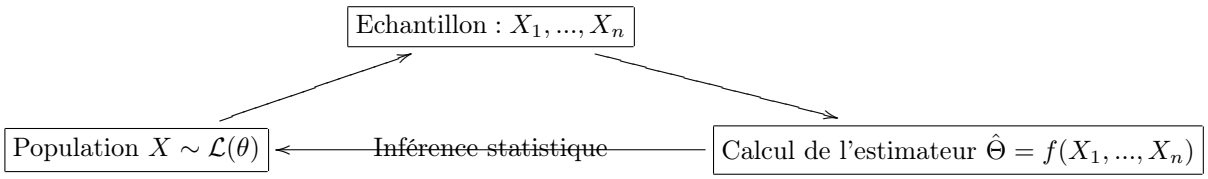
Chapitre 2

Estimateurs et intervalles de confiance

2.1 Principe général



1. Dans une population donnée, on s'intéresse à un caractère représenté par une v.a. X qui suit une loi de probabilité $\mathcal{L}(\theta)$ qui dépend d'un paramètre θ que l'on veut déterminer (par exemple $\theta = \mu =$ moyenne de X).

2. Pour celà, on prend un échantillon de la population, donc une réalisation de $X_1, ..., X_n$ qui sont des v.a. indépendantes et identiques à X.

3. On utilise un estimateur de $\hat{\Theta} = f(X_1, ..., X_n)$ pour obtenir une valeur approchée de θ (par exemple, pour $\theta = \mu =$, on utilise $\hat{\Theta} = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$).

4. On détermine ainsi, par inférence statistique, la caractéristique de la population totale.

On distingue ensuite, et selon les besoins, entre deux procédés :

- **L'estimation ponctuelle**, qui consiste à prendre la valeur donnée par l'estimateur comme valeur approchée de θ et à étudier l'erreur commise.

- **L'estimation par intervalle de confiance (IDC)**, qui consiste à trouver un intervalle dans lequel se trouve la valeur exacte de θ avec un niveau de confiance donné.

2.2 Estimation ponctuelle

2.2.1 Qualités d'un estimateur

Une des premières qualités demandées à un estimateur est d'être *sans biais*. En effet, l'erreur commise, en prenant la valeur estimée à la place de la valeur exacte est : $\hat{\Theta} - \theta = \left[\hat{\Theta} - E(\hat{\Theta})\right] + \left[E(\hat{\Theta}) - \theta\right]$.
Le premier terme correspond à la fluctuation naturelle d'une v.a. autour de sa moyenne, le deuxième terme est systématique et peut et doit être réduit.

1. **Biais :**
   On appelle *biais* d'un estimateur $\hat{\Theta}$ du paramètre θ, l'expression :
   $b(\hat{\Theta}) = E(\hat{\Theta}) - \theta$.
   Un estimateur est dit *sans biais* si $b(\hat{\Theta}) = 0$ et est dit *asymptotiquement sans biais* si $\lim\limits_{n\to\infty} b(\hat{\Theta}) = 0$.
2. **Convergence :**
   Le *risque quadratique* d'un estimateur ou *erreur quadratique moyenne* est donné par $R_\theta(\hat{\Theta}) = E\left((\hat{\Theta} - \theta)^2\right) = V(\hat{\Theta}) + b(\hat{\Theta})^2$.
   On dit d'un estimateur qu'il est *convergent* lorsque $\lim\limits_{n\to\infty} R_\theta(\hat{\Theta}) = 0$.
   Ceci est notamment le cas si l'estimateur est sans biais ou asymptotiquement sans biais et que sa variance tend vers 0.
3. **Efficacité :**
   Si $\hat{\Theta}_1$ et $\hat{\Theta}_2$ sont deux estimateurs sans biais du même paramètre θ, on dit que $\hat{\Theta}_1$ est *plus efficace* que $\hat{\Theta}_2$ si $V\left(\hat{\Theta}_1\right) < V\left(\hat{\Theta}_2\right)$.
   L'estimateur est dit *efficace* si sa variance est la plus petite parmi les estimateurs sans biais.
4. **Loi asymptotique :**
   L'étude probabilistique de l'erreur nécessite la connaissance de la loi suivie par l'estimateur ou à défaut sa loi asymptotique, càd la loi suivie pou n assez grand.
   Dans certains cas cette loi est connue pour tout n et dans d'autres cas, les théorèmes limites (LGN ou TCL) permettent de déterminer la loi asymptotique.

2.2.2 Estimateurs usuels

1. **Proportion empirique :**
   Lorsque X suit une loi de Bernouilli, $X \sim \mathcal{B}(p)$, et qu'on veut estimer le paramètre $\theta = p$, on utilise la fréquence empirique

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

   où $Y_n = \sum\limits_{i=1}^{n} X_i$ suit la loi binômiale $Y_n \sim \mathcal{B}(n, p)$.
   – Cet estimateur est **sans biais**, car $E(F_n) = \frac{E(Y_n)}{n} = p = \theta$.
   – Il est **convergent** car le risque quadratique
   $R_\theta(F_n) = V(F_n) = \frac{V(Y_n)}{n^2} = \frac{p(1-p)}{n}$ tend bien vers 0.

