Université Paris 1-Panthéon Sorbonne, UFR 02,
Licence de Sciences Economiques
STATISTIQUE, cours de Mme PRADEL

II- Estimation et prévision par intervalle

Définition

Pour un nombre η fixé entre 0 et 1, à toute réalisation $(x_1, ... x_n)$ de l'échantillon on associe un intervalle $I(x_1, ..., x_n)$ de telle façon que pour toute valeur θ du paramètre, la probabilité que l'intervalle aléatoire $I(X_1, ..., X_n)$ contienne θ soit supérieure ou égale à η :

$$\forall\ \theta,\ P_\theta\{I(X_1, ..., X_n) \ni \theta\} \geq \eta$$

- Le nombre η est le niveau de confiance de l'intervalle.
- $I(X_1, ... X_n)$ est un intervalle de confiance de niveau η.
- Parmi les intervalles de confiance de niveau η, on recherchera ceux de longueur minimum.

Une méthode générale d'estimation par intervalle

On choisit une statistique "bon" estimateur de θ, à partir de laquelle on construit une fonction des observations et de θ, notée $Z(X_1, ..., X_n,\theta)$, dont la loi est entièrement connue lorsque θ est donné.
Pour chaque θ, on calcule un intervalle $W_\theta = [a, b]$ pour Z, de probabilité η

$$\forall\ \theta \qquad P_\theta\{Z \in W_\theta\}=\eta$$

soit $$\forall\ \theta \qquad P_\theta\{a(\theta) \leq Z(X_1, ..., X_n,\theta) \leq b(\theta)\} = \eta$$

En résolvant ces inégalités par rapport à θ, si $a(\theta)$ et $b(\theta)$ sont des fonctions monotones de θ, on obtient un intervalle aléatoire qui a la probabilité η de recouvrir θ :

$$a(\theta) \leq Z(X_1, ..., X_n) \leq b(\theta) \Leftrightarrow A(X_1, ..., X_n) \leq \theta \leq B(X_1, ... X_n)$$

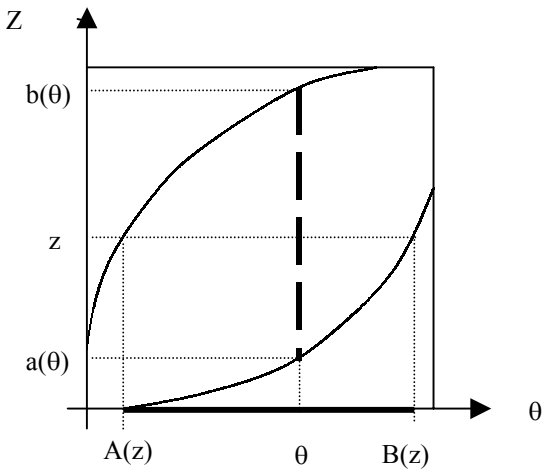
soit $$\forall\ \theta \qquad P_\theta\{A(X_1, ... X_n) \leq \theta \leq B(X_1, ..., X_n)\} = P_\theta\{a(\theta) \leq Z \leq b(\theta)\} = \eta$$

L'intervalle $[A(X_1, ..., X_n), B(X_1, ..., X_n)]$ est l'intervalle $I(X_1, ..., X_n)$ cherché.

Remarque

Si, la loi de Z étant discrète, il n'est pas possible d'obtenir un intervalle de probabilité exactement égale à η, on choisit W_θ de façon que sa probabilité soit supérieure à η, et la plus voisine possible de η (de façon à ce que sa longueur soit la plus petite possible).

Représentation graphique



Le graphique est fait dans le cas où Z est uniquement fonction des $X_1, ..., X_n$.

« Résoudre les inégalités » $a(\theta) \leq Z \leq b(\theta)$ par rapport à θ pour $Z = z$ observé revient alors à trouver l'ensemble des θ tels que $a(\theta) \leq z \leq b(\theta)$.

Remarquer que lorsque les bornes $a(\theta)$ et $b(\theta)$ sont croissantes en θ comme sur notre exemple, c'est la courbe des $b(\theta)$ qui conduit à la borne inférieure A(Z) de l'intervalle de confiance, et celle des $a(\theta)$ qui conduit à la borne supérieure B(Z).

