M4202C - Statistique inférentielle
Cours 1 - Estimation

2021/2022 - A. Ridard

A propos de ce document

- Pour naviguer dans le document, vous pouvez utiliser :
  - le menu (en haut à gauche)
  - l'icône en dessous du logo IUT
  - les différents liens
- Pour signaler une erreur, vous pouvez envoyer un message à l'adresse suivante : anthony.ridard@univ-ubs.fr

Plan du cours

1. Estimation ponctuelle
   - Estimateur et Loi Forte des Grands Nombres (LFGN)
   - Qualités d'un estimateur
2. Estimation par intervalle de confiance
   - Principe
   - Moyenne
   - Variance
   - Proportion (hors programme)

L'estimation consiste à donner une valeur approchée (ou un ensemble de valeurs plausibles) du paramètre **inconnu** m, σ^2 ou p, ceci à l'aide d'un échantillon de n observations issues de la population.

On considère alors :

- X la v.a. étudiée
- θ le paramètre inconnu à estimer : sa moyenne m ou sa variance σ^2
- $(X_1, ..., X_n)$ un échantillon

Définition (estimateur convergeant)

Un estimateur convergeant de θ est une fonction de $X_1, \ldots, X_n$ qui converge vers θ :

$$T_n = \phi(X_1, \ldots, X_n) \xrightarrow[n \to +\infty]{} \theta$$

En pratique, on retiendra que $T_n \simeq \theta$ pour n assez grand (supérieur ou égal à 30).

- Pour être plus rigoureux (pas nécessaire pour nous), il faudrait préciser le type de convergence :
  - s'il s'agit d'une convergence en probabilité [a], l'estimateur est dit convergeant
  - s'il s'agit d'une convergence presque sûre [b], l'estimateur est dit fortement convergeant
- Pour ne pas alourdir les notations, on notera T au lieu de T_n

a. $T_n \xrightarrow[n \to +\infty]{\mathscr{P}} \theta$ si et seulement si $\forall \epsilon > 0,\ \mathscr{P}(|T_n - \theta| \geq \epsilon) \xrightarrow[n \to +\infty]{} 0$

b. $T_n \xrightarrow[n \to +\infty]{p.s.} \theta$ si et seulement si $\forall c \in C,\ T_n(c) \xrightarrow[n \to +\infty]{} \theta$ pour une certaine partie C de Ω vérifiant $\mathscr{P}(C) = 1$

Dans toute la suite du cours, un estimateur désignera un estimateur (fortement) convergeant.

Théorème : LFGN

La moyenne empirique converge (presque sûrement) vers la moyenne théorique :

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{(p.s.)} m$$

En pratique, on retiendra que $\frac{1}{n}\sum_{i=1}^{n} X_i \simeq m$ pour n assez grand (supérieur ou égal à 30).

On en déduit :

Propriété : estimateurs usuels

- La moyenne empirique $\bar{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ est un estimateur de m.
  On le notera plus simplement $\bar{X}$.
- La variance empirique $S_n^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i-\bar{X})^2 = \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2\right) - \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i\right)^2$ est un estimateur de σ^2.
  On le notera plus simplement S^2.
- La fréquence empirique $F_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ est un estimateur de p où $X \sim \mathscr{B}(p)$.
  On le notera plus simplement F.

Les formules empiriques sont à comparer, cette fois, avec celles de la colonne 1 du tableau page 8 !

Définition (estimation ponctuelle)

Une estimation ponctuelle de θ est une réalisation d'un estimateur de θ.

Ne pas confondre l'estimateur $\bar{X}$ (en majuscule) qui est une v.a. et l'estimation $\bar{x}$ (en minuscule) qui est une valeur.

