STATISTIQUES

Alain Yves LeRoux
Bordeaux, Décembre 1999

1 Statistiques descriptives

On considère une variable aléatoire X, associée à une loi de probabilité dont on cherche à évaluer un paramètre. On dispose de différentes données issues de différents test expérimentaux, correspondant à autant d'épreuves effectuées dans la cadre de cette variable aléatoire X. Ce jeu de données constitue un échantillon, et il est fini par nature. On note $x_1, x_2, ...x_K$ les différents résultats obtenus au moins une fois après N tests ; N est l'effectif de l'échantillon. On a donc $K \leq N$. En notant n_j le nombre de fois où le résultat x_j a été acquis, on a

$$n_1 \, + \, n_2 \, + \, ... \, + \, n_K \, = \, N \, . \tag{1.1}$$

La quantité $\frac{n_j}{N}$ est la fréquence relative du résultat x_j. On introduit deux quantités, la moyenne m et la médiane μ définies par

$$m \, = \, \frac{1}{N} \sum_{j=1}^{K} \, n_j \, x_j \, \, , \tag{1.2}$$

$$Effectif(\{X \leq \mu\}) \, \leq \, \frac{N}{2} \, \leq \, Effectif(\{X \geq \mu\}) \tag{1.3}$$

Lorsque l'on regroupe deux échantillons d'effectifs respectifs N_1 et N_2, de moyennes respectives m_1 et m_2, de médianes respectives μ_1 et μ_2, on obtient un échantillon final d'effectif $N = N_1 + N_2$ et de moyenne

$$m \, = \, \frac{m_1 N_1 + m_2 N_2}{N_1 + N_2} \, , \tag{1.4}$$

mais on ne dispose pas de formule exploitable pour caractériser la médiane de l'échantillon final. Bien que cette notion de médiane soit particulièrement parlante, on préfèrera utiliser la moyennt m plus facile à manipuler par les outils mathématiques.

Uns fois la valeur moyenne évaluée, on s'intéresse à l'écart entre cette valeur moyenne et les valeurs de l'échantillon. On peut considérer l'écart moyen

$$e \, = \frac{1}{N} \sum_{j=1}^{K} \, n_j \, |x_j - m| \, ,$$

ou la variance

$$v \, = \, \frac{1}{N} \sum_{j=1}^{K} \, n_j \, |x_j - m|^2 \, . \tag{1.5}$$

Typeset by the TEX-preprocessor PreTEX

Nous allons préférer cette dernière, plus facile à manipuler lors du regroupement d'échantillons.

Le calcul pratique de la moyenne ou de la variance peut se faire à partir d'une valeur provisoire de la moyenne notée a. On obtient, pour la moyenne

$$m \;=\; a \;+\; \sum_{j=1}^{K} \frac{n_j}{N} \,(x_j - a)\,, \tag{1.6}$$

et pour la variance

$$v \;=\; \sum_{j=1}^{K} \frac{n_j}{N} \,(x_j - a)^2 \;-\; (m-a)^2\,. \tag{1.7}$$

L'écart type σ est définit comme la racine carrée de la variance : $\sigma = \sqrt{v}$. Il s'agit d'une quantité positive. Notons que si $v = 0$ ou $\sigma = 0$, on a nécessairement $x_j = m$ pour tout j. Ceci ne présente aucun intérêt statistique.

2 Estimation des paramètres

Une loi de probabilité dépend d'un ou de plusieurs paramètres, par exemple m et σ pour la loi normale $\mathcal{N}(m,\sigma)$, λ pour la loi de Poisson $\mathcal{P}(\lambda)$, etc.. On cherche à estimer ces différents paramètres à partir des valeurs d'un échantillon, pour une variable aléatoire donnée X. On effectue n tests indépendants de cette variable aléatoire X, que l'on peut assimiler à n variables aléatoires $X_1,\ X_2,\ ..\ X_n$ indépendantes, de même loi que X. On note respectivement $x_1,\ x_2,\ ..\ x_n$ les résultats effectifs de ces n différents tests. A la différence de la section précédente, on ne comptabilise pas ensemble les résultats identiques.

2.1 Estimation d'un seul paramètre

On note θ le paramètre considéré. L'estimation va conduire à une formule du type

$$\theta_n \;=\; F_n(x_1, x_2, ..x_n)\,, \tag{2.1}$$

où F_n est une fonction de plusieurs variables. En reprenant l'expression de cette formule non plus au niveau des valeurs x_j, mais au niveau des variables aléatoires X_j, on définit la quantité

$$T_n \;=\; F_n(X_1,\ X_2, ..X_n)\,, \tag{2.2}$$

qui est en fait une nouvelle variable aléatoire.

La variable aléatoire T_n est appelée un estimateur du paramètre θ, et la quantité θ_n est appelée une estimation du paramètre θ. On devrait plutot parler d'une suite d'estimateurs T_n et d'une suite d'estimation θ_n, mais la distinction n'est pas indispensable ici.

On note $E(T_n)$ la moyenne de T_n, et la différence

$$E(T_n) \;-\; \theta \tag{2.3}$$

est appelée le biais de l'estimateur T_n. Lorsque $E(T_n) \;=\; \theta$, on dit que l'estimateur T_n est sans biais. Lorsque $E(T_n) \longrightarrow \theta$, on dit que l'estimateur T_n est asymptotiquement sans biais.

L'estimateur T_n est correct (ou correct en probabilité) lorsque la suite de variables aléatoires T_n converge vers la variable aléatoire constante θ pour la convergence en probabilité, c'est à dire

$$\forall \epsilon > 0 \quad P(|X_n - \theta| \geq \epsilon) \longrightarrow 0 \quad si\ n \longrightarrow \infty \ .$$

Ce résultat est acquis dès que $E(T_n) \longrightarrow \theta$ et $v(T_n) \longrightarrow 0$.

