STATISTIQUE : ESTIMATION

Préparation à l'Agrégation Bordeaux 1

Année 2012 - 2013

Jean-Jacques Ruch

Table des Matières

CHAPITRE I

Estimation ponctuelle

En statistique, comme dans la théorie des probabilités le hasard intervient fortement. Mais dans la théorie des probabilités, on suppose la loi connue précisément et on cherche à donner les caractéristiques de la variable qui suit cette loi. L'objectif de la statistique est le contraire : à partir de la connaissance de la variable, que peut-on dire de la loi de cette variable ?

1. Définitions

Soit X une variable aléatoire dont la densité de probabilité $f(x,\theta)$ dépend d'un paramètre θ appartenant à $I \subset \mathbb{R}$. A l'aide d'un échantillon issu de X, il s'agit de déterminer au mieux la vraie valeur θ_0 de θ. On pourra utiliser deux méthodes :

- *estimation ponctuelle* : on calcule une valeur vraisemblable $\hat{\theta}$ de θ_0
- *estimation par intervalle* : on cherche un intervalle dans lequel θ_0 se trouve avec une probabilité élevée.

Définition 1. *Un* n-échantillon *de X est un n-uplet $(X_1, X_2, \ldots, X_n)$ tel que les X_k ont la même loi que X et sont indépendantes.*
Une réalisation de l'échantillon *est alors un n-uplet $(x_1, x_2, \ldots, x_n)$ de valeurs prises par l'échantillon.*

Définition 2. *Une* statistique *de l'échantillon est une variable aléatoire $\varphi(X_1, X_2, \ldots, X_n)$ où φ est une application de $\mathbb{R}^n$ dans $\mathbb{R}$.*
Un estimateur *T de θ est une statistique à valeurs dans I. Une* estimation *est la valeur de l'estimateur correspondant à une réalisation de l'échantillon.*

Exemple : $\overline{X_n} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$ est un estimateur de l'espérance mathématique.

Définition 3. *Le* biais *de l'estimateur T de θ est $\mathbb{E}[T]-\theta_0$. S'il est nul, on dit que T est un estimateur sans biais.*
L'estimateur T_n est asymptotiquement sans biais *si* $\lim \mathbb{E}[T_n] = \theta_0$.

On note souvent le biais $b_\theta(T)$.

Définition 4. *L'estimateur est dit* convergent *si la suite (T_n) converge en probabilité vers θ_0 :*

$$\forall \varepsilon > 0, \mathbb{P}(|T_n - \theta_0| > \varepsilon) \underset{n\to+\infty}{\longrightarrow} 0.$$

On parle d'estimateur fortement convergent *lorsqu'on a convergence presque sûre.*

D'après Bienaymé-Tchebychev pour qu'un estimateur asymptotiquement sans biais soit convergent il suffit que

$$\mathrm{Var}(T_n) \underset{n\to+\infty}{\longrightarrow} 0.$$

2. Critères de comparaison d'estimateurs

Un bon critère de comparaison est le *risque quadratique*.

Définition 5. *Soient T un estimateur de θ. Le risque quadratique est défini par*

$$R(T,\theta) = \mathbb{E}[(T-\theta)^2]$$

On peut alors comparer deux estimateurs.

Définition 6. *On dit que T_1 est un* meilleur estimateur *que T_2 si*

$$\forall \theta \in I, \quad R(T_1,\theta) \leq R(T_2,\theta)$$

et

$$\exists \theta \in I, \quad R(T_1,\theta) < R(T_2,\theta).$$

Un estimateur est dit *admissible* s'il n'existe pas d'estimateur meilleur.

L'erreur quadratique moyenne de T se décompose en deux termes, le carré du biais et la variance de T :

$$\mathbb{E}[(T-\theta)^2] = b_\theta^2(T) + \mathrm{Var}(T).$$

Cette décomposition permet de se ramener à une discussion sur la variance pour les estimateurs sans biais de θ.

Définition 7. *Soient T_1 et T_2 deux estimateurs sans biais de θ. On dit que T_1 est un* plus efficace *que T_2 si*

$$\forall \theta \in I, \quad \mathrm{Var}(T_1) \leq \mathrm{Var}(T_2)$$

et

$$\exists \theta \in I, \quad \mathrm{Var}(T_1) < \mathrm{Var}(T_2).$$

On parle d'estimateur à variance minimale si seul le premier point est vérifié, c'est-à-dire :

$$\mathrm{Var}(T_1) \leq \mathrm{Var}(T_2).$$

3. Exemples fondamentaux

Soit X une variable aléatoire telle que $\mathbb{E}[X] = m$ et $\mathrm{Var}(X) = \sigma^2$.

3.a. Estimation de m.

Théorème 8.

La moyenne empirique $\overline{X_n} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$ est un estimateur sans biais et convergent de m.

On a

$$\mathbb{E}[\overline{X_n}] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[X_k] = m \quad \text{et} \quad \mathrm{Var}(\overline{X_n}) = \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}[X_k] = \frac{\sigma^2}{n} \underset{n\to+\infty}{\longrightarrow} 0.$$

D'après la loi forte des grands nombres $\overline{X_n}$ est même fortemement convergent.

Il est possible de déterminer la loi asymptotique de la moyenne empirique.

Proposition 9. *Si n est assez grand on peut utiliser l'approximation normale (lorsque X admet un moment d'ordre 2)*

$$\overline{X_n} \overset{\mathcal{L}}{\sim} \mathcal{N}(m, \sigma^2/n).$$

C'est une conséquence du TCL qui nous assure que

$$\sqrt{n}(\overline{X_n} - m) \underset{n\to+\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0, \sigma^2).$$

3.b. Estimation de σ^2 en supposant m connu.

Théorème 10.

Lorsque m est connu

$$S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - m)^2$$

est un estimateur sans biais et convergent de σ^2.

On a

$$\mathbb{E}[S_n^2] = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}(X_k - m)^2\right] = \frac{1}{n}\sum_{k=1}^{n}\mathrm{Var}(X_k) = \sigma^2$$

Par ailleurs, les variables $(X_k - m)^2$ étant indépendantes :

$$\mathrm{Var}(S_n^2) = \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}((X_k - m)^2) = \frac{1}{n}\left(\mathbb{E}[(X-m)^4] - \mathbb{E}[(X-m)^2]^2\right) = \frac{1}{n}\left(\mu_4 - \sigma^4\right)$$

avec $\mu_k = \mathbb{E}((X-m)^k)$.

Donc S_n^2 est un estimateur convergent. La loi forte des grands nombres appliquée aux variables $(X_k - m)^2$ entraîne même la convergence presque sûre vers σ^2.