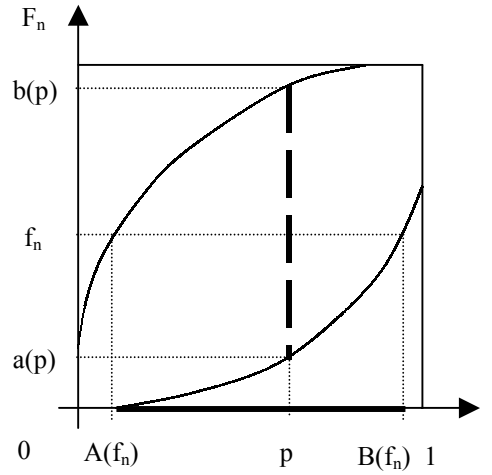
Remarque

Quand on aura vu la théorie des tests, pour chaque θ_o on prendra pour intervalle $[a(\theta_0), b(\theta_0)]$ la région d'acceptation de θ_0 pour tester $\theta = \theta_0$ contre $\theta \neq \theta_0$ au seuil 1 - η. L'intervalle de confiance apparaît alors comme l'ensemble des valeurs de θ qui auraient été acceptées au vu de l'échantillon observé au seuil (1-η).

Intervalle de confiance pour une proportion (paramètre d'une loi de BERNOULLI).

$X_1, ..., X_n$ est un n-échantillon d'une loi de Bernoulli B(1, p). Le "meilleur" estimateur de p est $F_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$

L'intervalle $V(p) = [a(p), b(p)]$ pour F_n , de probabilité η, est tel que : $P\{a(p) \le F_n \le b(p)\} = \eta$



Si n est assez grand (c'est-à-dire $np(1-p) > 15$) :

$$F_n \# N\left(p, \frac{p(1-p)}{n}\right).$$

L'intervalle de longueur minimum de probabilité η est centré autour de l'espérance mathématique p.
En lisant dans la table de la loi normale centrée réduite on lit t tel que $P\{|U| \le t\} = \eta$, et on en déduit :

$$P\left\{-t \le \frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}} \le t\right\} = \eta$$

D'où l'intervalle $W(p) = \left[p - t\sqrt{\frac{p(1-p)}{n}}; p + t\sqrt{\frac{p(1-p)}{n}}\right]$

La résolution numérique des inégalités conduit à $A(f_n) < p < B(f_n)$ où les deux bornes sont les racines d'une équation du second degré en p : $(n + t_\eta^2)\, p^2 - (2nF_n + t_\eta^2)\, p + nF_n^2 \le 0$.
En construisant point par point les courbes représentatives de

$$a(p) = p - t\sqrt{\frac{p(1-p)}{n}} \quad \text{et} \quad b(p) = p + t\sqrt{\frac{p(1-p)}{n}},$$

(ce sont des portions d'ellipses) on peut en déduire graphiquement pour f_n donné les limites de l'intervalle de confiance $I(x_1 ... x_n)$. On dispose d'abaques donnant ces courbes pour quelques valeurs de η, et quelques valeurs de n. On lit dans ce cas directement sur l'abaque l'intervalle correspondant du f_n observé.
Si on ne dispose pas d'abaque, on peut simplifier les calculs en remplaçant dans l'expression de a et b l'élément $\sqrt{\frac{p(1-p)}{n}}$ par une valeur approchée (seule la méthode approchée est au programme de licence).

a) On remplace $p(1-p)$ par son estimateur $F_n(1-F_n)$:

$$\forall p : P_p\left\{-t \le \frac{F_n - p}{\sqrt{\frac{F_n(I-F_n)}{n}}} \le t\right\} \cong \eta$$

$$\forall p : P_p\left\{F_n - t\sqrt{\frac{F_n(1-F_n)}{n}} \le p \le F_n + t\sqrt{\frac{F_n(1-F_n)}{n}}\right\} \cong \eta$$

$$I(X_1,..., X_n) = \left[F_n - t\sqrt{\frac{F_n(1-F_n)}{n}}, F_n + t\sqrt{\frac{F_n(1-F_n)}{n}}\right]$$ de niveau de confiance voisin de η.

b) On remplace $p(1-p)$ par sa valeur maximum 1/4.

$$\forall p : P_p\left\{-t \le \frac{F_n - p}{\sqrt{\frac{0{,}25}{n}}} \le t\right\} \ge \eta$$

$$\forall p : P_p\left\{F_n - t\frac{0{,}5}{\sqrt{n}} \le p \le F_n + t\frac{0{,}5}{\sqrt{n}}\right\} \ge \eta$$

$I(X_1,...,X_n) = \left[F_n - t\frac{0{,}5}{\sqrt{n}}, F_n + t\frac{0{,}5}{\sqrt{n}}\right]$ de niveau de confiance supérieur ou égal à η.