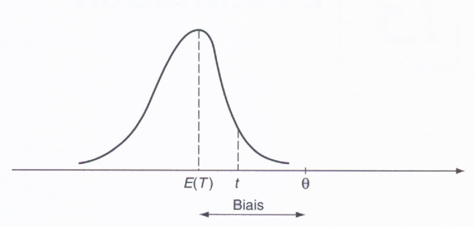
On considère ici T un estimateur de θ.

Définition (erreur - décomposition fluctuation d'échantillonnage / biais)

L'**erreur d'estimation** $T-\theta$ est une v.a. qui se décompose de la manière suivante :

$$T-\theta = \big(T-E(T)\big) + \big(E(T)-\theta\big)$$

- $T-E(T)$ est une erreur aléatoire : T varie autour de sa valeur centrale $E(T)$.
  Cette partie de l'erreur est appelée **fluctuation d'échantillonnage**.
- $E(T)-\theta$ est une erreur systématique : T varie autour de sa valeur centrale $E(T)$ et non autour de θ.
  Cette partie de l'erreur est appelée **biais**.



Il est souhaitable d'utiliser des estimateurs **sans biais** c'est à dire vérifiant $E(T) = \theta$

1. Montrer que $\bar{X}$ est un estimateur sans biais de m.
2. Montrer que S^2 est un estimateur biaisé [a] qui a tendance à sous-estimer σ^2
3. Montrer que S^2 est un estimateur **asymptotiquement sans biais** de σ^2 c'est à dire vérifiant $E(S^2) \xrightarrow[n \to +\infty]{} \sigma^2$.

a. On pourra d'abord montrer que $S^2 = \left[\frac{1}{n} \sum_{i=1}^{n} (X_i - m)^2 \right] - (\bar{X} - m)^2$

Propriété : variance empirique corrigée

La variance empirique corrigée $S^{*2} = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ est un estimateur sans biais de σ^2.

L'écart-type empirique corrigé S^* reste biaisé pour σ mais asymptotiquement sans biais.

Propriété : erreur quadratique moyenne - décomposition biais / variance

La précison de T est mesurée à l'aide de l'erreur quadratique moyenne $E\left((T-\theta)^2\right)$ qui se décompose sous la forme :

$$E\left((T-\theta)^2\right) = \left(E(T)-\theta\right)^2 + V(T)$$

L'objectif est de minimiser cette erreur quadratique moyenne.
Entre deux estimateurs sans biais du même paramètre, on choisira celui de plus petite variance [a].

a. L'Estimateur Sans Biais de Variance Minimale (ESBVM) sort du cadre de ce cours, tout comme l'Estimateur du Maximum de Vraisemblance (EMV) !

On considère $D = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$ et on admet qu'il s'agit d'un estimateur sans biais de σ^2.

1. Montrer que $V(D) = \frac{1}{n}(\mu_4 - \sigma^4)$ où $\mu_4 = E\Big((X-m)^4\Big)$ désigne le moment centré d'ordre 4 de X.
2. Montrer que $V(S^{*2}) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$.
3. Conclure.

1. Estimation ponctuelle
2. Estimation par intervalle de confiance

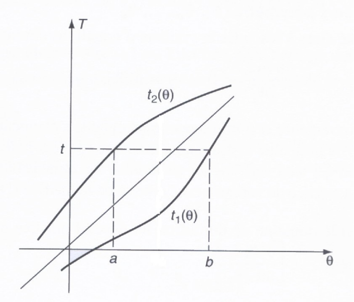
Plutôt que de fournir un renseignement du type $\theta \simeq c$, il est souvent plus intéressant de fournir un renseignement du type $a < \theta < b$ qui est certes moins précis, mais qui a l'avantage d'être accompagné d'une confiance.

Vannes iut Université Bretagne Sud

Soit T un estimateur de θ (le meilleur possible) dont on connaît la loi de probabilité qui est fonction de θ.

