III - Estimation et Prévision par intervalle

Exemple : soit $Y_1,...,Y_N$ un N-échantillon N[m ;1]

- Estimation de m par $\overline{Y}_N = \frac{Y_1 + ... + Y_N}{N}$
- pour chaque m :

$$\overline{Y}_N \approx N\left(m ; \frac{1}{N}\right), \; U = \frac{\overline{Y}_N - m}{\sqrt{1/N}} \approx N(0,1)$$

$$P\left\{-1,96 \leq \frac{\overline{Y}_N - m}{\sqrt{1/N}} \leq 1,96\right\} = 0,95$$

$$P\left\{\overline{Y}_N - \frac{1,96}{\sqrt{N}} \leq m \leq \overline{Y}_N + \frac{1,96}{\sqrt{N}}\right\} = 0,95$$

28

Exemple :

Supposons que les notes d'un élève de terminale au cours de l'année forment un échantillon de taille N d'une loi Normale d'espérance m et d'écart-type égal à 1.

Comment estimer sa valeur m ? nous avons vu que la meilleure estimation est la moyenne empirique des notes obtenues. Quelle précision peut-on attendre de cette valeur ? Cela n'est pas donné par l'estimation ponctuelle.

Pour chaque valeur possible de m, on peut construire un intervalle de probabilité, disons 95% , pour la moyenne empirique. Le même événement peut aussi s'écrire sous la forme d'un encadrement pour m.

L'intervalle obtenu est aléatoire, car ses bornes dépendent de la moyenne empirique des notes obtenues.

$$N = 16 \ : \ P\{\overline{Y}_N - 0{,}490 \le m \le \overline{Y}_N + 0{,}490\} = 0{,}95$$
$$N = 100 \ : \ P\{\overline{Y}_N - 0{,}196 \le m \le \overline{Y}_N + 0{,}196\} = 0{,}95$$
$$N = 1000 \ : \ P\{\overline{Y}_N - 0{,}062 \le m \le \overline{Y}_N + 0{,}062\} = 0{,}95$$

Ces intervalles sont dits

« intervalles de confiance 95% »

pour $\overline{y}_N = 8{,}543$, confiance 95% :

$N = 16 \quad : 8{,}053 \le m \le 9{,}033$

$N = 100 \quad : 8{,}347 \le m \le 8{,}739$

$N = 1000 : 8{,}481 \le m \le 8{,}605$

29

Nous voyons que lorsque le nombre N augmente, la longueur de l'intervalle diminue : l'information sur *m* est plus précise.

Si nous construisons, pour un N donné, les intervalles de confiance croissante, nous constatons que la valeur lue sur la table augmente : la longueur de l'intervalle augmente. Cette graduation fait peut-être mieux voir que les deux bornes ne sont pas des limites pour *m*, mais correspondent à une certaine confiance. Plus la confiance augmente, moins l'intervalle est précis.

Pour l'observation d'un échantillon donné, il correspond un nombre, réalisation de la statistique « moyenne des yn » : on calcule alors un intervalle numérique, réalisation de l'intervalle aléatoire. On ne peut plus dire que cet intervalle numérique a la probabilité 95% de recouvrir la valeur *m* (c'est au cours de multiples tirages d'échantillons de taille N que nous aurons 95% de chance d'obtenir un intervalle qui recouvre *m*).

Récapitulons :

- Pour une confiance η fixée entre 0 et 1 : $I(Y_1,...,Y_N) = [A(Y_1,...,Y_N), B(Y_1,...,Y_N)]$, intervalle aléatoire tel que pour toute loi P_θ :

  $P_\theta\{A(Y_1,...,Y_N) < \theta < B(Y_1,...,Y_N)\} \geq \eta$
- η est le niveau de confiance de l'intervalle
- $I(Y_1,...,Y_N)$ est un intervalle de confiance η
- Pour le construire, nous avons utilisé une fonction de $(Y_1,...,Y_N)$ et de θ dont la loi est complètement connue.

30

Soit la fonction $Z(y_1,...,y_n, \theta)$, dont la loi est complètement connue lorsque θ est connu.

Pour tout θ, on calcule un intervalle de probabilité η ;

$$\forall\ \theta,\ P_\theta\{a(\theta) < Z(y_1,...,y_n, \theta) < b(\theta)\} = \eta$$

On résout ces inégalités en θ (calcul ou graphique). On obtient généralement un intervalle aléatoire qui a la probabilité η de recouvrir θ :

$$\forall\theta\ \ P_\theta\{A(y_1,...,y_n) \leq \theta \leq B(y_1,...,y_n)\} = \eta$$

Si, la loi de Z étant discrète, il n'est pas possible d'obtenir un intervalle de probabilité exactement égale à η, on choisira l'intervalle [a, b] de probabilité supérieure à η et la plus proche possible de η (pour avoir la longueur la plus petite possible pour un niveau de confiance $\Delta\ \eta$).