Comme dans le cas de la moyenne empirique le TCL nous permet de déterminer la loi asymptotique de S_n^2 ; on a lorsque n est assez grand :

$$S_n^2 \overset{\mathcal{L}}{\sim} \mathcal{N}(\sigma^2, (\mu_4 - \sigma^4)/n).$$

3.c. Estimation de σ^2 lorsque m est inconnu.

En général on ne connaît pas m ; on le remplace par un estimateur et on introduit la variance empirique associée :

$$\overline{S_n^2} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X_n})^2.$$

Théorème 11.

La variance empirique $\overline{S_n^2}$ est un estimateur biaisé et convergent de σ^2. Il est asymptotiquement sans biais.

On a

$$\mathbb{E}[\overline{S_n^2}] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(X_k^2) - \mathbb{E}[\overline{X_n}^2] = \frac{1}{n}(n(m^2+\sigma^2)) - (m^2 + \frac{\sigma^2}{n}) = \frac{n-1}{n}\sigma^2.$$

D'autre part, on peut montrer que :

$$\mathrm{Var}(\overline{S_n^2}) = \frac{1}{n}\left(\mu_4 - \sigma^4\right) - \frac{2}{n^2}\left(\mu_4 - 2\sigma^4\right) + \frac{1}{n^3}\left(\mu_4 - 3\sigma^4\right) \to 0$$

avec $\mu_k = \mathbb{E}((X-m)^k)$. L'estimateur est donc convergent.

Le résultat précédent et le lemme de Slutsky (*Probabilité 2, Jean-Yves Ouvrard, p. 347*) permet de déterminer la loi asymptotique de $\overline{S_n^2}$:

$$\overline{S_n^2} \overset{\mathcal{L}}{\sim} \mathcal{N}(\sigma^2, (\mu_4 - \sigma^4)/n).$$

Théorème 12.

La variance empirique corrigée

$$\widehat{S_n^2} = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X_n})^2.$$

est un estimateur sans biais et convergent de σ^2.

Cela se montre facilement en remarquant que

$$\widehat{S_n^2} = \frac{n}{n-1}\overline{S_n^2}.$$

4. Cas particulier de la loi normale

On suppose dans ce paragraphe que X suit la loi normale $\mathcal{N}(m, \sigma^2)$. On sait que $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ suit alors la loi normale $\mathcal{N}(m, \sigma^2/n)$, ce qui confime que c'est un estimateur sans biais, convergent de m.
Les résultats obtenus au paragraphe précédent pour l'estimation de σ^2 sont encore valables ; en particulier on a :

$$\mathbb{E}(S_n^2) = \sigma^2 \quad \text{et } \mathrm{Var}(S_n^2) = \frac{2(n-1)}{n^2}\sigma^4$$

En effet, calculons μ_k

$$\begin{aligned}
\mu_k &= \mathbb{E}((X-m)^k) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}(x-m)^k \exp\left(-\frac{(x-m)^2}{2\sigma}\right)dx \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}(\sqrt{2}\sigma u)^k \exp\left(-u^2\right)\sqrt{2}\sigma du \quad \text{en posant } x = m\sqrt{2}\sigma u \\
&= 0 \quad \text{si } k \text{ est impair.}
\end{aligned}$$

Lorsque $k = 2p$ est pair on obtient

$$\begin{aligned}
\mu_{2p} &= \frac{2^p\sigma^{2p}}{\sqrt{\pi}}\int_{-\infty}^{+\infty} u^{2p}\exp\left(-u^2\right)du = \frac{2^{p+1}\sigma^{2p}}{\sqrt{\pi}}\int_{0}^{+\infty} u^{2p}\exp\left(-u^2\right)du \\
&= \frac{2^p\sigma^{2p}}{\sqrt{\pi}}\int_{0}^{+\infty} v^{p-1/2}\exp\left(-v\right)dv \quad \text{en posant } u = \sqrt{v} \\
&= \frac{2^p\sigma^{2p}}{\sqrt{\pi}}\Gamma(p+1/2) = \frac{(2p)!}{2^p(p!)}\sigma^{2p}
\end{aligned}$$

et donc

$$\mathrm{Var}(S_n^2) = \frac{1}{n}\left(\mu_4 - \sigma^4\right) - \frac{2}{n^2}\left(\mu_4 - 2\sigma^4\right) + \frac{1}{n^3}\left(\mu_4 - 3\sigma^4\right) = \frac{2(n-1)}{n^2}\sigma^4$$

Définition 13. *Soient $X_1, \dots, X_n$, n variables aléatoires indépendantes identiquement distribuées de loi $\mathcal{N}(0,1)$. La loi du χ^2 à n degrés de liberté est la loi de la variable aléatoire*

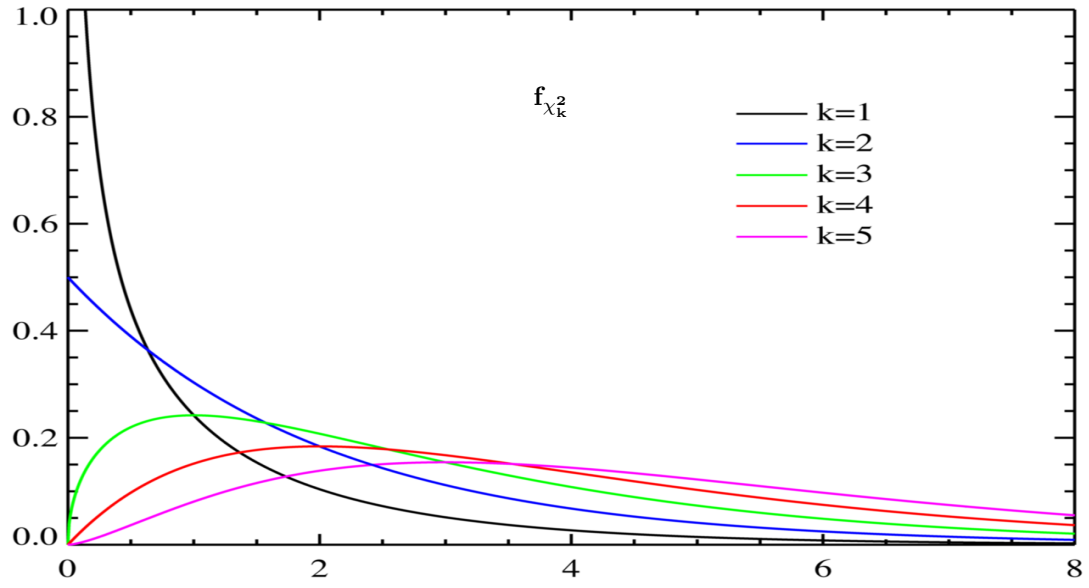
$$\chi^2_n = \sum_{k=1}^{n} X_k^2.$$

La densité de cette loi est donnée par :

$$f_{\chi^2_n}(u) = \frac{1}{2\Gamma(n/2)} \left(\frac{u}{2}\right)^{n/2-1} \exp\left(-\frac{u}{2}\right) 1_{u>0}$$

et sa fonction caractéristique par

$$\phi_{\chi^2_n}(t) = \frac{1}{(1-2it)^{n/2}}$$



Pour déterminer la densité on peut remarquer que : si U suit une loi $\mathcal{N}(0,1)$ alors on a pour $t > 0$

$$\mathcal{P}(U^2 \leq t) = \mathcal{P}(-t \leq U \leq t) = F_U(\sqrt{t}) - F_U(-\sqrt{t})$$

et par conséquent

$$f_{U^2}(t) = \frac{1}{2\sqrt{t}} f_U(\sqrt{t}) + \frac{1}{2\sqrt{t}} f_U(-\sqrt{t}) = \frac{1}{\sqrt{t}} f_U(\sqrt{t}) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{t}{2}\right)$$

Ensuite on obtient le résultat général par récurence.