– Sa loi asymptotique est donnée par le théorème de la limite centrale (TCL) : $F_n \xrightarrow{\mathcal{L}} \mathcal{N}(p, \frac{p(1-p)}{n})$.

2. **Moyenne empirique :**
Si X suit une loi de moyenne μ et de variance σ^2, on utilise la moyenne empirique $\overline{X_n}$ pour estimer $\theta = \mu$ avec

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

où $X_1, ..., X_n$ est un échantillon correspondant.
– C'est un estimateur **sans biais**. $E(\overline{X_n}) = \mu = \theta$.
– Il est **convergent** car le risque quadratique $R_\theta(\overline{X_n}) = V(\overline{X_n}) = \frac{\sigma^2}{n}$ tend bien vers 0.
– Le TCL donne : $\overline{X_n} \xrightarrow{\mathcal{L}} \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

3. **Variance empirique :**
Pour une v.a. X qui suit une loi de moyenne μ et de variance σ^2, lorsqu'on veut estimer la variance ($\theta = \sigma^2$), on utilise la variance empirique :

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X_n})^2$$

– $E(S_1^2) = \frac{n-1}{n}\sigma^2 \neq \theta$ c'est un estimateur **biaisé** même s'il est asymptotiquement sans biais.
C'est pour cette raison qu'on introduit la variance empirique corrigée :

$$S^2 = \frac{n}{n-1}S_1^2$$

qui est un estimateur sans biais de la variance.
– La variance empirique converge bien vers σ^2 en probabilité et presque sûrement, mais la convergence en moyenne quadratique n'a lieu que si X et donc $X_1, ..., X_n$ suivent une loi normale.
– On ne peut rien dire de la loi de S_1^2 ou S^2, sauf si $X \sim \mathcal{N}(\mu, \sigma^2)$ et donc aussi $X_1, ..., X_n$.
Dans ce cas la v.a. $\frac{n}{\sigma^2}S_1^2 = \frac{n-1}{\sigma^2}S^2$ suit la loi du khi-2 à $(n-1)$ degrés de liberté : χ^2_{n-1}.
Ceci permet d'obtenir $V(S^2) = \frac{2\sigma^4}{n-1}$ et donc d'établir la convergence de cet estimateur en moyenne quadratique.

2.3 Intervalles de confiance

2.3.1 Principe général

Le but est d'obtenir un intervalle $[a, b]$ dans lequel se trouve le paramètre avec un niveau de risque α donné.

$$P(a \leq \theta \leq b) = 1 - \alpha$$

α étant le niveau de risque pris, en général $\alpha = 10\ \%, 5\%$ ou 1%.
a et b sont obtenus en partageant le risque en $\alpha = \alpha_1 + \alpha_2$ avec $\alpha_1 = P(X < a)$ et $\alpha_2 = P(X > b)$ dans le cas d'un IDC bilatéral.
Cet intervalle bilatéral peut être symétrique ($\alpha_1 = \alpha_2$) ou dissymétrique.
Mais dans certains cas, on est amené à rechercher un IDC unilatéral de type $[a, +\infty[$ ou $]-\infty, b]$. C'est notamment le cas lorsque le paramètre θ est positif, cas de la variance.

2.3.2 IDC pour une proportion

On a vu que lorsque $X \sim \mathcal{B}(p)$, on utilisait la fréquence empirique F_n comme estimateur de $\theta = p$.
Pour n assez grand, le TCL nous permettait d'affirmer que $F_n \sim \mathcal{N}(\mu, \sigma_1^2)$ avec $\mu = p$ et $\sigma_1^2 = \frac{p(1-p)}{n}$
On en déduit que $Z = \dfrac{F_n - p}{\sigma_1}$ suit la loi normale centrée réduite $Z \sim \mathcal{N}(0, 1)$.
Pour obtenir un IDC symétrique au niveau de risque α, on utilise la valeur $z_{\alpha/2}$ lue dans la table de la fonction de répartition de la loi normale.
$z_{\alpha/2}$ vérifie $F_Z(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$ et aussi $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$.
Si f_n est la valeur estimée donnée par l'échantillon, on obient :

$$1-\alpha = P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = P\left(-z_{\alpha/2} \leq \frac{f_n - p}{\sigma_1} \leq z_{\alpha/2}\right) = P\left(f_n - z_{\alpha/2}\sigma_1 \leq p \leq f_n + z_{\alpha/2}\sigma_1\right)$$

L'intervalle de confiance recherché est donc :

$$I = \left[f_n - z_{\alpha/2}\sigma_1\ ;\ f_n + z_{\alpha/2}\sigma_1\right]$$

Problème : $\sigma_1 = \sqrt{\frac{p(1-p)}{n}}$ et notre intervalle dépend donc du paramètre p que l'on veut estimer ! ! !