.

Construction

• Trouver une variable aléatoire fonction des observations (et du paramètre), dont la loi est entièrement connue : $Z(Y_1,...,Y_N; \theta)$

• Pour chaque θ, calculer un intervalle de probabilité η pour Z :

$$P_\theta\{a(\theta) \le Z(Y_1,...,Y_N;\theta) \le b(\theta)\} = \eta$$

• Résoudre les inégalités en θ :

$$P_\theta\{A(Y_1,..,Y_N) \le \theta \le B(Y_1,..,Y_N)\} = \eta$$

31

Attention : quand on parle de la "probabilité que l'intervalle recouvre θ", il s'agit de la probabilité calculée en θ. C'est ce que signifie la notation P_θ.

La variable Z est construite à partir d'un bon estimateur de θ.

Autre exemple : soit un N-échantillon d'une loi exponentielle de densité $f(y;a) = (1/a)\exp(-y/a)$ pour $y > 0$: $E(Y) = a$, $V(Y) = a^2$. L'intervalle de confiance pour a est construit à partir de la moyenne empirique de l'échantillon, qui est le meilleur estimateur de a. Nous utilisons la loi approchée de cette moyenne empirique, pour N=100 et une confiance 90% :

$$\overline{Y}_{100} \approx N\left[a\,;\frac{a^2}{100}\right]$$

$$Z = \frac{\overline{Y}_{100} - a}{a/\sqrt{100}} \approx N[0;1]$$

$$P\left\{-1,645 < \frac{\overline{Y}_{100} - a}{a/10} < 1,645\right\} = 0,90$$

$$\Leftrightarrow P\left\{-1,645\ \frac{a}{10} < \overline{Y}_{100} - a < 1,645\ \frac{a}{10}\right\} = 0,90$$

$$\Leftrightarrow P\left\{\frac{\overline{Y}_{100}}{1,1645} < a < \frac{\overline{Y}_{100}}{0,8355}\right\} = 0,90$$

III.1. Espérance d'une loi Normale, variance connue

- $Y_1,\ldots, Y_N \approx$ i.i.d.N(m,σ^2), où σ^2 est connue.

$$\overline{Y}_N = \frac{Y_1 + \ldots + Y_N}{N} \approx N\left(m; \frac{\sigma^2}{N}\right) \qquad U = \frac{\overline{Y}_N - m}{\sigma/\sqrt{N}} \approx N[0;1]$$

$$P\left\{-u_\eta \le \frac{\overline{Y}_N - m}{\sigma/\sqrt{N}} \le u_\eta\right\} = \eta \Leftrightarrow P\left\{\overline{Y}_N - u_\eta \frac{\sigma}{\sqrt{N}} \le m \le \overline{Y}_N + u_\eta \frac{\sigma}{\sqrt{N}}\right\} = \eta$$

Application numérique : $\eta = 90\%$, $N = 100$, $\sigma = 2$

$u_\eta \frac{\sigma}{\sqrt{N}} = 1{,}645 \frac{2}{10} = 0{,}329$ $\quad P\{\overline{Y}_N - 0{,}329 \le m \le \overline{Y}_N + 0{,}329\} = 0{,}90$

observation $\overline{y} = 5{,}685 \Rightarrow$ intervalle réalisé : $5{,}356 \le m \le 6{,}014$

32

U est la variable centrée réduite construite à partir de ΣY/n en lui retranchant son espérance m et en divisant le tout par son écart type σ/√ n

Pour une probabilité η de 90%, on lit dans la table de la loi Normale centrée réduite :

P{-1,645 < U < 1,645} = 0.90

Donc $u_{0.90} = 1.645$

Remarquer que c'est la condition $Z(m) < u$ qui se résout en $A(Z) < m$: si nous voulions n'obtenir qu'une borne supérieure pour m, il faudrait calculer une borne inférieure pour Z.