Pour déterminer un intervalle de confiance pour θ au niveau $1-\alpha$ c'est à dire un intervalle $[a,b]$ avec une chance de contenir θ égale à $1-\alpha$ ou encore un risque de ne pas le contenir égal à α, il suffit de déterminer un intervalle de probabilité pour T au niveau $1-\alpha$ c'est à dire deux réels $t_1(\theta) < t_2(\theta)$ vérifiant :

$$P\Big(t_1(\theta) < T < t_2(\theta)\Big) = 1-\alpha$$



On lit alors l'intervalle de confiance $[a, b]$ selon l'horizontale issue de t.

- Si on augmente le niveau de confiance $1-\alpha$, les courbes s'écartent et donc l'intervalle grandit
- Si la taille de l'échantillon augmente, les courbes se rapprochent et donc l'intervalle diminue

Supposons $X \sim \mathcal{N}(m,\sigma)$ et estimons m.

Quand σ est connu

On utilise $\bar{X}$ le meilleur estimateur de m, et la fonction pivotale [1] $W = \dfrac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$ (savez-vous démontrer ce résultat ?).

L'intervalle de probabilité (à risques symétriques) pour W au niveau $1-\alpha$ est :

$$-u_{(1-\frac{\alpha}{2})} < W < u_{(1-\frac{\alpha}{2})}$$

où $u_{(1-\frac{\alpha}{2})}$ désigne le quantile d'ordre $(1-\frac{\alpha}{2})$ de la $\mathcal{N}(0,1)$ [2].

1. Il s'agit d'une fonction de $X_1, \ldots, X_n$ qui dépend de θ, mais dont la loi ne dépend pas de θ
2. C'est à dire le nombre pour lequel l'aire sous la courbe à gauche vaut $(1-\frac{\alpha}{2})$. Par exemple, $u_{0.975} = 1,96$.

L'intervalle de probabilité (à risques symétriques) pour $\bar{X}$ au niveau $1-\alpha$ est donc :

$$m - u_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} < \bar{X} < m + u_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}$$

L'intervalle de confiance [3] (bilatéral [4]) pour m au niveau $1-\alpha$ est alors :

$$\bar{x} - u_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} < m < \bar{x} + u_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}$$

3. Certains auteurs fournissent plutôt l'intervalle de confiance "aléatoire" : $\bar{X} - u_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} < m < \bar{X} + u_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}$

4. On ne considèrera pas l'intervalle de confiance unilatéral à droite (resp. gauche) qui regroupe du côté gauche (resp. droit) le risque α de ne pas contenir le paramètre et montrer ainsi que ce dernier est suffisamment grand (resp. petit)

Quand σ est inconnu

On utilise encore $\bar{X}$, mais cette fois la fonction pivotale $W = \dfrac{\bar{X} - m}{\frac{S^*}{\sqrt{n}}} \sim \mathcal{T}_{n-1}$ [5] .

L'intervalle de probabilité pour W au niveau $1-\alpha$ est :

$$-t_{(1-\frac{\alpha}{2})} < W < t_{(1-\frac{\alpha}{2})}$$

où $t_{(1-\frac{\alpha}{2})}$ désigne le quantile d'ordre $(1-\frac{\alpha}{2})$ de la $\mathcal{T}_{n-1}$.

5. Une Student à n degrés de liberté est définie par :

$$\mathcal{T}_n = \frac{U}{\sqrt{\frac{X}{n}}}$$

avec $U \sim \mathcal{N}(0,1)$ et $X \sim \chi^2_n$ indépendantes (cf. plus loin pour la définition de χ^2_n)

L'intervalle de probabilité pour $\bar{X}$ au niveau $1-\alpha$ est donc :

$$m - t_{(1-\frac{\alpha}{2})}\frac{S^*}{\sqrt{n}} < \bar{X} < m + t_{(1-\frac{\alpha}{2})}\frac{S^*}{\sqrt{n}}$$