Théorème 14.

Soit $(X_1, \dots, X_n)$ un $n-$échantillon de le loi $\mathcal{N}(0,1)$. Les variables aléatoires

$$\sqrt{n}\,\overline{X}_n \quad \text{et} \quad \sum_{k=1}^{n} (X_k - \overline{X}_n)^2 = n\overline{S_n^2} = (n-1)\widehat{S_n^2}$$

sont indépendantes et suivent respectivement la loi normale réduite et la loi du χ^2 à $(n-1)$ degrés de liberté.

DÉMONSTRATION. Montrons que $\overline{X}_n$ et $\sum_{k=1}^{n}(X_k-\overline{X}_n)^2$ sont indépendantes. On a

$$\begin{pmatrix} \overline{X}_n \\ X_1-\overline{X}_n \\ \vdots \\ X_n-\overline{X}_n \end{pmatrix} = A \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \quad \text{où} \quad A = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{n} \\ -\frac{1}{n} & \cdots & -\frac{1}{n} & \frac{n-1}{n} \end{pmatrix}$$

Le vecteur aléatoire $\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ est gaussien de loi $\mathcal{N}(0, I_n)$ où I_n est la matrice identité d'ordre n.

Par conséquent, le vecteur $\begin{pmatrix} \overline{X}_n \\ X_1-\overline{X}_n \\ \vdots \\ X_n-\overline{X}_n \end{pmatrix} = A \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ est également gaussien de loi $\mathcal{N}(0, AI_nA^t) =$ $\mathcal{N}(0, AA^t)$. Or

$$AA^t = \begin{pmatrix} \frac{1}{n} & 0 & 0 & \cdots & 0 \\ 0 & \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ 0 & -\frac{1}{n} & \frac{n-1}{n} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -\frac{1}{n} \\ 0 & -\frac{1}{n} & \cdots & -\frac{1}{n} & \frac{n-1}{n} \end{pmatrix}$$

donc la variable $\overline{X}_n$ est indépendante du vecteur $\begin{pmatrix} X_1-\overline{X}_n \\ \vdots \\ X_n-\overline{X}_n \end{pmatrix}$ et donc de $\sum_{k=1}^{n}(X_k-\overline{X}_n)^2$.

Comme $\overline{X}_n$ suit la loi $\mathcal{N}(0, 1/n)$ on en déduit que $\sqrt{n}\overline{X}_n$ suit la loi $\mathcal{N}(0,1)$.

Montrons $n\overline{S_n^2} = \sum_{k=1}^{n}(X_k-\overline{X}_n)^2$ suit la loi du χ^2 à $(n-1)$ degrés de liberté. On a

$$\begin{pmatrix} X_1-\overline{X}_n \\ \vdots \\ X_n-\overline{X}_n \end{pmatrix} = B \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \quad \text{où} \quad B = \begin{pmatrix} \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{n} \\ -\frac{1}{n} & \cdots & -\frac{1}{n} & \frac{n-1}{n} \end{pmatrix}$$

Comme B est une matrice symétrique, il existe une matrice orthogonale U et une matrice diagonale D telle que

$$B = UDU^t$$

Or les valeurs propres de B sont :
- la valeur propre simple 0 dont le sous-espace propre associé a pour équation $x_1 = \cdots = x_n$;
- la valeur propre d'ordre $(n-1)$ égale à 1 dont le sous-espace propre associé a pour équation $x_1 + x_2 + \cdots + x_n = 0$

En ordonnant convenablement la base de vecteurs propres on peut choisir

$$D = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix}$$

On a $Y = BX = UDU^tX$ et

$$\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = Y^tY = X^tUDU^tUDU^tX = (U^tX)^tD(U^tX)$$

Or le vecteur aléatoire $Z = U^tX$ est gaussien de loi $\mathcal{N}(0, U^tI_nU) = \mathcal{N}(0, U^tU) = \mathcal{N}(0, I_n)$. D'où

$$\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = Z^tDZ = \sum_{k=1}^{n-1} Z_i^2$$

qui suit la loi du χ^2 à $(n-1)$ degrés de liberté. □

On en déduit immédiatement que si $(X_1, \ldots, X_n)$ est un échantillon d'une variable aléatoire $\mathcal{N}(m, \sigma^2)$ la variable aléatoire

$$\frac{1}{\sigma^2}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2$$

suit la loi du χ^2 à $(n-1)$ degrés de liberté.
Il suffit de poser $Y_k = X_k - m/\sigma$. Alors, comme on a

$$\overline{Y}_n = \frac{\overline{X}_n - m}{\sigma} \quad \text{et} \quad \frac{1}{\sigma^2}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = \sum_{k=1}^{n}(Y_k - \overline{Y}_n)^2$$

le résultat découle de ce qui précède.

5. Construction d'estimateur par la méthode du maximum de vraisemblance

5.a. Cas discret. On suppose donnée une observation X tirée selon une loi $\mathbb{P}_\theta$, $\theta \in \Theta$. On supposera ici que $\mathbb{P}_\theta$ est discrète et on pose :

$$f_\theta(x) = \mathbb{P}_\theta(X = x).$$

On appelle alors *fonction de vraisemblance* la fonction $L_X(\theta) = f_\theta(X)$. Quand on dispose d'un $n-$échantillon $(X_1, \ldots, X_n)$ de loi $\mathbb{P}_\theta$, la vraisemblance s'écrit alors

$$L_{X_1,\ldots,X_n}(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

Lorsque la fonction de vraisemblance admet un unique maximum atteint en $\hat{\theta} = g_n(X_1, \ldots, X_n)$, on peut utiliser cette valeur pour estimer θ. On dit alors que

$$T = g_n(X_1, \ldots, X_n)$$

est l'*estimateur par maximum de vraisemblance* de θ.

Cet estimateur est naturel puisqu'il conduit à privilégier la valeur de θ la "plus probable" au vu de l'observation. Il possède en général de bonnes propriétés. L'inconvénient est que ce maximum peut ne pas exister ou ne pas être unique et il peut être difficile à exhiber.