III.2. Intervalle de confiance pour une proportion théorique

- $Y_1,..., Y_N \approx$ i.i.d. B(*1*,*p*)
- *p* estimé par la fréquence empirique

$$F_N = (Y_1 + ... + Y_N)/N$$

$$Z = \frac{F_N - p}{\sqrt{\frac{F_N(1-F_N)}{N}}} \# N(0,1) \quad \text{dès que} \quad np(1-p) \geq 15$$

$$P\left\{ F_N - \frac{t_\eta}{\sqrt{N}}\sqrt{F_N(1-F_N)} \leq p \leq F_N + \frac{t_\eta}{\sqrt{N}}\sqrt{F_N(1-F_N)} \right\} \cong \eta$$

33

Intervalle de probabilité pour la fréquence observée :

$P\{a(p) \leq F_n \leq b(p)\} \geq \eta$

1°) Lorsque $npq \leq 5$, il faut calculer $a(p)$ et $b(p)$ point par point, en utilisant la loi Binomiale: c'est une loi discrète . Il existe des graphiques représentant les courbes $a(p)$ et $b(p)$ en fonction de p.

2°) Si $npq > 5$, on utilise la loi normale approchée de F_n

L'intervalle de probabilité η pour la variable normale centrée réduite est

$P\{- t_\eta \leq U \leq t_\eta\} = \eta$ (symétrique autour de 0 car la loi est symétrique).

On en déduit l'intervalle de probabilité pour F_n.

p intervient de façon non linéaire dans les bornes $a(p)$ et $b(p)$.

- On peut résoudre analytiquement les inégalités (ce n'est pas au programme de la licence) :

$n(F_n - p)^2 \leq t_\eta^2 p(1 - p)$

$(n + t_\eta^2)p^2 - (2nF_n + t_\eta^2)p + nF_n^2 \leq 0$: p est entre les racines de l'équation.

- On peut aussi remplacer $p(1-p)$ par 1/4, car $0 < p < 1$ entraîne

$0 < p(1-p) < 1/4$,

et l'intervalle obtenu est de probabilité $> \eta$. $\quad P\left\{ p - \frac{t_\eta}{2\sqrt{n}} \leq F_n \leq p + \frac{t_\eta}{2\sqrt{n}} \right\} \geq \eta$

3°) Si $npq > 15$, alors $p(1-p)$ peut être remplacé par $F_n(1-F_n)$ dans la variance de F_n. Les bornes obtenues sont linéaires en p, et la résolution analytique est immédiate.

Exemple

- Sur 150 entreprises observées, 30 ont dû cesser leurs activités au cours de la 1ère année.
- Donner un intervalle de confiance 0,95 pour la probabilité de cessation totale d'une entreprise au cours de sa première année.

$$P\{1,96 \le U \le 1,96\} = 0,95 \; ; \; \frac{1,96}{\sqrt{150}} = 0,160$$

$$P\left\{F_N - 0,160\sqrt{F_N(1-F_N)} \le p \le F_N + 0,160\sqrt{F_N(1-F_N)}\right\} \cong 0,95$$

$$\text{observation : } f_N = \frac{30}{150} = 0,20 \quad \Rightarrow \quad 0,1360 < p < 0,2640$$

$$\text{Vérification : } 150 * 0,136 * 0,864 = 17,6 > 15$$

34

Seul le calcul approché est au programme de la licence. L'intervalle de confiance égale à environ 95% est donc calculé en remplaçant la variance $p(1-p)$ par son estimation..

Les quantités aléatoires dans les bornes de l'intervalle sont celles qui dépendent de la fréquence observée F_N. Les autres termes sont calculables **avant** l'observation de l'échantillon : on obtient ainsi l'expression exacte de l'intervalle aléatoire, qui a la probabilité 95% de recouvrir la vraie valeur de p.

Une fois l'échantillon tiré, on calcule numériquement la réalisation de l'intervalle aléatoire : il n'y a plus de proba pour l'événement

$$0,136 < p < 0,264$$

puisque plus rien n'est aléatoire. L'événement est vrai ou faux (proba 1 ou 0) : nous ne le saurons généralement jamais…

Il faut pour finir vérifier que la condition d'approximation par la loi Normale est bien satisfaite en tout point de l'intervalle. La valeur la moins favorable est celle qui est la plus éloignée de 0,5 : ici, c'est 0,136. L'approximation est acceptable.

III.3. Espérance d'une loi Normale, variance inconnue

$Y_1,..., Y_N \approx$ i.i.d.N(m,σ^2), σ^2 inconnue.

$$\overline{Y_N} = \frac{Y_1 + ... + Y_N}{N} \qquad U = \frac{\overline{Y_N} - m}{\sigma/\sqrt{N}} \approx N(0;1)$$

$$S^2 = \frac{\sum_{n=1}^{N} \left(Y_n - \overline{Y_N}\right)^2}{N-1} \qquad (N-1)\frac{S^2}{\sigma^2} \approx \text{CHI - DEUX}(N-1)$$

$$S \text{ et } U \text{ indépendants} \Rightarrow T = \frac{\overline{Y_N} - m}{S/\sqrt{N}} \approx \text{STUDENT}(N-1)$$

(voir définitions des lois associées à la loi Normale, notes de cours « lois03.pdf »)

35

La statistique "moyenne empirique" ne peut être utilisée lorsque la variance n'est pas connue.