2.2 Méthode du maximum de vraisemblance

Dans le cas d'une variable discrète, où la probabilité d'avoir $X_j = x_j$ est une expression dépendant de θ et x_j, c'est à dire

$$P(X_j = x_j) \ = \ h(\theta, x_j) \ ,$$

en notant $h(\theta, x_j)$ cette expression, on considère la quantité

$$P \ = \ P(X_1 = x_1,\ X_2 = x_2,\ ..\ X_n = x_n) \ = \ \prod_{j=1}^{n} h(\theta, x_j) \ , \tag{2.4}$$

qu'on évalue ainsi grâce à l'hypothèse d'indépendance des variables aléatoires X_j. On choisit l'estimation θ_n qui rend cette quantité maximale pour le paramètre θ. Comme $P > 0$, et que la fonction ln est strictement croissante, on a le même résultat en dérivant $ln(P)$ plutot que P. Ceci offre l'avantage de remplacer le produit par une somme, plus facile à manipuler dans une opération de dérivation. Une condition nécessaire exige que la dérivée par rapport en θ, en $\theta = \theta_n$ réalisant un extrémum, soit nulle :

$$\frac{d}{d\theta}\left(ln(P)\right) \ = \ 0 \quad \Longleftrightarrow \quad \sum_{j=1}^{n} \frac{\partial h}{\partial \theta}\left(\theta_n, x_j\right) \ = \ 0 \ , \tag{2.5}$$

ce qui est une équation en θ_n, que l'on résout par une expression de la forme

$$\theta_n \ = \ F_n(x_1, x_2, ..x_n) \ ,$$

qui constitue une estimation θ_n de θ. Il en résulte, en utilisant la même expression, l'introduction d'un estimateur $T_n = F_n(X_1, X_2, ..X_n)$.

Exemple : on considère une variable aléatoire de Bernoulli, de paramètre p à estimer. Ainsi

$$h(p, x_j) \ = \ \begin{cases} p & si\ x_j = 1 \\ (1-p) & si\ x_j = 0 \end{cases} \tag{2.6}$$

On obtient, en notant $k\ (\leq n)$ le nombre de fois où la valeur 1 a été réalisé par l'ensemble des variables X_j,

$$P \ = \ p^k\ (1-p)^{n-k} \ ,$$

et

$$ln(P) \ = \ k\ ln(p) \ + \ (n-k)\ ln(1-p) \ .$$

Donc

$$\frac{d}{dp}\ ln(P) \ = \ \frac{k}{p} \ - \ \frac{n-k}{1-p} \ ,$$

qui est nulle pour

$$p \;=\; p_n \;=\; \frac{k}{n} \;\;=\; \frac{1}{n}\sum_{j=1}^{n} X_j. \tag{2.7}$$

Comme de plus $\frac{d^2}{dp^2}(lnP) \;=\; -\,(\frac{k}{p^2} + \frac{n-k}{(1-p)^2}$ est toujours négatif, on est assuré qu'il s'agit bien d'un maximum.

On a obtenu ainsi l'estimateur

$$\overline{X}_n \;=\; \frac{1}{n}\sum_{j=1}^{n} X_j \;, \tag{2.8}$$

du paramètre p. Notons que p est aussi la moyenne de X, et ainsi $\overline{X}_n$ est un estimateur sans biais de la moyenne.

Un autre exemple : On considère une variable aléatoire suivant la loi de Poisson $\mathcal{P}(\lambda)$. Les valeurs de tests x_j sont positives ou nulles. On trouve

$$P \;=\; \prod_{j=1}^{n} \frac{\lambda^{x_j}}{x_j!}\, e^{-\lambda}, \quad ln(P) \;=\; -n\lambda \;+\; \sum_{j=1}^{n} x_j\, ln(\lambda) \;-\; \sum_{j=1}^{n} ln(x_j!) \tag{2.9}$$

puis

$$\frac{d}{d\lambda}\, ln(P) \;=\; -n \;+\; \frac{1}{\lambda}\sum_{j=1}^{n} x_j \;\;= 0 \;\;\Longleftrightarrow\;\; \lambda \;=\; \frac{1}{n}\sum_{j=1}^{n} x_j \;,$$

$$\frac{d^2}{dp^2}\,(ln(P)) \;=\; -\,\frac{1}{\lambda^2}\sum_{j=1}^{n} x_j \;<\; 0 \;\;.$$

Au maximum, on retrouve le même estimateur $\overline{X}_n$ défini en (2.8). Il s'agit encore d'un estimateur sans biais de la moyenne.

Dans le cas d'une variable aléatoire continue, de densité $f(x)$ $(= f(\theta, x)$ en fait), on prend

$$P \;=\; \prod_{j=1}^{n} f(\theta, x_j) \;\;. \tag{2.10}$$

2.3 Estimation de deux paramètres

On note θ_1 et θ_2 ces deux paramètres. Comme précédemment, on introduit une fonction $h(\theta_1, \theta_2, x_j)$ pour chaque valeur x_j et on pose

$$P \;=\; P(\theta_1, \theta_2) \;=\; \prod_{j=1}^{n} h(\theta_1, \theta_2, x_j) \;, \tag{2.11}$$

dont on cherche les maxima. On les obtient en dérivant $ln(P)$ par rapport à chacune des deux variables θ_1 et θ_2. Pour un maximum, on rappelle qu'il faut réunir les conditions suivantes

$$\frac{\partial}{\partial\theta_1}\,(lnP) = 0 \;,\; \frac{\partial}{\partial\theta_2}(lnP) = 0, \;\left(\;\frac{\partial^2}{\partial\theta 1 \partial\theta_2}(lnP)\right)^2 \;-\; \frac{\partial^2}{\partial\theta_1^2}(lnP)\frac{\partial^2}{\partial\theta_2^2}(lnP) \;<\; 0 \;,$$

avec $\frac{\partial^2}{\partial\theta_1^2}(lnP) < 0$.