L'intervalle de confiance pour m au niveau $1-\alpha$ est alors :

$$\bar{x} - t_{(1-\frac{\alpha}{2})}\frac{s^*}{\sqrt{n}} < m < \bar{x} + t_{(1-\frac{\alpha}{2})}\frac{s^*}{\sqrt{n}}$$

Quand l'échantillon n'est plus gaussien mais de grande taille

On utilise le Théorème Central Limite :

Théorème : TCL

Toujours en notant $m = E(X)$ et $\sigma^2 = V(X)$, on a :

$$\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n \to +\infty]{\mathscr{L}} \mathscr{N}(0,1)$$

En pratique, on retiendra que la loi de $\dfrac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}$ est proche de la gaussienne centrée réduite pour $n \geq 30$.

Quand l'échantillon n'est plus gaussien mais de grande taille

Quand σ est connu, le TCL nous assure :

$$W = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,1)$$

Autrement dit, la fonction asymptotiquement pivotale $W = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}$ suit approximativement une $\mathcal{N}(0,1)$ et donc ...

Quand σ est inconnu, le TCL accompagné du théorème de Slutsky nous assure :

$$W = \frac{\bar{X} - m}{\frac{S^*}{\sqrt{n}}} \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,1)$$

Autrement dit, la fonction asymptotiquement pivotale $W = \frac{\bar{X} - m}{\frac{S^*}{\sqrt{n}}}$ suit approximativement une $\mathcal{N}(0,1)$ et donc ...

Les **intervalles de confiance** qui en découlent sont qualifiés d'**asymptotiques**.

Un artisan qui fabrique des objets de maroquinerie souhaite estimer le nombre moyen m de porte-cartes vendus quotidiennement. En notant ses ventes sur 36 jours, il obtient une moyenne de 120 et un écart-type corrigé de 17.

Donner un intervalle de confiance pour m au niveau 95% dans les deux cas suivants :

1. Si le nombre de porte-cartes est gaussien.
2. Sans l'hypothèse de normalité.

Supposons $X \sim \mathcal{N}(m,\sigma)$ et estimons σ^2.

Quand m est connue

On utilise D le meilleur estimateur de σ^2, et la fonction pivotale $W = \frac{nD}{\sigma^2} \sim \chi_n^2$ [6] .

L'intervalle de probabilité pour W au niveau $1-\alpha$ est :

$$k_{\frac{\alpha}{2}} < \frac{nD}{\sigma^2} < k_{(1-\frac{\alpha}{2})}$$

où $k_{\frac{\alpha}{2}}, k_{(1-\frac{\alpha}{2})}$ désignent les quantiles d'ordre $\frac{\alpha}{2}$ et $(1-\frac{\alpha}{2})$ de la χ_n^2.

6. Un chi 2 à n degrés de liberté est défini par :

$$\chi_n^2 = \sum_{i=1}^{n} U_i^2$$

avec les $U_i \sim \mathcal{N}(0,1)$ indépendantes

L'intervalle de probabilité pour D au niveau $1-\alpha$ est donc :

$$k_{\frac{\alpha}{2}} \frac{\sigma^2}{n} < D < k_{(1-\frac{\alpha}{2})} \frac{\sigma^2}{n}$$

L'intervalle de confiance pour σ^2 au niveau $1-\alpha$ est alors :

$$\boxed{\frac{nd}{k_{(1-\frac{\alpha}{2})}} < \sigma^2 < \frac{nd}{k_{\frac{\alpha}{2}}}}$$

Quand m est inconnue

On utilise cette fois S^2 comme estim. de σ^2, et la fonction pivotale $W = \frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$.

L'intervalle de probabilité pour W au niveau $1-\alpha$ est :

$$k_{\frac{\alpha}{2}} < \frac{nS^2}{\sigma^2} < k_{(1-\frac{\alpha}{2})}$$

où $k_{\frac{\alpha}{2}}, k_{(1-\frac{\alpha}{2})}$ désignent cette fois [7] les quantiles d'ordre $\frac{\alpha}{2}$ et $(1-\frac{\alpha}{2})$ de la χ^2_{n-1}.