On remplace alors σ par s, en utilisant la variance empirique non biaisée.

La variable T obtenue ne dépend que de *m* et des observations, et a une loi dite "de Student à *N*-1 degrés de liberté".

Cette loi est tabulée, elle ressemble à la loi normale, symétrique et unimodale, mais elle est plus aplatie : l'intervalle symétrique correspondant à un niveau donné est plus large que pour la loi Normale.

Lorsque N tend vers l'infini, la loi de Student tend vers la loi Normale.

Pourquoi « N-1 » ? ce nombre représente le degré de liberté de la statistique S^2 .

En effet, S^2 est la somme des carrés de N termes, mais la somme de ces termes est nulle : ils sont donc liés par une relation (et une seule) et n'ont plus que (N-1) degrés de liberté.

Nous retrouverons cette notion de degrés de liberté dans le cas des modèles linéaires (ici, il n'y a pas de variable explicative autre que la constante).

Student à k degrés de liberté

- Par définition :

$$\begin{array}{l} U \approx N[0;1] \\ Z \approx CHI-DEUX(k) \\ U, Z \text{ indépendants} \end{array} \Rightarrow T = \frac{U}{\sqrt{Z/k}} \approx \text{STUDENT(k)}$$

- loi symétrique, en cloche comme la loi normale, mais plus aplatie.
- STUDENT(k) # N[0;1] lorsque k est grand

36

La loi de Chi-deux est, elle aussi, une loi liée à la loi normale :

Un « chi-deux à k degrés de liberté » est, par définition, la loi de la somme des carrés de k variables N[0;1] indépendantes. Cette loi se retrouve également dans certaines sommes de carrés issues de N variables N[0;1] indépendantes. Cela fera l'objet d'un théorème général dans le cas du modèle linéaire standard normal.

La loi de chi-deux est une loi dont la densité est >0 pour x > 0, unimodale.

Si Z ~ CHI-DEUX(k), alors

$E(Z) = k$

$V(Z) = 2k$

La définition de la loi de Student est étudiée pour justement donner la loi de la statistique obtenue en remplaçant σ^2 inconnu par son estimation sans biais.

Ce qui est moins trivial est que cette loi sera également obtenu dans le cadre du modèle linéaire standard normal, lorsque l'espérance *m* dépend linéairement de K paramètres. Le degré de liberté de la variance estimée sera alors N-K au lieu de N-1.

Exemple

Application numérique : $\eta = 90\%$, N = 100, $S^2 = 3{,}61$

$N = 100 \Rightarrow$ STUDENT(99) $\cong$ N[0;1]

$t_\eta \frac{1}{\sqrt{N}} = 1{,}65 \frac{1}{10} = 0{,}165$

formule EXCEL : loi.student.inverse(0,90;99) = 1,66

$$P\left\{\overline{Y}_N - 0{,}165\,.S \le m \le \overline{Y}_N + 0{,}165\,.S\right\} = 0{,}90$$

observation :

$\overline{y} = 5{,}685$ et $s^2 = 3{,}61 \Rightarrow$ intervalle réalisé : $5{,}37 \le m \le 6{,}00$

37

L'intervalle est calculé **avant l'observation** de l'échantillon : la confiance et la taille de l'échantillon sont connues à l'avance, mais les statistiques « moyenne empirique » et « variance empirique » sont aléatoires (donc ne sont connues qu'après le tirage de l'échantillon, et elles changent si on fait un deuxième tirage).

Après observation de l'échantillon, on peut calculer la réalisation correspondante de l'intervalle de confiance en remplaçant les statistiques d'échantillon par les valeurs numériques correspondant au tirage effectué. Il n'y a plus de proba…

Intervalles de confiance pour l'espérance d'une loi Normale

effectif = 8 espérance = 51 moyenne = 50,05
dl = 7 σ = 1,5 S = 2,25

u(η)	A(Z)	B(Z)	longueur
95%			
1,960	49,011< m < 51,089		2,079
98%			
2,326	48,816 < m < 51,284		2,467
95%			
1,645	m < 50,922		
98%			
2,054	m < 51,139		

t(η)	A(Z)	B(Z)	longueur
95%			
2,365	48,796 < m < 51,304		2,508
98%			
2,998	48,460 < m < 51,640		3,180
95%			
1,895	m < 51,055		
98%			
2,517	m < 51,385		

38

Les vraies valeurs des paramètres *m* et σ^2 sont connues ici, car il s'agit d'une simulation, les 8 valeurs étant tirées d'une loi Normale d'espérance 51 et d'écart-type 1.5. On fait comme si on ne connaissait pas *m*, et on envisage successivement les deux cas, variance connue ou inconnue.