Intervalle de confiance pour l'espérance mathématique d'une loi normale

1°) Lorsque la variance σ^2 est connue

L'estimateur de m est $\bar{X}$ qui suit une loi $N(m, \frac{\sigma^2}{n})$. L'intervalle V_m = [a(m), b(m)] pour $\bar{X}$, de probabilité η est tel que P{a(m) ≤ $\bar{X}$ ≤ b(m)} = η.
On lit dans la table de la loi normale N(0,1) pour n donné le nombre u tel que

$$P\left\{-u \le \frac{X - m}{\sigma/\sqrt{n}} \le u\right\} = \eta$$

soit ∀m,

$$P\left\{m - u\frac{\sigma}{\sqrt{n}} \le \bar{X} \le m + u\frac{\sigma}{\sqrt{n}}\right\} = \eta$$

on en déduit l'intervalle de confiance η pour m :

$$I = \left[\bar{X} - u\frac{\sigma}{\sqrt{n}}; \bar{X} + u\frac{\sigma}{\sqrt{n}}\right]$$

Pour η = 95%, par exemple, u = 1.96
Puisque $a(m)$ et $b(m)$ sont ici des fonctions linéaires de m, leurs représentations sont des droites et la résolution des inéquations est faite directement sans nécessiter de graphique.

2°) Lorsque la variance σ^2 est inconnue

En posant $S^2 = \frac{\sum_1^n (X_t - \bar{X})^2}{n-1}$, on sait que $\frac{\bar{X} - m}{S/\sqrt{n}}$ suit une loi de STUDENT à n-1 degrés de liberté.

On lit dans la table de la loi de Student, pour η fixé, le nombre t tel que $F\left\{-t \le \frac{\bar{X} - m}{S/\sqrt{n}} \le t\right\} = \eta$

L'intervalle de confiance pour m est donc $I = \left[\bar{X} - t\frac{S}{\sqrt{n}}, \bar{X} + t\frac{S}{\sqrt{n}}\right]$

Intervalle de prévision d'une variable aléatoire X

Comme pour la prévision ponctuelle, on traite le cas où la variable à prévoir n'est pas corrélée avec les observations qui en ont été faites.
On désire calculer un intervalle de probabilité η donnée pour $Y_o \approx N(m ; \sigma^2)$

- si m et σ^2 sont connus : intervalle de probabilité, dont les bornes sont non aléatoires
- si m et σ^2 sont estimés : intervalle de prévision, dont les bornes seront aléatoires.

La seule intervention de l'information dans la prévision sera de permettre l'estimation des paramètres définissant la loi de Y_o.

Prévision d'une variable normale $Y_o \approx N(m ; \sigma^2)$

$$U = \frac{Y_o - m}{\sigma} \approx N(0;1)$$

Pour un niveau de confiance η, $P\{|U| \le u_\eta\} = \eta$

$$P\left\{-u_\eta \le \frac{Y_o - m}{\sigma} \le u_\eta\right\} = \eta \qquad \Leftrightarrow \qquad P\left\{m - u_\eta\sigma \le Y_o \le m + u_\eta\sigma\right\} = \eta$$

L'intervalle de probabilité η peut ainsi être interprété comme un intervalle de prévision de niveau η pour Y_o

Espérance et variances inconnues pour $Y_o \approx N(m ; \sigma^2)$

Les paramètres m et σ^2 sont estimés à partir d'un n-échantillon de la loi $N(m ; \sigma^2)$, par

$$\overline{Y} = \frac{1}{n}\sum_1^n Y_i \quad \text{et} \quad S^2 = \frac{\sum_1^n (Y_i - \overline{Y})^2}{n-1},$$

La variable U ne peut plus être utilisée, mais si on cherche à remplacer m et σ^2 par leurs estimateurs, on peut définir une variable de loi connue :

(a) l'indépendance de Y_o et de l'échantillon $(Y_1,.., Y_n)$ entraîne que :

$$Y_o - \overline{Y} \approx N\left(0; \sigma^2 + \frac{\sigma^2}{n}\right) \Rightarrow U = \frac{Y_o - \overline{Y}}{\sigma\sqrt{1 + \frac{1}{n}}} \approx N(0;1)$$

(b) la variance empirique S^2 est indépendante de Y_o comme de la moyenne empirique :

$$\frac{(n-1)S^2}{\sigma^2} \approx \chi^2(n-1) \text{ indépendante de } U$$

(c) la variable Z suit un Student à $(n-1)$ degrés de liberté :

$$Z = \frac{Y_o - \overline{Y}}{S\sqrt{1 + \frac{1}{n}}} \approx T(n-1)$$

Pour un niveau de confiance η, on lit dans la table de la loi de Student la valeur t_η ayant la probabilité η d'être dépassée en valeur absolue par une variable de Student à $(n-1)$ degrés de liberté. On en déduit l'intervalle de prévision de niveau η pour Y_o.

$$P\left\{\overline{Y} - t_\eta S\sqrt{1 + \frac{1}{n}} \le Y_o \le \overline{Y} + t_\eta S\sqrt{1 + \frac{1}{n}}\right\} = \eta$$

Prévision de la variable expliquée dans le cas d'un modèle linéaire standard normal

Le modèle postulé ici est :

$Y_o \approx N(a\, x_o + b ; \sigma^2)$

Pour $i = 1,\ldots, n$: $Y_i \approx$ i.i.d. $N(a\, x_i + b ; \sigma^2)$ indépendantes de Y_o.