En pratique, la recherche de ce maximum se fait par dérivation de L relativement à θ. On peut de manière équivalente maximiser le logarithme de la vraisemblance (la fonction logarithme étant croissante, maximiser la vraisemblance et la log-vraisemblance revient au même, mais souvent les calculs sont plus simples).

Exemple : Estimation du paramètre d'une loi de Bernoulli.

Ici on suppose $\Theta =]0, 1[$ et les X_i suivent une loi de Bernoulli de paramètre $\theta \in \Theta$. On a

$$f_\theta(x) = \mathbb{P}_\theta(X = x) = \begin{cases} \theta & \text{si } x = 1 \\ 1-\theta & \text{si } x = 0 \\ 0 & \text{sinon} \end{cases}$$

Posons $S_n = X_1 + \cdots + X_n$. Ainsi S_n est le nombre de 1 dans l'échantillon et $n - S_n$ le nombre de 0. La vraisemblance et la log-vraisemblance s'écrivent alors :

$$L_{X_1,\ldots,X_n}(\theta) = \theta^{S_n}(1-\theta)^{n-S_n} \qquad lnL_{X_1,\ldots,X_n}(\theta) = S_n ln\theta + (n - S_n)ln(1-\theta).$$

Un calcul montre alors que le maximum est atteint en

$$\hat{\theta} = n^{-1}S_n.$$

Par conséquent l'estimateur de θ par maximum de vraisemblance est

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

qui est également l'estimateur de la moyenne.

5.b. Cas à densité. On suppose donnée une observation X tirée selon une loi $\mathbb{P}_\theta$, $\theta \in \Theta$. On supposera ici que $\mathbb{P}_\theta$ admet une densité par rapport à la mesure de Lebesgue notée f_θ.

On appelle alors *fonction de vraisemblance* la fonction $L_X(\theta) = f_\theta(X)$. Quand on dispose d'un $n-$échantillon $(X_1, \ldots, X_n)$ on a la vraisemblance

$$L_{X_1,\ldots,X_n}(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

Ensuite, on procède comme dans le cas discret.

Exemple : On cherche à estimer le paramètre θ inconnu d'une loi exponentielle. La vraisemblance s'écrit :

$$L_{X_1,\ldots,X_n}(\theta) = \exp\left(-\sum_{i=1}^{n} X_i/\theta\right)\theta^{-\sum_{i=1}^{n} X_i}$$

Le maximum est atteint en un unique point $\hat{\theta}_n = \overline{X}_n$.

CHAPITRE II

Estimation par intervalle

1. Définition d'une région de confiance

Soit $\alpha \in]0,1[$ un *niveau de risque* fixé par le statisticien.

Définition 1. *Une* région de confiance *de θ de* niveau de confiance $1-\alpha$ *est un ensemble (dépendant de l'observation mais pas du paramètre inconnu θ), $C(X) \subset \Theta$, telle que*

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta\left(\theta \in C(X)\right) \geq 1-\alpha$$

On dit alors qu'on a une région par excès. Dans le cas où on a égalité on parle de niveau exactement égal à $1-\alpha$.

Lorsqu'on a $X = (X_1, \ldots, X_n)$, on parle de *région de confiance asymptotique* de niveau $1-\alpha$, si

$$\forall \theta \in \Theta, \quad \lim_{n \to +\infty} \mathbb{P}_\theta\left(\theta \in C(X_1, \ldots, X_n)\right) \geq 1-\alpha$$

Les valeurs usuelles de α sont 1%, 5% ou 10%. Dans le cas unidimensionnel, la plupart du temps, une région de confiance s'écrit sous la forme d'un intervalle (unilatère ou bilatère). Un intervalle de confiance de niveau de confiance 95% a une probabilité au moins égale à $0,95$ de contenir la vraie valeur inconnue θ. Par passage au complémentaire, le niveau de risque α correspondant à une majoration de la probabilité que la vraie valeur du paramètre θ ne soit pas dans $C(X)$. A niveau de confiance fixé, une région de confiance est d'autant meilleure qu'elle est de taille petite. Avant d'aller plus loin, rappelons la notion de quantile d'une loi de probabilité.

Définition 2. *Soit $\alpha \in]0,1[$. On appelle* quantile *d'ordre α d'une loi de probabilité $\mathbb{P}$, la quantité*

$$z_\alpha = \inf\left\{x,\ \mathbb{P}(]-\infty, x])\geq \alpha\right\}.$$

Par exemple pour la loi $\mathcal{N}(0,1)$, le quantile d'ordre $97,5\%$ est 1.96, et celui d'ordre 95% est 1.645.

2. Construction de régions de confiance

Une première méthode consiste à appliquer l'inégalité de Bienaymé-Tchebychev. Rappelons que si X est une variable aléatoire ayant un moment d'ordre 2, alors

$$\forall \varepsilon > 0, \mathbb{P}(|X - \mathbb{E}(X)| > \varepsilon) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}$$

Appliquons cette inégalité dans le cas de variables aléatoires idépendantes $X_1, \ldots, X_n$ identiquement distribuées de loi de Bernoulli $\mathcal{B}(\theta)$, où l'on souhaite estimer θ à l'aide de $\overline{X}_n$. On a

$$\forall \varepsilon > 0, \mathbb{P}(|\overline{X}_n - \theta| > \varepsilon) \leq \frac{\theta(1-\theta)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

On obtient ainsi une région de confiance de niveau $1-\alpha$ en considérant

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

Pour $\alpha = 5\%$ et $n = 100$, la précision de l'intervalle est 0.22. Il faut noter que la majoration obtenue par l'application de l'inégalité de Bienaymé-Tchebychev n'est pas très précise .
On obtient un meilleur résultat en utilisant l'inégalité de Hoeffding (*Ouvrad, Probabilité 2, page 132*).

Proposition 3. *Soit $(X_1, \dots, X_n)$ une suite de variables aléatoires indépendantes telles que pour tout i, $a_i \leq X_i \leq b_i$ p.s., alors pour tout $\varepsilon > 0$, en posant $S_n = \sum_{i=1}^n X_i$,*

$$\mathbb{P}(S_n - \mathbb{E}(S_n) \geq \varepsilon) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \qquad \mathbb{P}(S_n - \mathbb{E}(S_n) \leq -\varepsilon) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

et

$$\mathbb{P}(|S_n - \mathbb{E}(S_n)| \geq \varepsilon) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Appliquons cette inégalité à l'exemple précédent. On a :

$$\forall \varepsilon > 0,\ \mathbb{P}(|\overline{X}_n - \theta| \geq \varepsilon) \leq \mathbb{P}(|S_n - n\theta| \geq n\varepsilon) \leq 2\exp(-2n\varepsilon^2).$$

On obtient ainsi l'intervalle de confiance de niveau $1 - \alpha$ suivant :

$$\left[\overline{X}_n - \sqrt{\frac{1}{2n}\ln(2/\alpha)}, \overline{X}_n + \sqrt{\frac{1}{2n}\ln(2/\alpha)}\right].$$

Pour $\alpha = 5\%$ et $n = 100$, la précision de l'intervalle est 0.14 et 0.23 avec la première méthode.