Pour s'en sortir, on a trois techniques :

1. On remplace p par sa valeur estimée f_n. Et donc σ_1 devient $\sqrt{\frac{f_n(1-f_n)}{n}}$.
2. Ou bien on remplace l'expression $p(1-p)$ dans σ_1 par la valeur maximale de $x(1-x)$, qui est réalisée pour $x = \frac{1}{2}$.
   Autrement dit on prend $p = \frac{1}{2}$, quitte à obtenir un intervalle trop large.
3. Troisième technique : on utilise les abaques données avec les tables statistiques. Celle qui vous été distribuée correspond à $\alpha = 5\ \%$ et à des échantillons de taille allant de $n = 5$ à $n = 100$.

2.3.3 IDC pour une moyenne

L'estimateur utilisé est évidemment $\overline{X_n}$. Sa loi asymptotique dépend du fait que la variance soit connue ou non.
Le calcul va dépendre de la taille de l'échantillon n, de la loi suivie par X et de la connaissance ou non de la variance de X.

1. 1^{er} **cas : X suit une loi quelconque, $n > 30$ et σ connue :**
   Dans ce cas on peut considérer que : $\overline{X_n} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ ou encore que

$$\frac{\overline{X_n} - \mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0, 1)$$

Si $\bar{x}_n$ est la valeur estimée donnée par l'échantillon, et pour un niveau de risque α, on obtient comme intervalle de confiance :

$$I = \left[\bar{x}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \ ; \ \bar{x}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

2. 2^{eme} **cas : X suit une loi normale, n quelconque et σ connue :**
   Même procédé que pour le cas 1.
3. 3^{eme} **cas : X suit une loi normale, n quelconque et σ inconnue :**
   Dans ce cas, on remplace σ^2 par son estimateur corrigé S^2 et on utilise :

   $$\frac{\overline{X_n} - \mu}{S} \sqrt{n} \sim T_{n-1}$$

   loi de Student à $(n-1)$ degrés de liberté qui est symétrique tout comme la loi normale.
   Pour un niveau de risque α donné, à l'aide des valeurs estimées $\bar{x}_n, s$ et de la valeur $t_{\alpha/2}$ équivalent de $z_{\alpha/2}$ pour la loi de Student, on obtient l'IDC suivant :

   $$I = \left[\bar{x}_n - t_{\alpha/2} \frac{s}{\sqrt{n}} \ ; \ \bar{x}_n + t_{\alpha/2} \frac{s}{\sqrt{n}} \right]$$

4. **Autres cas**
   Dans tous les autres cas, on ne peut rien dire.

2.3.4 IDC pour une variance

On ne peut rien dire des lois suivies par les deux estimateurs de la variance, sauf dans le cas gaussien.
Lorsque X et donc $X_1, X_2, ..., X_n$ suivent une loi $\mathcal{N}(\mu, \sigma^2)$, on distingue deux cas :

1. 1^{er} **cas : moyenne μ connue :**

   L'estimateur utilisé pour la variance est alors : $V_1 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ et :

   $$n \frac{V_1}{\sigma^2} \sim \chi^2_n$$

   loi du khi-2 à n degrés de liberté.
   Pour un risque de niveau $\alpha = \alpha_1 + \alpha_2$, l'IDC est obtenu à l'aide des valeurs : v_1 valeur estimée de V_1 et des valeurs k_1 et k_2 relevées sur la table du χ^2_n grâce aux relations : $\alpha_1 = P(\chi^2_n < k_1)$ et $\alpha_2 = P(\chi^2_n > k_2)$

   $$1 - \alpha = P(k_1 \leq \chi^2_n \leq k_2) = P\left(k_1 \leq n \frac{v_1}{\sigma^2} \leq k_2 \right) = P\left(n \frac{v_1}{k_2} \leq \sigma^2 \leq n \frac{v_1}{k_1} \right)$$

   L'intervalle de confiance pour σ^2 est donc :

   $$\sigma^2 \in \left[n \frac{v_1}{k_2} \ \ ; \ \ n \frac{v_1}{k_1} \right]$$

   Un passage à la racine carrée permet alors d'obtenir l'IDC pour σ.

2. 2^{eme} **cas : moyenne μ inconnue :**

   L'estimateur utilisé pour la variance est alors : $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$ ou $S^2 = \frac{n}{n-1}S_1^2$, on alors :

$$n\frac{S_1^2}{\sigma^2} = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

   loi du khi-2 à $(n-1)$ degrés de liberté.
   Pour un risque de niveau $\alpha = \alpha_1 + \alpha_2$, l'IDC est obtenu à l'aide des valeurs : s valeur estimée de S et des valeurs k_1 et k_2 relevées sur la table du χ^2_{n-1} grâce aux relations : $\alpha_1 = P(\chi^2_{n-1} < k_1)$ et $\alpha_2 = P(\chi^2_{n-1} > k_2)$
   On obtient alors comme intervalle de confiance pour σ^2 :

$$\sigma^2 \in \left[(n-1)\frac{s}{k_2} \;\;;\;\; (n-1)\frac{s}{k_1}\right]$$

   Un passage à la racine carrée permet alors d'obtenir l'IDC pour σ.

Le cas d'un intervalle unilatéral se traite de la même manière.