Dans la première moitié gauche du tableau, les résultats concernent le cas où σ^2 est connue, alors que dans la moitié droite figurent ceux du cas où σ^2 n 'est pas connue, mais estimée par $(2.25)^2$.

Remarques :

• la longueur de l'intervalle de confiance est connue dans le premier cas, et elle est aléatoire dans le second cas.

• Le manque d'information se traduit par une plus grande incertitude : pour un même niveau de confiance, les valeurs de u(η) de la loi Normale sont plus petites que les valeurs de t(η) de la loi de Student à 7 degrés de liberté (échantillon de taille 8).

• Avec une confiance 95%, la borne supérieure calculée pour *m* est 50.92 : nous constatons que la réalisation de l'intervalle ne recouvre pas la vraie valeur de *m* (51). Cela ne remet pas en cause les qualités de l'intervalle ainsi construit : nous avons accepté une proba de 5% pour que cet intervalle ne recouvre pas *m*. Notre tirage fait partie de ces 5%. Too bad…

• Plus on augmente la confiance, plus la longueur de l'intervalle augmente, et donc la précision du renseignement diminue. (la seule façon de ne pas se tromper, c'est de ne rien dire, mais cela ne tient aucun compte de l'information que constitue l'observation de 8 tirages de y). Il est toujours indispensable de préciser quel est le niveau de confiance de l'intervalle construit

III.4. Modèle linéaire standard normal : intervalle de confiance pour un coefficient

$Y_1,\ldots,Y_N$ indépendants $Y_n \approx \mathrm{N}\left(a\,x_n + b, \sigma^2\right)$

$$\text{MCO} : \hat{\beta} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} \approx N\left[\begin{pmatrix} a \\ b \end{pmatrix} ; \sigma^2 \begin{pmatrix} m_{aa} & m_{ab} \\ m_{ab} & m_{bb} \end{pmatrix} \right]$$

$$\frac{\sum_{n=1}^{N}\left(Y_n - \hat{a}x_n - \hat{b}\right)^2}{\sigma^2} = \frac{SCR}{\sigma^2} \approx \chi^2_{(N-2)}$$

2 = dimension du paramètre (a,b) estimé par les MCO dans $E(Y_n)$

$\hat{\beta}$ et SCR sont indépendants en probabilité

Ces résultats ne sont valables que dans le cas du modèle linéaire standard Normal. Notamment, s'il y a hétéroscédasticité ou auto corrélation des perturbations (Y_i - ax_i - b), alors ces propriétés ne sont plus vérifiées.

La statistique utilisée pour construire l'intervalle de confiance est la variable centrée "réduite" où la variance est estimée sans biais par SCR/(N-2).

Le degré de liberté est maintenant (N-2) : c'est un résultat que nous ne démontrerons pas, mais pour le retenir, notons qu'il y a deux paramètres estimés, les N termes d'erreurs sont liés par les 2 équations normales qui ont servi à calculer les estimations MCO de a et b, et n'ont plus que (N-2) degrés de liberté :

$$\sum_{n=1}^{N}\left(Y_n - \hat{a}x_n - \hat{b}\right)x_n = 0 \quad \text{et} \quad \sum_{n=1}^{N}\left(Y_n - \hat{a}x_n - \hat{b}\right) = 0$$

pour N = 25, η = 95% :

Normale u = 1,960 (variance connue).

Student(24) : t = 2,064 (variance estimée, 1 paramètre estimé dans *m*).

Student(23) : t = 2,069 (variance estimée, 2 paramètres estimés dans *m*).

Pour une confiance donnée, l'imprécision augmente lorsque l'information diminue.

on en déduit un estimateur de la variance σ^2 et donc des variances des estimateurs de a et b

$$\hat{\sigma}^2 = \frac{SCR}{N-2}\ ;\ S_a^2 = \hat{\sigma}^2 m_{aa}\ ;\ S_b^2 = \hat{\sigma}^2 m_{bb}$$

(Rappel : $m_{aa} = \frac{1}{\sum_{n=1}^{N} (x_n - \bar{x}_N)^2}$ et $m_{bb} = \frac{\frac{1}{N}\sum_{n=1}^{N} x_n^2}{\sum_{n=1}^{N} (x_n - \bar{x}_N)^2}$)

et on en tire :

$$T_a = \frac{\hat{a} - a}{S_a} = \frac{\hat{a} - a}{\hat{\sigma}\sqrt{m_{aa}}} \approx STUDENT(N-2)$$

$$T_b = \frac{\hat{b} - b}{S_b} = \frac{\hat{b} - b}{\hat{\sigma}\sqrt{m_{bb}}} \approx STUDENT(N-2)$$

40

Nous avons vu que la matrice de variance-covariance de (â, ^b) est $\sigma^2 (X\,'X)^{-1}$.

nous constatons donc que les variances de ces statistiques sont proportionnelles à la variance σ^2 , les coefficients m_{aa} et m_{bb} étant les termes de la diagonale principale de la matrice $(X\,'X)^{-1}$.