Il peut s'avérer plus pratique de chercher un intervalle de confiance asymptotique.

Supposons que nous cherchions un intervalle de confiance pour un paramètre θ à partir d'un échantillon de taille n de loi $\mathbb{P}_\theta$. Lorsqu'on dispose de suffisamment de données et pour les modèles les plus classiques, le théorème central limite s'avère être un très bon outil, pour obtenir un intervalle de confiance asymptotique. Par exemple si on souhaite estimer la moyenne d'une variable aléatoire dont on connait la variance $\sigma^2 = 1$. On prend un n-échantillon $(X_1, \dots, X_n)$. L'application du TCL donne :

$$\sqrt{n}(\overline{X}_n - m) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

On obtient alors l'intervalle de confiance asymptotique de niveau α suivant

$$\left[\overline{X}_n - \frac{q_{1-\alpha/2}}{\sqrt{n}}, \overline{X}_n + \frac{q_{1-\alpha/2}}{\sqrt{n}}\right]$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$.
Ce n'est pas toujours aussi évident. Si on part d'une variable aléatoire de Bernoulli dont on veut estimer le paramètre θ. En considérant l'estimateur du maximum de vraisemblance $\overline{X}_n$, le TCL donne :

$$\sqrt{n}(\overline{X}_n - \theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \theta(1-\theta))$$

Ici la loi limite dépend de θ ce qui est gênant pour construire un intervalle de confiance. Dans ce cas, on peut surmonter ce problème, en remarquant que $\theta(1 - \theta) \leq 0.25$. On obtient donc un intervalle de confiance asymptotique :

$$\left[\overline{X}_n - \frac{q_{1-\alpha/2}}{2\sqrt{n}}, \overline{X}_n + \frac{q_{1-\alpha/2}}{2\sqrt{n}}\right].$$

Dans le cas où on considère $(X_1, \dots, X_n)$ un échantillon de loi de Poisson de paramètre $\theta > 0$ à estimer, le TLC donne :

$$\sqrt{n}(\overline{X}_n - \theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \theta)$$

Des outils plus élaborés doivent être utilisés pour construire un intervalle de confiance si on ne connaît pas de majorant de θ. Le *lemme de Slutsky* permet de surmonter certaines difficultés comme le montre l'exemple suivant.

Si on reprend l'exemple, en utilisant les propriétés de forte consistance d'estimateur, on obtient :

$$\frac{\sqrt{n}(\overline{X}_n - \theta)}{\sqrt{\overline{X}_n}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

mais aussi avec

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

qui est un estimateur fortement consistant de la variance

$$\frac{\sqrt{n}(\overline{X}_n - \theta)}{\sqrt{\overline{S_n^2}}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

3. Exemples classiques d'estimation par intervalle

Soit X une variable aléatoire de loi normale $\mathcal{N}(m, \sigma^2)$.

3.a. Estimation de la moyenne quand la variance est connue.

Théorème 4.

Lorsque σ^2 est connu un intervalle de confiance au niveau $1-\alpha$ de m est

$$\left[\overline{X}_n - q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ ($F(q_{1-\alpha/2}) = 1-\alpha/2$) de la loi normale centrée réduite.

Démonstration. On sait que $(\overline{X}_n - m)/\sigma$ suit la loi $\mathcal{N}(0,1)$. Par conséquent on a

$$\frac{|\overline{X}_n - m|}{\sigma} \in \left[-q_{1-\alpha/2}, q_{1-\alpha/2}\right] \iff m \in \left[\overline{X}_n - q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

□

Exemple : puisque $q_{0,975} = 1.96$ l'intervalle de confiance de m au niveau 95% est

$$\left[\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

ce qui signifie que sur un grand nombre d'expériences cet intervalle contiendra effectivement m dans 95% des cas en moyenne.

3.b. Estimation de la moyenne quand la variance est inconnue.

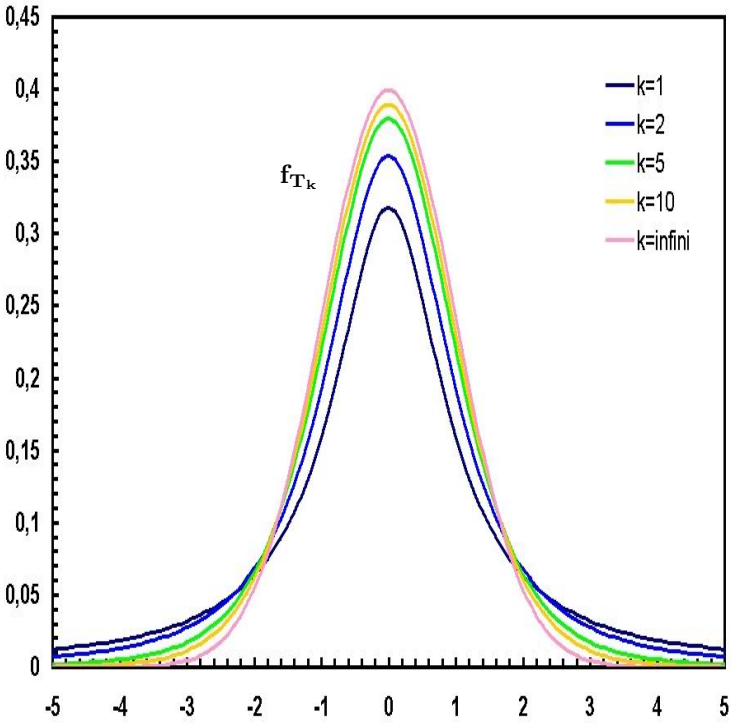
Définition 5. *Soient X et Y deux variables aléatoires indépendantes suivant respectivement la loi normale centrée réduite et la loi du χ^2 à n degrés de liberté. La variable aléatoire*

$$T = \frac{X}{\sqrt{Y/n}}$$

suit la loi de Student à n degrés de liberté. La densité de cette loi est donnée par :

$$f_T(u) = \frac{1}{\sqrt{n\pi}}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}\frac{1}{(1+u^2/n)^{(n+1)/2}}$$

Cette variable n'a pas d'espérance pour $n = 1$ et pas de variance pour $n \leq 2$. Sinon on a $\mathbb{E}(T) = 0$ et $\mathrm{Var}(T) = n/(n-2)$.



Théorème 6.

