante :**

$$\bar{x} - t\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{x} + t\frac{\sigma}{\sqrt{n}}$$

Remarques :

a) On peut noter que plus le seuil de risque α augmente plus t diminue et donc aussi la longueur de l'intervalle de confiance qui est $2t\frac{\sigma}{\sqrt{n}}$.

Par conséquent, **à n fixé**, une amélioration de la précision, c.à.d. une diminution de la longueur de l'intervalle de confiance, ne peut s'obtenir qu'au prix d'une augmentation du risque.

b) Le risque α (risque de voir la vraie valeur de m ne pas appartenir à l'intervalle) a une signification statistique : si l'on répète un assez grand nombre de fois le tirage de l'échantillon, en moyenne, on aura 100(1-α)% de chance que l'estimation de m soit réussie, c'est-à-dire que l'évènement de la relation (2), à savoir : "l'intervalle de confiance recouvre la valeur inconnue de m", soit réalisé. On peut dire aussi que cet évènement à une probabilité 1 - α de se réaliser.

III. 1.2. L'écart-type σ est inconnu

En pratique, il est rare de connaître σ lorsque m est inconnu. Dans ce cas l'intervalle de confiance pour m, donné ci-dessus ne convient plus car il contient l'inconnue supplémentaire σ.

On considère alors la variance empirique corrigée : $S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

S'^2 est un estimateur sans biais et convergent de σ^2. Ceci suggère de remplacer σ par S' de sorte que l'intervalle de confiance pour le paramètre m est :

$$\left[\bar{X} - \frac{t\,S'}{\sqrt{n}} \quad \bar{X} + \frac{t\,S'}{\sqrt{n}}\right]$$

Cependant, il faut noter que le calcul de cet intervalle repose sur la statistique :

$$T = \frac{\bar{X} - m}{S'/\sqrt{n}}$$

lorsque $n \geq 30$ on peut considérer cette v.a. comme une v.a. de loi normale $\mathcal{N}(0, 1)$, lorsque $n < 30$ cette v.a suit une **loi de Student** à $\nu = n - 1$ degrés de liberté.
La loi de Student est tabulée : la table fournit directement, en fonction de n et α la valeur positive t vérifiant : $P(-t \leq T \leq t) = 1 - \alpha$.

Autrement dit, **lorsque σ est inconnu, l'intervalle de confiance de niveau 1 - α pour m est donné par :**

$$\left[\bar{X} - t\frac{S'}{\sqrt{n}} \quad \bar{X} + t\frac{S'}{\sqrt{n}}\right]$$

où t est défini par $P(-t \leq T \leq t) = 1 - \alpha$ et la valeur de t est lue dans la table des fractiles de la loi $\mathcal{N}(0, 1)$ si $n \geq 30$, dans la table de la loi de Student sinon.

CamScanner

Si maintenant on dispose d'un échantillon d'observations $(x_1, x_2, ..., x_n)$ de la v.a. X i.e. une fois connues les valeurs $\bar{x}$ et s' réalisations respectivement des v.a. $\bar{X}$ et S' on calcule une réalisation de cet intervalle de confiance ou **fourchette d'estimation pour le paramètre m au risque α soit :**

$$\bar{x} - t\frac{s'}{\sqrt{n}} \leq m \leq \bar{x} + t\frac{s'}{\sqrt{n}}.$$

Exemple :

Si dans un échantillon de taille n = 30 on a observé $\bar{x} = 55$, $s'^2 = 12$ et si le risque α est choisi égal à 0,10, on obtient alors par lecture dans la table de la loi de Student t =1,699. On en déduit l'intervalle de confiance $\left[\bar{X} - \frac{1,7\,S'}{\sqrt{30}} \; ; \; \bar{X} + \frac{1,7\,S'}{\sqrt{30}}\right]$.

Si on utilise les valeurs observées pour les statistiques $\bar{X}$ et S'^2 on obtient alors la fourchette d'estimation pour m au risque 10% suivante : $53,93 \leq m \leq 56,07$.

Remarque : Pour une même valeur de α la valeur de t obtenue par la loi de Student est plus grande que celle obtenue par la loi normale centrée réduite et diminue quand ν (ou n) augmente.