En fait, ces calculs sont faits par un logiciel, qui donne de manière classique les écart-types estimés de chacun des coefficients. Il suffit donc d'utiliser ces écart-types s_a et s_a.

Le degré de liberté d'un student est celui du chi-deux figurant au dénominateur : ici, il s'agit de l'estimateur, SCR/(N-2) de σ^2 : il y a N résidus élevés au carré, mais ces N résidus sont liés par les 2 relations qui ont défini l'estmateur MCO. La SCR n'a plus que N-2 degrés de liberté.

La formule est parfaitement généralisable au cas d'un nombre K de variables explicatives : on obtient des Student à (N-K) degrés de liberté (ici encore, la SCR est la somme des carrés des N résidus, liés par les K équations définissant l'estimateur MCO : restent N-K degrés de liberté)

Intervalle de confiance pour a :

$$T_a = \frac{\hat{a} - a}{S_a} = \frac{\hat{a} - a}{\hat{\sigma}\sqrt{m_{aa}}} \approx STUDENT(N-2)$$

Lecture de t_η dans la table de Student (N-2)

$$P\left\{-t_\eta < \frac{\hat{a} - a}{S_a} < t_\eta\right\} = \eta$$

Intervalle de confiance η :

$$P\left\{\hat{a} - t_\eta S_a < a < \hat{a} + t_\eta S_a\right\} = \eta$$

41

Le reste du calcul est identique aux calculs précédents, et fournit les intervalles de confiance η% :

$$\hat{a} - t_\eta S_a \le a \le \hat{a} - t_\eta S_a$$

$$\hat{b} - t_\eta S_b \le b \le \hat{b} - t_\eta S_b$$

Les valeurs t_η sont les mêmes pour tous les coefficients : elles ne dépendent que du niveau de confiance choisi.

III.5. Prévision par intervalle d'une variable aléatoire Normale

- Comme pour la prévision ponctuelle, on suppose ici que la variable à prévoir est indépendante des observations qui en ont été faites.
- On désire calculer un intervalle de probabilité η donnée pour $Y_o \sim N(m ; \sigma^2)$
  m et σ^2 connus : intervalle de probabilité
  m et σ^2 estimés : intervalle de prévision

Intervalle de probabilité 95% pour *m* et σ^2 connus :

$$P(A \le Y_o \le B) = P\left(\frac{A-m}{\sigma} \le \frac{Y_o - m}{\sigma} \le \frac{B-m}{\sigma}\right)$$

L 'intervalle de longueur minimum est symétrique autour de l 'espérance mathématique. Pour la variable Normale centrée réduite, on obtient :

$$P(-1.96 \le U \le 1.96) = 95\%$$

On en déduit :

$$P\left(-1.96 \le \frac{Y_o - m}{\sigma} \le 1.96\right) = 95\%$$

et donc

$$P(m - 1.96\sigma \le Y_o \le m + 1.96\sigma) = 95\%$$

III.5a. Espérance et variance inconnues

- $Y_o \approx N(m; \sigma^2)$
- $\overline{Y}_N = \frac{1}{N}\sum_{n=1}^{N} Y_n \approx N\left(m; \frac{\sigma^2}{N}\right)$
- les deux variables sont indépendantes

$$\Rightarrow Y_o - \overline{Y}_N \approx N\left(0; \sigma^2\left[1 + \frac{1}{N}\right]\right)$$

43

C 'est pour Y_o que nous cherchons un intervalle de prévision : nous avons les deux paramètres *m* et σ^2 inconnus à éliminer :

m disparaît en effectuant la différence entre Y_o et m^ : c'est l'erreur de prévision ponctuelle que l'on fait lorsqu'on attribue à Y_o la valeur estimée de son espérance mathématique.

Les deux variables Y_o et sa prévision sont indépendantes en raison de l'hypothèse d'indépendance entre Y_o et les variables $Y_1, \ldots Y_n$.

Pour éliminer la variance inconnue, on utilisera la variance empirique.