Lorsque σ^2 est inconnu un intervalle de confiance au niveau $1-\alpha$ de m est

$$\left[\overline{X}_n - t_{n-1,1-\alpha/2}\frac{\sqrt{\widehat{S_n^2}}}{\sqrt{n}}, \overline{X}_n + t_{n-1,1-\alpha/2}\frac{\sqrt{\widehat{S_n^2}}}{\sqrt{n}}\right]$$

où $t_{n-1,1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-1$ degrés de liberté.

Cela provient du résultat précédent et de l'estimation de σ^2 par $\widehat{S_n^2}$.
Exemple : pour $n = 10$, avec un niveau de confiance de 95% et un intervalle symétrique on obtient l'intervalle

$$\left[\overline{X}_n - 2,26\frac{\sqrt{\widehat{S_n^2}}}{\sqrt{n}}, \overline{X}_n + 2,26\frac{\sqrt{\widehat{S_n^2}}}{\sqrt{n}}\right]$$

L'intervalle de confiance est plus grand que celui obtenu lorsqu'on connaît la variance.

3.c. Estimation de la variance quand la moyenne est connue.

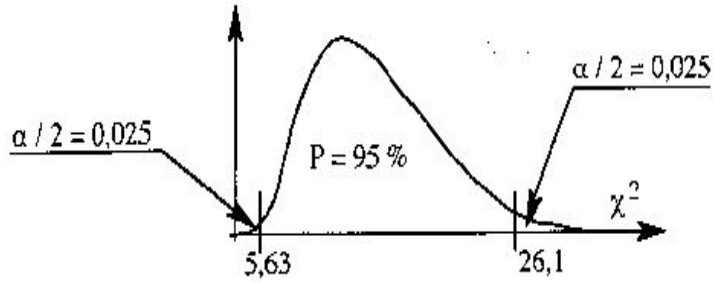
Théorème 7.

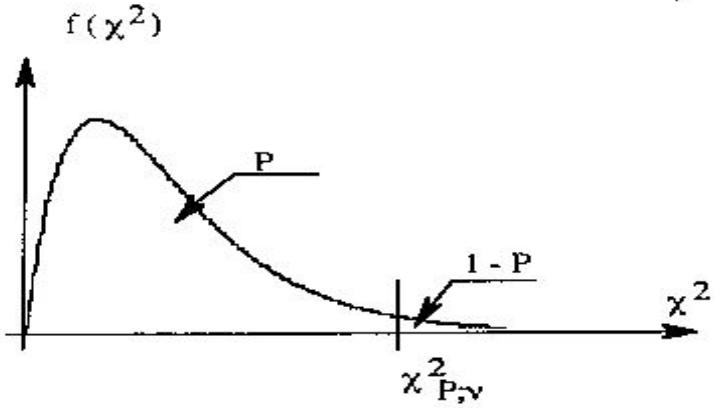
Lorsque m est connu un intervalle de confiance au niveau $1-\alpha$ de σ^2 est

$$\left[\frac{1}{u_2}\sum_{k=1}^{n}(X_k-m)^2, \frac{1}{u_1}\sum_{k=1}^{n}(X_k-m)^2\right]$$

où u_1 et u_2 sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi du χ^2 à n degrés de liberté.

Exemples d'intervalles bilatères et unilatères pour la loi du χ^2 :





Démonstration. Si X_k suit une loi $\mathcal{N}(m,\sigma^2)$ alors $(X_k-m)/\sigma$ suit une loi $\mathcal{N}(0,1)$ et par conséquent $\sum_{k=1}^{n}(\frac{X_k-m}{\sigma})^2$ suit une loi du χ^2 à n degrés de liberté. On définit alors u_1 et u_2 tels que

$$\mathbb{P}(\chi^2 \leq u_1) = \frac{\alpha}{2} \quad \text{et } \mathbb{P}(\chi^2 \geq u_2) = \frac{\alpha}{2},$$

et donc on a

$$\mathbb{P}\left(u_1 \leq \frac{1}{\sigma^2}\sum_{k=1}^{n}(X_k-m)^2 \leq u_2\right) = 1-\alpha.$$

D'où on en déduit le résultat. □

3.d. Estimation de la variance quand la moyenne est inconnue.

Théorème 8.

Lorsque m est inconnu un intervalle de confiance au niveau $1-\alpha$ de σ^2 est

$$\left[\frac{1}{u_2}\sum_{k=1}^{n}(X_k-\overline{X}_n)^2, \frac{1}{u_1}\sum_{k=1}^{n}(X_k-\overline{X}_n)^2\right]$$

où u_1 et u_2 sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi du χ^2 à $n-1$ degrés de liberté.

DÉMONSTRATION. On estime m par $\overline{X}_n$, puis

$$\sum_{k=1}^{n}\left(\frac{X_k-\overline{X}_n}{\sigma}\right)^2$$

suit une loi du χ^2 à $n-1$ degrés de liberté. Ensuite on procède comme dans la preuve précédente. □

Lorsqu'on s'intéresse à l'écart-type on prend les racines carrées des bornes des intervalles obtenus pour la variance.

4. Comparaison de moyennes et de variances

Soient $(X_1, X_2, \ldots, X_{n_1})$ un échantillon d'une population suivant la loi normale $\mathcal{N}(m_1, \sigma_1^2)$ et $(Y_1, Y_2, \ldots, Y_{n_2})$ un échantillon d'une population suivant la loi normale $\mathcal{N}(m_2, \sigma_2^2)$; ces deux échantillons sont supposés indépendants. Nous souhaitons comparer les moyennes, m_1 et m_2, et les variances, σ_1^2 et σ_2^2, à l'aide de ces échantillons. Pour cela nous allons construire des intervalles de confiance pour $m_1 - m_2$ et pour σ_1^2 et σ_2^2.

4.a. Intervalle de confiance de la différence de deux moyenne.

On pose

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \quad \overline{Y} = \frac{1}{n_1}\sum_{i=1}^{n_2} Y_i, \quad D = \overline{X} - \overline{Y}$$

Proposition 9. *L'estimateur D est un estimateur sans biais $m_1 - m_2$.*

DÉMONSTRATION. Par définition $\overline{X}$ suit une loi $\mathcal{N}(m_1, \sigma_1^2/n_1)$ et $\overline{Y}$ suit une loi $\mathcal{N}(m_2, \sigma_2^2/n_2)$ et par conséquent D suit la loi $\mathcal{N}(m_1 - m_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$, d'où le résultat. □

Théorème 10.

Si σ_1 et σ_2 sont connues, un intervalle de confiance de $m_1 - m_2$ au niveau $1-\alpha$ est

$$\left[D - q_{1-\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}, D + q_{1-\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right]$$

où $q_{1-\alpha/2}$ représente le fractile d'ordre $1-\alpha/2$ de la loi normale centrée réduite.