III. 2. INTERVALLE DE CONFIANCE POUR LA MOYENNE D'UNE LOI QUELCONQUE

Si l'on ne fait plus l'hypothèse de normalité sur X les résultats du III.1. ne sont plus applicables. Néanmoins, si la taille de l'échantillon est assez grande le théorème central limite montre que la v.a. : $U = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$ suit approximativement la loi $\mathcal{N}(0, 1)$, ce qui implique en particulier que pour tout t positif la probabilité $P(-t \leq U \leq t)$ est proche de $F_0(t) - F_0(-t) = 2\,F_0(t) - 1$. On pourra donc considérer que (2) est vérifiée de façon approchée et construire l'intervalle pour m comme précédemment, lorsque σ est connu.

Lorsque σ est inconnu, la variable aléatoire :

$$T = \frac{\bar{X} - m}{S'/\sqrt{n}} = \frac{\bar{X} - m}{S/\sqrt{n-1}}$$

peut être considérée comme normale quand n est assez grand ce qui permet de construire un intervalle de confiance pour m vérifiant de façon approchée la relation (2).

En pratique, si le paramètre σ est connu l'échantillon sera considéré comme grand dès que $n \geq 50$; si σ est inconnu on exigera $n \geq 100$.

III.3. INTERVALLE DE CONFIANCE POUR UNE PROPORTION

On s'intéresse à un caractère qualitatif à deux modalités codées 1 et 0 .
La répartition du caractère dans la population est caractérisée par la proportion p des individus qui possèdent la modalité codée 1. On souhaite estimer p en construisant un intervalle de confiance ayant une forte probabilité de contenir la valeur exacte de p.

A cette fin, on choisit au hasard n individus dans la population et l'on note X_i la valeur prise par le caractère sur le $i^{ème}$ individu de l'échantillon (i=1, ..., n). On pourra admettre que la suite $(X_1, ..., X_n)$ ainsi obtenue peut être considérée comme un n-échantillon de la v.a. X suivant une loi de Bernoulli de paramètre p.

Alors la fréquence empirique : $F_n = \frac{X_1+\ldots+X_n}{n}$ est un estimateur sans biais et convergent de p, et l'on a : $E(F_n) = p$ et $V(F_n) = \frac{p(1-p)}{n} = \frac{pq}{n}$ où l'on a posé $1-p = q$.

Cet estimateur va nous permettre de construire un intervalle de confiance pour p. Nous n'envisageons que le **cas où n est suffisamment grand** et où l'on peut approcher la loi de F_n par une loi normale.

En effet, si la taille de l'échantillon est suffisamment grande, le TCL permet d'approcher la loi de F_n par une loi normale $\mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$ (où $q = 1 - p$).

Calcul d'un intervalle de confiance pour p :

On se restreint au cas où l'échantillon est grand et où l'estimateur F_n suit une loi approximativement normale $\mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$

On prendra alors un intervalle de la forme :

$$\left[F_n - t\sqrt{\frac{pq}{n}},\ F_n + t\sqrt{\frac{pq}{n}}\right]$$

où le nombre positif t est le fractile d'ordre $1 - \alpha/2$ de la loi normale $\mathcal{N}(0,1)$.

En effet, si F_n suit approximativement la loi $\mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$ alors la variable $U = \frac{F_n - p}{\sqrt{pq/n}}$ suit approximativement une loi $\mathcal{N}(0,1)$ et si t est le fractile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0,1)$, autrement dit, si t vérifie $t = F_0^{-1}(1-\alpha/2)$ on a :

$$P(|U| < t) = 2F_0(t) - 1 = 1 - \alpha \quad \text{et donc aussi} \quad P\left(\frac{|F_n - p|}{\sqrt{pq/n}} < t\right) = 1 - \alpha.$$

Soit encore : $P(|F_n - p| < t\sqrt{pq/n}) = P(-t\sqrt{pq/n} < F_n - p < t\sqrt{pq/n}) = 1 - \alpha$.

Par conséquent : $P(F_n - t\sqrt{pq/n} < p < F_n + t\sqrt{pq/n}) = 1 - \alpha$.