• Estimation de la variance σ^2

$$S^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(Y_n - \overline{Y}_N\right)^2 \; ; \quad \frac{N-1}{\sigma^2}S^2 \approx \textit{CHI-DEUX(N-1)}$$

• L'indépendance de Y_o et $\overline{Y}_N$ avec S^2 entraîne :

$$\frac{Y_o - \overline{Y}_N}{S\sqrt{1 + \frac{1}{N}}} \approx STUDENT\left(N-1\right)$$

44

La variance empirique est indépendante de la moyenne empirique dans un échantillon de loi Normale

S^2 est fonction des $Y_1,..,Y_n$ qui sont indépendantes de Y_o.

Calcul de l'intervalle de prévision, η choisi

- lecture de t_η dans la table de la loi de Student à N-1 degrés de liberté :

$$P\left(-t_\eta \leq \frac{Y_o - \overline{Y}_N}{S\sqrt{1 + \frac{1}{N}}} \leq t_\eta \right) = \eta$$

- Résolution des inégalités :

$$P\left(\overline{Y}_N - t_\eta S\sqrt{1 + \tfrac{1}{N}} \leq Y_o \leq \overline{Y}_N + t_\eta S\sqrt{1 + \tfrac{1}{N}}\right) = \eta$$

45

Si nous avions connu la variance, nous aurions pu nous contenter de la loi Normale, et obtenu comme intervalle de prévision :

$$P\left(-u_\eta \leq \frac{Y_o - \overline{Y}_n}{\sigma\sqrt{1 + \frac{1}{n}}} \leq u_\eta \right) = \eta$$

$$P\left(\overline{Y}_n - u_\eta \sigma\sqrt{1 + \tfrac{1}{n}} \leq Y_o \leq \overline{Y}_n + u_\eta \sigma\sqrt{1 + \tfrac{1}{n}}\right) = \eta$$

Exemple

Application numérique : $\eta = 90\%$, N = 100, $S^2 = 3{,}61$

$N = 100 \Rightarrow$ STUDENT(99) $\cong$ N[0;1]

$t_\eta\left(1+\frac{1}{\sqrt{N}}\right) = 1{,}65 * 1{,}1 = 1{,}815$

formule EXCEL : loi.student.inverse(0,90;99) = 1,66

$$P\left\{\overline{Y}_N - 1{,}815\,.S \le Y_o \le \overline{Y}_N + 1{,}815\,.S\right\} = 0{,}90$$

observation :

$\overline{y} = 5{,}685$ et $s^2 = 3{,}61 \Rightarrow$ intervalle réalisé : $2{,}235 \le Y_o \le 9{,}132$

comparer avec intervalle pour l'espérance m .

46

La variance de l'erreur de prévision est supérieure à la variance de l'estimation ponctuelle de m , puisqu'il s'y ajoute la variance propre de Y_o.

Les quantités aléatoires dans les bornes de l'intervalle sont la moyenne empirique et S.

Ici encore lorsqu'on calcule la réalisation correspondant aux observations faites, il n'y a plus de proba.

III.5b. Prévision de la variable expliquée dans un modèle linéaire

- modèle :

$Y_o, Y_1, \ldots, Y_N$ indépendants, $Y_n \approx N(a\,x_n + b, \sigma^2)$

- estimation MCO de (a,b) et de σ^2 :

$$Y_o \approx N(ax_o + b, \sigma^2)$$

$$\hat{a}x_o + \hat{b} \approx N\left[ax_o + b, \sigma^2\left(x_o^2 m_{aa} + 2x_o m_{ab} + m_{bb}\right)\right]$$

$$Y_o - \hat{a}x_o - \hat{b} \approx N\left[0, \sigma^2\left(1 + x_o^2 m_{aa} + 2x_o m_{ab} + m_{bb}\right)\right]$$

47

Il ne faut pas oublier la covariance entre â et ^b.

Nous avons déjà calculé cette variance de l'erreur : c'est le RQM associé à la prévision.

matriciellement :

$E(Y_o) = (x_o\ 1)\beta = c_o'\,\beta$ estimée par $c_o'\hat{\beta} \sim N\left(c_o'\beta, \sigma^2 c_o'(X'X)^{-1}c_o\right)$

$$Y_o - c_o'\hat{\beta} \sim N\left(0, \sigma^2\left[1 + c_o'(X'X)^{-1}c_o\right]\right)$$

Pour la prévision de Y_o, l'erreur suit une loi normale centrée (d'espérance nulle) : la prévision est sans biais.

La précision de cette prévision (mesurée par la variance de l'erreur) dépend de la variance propre de Y_o et des variances et covariance des estimateurs de a et b.