Démonstration. Un intervalle de confiance de $m_1 - m_2$ au niveau $1-\alpha$ est $[D-a, D+b]$ si

$$\begin{aligned} 1-\alpha &= \mathbb{P}(D-a \leq m_1 - m_2 \leq D+b) = \mathbb{P}(-b \leq D-(m_1-m_2) \leq a) \\ &= \mathbb{P}\left(-\frac{b}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \leq \frac{D-(m_1-m_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \leq \frac{a}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{u_1}^{u_2} \exp(-x^2/2) dx \end{aligned}$$

On en déduit l'intervalle annoncé. □

En général les variances ne sont pas connues. Il peut alors se présenter deux cas.
Posons :

$$\widehat{S^2_{n_1,X}} = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2 \quad \widehat{S^2_{n_2,Y}} = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$$

et

$$\widehat{S^2_{X,Y}} = \frac{1}{n_1+n_2-2}\left(\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2 \right) = \frac{(n_1-1)\widehat{S^2_{n_1,X}} + (n_2-1)\widehat{S^2_{n_2,Y}}}{n_1+n_2-2}$$

Théorème 11.

Si les variances σ_1 et σ_2 sont inconnues mais égales, un intervalle de confiance de $m_1 - m_2$ au niveau $1-\alpha$ est

$$\left[D - t_{n_1+n_2-2,1-\alpha/2}\sqrt{S^2_{X,Y}/n_1 + S^2_{X,Y}/n_2}, D + t_{n_1+n_2-2,1-\alpha/2}\sqrt{\widehat{S^2_{X,Y}}/n_1 + \widehat{S^2_{X,Y}}/n_2} \right]$$

où $t_{n_1+n_2-2,1-\alpha/2}$ représente le fractile d'ordre $1-\alpha/2$ de la loi de Student à n_1+n_2-2 degrés de libertés.

On aurait pu remplacer σ_1^2 et σ_2^2 par $\widehat{S^2_{n_1,X}}$ et $\widehat{S^2_{n_2,Y}}$, mais en général on préfère prendre un estimateur basé sur la réunion des deux échantillons. On a déjà vu que :

$$Z_1 = \sum_{i=1}^{n_1}\left(\frac{X_i - \overline{X}}{\sigma_1}\right)^2 \quad \text{et } Z_2 = \sum_{i=1}^{n_2}\left(\frac{Y_i - \overline{Y}}{\sigma_2}\right)^2$$

suivent respectivement une loi du χ^2 à n_1-1 et n_2-1 degrés de liberté. Par conséquent, comme Z_1 et Z_2 sont indépendantes, Z_1+Z_2 suit une loi du χ^2 à n_1+n_2-2 degrés de liberté. On obtient alors, en posant $\sigma_1^2 = \sigma_2^2 = \sigma^2$,

$$\mathbb{E}(Z_1+Z_2) = n_1+n_2-2 = \frac{1}{\sigma^2}\mathbb{E}\left[\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2 \right]$$

c'est-à-dire que $\widehat{S^2_{X,Y}}$ est un estimateur sans biais de σ^2.

Démonstration. On va remplacer l'écart-type de D, $\sqrt{\widehat{S^2_{X,Y}}/n_1 + \widehat{S^2_{X,Y}}/n_2}$ par $\sqrt{S^2_{X,Y}/n_1 + S^2_{X,Y}/n_2}$.
On a

$$\begin{aligned} \alpha &= \mathbb{P}(D-a \leq m_1 - m_2 \leq D+b) = \mathbb{P}(-b \leq D-(m_1-m_2) \leq a) \\ &= \mathbb{P}\left(-\frac{b}{\sqrt{\widehat{S^2_{X,Y}}/n_1 + \widehat{S^2_{X,Y}}/n_2}} \leq \frac{D-(m_1-m_2)}{\sqrt{\widehat{S^2_{X,Y}}/n_1 + \widehat{S^2_{X,Y}}/n_2}} \leq \frac{a}{\sqrt{\widehat{S^2_{X,Y}}/n_1 + \widehat{S^2_{X,Y}}/n_2}} \right) \end{aligned}$$

On pose

$$T = \frac{D-(m_1-m_2)}{\sqrt{\widehat{S^2_{X,Y}}/n_1 + \widehat{S^2_{X,Y}}/n_2}} = \frac{\frac{D-(m_1-m_2)}{\sqrt{\sigma^2/n_1+\sigma^2/n_2}}}{\sqrt{\frac{\widehat{S^2_{X,Y}}}{\sigma^2}}}$$

Le numérateur suit une loi normale centrée réduite. Au dénominateur on a $\sqrt{\frac{\widehat{S^2_{X,Y}}}{\sigma^2}} = \frac{U}{n_1+n_2-2}$ où U suit une loi du χ^2 à n_1+n_2-2 degrés de liberté. On en déduit que T suit une loi de Student à n_1+n_2-2 degrés de liberté, et donc on obtient le résultat. □

Lorsqu'on σ_1 et σ_2 sont inconnues mais non nécessairement égales, on utilise la méthode approchée suivante.

Théorème 12.

Si les échantillons ont des tailles importantes et égales à $n_1 = n_2 = n > 30$, un intervalle de confiance de $m_1 - m_2$ au niveau $1-\alpha$ est

$$\left[D - q_{1-\alpha/2}\sqrt{(\widehat{S^2_{n_1,X}} + \widehat{S^2_{n_2,Y}})/n}, D + q_{1-\alpha/2}\sqrt{(\widehat{S^2_{n_1,X}} + \widehat{S^2_{n_2,Y}})/n}\right]$$

où $q_{1-\alpha/2}$ représente le fractile d'ordre $1-\alpha/2$ de la loi normale centrée réduite.

Il suffit de remarquer que la variable $\frac{D-(m_1-m_2)}{\sqrt{(\widehat{S^2_{n_1,X}} + \widehat{S^2_{n_2,Y}})/n}}$ suit sensiblement une loi normale centrée réduite.

4.b. Intervalle de confiance du rapport de deux variances.

Théorème 13.

Un intervalle de confiance au niveau $1-\alpha$ de σ_1^2/σ_2^2 est

$$\left[\frac{1}{f_{n_1-1,n_2-1,1-\alpha/2}}\frac{\widehat{S^2_{n_1,X}}}{\widehat{S^2_{n_2,Y}}}, \frac{1}{f_{n_1-1,n_2-1,1-\alpha/2}}\frac{\widehat{S^2_{n_1,X}}}{\widehat{S^2_{n_2,Y}}}\right]$$

où les fractiles sont ceux de loi de Fisher-Snedecor $\mathcal{F}(n_1-1, n_2-1)$.

Le résultat s'obtient par les mêmes méthodes que pour les théorèmes précédents.