Exemple : la loi normale $\mathcal{N}(m,\sigma)$. On aura $\theta_1 = m,\ \theta_2 = \sigma$, et

$$h(m,\sigma,x_j) \ = \ \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-m)^2}{2\sigma^2}} \quad ,$$

puis

$$P = P(m,\sigma) = \frac{1}{\sigma^n\,(2\pi)^{\frac{n}{2}}}\, e^{-\frac{1}{2\sigma^2}\left((x_1-m)^2+\ldots+(x_n-m)^2\right)} \quad ,$$

et

$$ln(P) \ = \ -\,n\,ln(\sigma) \ - \ \frac{n}{2}\,ln(2\pi) \ - \ \frac{1}{2\sigma^2}\left((x_1-m)^2+\ldots+(x_n-m)^2\right) .$$

On dérive par rapport à m :

$$\frac{\partial}{\partial m}(ln(P)) \ = \ \frac{1}{\sigma^2}\left((x_1-m)+(x_2-m)+\ldots(x_n-m)\right) \ = \ \frac{x_1+x_2+\ldots x_n \ - \ nm}{\sigma^2} \quad ,$$

qui est nulle pour $m \ = \ \frac{1}{n}(x_1+\ldots x_n)$, et on retrouve le même estimateur $\overline{X}_n$ pour la moyenne.

On dérive ensuite par rapport à σ, pour obtenir

$$\frac{\partial}{\partial\sigma}(ln(P)) \ = \ \frac{1}{\sigma^3}\left((x_1-m)^2+\ldots(x_n-m)^2\right) \ - \ \frac{n}{\sigma} \quad ,$$

qui est nulle pour

$$\sigma^2 \ = \ \frac{1}{n}\sum_{j=1}^{n}\ (x_j-m)^2 \ .$$

On vérifie qu'il s'agit bien d'un maximum :

$$\left(\frac{\partial^2}{\partial m\partial\sigma}lnP\right)^2 - \left(\frac{\partial^2}{\partial m^2}lnP\right)\left(\frac{\partial^2}{\partial\sigma^2}lnP\right) \ = \ \frac{-2n^2}{\sigma^4} \ < \ 0 \ ; \ \frac{\partial^2}{\partial m^2}lnP = -\frac{n}{\sigma^2} \ < \ 0 \ .$$

On a obtenu un estimateur de la variance

$$U_n \ = \ \frac{1}{n}\sum_{j=1}^{n}\ (X_j-m)^2 \ . \tag{2.12}$$

2.4 Résultats

Nous avons obtenu les estimateurs suivants :

Estimateur de la moyenne

$$\overline{X}_n \ = \ \frac{1}{n}\sum_{j=1}^{n}\ X_j \tag{2.13}$$

Estimateur de la variance, la moyenne m étant connue

$$U_n \ = \ \frac{1}{n}\sum_{j=1}^{n}\ (X_j-m)^2 \ . \tag{2.14}$$

Estimateur de la variance, la moyenne m étant inconnue

$$W_n \;=\; \frac{1}{n}\sum_{j=1}^{n}\;(X_j-\overline{X}_n)^2\;. \tag{2.15}$$

où on a simplement remplacé m par $\overline{X}_n$.

Proposition 2.1 *Soit X une variable aléatoire de moyenne m et admettant un moment d'ordre deux. Alors $\overline{X}_n$ défini par (2.13) est un estimateur sans biais et correct de la moyenne.*

Démonstration : On a $E(\overline{X}_n)=\frac{1}{n}\sum_{j=1}^{n}E(X_j)=\frac{nm}{n}=m$. L'estimateur est sans biais.

De plus $v(\overline{X}_n)=\frac{1}{n^2}\sum_{j=1}^{n}v(X_j)=\frac{nv(X)}{n^2}=\frac{v(X)}{n}$ qui tend vers zéro. L'estimateur est correct.

Proposition 2.2 *Soit X une variable aléatoire de moyenne m, de variance $v(X)$, et admettant un moment d'ordre 4. Alors U_n défini en (2.14) est un estimateur sans biais, correct de la variance, W_n défini en (2.15) est un estimateur biaisé, asymptotiquement sans biais, correct de la variance, et l'estimateur*

$$V_n=\;\frac{1}{n-1}\sum_{j=1}^{n}\;(X_j-\overline{X}_n)^{\,2} \tag{2.16}$$

est un estimateur sans biais, correct de la variance.

Démonstration On a $E(U_n)=\frac{1}{n}\sum_{j=1}^{n}E(X_j-m)^2=\frac{nv(X)}{n}=v(X)$, et donc Un est un estimateur sans biais. De plus

$$v(U_n)\;=\;\frac{1}{n^2}\sum_{j=1}^{n}v((X_j-m)^2)\;=\;\frac{1}{n^2}\sum_{j=1}^{n}E\Big(((X_j-m)^2-v(X))^2\Big)\;=\;\frac{nA(X)}{n^2}\;=\;\frac{A(X)}{n},$$

en posant $A(X)=E\Big(((X-m)^2-v(X))^2\Big)$ qui existe car X admet un moment d'ordre 4 et qui est indépendant de n. Donc $\frac{A(X)}{n}\longrightarrow 0$, et l'estimateur U_n est correct.

Par ailleurs

$$E(W_n)=\frac{1}{n}\sum_{j=1}^{n}E((X_j-\overline{X}_n)^2)=\frac{1}{n}\sum_{j=1}^{n}E\Big(((X_j-m)-(\overline{X}_n-m))^2\Big),$$

d'où

$$E(W_n)=\frac{1}{n}\sum_{j=1}^{n}\Big(E((X_j-m)^2)-2E(X_j-m)E(\overline{X}_n-m)+E((\overline{X}_n-m)^2)\Big).$$

Comme $E(\overline{X}_n-m)=0,\;E((X_j-m)^2)=v(X)$ et $E((\overline{X}_n-m)^2)=v(\overline{X}_n)=\frac{nv(X)}{n^2}=\frac{v(X)}{n}$, il vient

$$E(W_n)\;=\;v(X)\;-\;v(\overline{X}_n)\;=\;\frac{n-1}{n}\,v(X). \tag{2.17}$$

On en déduit quz $E(W_n)\longrightarrow v(X)$. On montre que l'estimateur W_n est correct en le comparant à U_n, en effet : $W_n=U_n-(\overline{X}_n-m)^2$, et en remarquant que $(\overline{X}_n-m)^2$ tend vers zéro en probabilité.