A cela s'ajoute l'incertitude sur la variance σ^2. C'est elle qui oblige à utiliser la loi de Student au lieu de la loi Normale : il faut éliminer la variance inconnue, en utilisant la SCR.

$$\frac{SCR}{\sigma_2} \approx \text{CHI - DEUX}(N-2) \quad \text{indépendte de } Y_o \text{ et de } (\hat{a}, \hat{b})$$

$$\frac{Y_o - \hat{a}x_o - \hat{b}}{\hat{\sigma}\sqrt{1 + x_o^2 m_{aa} + 2x_o m_{ab} + m_{bb}}} \sim T(N-2)$$

$$\Rightarrow \boxed{P\left[\hat{a}x_o + \hat{b} - t_\eta S_e \le Y_o \le \hat{a}x_o + \hat{b} + t_\eta S_e\right] = \eta}$$

$$\text{en posant : } S_e = \hat{\sigma}\sqrt{1 + x_o^2 m_{aa} + 2x_o m_{ab} + m_{bb}}$$

48

$$Y_o - c_o\hat{\beta} \sim N(0, \sigma_{\hat{e}}^2) \quad \text{en posant } \sigma_2\left[1 + c_o(X'X)^{-1}c_o\right] = \sigma_{\hat{e}}^2$$

$$s_{\hat{e}}^2 = \hat{\sigma}_2\left[1 + c_o(X'X)^{-1}c_o\right] = \frac{SCR}{n-2}\left[1 + c_o(X'X)^{-1}c_o\right]$$

$$\frac{Y_o - c_o\hat{\beta}}{s_e} \sim T(n-2) \quad \text{Student à (n - 2) dl}$$

$$P\left(c_o\hat{\beta} - t_\eta s_e \le Y_o \le c_o\hat{\beta} + t_\eta s_e\right) = \eta$$

Numériquement, lorsque les résultats numériques sont fournis par un logiciel de régression, il faut demander la matrice de variance-covariance des coefficients pour obtenir la covariance. Ce sont directement ces variances et covariances qui sont utilisées pour calculer s_e :

$$s_{\hat{e}}^2 = \hat{\sigma}_2 + x_o^2\hat{V}(\hat{a}) + 2x_o\hat{C}ov(\hat{a}, \hat{b}) + \hat{V}(\hat{b})$$

$$s_{\hat{e}}^2 = \frac{SCR}{N-2} + c_o\hat{V}(\hat{\beta})c_o$$

Exemple

- $N = 15$ $\quad (X'X)^{-1} = \begin{bmatrix} 0{,}00271137 & -0{,}02327467 \\ -0{,}02327467 & 0{,}24524639 \end{bmatrix}$

$$\hat{y}_n = 0{,}909\, x_n + 2{,}173 \;,\; n = 1,..,15; \quad SCR = 14{,}28$$

- Intervalle de confiance 95% pour Y_o correspondant à $x_o = 12$

$15 - 2 = 13, \eta = 0{,}95 : t_\eta = 2{,}160$ et $\hat{\sigma} = \sqrt{\frac{SCR}{13}}$

$$S_e = \hat{\sigma}\sqrt{1 + (12)^2 m_{aa} + 2 * 12\, m_{ab} + m_{bb}} = 1{,}07689\ \hat{\sigma}$$

$$\Rightarrow P\left[12\,\hat{a} + \hat{b} - 2{,}326\,\hat{\sigma} \le Y_o \le 12\,\hat{a} + \hat{b} + 2{,}326\,\hat{\sigma}\right] = 0{,}95$$

49

Nous reprenons ici l'exemple traité dans le chapitre 2 : nous y ajoutons l'information sur la taille de l'échantillon (qui n'intervenait pas dans la prévision ponctuelle) et la SCR qui sera utilisée pour estimer la variance σ^2 des perturbations.

Les quantités aléatoires restant dans l'expression de l'intervalle de confiance sont les estimateurs des paramètres a et b, et l'estimateur de la variance σ^2 .

Exemple (suite)

- réalisation correspondant aux observations :

$$\hat{y}_o = 12 * 0{,}909 + 2{,}173 = 10{,}908$$

$$\hat{\sigma}^2 = \frac{14{,}28}{15-2} = 1{,}0985 = (1{,}0481)^2$$

$$10{,}908 - 2{,}326 * 1{,}0481 < y_o < 10{,}908 + 2{,}326 * 1{,}0481$$

$$8{,}47 < y_o < 13{,}35$$

est la réalisation de l'intervalle de confiance 95%.

La réalisation de l'intervalle de prévision est calculée pour les valeurs observées. Il n'y a plus de proba.