La loi de Fisher-Snedecor peut être obtenue comme le quotient de deux lois du χ^2 indÕpendantes :

$$F(n_1-1, n_2-1) = \frac{\chi^2_{n_1-1}/(n_1-1)}{\chi^2_{n_2-1}/(n_2-1)}$$

5. Estimation d'une proportion

Dans une certaine population, la proportion d'individus ayant une propriété donnée est égale à p. Soit X le nombre d'individus d'un échantillon de taille n ayant la propriété.

5.a. Estimation ponctuelle.

Théorème 14.

Un estimateur sans biais et consistant de p est :

$$T = \frac{X}{n}$$

En effet, le nombre X d'individus de l'échantillon ayant la propriété suit la loi binomiale $B(n,p)$. On a :

$$\mathbb{E}(T) = \frac{\mathbb{E}(X)}{n} = p \quad \text{et} \quad \mathrm{Var}(T) = \frac{\mathrm{Var}(X)}{n^2} = \frac{p(1-p)}{n} \to 0$$

5.b. Estimation par intervalle.

On ne sait pas déterminer exactement un intervalle de confiance. On utilise des solutions approchées, qui fonctionnent lorsqu'on dispose d'échantillon de grande taille. Ainsi, lorsque n est grand ou/et p voisin de $0,5$ on peut approcher la loi binomiale par une loi normale.
Rappel : Soit une suite de variables aléatoires Z_n suivant la loi binomiale $B(n,p)$; la suite des variables réduites $Z_n^* = \dfrac{Z_n - np}{\sqrt{np(1-p)}}$ converge en loi vers la loi normale centrée réduite, et on a :

$$\mathbb{P}(a \leq Z_n^* \leq b) \longrightarrow \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(-x^2/2\right) dx$$

Théorème 15.

Un intervalle de confiance approché de p au niveau $1-\alpha$ est donnée par

$$\left[T - q_{1-\alpha/2}\sqrt{\frac{T(1-T)}{n}}, T + q_{1-\alpha/2}\sqrt{\frac{T(1-T)}{n}} \right]$$

où $q_{1-\alpha/2}$ représente le fractile d'ordre $1-\alpha/2$ de la loi normale centrée réduite.

Démonstration. D'après le rappel,

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{T-p}{\sqrt{p(1-p)/n}}$$

suit approximativement la loi normale centrée réduite. L'intervalle $[T-a, T+b]$ est un intervalle de confiance de p au niveau α si :

$$\begin{aligned} \alpha &= \mathbb{P}(T-a < p < T+b) = \mathbb{P}\left(-\frac{b}{\sqrt{p(1-p)/n}} < \frac{T-p}{\sqrt{p(1-p)/n}} < \frac{a}{\sqrt{p(1-p)/n}} \right) \\ &\simeq \Phi\left(\frac{a}{\sqrt{p(1-p)/n}} \right) - \Phi\left(-\frac{b}{\sqrt{p(1-p)/n}} \right) \quad \text{où} \quad \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-x^2/2\right) dx \end{aligned}$$

En choisissant un intervalle symétrique on obtient la quantile $q_{1-\alpha/2}$. Comme $\sqrt{p(1-p)}$ n'est pas connu, on obtient alors un intervalle de confiance approché en remplaçant p par l'estimateur T. □

On peut donner une approximation de l'intervalle de confiance un peu plus précise.

Un intervalle de confiance approché de p au niveau $1-\alpha$ est donnée par

$$\left[\frac{1}{1+\frac{q^2_{1-\alpha/2}}{n}}\left(T+\frac{q^2_{1-\alpha/2}}{2n}-\Delta\right), \frac{1}{1+\frac{q^2_{1-\alpha/2}}{n}}\left(T+\frac{q^2_{1-\alpha/2}}{2n}+\Delta\right)\right]$$

où $\Delta = \frac{q_{1-\alpha/2}}{\sqrt{n}}\sqrt{T(1-T)+\frac{q^2_{1-\alpha/2}}{4n}}$ et $q_{1-\alpha/2}$ représente le fractile d'ordre $1-\alpha/2$ de la loi normale centrée réduite.

DÉMONSTRATION. On a :

$$-q_{1-\alpha/2} < \frac{T-p}{\sqrt{\frac{p(1-p)}{n}}} < q_{1-\alpha/2} \quad \Longleftrightarrow \quad (T-p)^2 < q^2_{1-\alpha/2}\frac{p(1-p)}{n}$$

$$\Longleftrightarrow \quad p^2\left(1+\frac{q^2_{1-\alpha/2}}{n}\right) - 2p\left(T+\frac{q^2_{1-\alpha/2}}{2n}\right) + T^2 < 0$$

Le paramètre p doit donc être compris entre les racines de l'équation du second degré. On vérifie aisément qu'elle a deux racines réelles appartenant à l'intervalle $[0,1]$. D'où, on obtient l'intervalle de confiance indiqué. □

5.c. Méthode du Bootstrap.

A partir d'un échantillon $X = (X_1, X_2, \ldots, X_n)$ on détermine un estimateur ponctuel $s(X)$ d'un paramètre θ. Sauf dans quelques cas particuliers ($s(X) = \overline{X}_n$ par exemple) le calcul de la variance n'est pas aisé, ce qui rend problématique la détermination d'intervalles de confiance pour ?. En 1979 une nouvelle méthode a été développé. Cette méthode s'appuie sur des concepts simples permettant, à partir d'une réalisation $(x_1, x_2, \ldots, x_n)$ de l'échantillon, d'obtenir une estimation de la variance de $s(X)$ et un intervalle de confiance pour θ.

On considère que la réalisation de l'échantillon $(x_1, x_2, \ldots, x_n)$ est représentative de la population et on tire parmi les x_k, au hasard et avec remise, un échantillon bootstrapé $X^* = (X_1^*, X_2^*, \ldots, X_n^*)$; en pratique on tire n nombres au hasard entre 1 et n et on associe au nombre tiré k la valeur X_k . Sur cet échantillon bootstrapé on peut calculer un estimateur $s(X^*)$ par le même algorithme que celui qui donne $s(X)$.

On répète le tirage un grand nombre de fois, B, ce qui donne une population de valeurs de $s(X^*)$ $S = s_1, s_2, \ldots, s_B$ que l'on peut représenter par un histogramme. Sur cette population on peut calculer une estimation de la moyenne et de l'écart-type :

$$\overline{s} = \frac{1}{B}\sum_{k=1}^{B} s_k, \quad \sqrt{\frac{1}{B-1}\sum_{k=1}^{B}(s_k-\overline{s})^2}$$

La population S peut être triée par valeurs croissantes ce qui permet de déterminer un intervalle de confiance en gardant une certaine proportion des valeurs centrales. Par exemple si $B = 1000$ et si les valeurs triées de S sont $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{1000}$, l'intervalle de confiance à 95% est $[\alpha_{25}, \alpha_{975}]$.