En remarquant que $V_n = \frac{n}{n-1} W_n$, il vient

$$E(V_n) \;=\; \frac{n}{n-1}\, E(W_n) \;=\; \frac{n}{n-1}\,\frac{n-1}{n}\, v(X) \;=\; v(X)\,,$$

ce qui établit que l'estimateur V_n est sans biais. On montre qu'il est correct en le comparant à W_n.

3 Intervalles de confiance

Soit X une variable aléatoire, correspondant à une loi de probabilité faisant intervenir un paramètre θ inconnu. On considère un estimateur T_n de θ. On veut construire un intervalle de confiance pour ce paramètre θ, c'est à dire, étant donné $\alpha \in]0,1[$, un intervalle $]a,b[$ tel que

$$P(a < \theta < b) \;\geq\; 1-\alpha\,. \tag{3.1}$$

Bien entendu, plus le seuil α est petit, plus l'intervalle $]a,b[$ est grand. On choisit en fait a et b tels que

$$P(T_n \leq a) \;\leq\; \frac{\alpha}{2}\;,\; a\; maximal \quad P(T_n \geq b) \;\leq\; \frac{\alpha}{2}\;,\; b\; minimal \tag{3.2}$$

Pour une variable continue, on obtient des égalités, et si on note F_n la fonction de répartion de la loi associée à T_n, on a

$$F_n(a) \;=\; \frac{\alpha}{2} \quad , \quad F_n(b) \;=\; 1 \;-\; \frac{\alpha}{2}\;. \tag{3.3}$$

Ces valeurs peuvent en général être obtenues à partir de tables qui donnent les valeurs de quelques fonctions de répartition de référence. Si H est une telle fonction de référence, on note en général, pour $\beta \in]0,1[$,

$$H(t_\beta) \;=\; \beta\,,$$

et ainsi a se déduit de $t_{\frac{\alpha}{2}}$, et b de $t_{1-\frac{\alpha}{2}}$. Si de plus la dérivée de H (c'est à dire la densité) est paire, on a $t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$. Les passages de $t_{\frac{\alpha}{2}}$ à a et de $t_{1-\frac{\alpha}{2}}$à b correspond en général à des opérations algébriques triviales. Les exemples suivants sont des exemples caractéristiques de ces types d'opération.

3.1 Intervalle de confiance pour la moyenne

On suppose que X suit une loi normale $\mathcal{N}(m,\sigma)$, que l'on connait σ mais pas m. On cherche un intervalle de confiance pour m.
On dispose de résultats $x_1, x_2, .., x_n$ de n tests correspondants à n valeurs des variables aléatoires $X_1, X_2, .., X_n$, indépendantes et de même loi que X. On prend l'estimateur $\overline{X}_n = \frac{1}{n}(X_1 + ...X_n)$, qui suit la loi normale $\mathcal{N}(m, \frac{\sigma}{\sqrt{n}})$, compte tenu de l'indépendance des X_j.

On se donne le seuil α et on recherche d'abord la valeur b, qui doit vérifier

$$\frac{\sqrt{n}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{b} e^{-\frac{n(x-m)^2}{2\sigma^2}}\, dx \;=\; 1 \;-\; \frac{\alpha}{2}\,. \tag{3.4}$$

On se ramene à la loi normale réduite, dont on note H la fonction de répartition, et pour laquelle des tables sont disponibles, par le changement de variable

$$\xi \;=\; \frac{x-m}{\sigma}\,\sqrt{n}\,, \quad (dx \;=\; \frac{\sigma}{\sqrt{n}}\,d\xi)$$

et il vient, en posant $B = \frac{b-m}{\sigma}\sqrt{n}$,

$$H(B) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{B} e^{-\frac{\xi^2}{2}}\,d\xi \;=\; 1 \;-\; \frac{\alpha}{2} \;(= H(t_{1-\frac{\alpha}{2}})). \tag{3.5}$$

Ainsi $B = t_{1-\frac{\alpha}{2}}$, valeur lue dans une table de la fonction de répartition H de la loi normale réduite. On en déduit

$$b \;=\; m \;+\; t_{1-\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}}\,. \tag{3.6}$$

On procède de façon identique pour obtenir a qui doit vérifier

$$\frac{\sqrt{n}}{\sigma\sqrt{2\pi}}\int_{-\infty}^{a} e^{-\frac{n(x-m)^2}{2\sigma^2}}\,dx \;=\; \frac{\alpha}{2}\,. \tag{3.7}$$

On se ramene à la loi normale réduite par le même changement de variable (3.6), en posant $A = \frac{a-m}{\sigma}\sqrt{n}$:

$$H(A) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{A} e^{-\frac{\xi^2}{2}}\,d\xi \;=\; \frac{\alpha}{2} \;(= H(t_{\frac{\alpha}{2}})). \tag{3.8}$$

Ainsi $A = t_{\frac{\alpha}{2}} \;=\; -t_{1-\frac{\alpha}{2}}$ car de la densité de la loi normale réduite est paire, et $a = m \;-\; t_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$.

On pose $\overline{m} = \frac{1}{n}(x_1 + x_2 + ...x_n)$. On a l'encadrement

$$m \;-\; t_{1-\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}} \;\leq\; \overline{m} \leq\; m \;+\; t_{1-\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}} \tag{3.9}$$

avec une probabilité égale à $1-\alpha$. On en déduit que l'on a aussi, avec la même probabilité $1-\alpha$, l'encadrement

$$\overline{m} \;-\; t_{1-\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}} \;\leq\; m \;\leq\; \overline{m} \;+\; t_{1-\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}} \tag{3.10}$$

qui constitue l'intervalle de confiance au seuil α. On a démontré le résultat suivant.

Proposition 3.1 *Soit $\alpha \in]0,1[$, et $X_1, X_2, ...X_n$ des variables aléatoires indépendantes, de même loi normale $\mathcal{N}(m,\sigma)$. Alors l'intervalle*

$$]\;\overline{X}_n \;-\; t_{1-\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}}\,,\;\; \overline{X}_n \;+\; t_{1-\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}}\;[\tag{3.11}$$

est un intervalle de confiance de seuil α pour la moyenne m. La valeur de $t_{1-\frac{\alpha}{2}}$ est à lire dans une table de la fonction de répartition H de la loi normale réduite, telle que $H(t_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$.