On suppose également que les x_i ne sont pas toutes égales, et que a et b sont des réels quelconques a priori.

Le principe des calculs faits dans le cas précédent est encore valable, mais ici l'espérance mathématique de Y_o dépend de deux paramètres, qui sont estimés par la méthode des MCO (meilleur estimateur dans le cas d'un modèle linéaire standard comme c'est ici le cas). On sait alors que :

$$E(\hat{a}) = a \quad , E(\hat{b}) = b$$

$$V(\hat{a}) = \sigma^2 V_{11}\,,\ V(\hat{b}) = \sigma^2 V_{22}\,,\ \operatorname{cov}(\hat{a}, \hat{b}) = \sigma^2 V_{12}\,,$$

où les coefficients V_{11}, V_{12} et V_{22} sont des fonctions des $(x_1,\ldots, x_n)$, et la variance σ^2 est estimée par la variance résiduelle :

$$\hat{\sigma}^2 = \frac{SCR}{n-2} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{a}x_i - \hat{b}\right)^2}{n-2}$$

(a) $E\left(Y_o - \hat{a}x_o - \hat{b}\right) = ax_o + b - ax_o - b = 0$ et

$$V\left(Y_o - \hat{a}x_o - \hat{b}\right) = \sigma^2 + x_o^2 V(\hat{a}) + V\left(\hat{b}\right) + 2x_o \operatorname{cov}\left(\hat{a},\hat{b}\right) = \sigma^2\left[1 + x_o^2 V_{11} + V_{22} + 2x_o V_{12}\right]$$

l'indépendance de Y_o et de l'échantillon (Y_1,.., Y_n) entraîne que :

$$Y_o - \hat{a}x_o - \hat{b} \approx N\left(0;\sigma^2\left[1 + x_o^2 V_{11} + V_{22} + 2x_o V_{12}\right]\right) \Rightarrow U = \frac{Y_o - \hat{a}x_o - \hat{b}}{\sigma\sqrt{1 + x_o^2 V_{11} + V_{22} + 2x_o V_{12}}} \approx N(0;1)$$

(b) la variance résiduelle est indépendante de Y_o comme des estimateurs MCO de *a* et *b*, et elle n'a plus que *n*-2 degrés de liberté :

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \approx \chi^2(n-2) \text{ indépendante de } U$$

(c) la variable *Z* suit un Student à (*n*-2) degrés de liberté (voir définitions des lois dans lois03.pdf)

$$Z = \frac{Y_o - \hat{a}x_o - \hat{b}}{\hat{\sigma}\sqrt{1 + x_o^2 V_{11} + V_{22} + 2x_o V_{12}}} \approx T(n-2)$$

Pour un niveau de confiance η, on lit dans la table de la loi de Student la valeur t_η ayant la probabilité η d'être dépassée en valeur absolue par une variable de Student à (*n*-2) degrés de liberté. On en déduit l'intervalle de prévision de niveau η pour Y_o.

$$P\left\{\hat{a}x_o + \hat{b} - t_\eta\hat{\sigma}\sqrt{1 + x_o^2 V_{11} + V_{22} + 2x_o V_{12}} \le Y_o \le \hat{a}x_o + \hat{b} + t_\eta\hat{\sigma}\sqrt{1 + x_o^2 V_{11} + V_{22} + 2x_o V_{12}}\right\} = \eta$$

Si on note $\hat{\sigma}_a^2 = \hat{\sigma}^2 V_{11}$, $\hat{\sigma}_b^2 = \hat{\sigma}^2 V_{22}$ et $\hat{\sigma}_{ab} = \hat{\sigma}^2 V_{12}$ les variances et covariances estimées des estimateurs MCO de *a* et *b*, on peut aussi écrire :

$$P\left\{\hat{a}x_o + \hat{b} - t_\eta\sqrt{\hat{\sigma}^2 + x_o^2\hat{\sigma}_a^2 + \hat{\sigma}_b^2 + 2x_o\hat{\sigma}_{ab}} \le Y_o \le \hat{a}x_o + \hat{b} + t_\eta\sqrt{\hat{\sigma}^2 + x_o^2\hat{\sigma}_a^2 + \hat{\sigma}_b^2 + 2x_o\hat{\sigma}_{ab}}\right\} = \eta$$

Ces différentes estimations sont fournies par les logiciels de régression utilisés pour le calcul des MCO, excepté peut-être pour la covariance estimée des estimateurs de *a* et *b* (il faut alors expressément en demander l'affichage).