7. On ne précise pas le degré de liberté pour ne pas alourdir les notations donc attention...

L'intervalle de probabilité pour S^2 au niveau $1-\alpha$ est donc :

$$k_{\frac{\alpha}{2}} \frac{\sigma^2}{n} < S^2 < k_{(1-\frac{\alpha}{2})} \frac{\sigma^2}{n}$$

L'intervalle de confiance pour σ^2 au niveau $1-\alpha$ est alors :

$$\boxed{\frac{ns^2}{k_{(1-\frac{\alpha}{2})}} < \sigma^2 < \frac{ns^2}{k_{\frac{\alpha}{2}}}}$$

Quand l'échantillon n'est plus gaussien mais de grande taille

Si la loi de X est unimodale pas trop dissymétrique , on utilise la fonction asympt. pivotale $W = \dfrac{S^{*2} - \sigma^2}{\sqrt{\frac{2S^{*4}}{n-1}}} \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,1)$

En mesurant la quantité d'alcool (gr/l) contenue dans 10 cidres doux du marché, on obtient :

$$5,42 - 5,55 - 5,61 - 5,91 - 5,93 - 6,15 - 6,20 - 6,79 - 7,07 - 7,37$$

Estimer la variance par intervalle de confiance au niveau 95% en supposant la quantité d'alcool gaussienne.

Supposons l'échantillon de grande taille et estimons p.

Le nombre d'individus nF possédant le caractère étudié dans l'échantillon suit une loi binomiale $\mathscr{B}(n,p)$ donc, si n est grand, l'approximation d'une binomiale par une gaussienne [8] fournit :

$$nF \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}\left(np, \sqrt{np(1-p)}\right)$$

ou encore :

$$F \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

8. Le TCL s'exprime aussi sous la forme :

$$\bar{X} \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}\left(m, \frac{\sigma}{\sqrt{n}}\right)$$

ou encore :

$$\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}\left(nm, \sqrt{n\sigma^2}\right)$$

Dans le cas où $X \sim \mathscr{B}(p)$, on obtient donc l'approximation d'une binomiale par une gaussienne :

$$\mathscr{B}(n,p) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}\left(np, \sqrt{np(1-p)}\right)$$

On utilise alors F le meilleur estimateur de p, et la fonction asymptotiquement pivotale $W = \dfrac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,1)$.

L'intervalle de probabilité (asymptotique) pour F au niveau $1-\alpha$ est donc :

$$p - u_{(1-\frac{\alpha}{2})}\sqrt{\frac{p(1-p)}{n}} < F < p + u_{(1-\frac{\alpha}{2})}\sqrt{\frac{p(1-p)}{n}}$$

L'intervalle de confiance (asymptotique) pour p au niveau $1-\alpha$ est alors[9] :

$$\boxed{f - u_{(1-\frac{\alpha}{2})}\sqrt{\frac{f(1-f)}{n}} < p < f + u_{(1-\frac{\alpha}{2})}\sqrt{\frac{f(1-f)}{n}}}$$

- Il s'agit d'un intervalle de confiance asymptotique

9. Admis (utilisation des coniques)

Un échantillon de 100 votants choisis au hasard parmi tous les votants d'une circonscription a montré que 55% d'entre eux étaient favorables à un certain candidat.

1. Estimer la proportion de votants favorables à ce candidat par intervalle de confiance au niveau 95%.
2. Déterminer la taille de l'échantillon minimal pour assurer, au niveau 95%, une incertitude [a] n'excédant pas 0,02.

a. Il s'agit de la demi-longueur de l'intervalle de confiance $f \pm u_{(1-\frac{\alpha}{2})}\sqrt{\frac{f(1-f)}{n}}$ c'est à dire $u_{(1-\frac{\alpha}{2})}\sqrt{\frac{f(1-f)}{n}}$