Remarque 3.2 *On a supposé σ connu, et on remarque que la largeur de l'intervalle se réduit lorsque n devient grand*

Lorsque σ est inconnu, on utilise un estimateur de la variance, par exemple V_n défini en (2.16), et on pose $\sigma_n = \sqrt{V_n}$. On obtient ainsi l'intervalle de confiance de seuil α

$$\left] \overline{X}_n - t_{1-\frac{\alpha}{2}} \frac{\sigma_n}{\sqrt{n}} , \ \overline{X}_n + t_{1-\frac{\alpha}{2}} \frac{\sigma_n}{\sqrt{n}} \right[\tag{3.12}$$

mais cette méthode n'est pas fiable pour de petites valeurs de n, et on la réservera pour des valeurs $n \geq 30$. Pour de plus petites valeurs de n, à savoir $2 \leq n \leq 30$, et toujours lorsque σ est inconnu, on préfère utiliser la variable aléatoire

$$T_{n-1} = \frac{(\overline{X}_n - m)\sqrt{n}}{\sigma_n} , \tag{3.13}$$

qui suit une loi de Student à $n-1$ degrés de liberté. Cette loi est tabulée, sa densité est donnée par

$$f_{n-1}(t) = \frac{\Gamma(\frac{n}{2})}{\sqrt{\pi}\,\Gamma(\frac{n-1}{2})} \ \frac{1}{(1 + \frac{t^2}{n-1})^{\frac{n}{2}}} . \tag{3.14}$$

La variable T_{n-1} présente l'avantage de ne pas dépendre de σ. Des tables de la loi de Student sont disponibles pour des valeurs de n allant de 2 à 30, et même au delà, bien que la variation soit alors réduite et les valeurs très proches de celles obtenues pour la loi normale réduite. Notons également que pour $n = 2$, T_1 n'a pas de moyenne, et que les autres T_{n-1} ont une moyenne nulle. De plus T_2 n'a pas de variance. La densité est une fonction paire, d'où $t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$.
L'intervalle de confiance a la même expression que (3.12), à cette différence près que la valeur $t_{1-\frac{\alpha}{2}}$ est lue dans une table de la loi de Student.

On remarque que pour α fixé, les valeurs de $t_{1-\frac{\alpha}{2}}$ lues dans une table de loi normale sont plus petites que celles lues dans une table de loi de Student, notamment pour les petites valeurs de n. Cela ne veut pas du tout dire que cette dernière est moins précise ; c'est au contraire la première qui n'est pas fiable dans ce cas.

3.2 Intervalle de confiance pour la variance

Dans le même cadre que précédemment, on prend l'estimateur V_n défini en (2.16), puis l'estimateur

$$\chi^2_n = (n-1)\frac{V_n}{\sigma^2} , \tag{3.15}$$

qui suit une loi du χ^2 à $n-1$ degrés de libertés. Il s'agit d'une loi exponentielle particulière, donc à valeurs positives (et donc pas d'espoir de parité..), de densité

$$\varphi_{n-1}(x) = \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}\, x^{\frac{n-3}{2}}\, e^{-\frac{x}{2}} \qquad (x > 0,\ n \geq 2) \tag{3.16}$$

La détermination de a et b se fait de la façon suivante. La valeur de a est telle que, en notant Φ_{n-1} la fonction de répartition de la loi χ^2 à $n-1$ degrés de liberté,

$$\Phi_{n-1}(a) = \int_0^a \varphi_{n-1}(x)\, dx = \frac{\alpha}{2} , \tag{3.17}$$

d'où $a = t_{\frac{\alpha}{2}}$ tel que $\Phi_{n-1}(t_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$. De la même façon $b = t_{1-\frac{\alpha}{2}}$ tel que $\Phi_{n-1}(t_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$. On obtient l'encadrement

$$t_{\frac{\alpha}{2}} \leq (n-1)\frac{V_n}{\sigma^2} \leq t_{1-\frac{\alpha}{2}} \quad ,$$

d'où l'intervalle de confiance de seuil α

$$(n-1)\frac{V_n}{t_{1-\frac{\alpha}{2}}} \leq \sigma^2 \leq (n-1)\frac{V_n}{t_{\frac{\alpha}{2}}} \quad . \tag{3.18}$$

Ces valeurs de $t_{\frac{\alpha}{2}}$ et $t_{1-\frac{\alpha}{2}}$ seront lues dans une table du χ^2 à $n-1$ degrés de liberté.

Pour $n > 30$, on peut considérer que $\sqrt{2\chi_n^2} - \sqrt{2n-1}$ est proche d'une loi normale réduite et utiliser une table de la loi normale réduite.

4 Estimation d'une probabilité

On suppose maintenant que les X_j sont n épreuves de Bernoulli de paramètre p, indépendantes. Si k ($\leq n$) est le nombre de réalisations effectives (lorsque $X_j = 1$), on a l'estimation $\overline{p} = \frac{k}{n}$, et l'estimateur $\overline{X}_n = \frac{X}{n}$ où X est une variable binomiale $\mathcal{B}(n,p)$.

Le seuil α étant choisi, il est difficile de déterminer a et b dans la mesure où X ne prend que des valeurs entières entre 0 et n. On retient un encadrement

$$\frac{s_{\frac{\alpha}{2}}}{n} \leq \overline{p} \leq \frac{s_{1-\frac{\alpha}{2}}}{n} \quad , \tag{4.1}$$

avec $p(\overline{X}_n \leq \frac{s_{\frac{\alpha}{2}}}{n}) \leq \frac{\alpha}{2}$, $s_{\frac{\alpha}{2}}$ maximal et $p(\overline{X}_n \leq \frac{s_{1-\frac{\alpha}{2}}}{n}) \leq 1 - \frac{\alpha}{2}$, $s_{1-\frac{\alpha}{2}}$ minimal.