On remarque que bien que p ne soit qu'une moyenne théorique particulière, on n'est cependant pas dans la situation étudiée en III.1.1 car l'écart-type de F_n dépend de p, qui est inconnu.

Deux méthodes rapides mais quelque peu approximatives permettent de lever cette difficulté :

(1) La première consiste à **estimer pq par $F_n(1-F_n)$** qui est un estimateur asymptotiquement sans biais et convergent de ce paramètre. En effet on voit que $F_n(1-F_n)$ n'est autre que la variance empirique de l'échantillon $(X_1, \ldots, X_n)$.

L'intervalle de confiance pour p est alors :

$$\left[F_n - t\sqrt{\frac{F_n(1-F_n)}{n}},\ F_n + t\sqrt{\frac{F_n(1-F_n)}{n}}\right]$$

CamScanner

(2) La seconde méthode est celle de l'**intervalle de confiance par excès**. Elle est basée sur le fait que, pour tout p dans [0, 1], on peut majorer la quantité pq = p(1-p) par 1/4. On a donc, pour tout n : $\sqrt{\frac{pq}{n}} \leq \frac{1}{2\sqrt{n}}$ et **l'intervalle de confiance par excès sera :**

$$\left[F_n - \frac{t}{2\sqrt{n}}, F_n + \frac{t}{2\sqrt{n}} \right]$$

En effet, puisque l'événement : $\left[F_n - \frac{t}{\sqrt{pq/n}} \leq p \leq F_n + \frac{t}{\sqrt{pq/n}} \right]$

implique l'événement : $\left[F_n - \frac{t}{2\sqrt{n}} \leq p \leq F_n + \frac{t}{2\sqrt{n}} \right]$

on en déduit que l'intervalle aléatoire trouvé vérifie :

$$P\left(p \in \left[F_n - \frac{t}{2\sqrt{n}} \ , \ F_n + \frac{t}{2\sqrt{n}} \right] \right) \geq 1 - \alpha$$

d'où le nom de cette méthode.

IV. DETERMINATION DE LA TAILLE D'ECHANTILLON POUR OBTENIR UNE PRECISION ET UNE CREDIBILITE FIXEES

Plusieurs exemples de construction d'intervalles de confiance, ont été vus dans les paragraphes précédents. Connaissant la valeur du risque α et la taille n de l'échantillon, les bornes de l'intervalle et donc aussi sa longueur, sont alors entièrement connus. Par voie de conséquense la **précision est entièrement déterminée par la connaissance de α et n.**

En pratique, une question importante se pose souvent : comment choisir la **taille de l'échantillon** pour obtenir une précision et une crédibilité imposées.

La **crédibilité** est mesurée par le niveau de confiance 1- α.
La **précision**, peut être exprimée en valeur absolue ou en valeur relative. Pour fixer les idées, supposons la donnée en valeur relative, c'est-à-dire par un nombre k strictement compris entre 0 et 1. Alors, si θ est le paramètre à estimer, la demi-longueur de l'intervalle de confiance devra être au plus égale à $k|\theta|$. Autrement dit, en reprenant les notations du I, on exigera :

(5) $$B(X_1), \ldots, X_n) - A(X_1, . X_n) \leq 2k|\theta|$$

Ceci suppose que l'on accepte la valeur $\frac{a+b}{2}$ comme vraie valeur du paramètre θ.
En pratique le prélèvement d'un échantillon est d'autant plus coûteux que sa taille est grande, on cherchera donc **le plus petit entier n** pour lequel la condition (5) est vérifiée.

Deux exemples vont illustrer ces considérations.

exemple 1 : Estimation de la moyenne d'une v.a. normale

Soit X un v.a. de loi $\mathcal{N}(m, \sigma)$ Lorsque l'écart-type σ est connu et le paramètre m inconnu, l'intervalle de confiance pour m est de la forme :

$$\left[\bar{X}_n - \frac{t\sigma}{\sqrt{n}} \quad \bar{X}_n + \frac{t\sigma}{\sqrt{n}} \right]$$ où t est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{N}(0, 1)$.