Pour n grand, ($n \geq 30$), une loi $\mathcal{B}(n,p)$ peut être approchée par la loi normale $\mathcal{N}(np, \sqrt{np(1-p)})$, et donc on peut approcher $\overline{X}_n$ par une loi normale $\mathcal{N}(p, \sqrt{\frac{p(1-p)}{n}})$. On se retrouve alors dans le cadre d'un exemple précédent, et lire $t_{1-\frac{\alpha}{2}}$ dans une table de la loi normale réduite, et obtenir l'encadrement

$$p - t_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq \overline{p} \leq p + t_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \tag{4.2}$$

puis l'intervalle de confiance de seuil α

$$\overline{p} - t_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \overline{p} + t_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \quad . \tag{4.3}$$

5 Coefficient de corrélation

On considère deux variables aléatoires X_1 et X_2, de moyennes respectives $m_1 = E(X_1)$, $m_2 = E(X_2)$.

Définition 5.1 *La covariance de X_1 et X_2 est donnée par*

$$Cov(X_1, X_2) \;=\; E((X_1 - m_1)(X_2 - m_2)) \;. \tag{5.1}$$

Proposition 5.2 *La covariance a les propriétés suivantes*

$$E(X_1 X_2) \;=\; m_1 m_2 \;+\; Cov(X_1, X_2) \tag{5.2}$$

$$Cov(X, X) \;=\; Var(X) \;,$$

$$\forall \alpha, \beta, \gamma, \delta \in \mathbb{R} \;\; Cov(\alpha X_1 + \gamma, \beta X_2 + \delta) \;=\; \alpha\beta \; Cov(X_1, X_2) \tag{5.3}$$

et si X_1 et X_2 sont indépendantes, $Cov(X_1, X_2) = 0$.

Démonstration : On a
$Cov(X_1, X_2) = E(X_1 X_2) - m_1 E(X_2) - m_2 E(X_1) + m_1 m_2 = E(X_1 X_2) - m_1 m_2$,
d'où (5.2), et si $X_1 = X_2 (= X)$, $Cov(X, X) = E(X^2) - E(X)^2 = Var(X)$.
Ensuite, $Cov(\alpha X_1 + \gamma, \beta X_2 + \delta) = E((\alpha X_1 + \gamma - \alpha m_1 - \gamma)(\beta X_2 + \delta - \beta m_2 - \delta))$
$= \alpha\beta E((X_1 - m_1)(X_2 - m_2)) = \alpha\beta Cov(X_1, X_2)$, d'où (5.3).
Enfin, si X_1 et X_2 sont indépendantes, $E(X_1, X_2) = m_1 m_2$, d'où $Cov(X_1, X_2) = 0$.

Définition 5.3 *Le coéfficient de corrélation de X_1 et X_2 est défini par*

$$\rho(X_1, X_2) \;=\; \frac{Cov(X_1, X_2)}{\sigma(X_1)\sigma(X_2)} \;. \tag{5.4}$$

Théorème 5.4 *On a toujours*

$$-1 \;\leq\; \rho(X_1, X_2) \;\leq\; 1 \;. \tag{5.5}$$

Démonstration : On pose $Y_1 = \dfrac{X_1 - m_1}{\sigma(X_1)}$, $Y_2 = \dfrac{X_2 - m_2}{\sigma(X_2)}$, d'où

$$\rho(X_1, X_2) \;=\; E(Y_1, Y_2) \;.$$

Les variables aléatoires Y_1, Y_2 sont réduites, donc $E(Y_1^2) = E(Y_2^2) = 1$. De plus
$0 \;\leq\; E((Y_1 - Y_2)^2) = E(Y_1^2) + E(Y_2^2) - 2E(Y_1 Y_2) = 2(1 - E(Y_1 Y_2))$
$0 \;\leq\; E((Y_1 + Y_2)^2) = E(Y_1^2) + E(Y_2^2) + 2E(Y_1 Y_2) = 2(1 + E(Y_1 Y_2))$
d'où $-1 \leq E(Y_1 Y_2) \leq 1$, et (5.5).

Proposition 5.5 *On considère n variables aléatoires $X_1, X_2 ... X_n$ et n réels $\alpha_1, \alpha_2 ... \alpha_n$, et on pose $X = \sum_{j=1}^{n} \alpha_j X_j$, $m_j = E(X_j)$, $\sigma_j^2 = Var(X_j)$. Alors*

$$E(X) = \sum_{j=1}^{n} \alpha_j \, m_j \;, \quad Var(X) \;=\; \sum_{j=1}^{n} \alpha_j^2 \, \sigma_j^2 \;+\; 2 \sum_{j<k} \alpha_j \alpha_k \, Cov(X_j, X_k) \tag{5.6}$$

Démonstration : Evidente pour $E(X)$, et il suffit de développer

$Var(X) = E((\sum_{j=1}^{n} (\alpha_j X_j \;-\; \alpha_j m_j))^2)$ pour conclure.

5.1 Estimation de la covariance

On considère deux variables aléatoires X et Y. On dispose de deux échantillons de même taille n, coreespondants aux résultats respectifs $x_j,\ y_j$ des variables aléatoires $X_1, X_2, ..X_n$ et $Y_1, Y_2, ; ; Y_n$ telles que

$$\forall j\ \forall k \neq j\ X_j \textit{ et } X_k,\ Y_j \textit{ et } Y_k,\ X_j \textit{ et } Y_k \textit{ sont indépendantes.}$$

On pose

$$C_n \ = \ \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)(Y_j - \overline{Y}_n)\ ,\ K_n \ = \ \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)(Y_j - \overline{Y}_n)\ . \tag{5.7}$$

Proposition 5.6 *Les quantités C_n et K_n sont des estimateurs de la covariance; l'estimateur K_n est sans biais, l'estimateur C_n est asymptotiquement sans biais.*

Démonstration : On a

$$C_n \ = \ \frac{1}{n}\sum_{j=1}^{n} X_j Y_j \ - \ \overline{X}_n\overline{Y}_n \ = \ \frac{1}{n}\sum_{j=1}^{n} X_j Y_j \ - \ \frac{1}{n^2}\sum_{j\neq k} X_j Y_k \ - \ \frac{1}{n^2}\sum_{j=1}^{n} X_j Y_j,$$

d'où

$$C_n \ = \ \frac{n-1}{n^2}\sum_{j} X_j Y_j \ - \ \frac{1}{n^2}\sum_{j\neq k} X_j Y_k\ , \tag{5.8}$$

et

$$E(C_n) \ = \ \frac{(n-1)n}{n^2}\,E(XY) \ - \ \frac{n(n-1)}{n^2}\,E(X)E(Y) \ = \ \frac{n-1}{n}\,Cov(X,Y)$$

qui tend vers $Cov(X,Y)$ si n tend vers l'infini, et C_n est bien asymptotiquement sans biais.
Par ailleurs,

$$K_n \ = \ \frac{n}{n-1}\,C_n\ ,$$

donc $E(K_n) = \frac{n}{n-1}E(C_n) = Cov(X,Y)$, qui est bien un estimateur sans biais.

On utilisera l'estimation

$$k_n \ = \ \frac{1}{n-1}\left(\sum_{j=1}^{n} x_j y_j \ - \ \frac{1}{n}\,(\sum_{j=1}^{n} x_j)\,(\sum_{j=1}^{n} y_j)\right)\ , \tag{5.9}$$

pour évaluer la covariance.

5.2 Estimation du coéfficient de corrélation

On utilisera l'estimation suivante du coefficient de corrélation

$$\rho_n \ = \ \frac{k_n}{\sigma_n\,\tau_n} \ = \ \frac{\sum_{j=1}^{n} x_j y_j \ - \ \frac{1}{n}\,(\sum_{j=1}^{n} x_j)(\sum_{j=1}^{n} y_j)}{\sqrt{(\sum_{j=1}^{n} x_j^2) - \frac{1}{n}(\sum_{j=1}^{n} x_j)^2)\ (\sum_{j=1}^{n} y_j^2 - \frac{1}{n}(\sum_{j=1}^{n} y_j)^2)}} \tag{5.10}$$

où

$$\sigma_n \ = \ \sqrt{\sum_{j=1}^{n} x_j^2 \ - \ \frac{1}{n}(\sum_{j=1}^{n} x_j)^2}\quad ,\quad \tau_n = \sqrt{\sum_{j=1}^{n} y_j^2 \ - \ \frac{1}{n}\,(\sum_{j=1}^{n} y_j)^2}\ ,$$

et l'estimateur associé

$$R_n \;=\; \frac{K_n}{S_n\,T_n} \;=\; \frac{\sum_{j=1}^n X_j Y_j \;-\; \frac{1}{n}\,(\sum_{j=1}^n X_j)(\sum_{j=1}^n Y_j)}{\sqrt{(\sum_{j=1}^n X_j^2) - \frac{1}{n}(\sum_{j=1}^n X_j)^2)\;(\sum_{j=1}^n Y_j^2 - \frac{1}{n}(\sum_{j=1}^n Y_j)^2)}} \tag{5.11}$$

où

$$S_n \;=\; \sqrt{\sum_{j=1}^n X_j^2 \;-\; \frac{1}{n}(\sum_{j=1}^n X_j)^2} \quad , \quad T_n = \sqrt{\sum_{j=1}^n Y_j^2 \;-\; \frac{1}{n}\,(\sum_{j=1}^n Y_j)^2} \quad .$$

On utilise la transformation de Fisher

$$Z \;=\; \frac{1}{2}\,ln\!\left(\frac{1+\rho}{1-\rho}\right) \;\; (= Argth(\rho)), \tag{5.12}$$

pour définir

$$Z_n \;=\; \frac{1}{2}\,ln\!\left(\frac{1+R_n}{1-R_n}\right). \tag{5.13}$$

On démontre que

$$E(Z_n) \;=\; \frac{1}{2}\,ln\!\left(\frac{1+\rho}{1-\rho}\right) \;+\; \frac{\rho}{2(n-1)} \quad , \tag{5.14}$$

$$Var(Z_n) \;=\; \frac{1}{n-3} \quad , \quad \sigma(Z_n) \;=\; \frac{1}{\sqrt{n-3}} \quad , \tag{5.15}$$

et que la quantité réduite $\dfrac{Z_n - E(Z_n)}{\sigma(Z_n)}$ converge en loi vers une loi normale réduite lorsque n tend vers l'infini. On observe que la variance $Var(Z_n)$ est indépendante de ρ, que Z_n est un estimateur de Z asymptotiquement sans biais ($E(Z_n) \longrightarrow Z$) et que la convergence vers la loi normale réduite est très rapide en pratique.

5.3 Intervalle de confiance pour le coefficient de corrélation

Soit $\alpha \in]0,1[$ un seuil déterminé, et F la fonction de répartition de la loi normale réduite. On note $t_{1-\frac{\alpha}{2}}$ la valeur telle que $F(t_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$. Alors, pour n assez grand, on a

$$-t_{1-\frac{\alpha}{2}} \;\leq\; \frac{Z_n - E(Z_n)}{\sigma(Z_n)} \;\leq\; t_{1-\frac{\alpha}{2}} \quad ,$$

avec une probabilité supérieure ou égale à $1-\alpha$. Si $\overline{z}_n$ est une estimation de Z_n, on a, toujours avec une probabilité au moins égale à $1-\alpha$,

$$Z \;+\; \frac{\rho}{2(n-1} \;-\; \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \;\leq\; \overline{z}_n \;\leq\; Z \;+\; \frac{\rho}{2(n-1)} \;+\; \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\,, \tag{5.16}$$

d'où

$$\overline{z}_n \;-\; \frac{\rho}{2(n-1)} \;-\; \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \;\leq\; Z \;\leq\; \overline{z}_n \;-\; \frac{\rho}{2(n-1)} \;+\; \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\,, \tag{5.17}$$

et comme $\frac{\rho}{2(n-1)}$ est petit devant $\frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}$, on le néglige souvent en pratique. Ceci rend (5.17) indépendante de ρ. Il reste donc

$$\overline{z}_n \; - \; \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \; \leq \; Z \; \leq \; \overline{z}_n \; + \; \frac{t_{1-\frac{\alpha}{2}}}{\sqrt{n-3}} \, , \tag{5.18}$$

qui constitue notre intervalle de confiance au seuil α. On obtient un intervalle de confiance pour ρ par la relation $\rho = th(Z)$.

6 La régression linéaire

On reprend les variables aléatoires X_j et Y_j précédentes, avec leurs résultats x_j et y_j.On suppose qu'au moins deux parmi les valeurs x_j sont différentes. On pose, en reprenant des notations précédentes

$$\overline{x}_n \; = \; \frac{1}{n} \sum_{j=1}^{n} x_j \, , \;\; \overline{y}_n \; = \; \frac{1}{n} \sum_{j=1}^{n} y_j \, , \tag{6.1}$$

$$\sigma_n^2 \; = \; \frac{1}{n-1} \left(\sum_{j=1}^{n} x_j^2 \; - \; n \overline{x}_n^2 \right) \, , \; \tau_n^2 \; = \; \frac{1}{n-1} \left(\sum_{j=1}^{n} y_j^2 \; - \; n \; \overline{y}_n^2 \right) \, , \tag{6.2}$$

$$k_n \; = \; \frac{1}{n-1} \left(\sum_{j=1}^{n} x_j y_j \; - \; n \overline{x}_n \overline{y}_n \right) \, , \;\; \rho_n \; = \; \frac{k_n}{\sigma_n \, \tau_n} \, . \tag{6.3}$$

On cherche à établir une relation du type

$$Y \; = \; \alpha \, X \; + \; \beta \, , \tag{6.4}$$

en évaluant les coefficients α et β de façon optimale, en fonction des données disponibles, x_j et y_j. On considére la somme des carrés

$$J(\alpha, \beta) \; = \; \sum_{j=1}^{n} \; (y_j \; - \; \alpha x_j \; - \; \beta)^2 \, , \tag{6.5}$$

qui est une fonction positive définie sur $\mathbb{R}^2$, et on recherche le couple (α, β) réalisant le minimum de J. Pour cela, on calcule les dérivées partielles, et on écrit qu'elles sont nulles en un extrémum. On a

$$\frac{\partial J}{\partial \alpha}(\alpha, \beta) \; = \; 2 \; (\; \alpha \sum_{j=1}^{n} x_j^2 \; + \; \beta \sum_{j=1}^{n} x_j \; - \; \sum_{j=1}^{n} x_j y_j \;) \tag{6.6}$$

$$\frac{\partial J}{\partial \beta}(\alpha, \beta) \; = \; 2 \; (\; \alpha \sum_{j=1}^{n} x_j \; + \; n \; \beta \; - \; \sum_{j=1}^{n} y_j \;) \tag{6.7}$$

et si (α, β) correspond à un extrémum,

$$\alpha \sum_{j=1}^{n} x_j^2 \; + \; \beta \sum_{j=1}^{n} x_j \; = \; \sum_{j=1}^{n} x_j y_j \, , \tag{6.8}$$

$$\alpha \sum_{j=1}^{n} x_j \ + \ n \ \beta \ = \ \sum_{j=1}^{n} y_j \ . \tag{6.9}$$

Avec les notations précédentes, ceci s'écrit aussi

$$((n-1)\sigma_n^2 + n\overline{x}_n^2) \ \alpha \ + \ n\overline{x}_n \ \beta \ = \ (n-1)k_n \ + \ n\overline{x}_n\overline{y}_n \tag{6.10}$$

$$n\overline{x}_n \ \alpha \ + \ n \ \beta \ = \ n\overline{y}_n \ , \tag{6.11}$$

qu'on résoud immédiatement pour obtenir

$$\alpha \ = \ \frac{k_n}{\sigma_n^2} \ , \ \beta \ = \ \overline{y}_n \ - \ \overline{x}_n \ \frac{k_n}{\sigma_n^2} \ .$$

On vérifie qu'il s'agit bien d'un minimum en calculant :

$$A = \frac{\partial^2 J}{\partial \alpha^2}(\alpha, \beta) = 2\sum_{j=1}^{n} x_j^2 \ , \ B = \frac{\partial^2 J}{\partial \alpha \partial \beta}(\alpha, \beta) = 2\sum_{j=1}^{n} x_j \ , \ C = \frac{\partial^2 J}{\partial \beta^2}(\alpha, \beta) = 2n,$$

puis en évaluant $B^2 - AC = 4((\sum_{j=1}^{n} x_j)^2 - n\sum_{j=1}^{n} x_j^2) = 4(n^2\overline{x}_n^2 - n(n-1)\sigma_n^2 - n^2\overline{x}_n^2) = -4n(n-1)\sigma_n^2$ qui est toujours strictement négatif (sinon $\sigma_n = 0$ et tous les x_j sont confondus, ce que l'on a exclu). Il y a donc un extrémum, et par ailleurs $A > 0$, il s'agit bien d'un minimum.

Revenons à l'écriture de la droite ainsi obtenue, dite droite de régression linéaire. Son équation est de la forme

$$Y \ = \ \frac{k_n}{\sigma_n^2} \ X \ + \ \overline{y}_n \ - \ \overline{x}_n \ \frac{k_n}{\sigma_n^2} \ , \tag{6.12}$$

qu'on peut, en multipliant par $\frac{1}{\tau_n}$, transformer en

$$\frac{Y - \overline{y}_n}{\tau_n} \ = \ \frac{k_n}{\sigma_n \ \tau_n} \ \frac{X - \overline{x}_n}{\sigma_n} \tag{6.13}$$

ou encore, en faisant apparaitre l'estimation du coefficient de corrélation,

$$\frac{Y - \overline{y}_n}{\tau_n} \ = \ \rho_n \ \frac{X - \overline{x}_n}{\sigma_n} \ . \tag{6.14}$$

Cette dernière équation ne fait intervenir que des quantités réduites et le coefficient de corrélation.