Outils de calcul probas-stat 3

par :

Faicel Chamroukhi

chamroukhi@unicaen.fr

http://chamroukhi.univ-tln.fr

Table des matières

Chapitre 1

Estimation de paramètres

1.1 Introduction

Supposons que, pour étudier et caractériser un phénomène physique, naturel ou autre, nous avons choisi et adopté un modèle probabiliste paramétrique représenté par une fonction de densité de probabilité $f(x;\theta)$ (ou une fonction de masse de probabilité $P(x;\theta)$ dans le cas discret). L'explication de ce phénomène nécessite donc l'estimation de ce modèle probabiliste à partir des données que l'on a observées (l'échantillon que l'on a à notre disposition). Ceci consiste donc à estimer le(s) paramètre(s) θ de ce modèle à partir des données observées $(x_1, \ldots, x_n)$ que l'on va supposer indépendantes dans le cadre de ce cours. Nous considérerons d'abord le cas d'un seul paramètre θ à estimer pour plus de clarté et simplicité et notons par $f(x;\theta)$ la densité ayant θ comme vrai paramètre mais qui est inconnu et que l'on cherche à estimer.

Le problème d'estimation de paramètres est donc celui de déterminer une fonction appropriée des données $(x_1, \ldots, x_n)$, que nous noterons $h(x_1, \ldots, x_n)$ qui donne la "meilleure" estimation de θ au sens de critère d'optimalité que nous verrons.

Nous avons donc

$$\hat{\theta} = h(x_1, \ldots, x_n)$$

et plus généralement, sous forme de variable aléatoire (car en effet pour des nouvelles réalisation des X_j, la valeur de $\hat{\theta}$ change) :

$$\hat{\Theta} = h(X_1, \ldots, X_n).$$

Cette statistique à déterminer s'appelle *un estimateur*.

1.2 Critères de qualité pour les estimations

Nous verrons maintenant un certain nombre de critères selon lesquels la qualité d'une estimation peut être évaluée. Ces critères définissent en général des propriétés souhaitables pour un estimateur et fournissent un guide par lequel la qualité d'un estimateur peut être comparée à celle d'un autre.

Notre objectif est de déterminer une statistique

$$\hat{\Theta} = h(X_1, \ldots, X_n)$$

qui fournit une bonne estimation de θ. Cette statistique à déterminer s'appelle *un estimateur* de θ, pour lequel des propriétés comme la moyenne, la variance ou la distribution fournissent une mesure de qualité pour cet estimateur. Une fois nous avons observé un échantillon de valeurs $(x_1, \ldots, x_n)$ la valeur de l'estimateur

$$\hat{\theta} = h(x_1, \ldots, x_n)$$

qui est une valeur numérique, est appelé *estimation* du paramètre θ.

Propriété 1.2.1. Absence de biais. *Un estimateur $\hat{\Theta}$ de θ est dis sans biais si*

$$\mathbb{E}[\hat{\Theta}] = \theta, \tag{1.1}$$

Ceci est clairement une propriété désirée pour $\hat{\Theta}$, qui consiste à dire que, en moyenne, on espère que $\hat{\Theta}$ est égal à la valeur du vrai paramètre θ.

Il est naturel que, si $\hat{\Theta} = h(X_1, \ldots, X_n)$ est à qualifier comme un bon estimateur de θ , non seulement sa moyenne doit être très proche du vrai paramètre θ mais aussi il faudrait qu'il y ait une grande probabilité que toute valeur $\hat{\theta}$ soit très proche de θ. Cela revient à sélectionner un estimateur de façon à ce que non seulement il soit sans biais mais aussi sa variance soit la plus petite possible. Ainsi, la seconde qualité désiré d'un estimateur et celle de *variance minimale*.

Propriété 1.2.2. Variance minimale. *Soit $\hat{\Theta}$ un estimateur sans biais de θ. Il est dit à variance minimale pour θ si, pour tout autre estimateur sans biais $\Theta^\star$ de θ, à partir du même échantillon, on a :*

$$\mathrm{var}(\hat{\Theta}) \leq \mathrm{var}(\Theta^\star)\cdot \tag{1.2}$$

Étant donné deux estimateurs sans biais pour un paramètre donné, celui ayant la variance plus faible est préférable, car une plus petite variance implique que les valeurs observées de l'estimateur (les estimations) ont tendance à être plus proche de sa moyenne qui est la valeur du vrai paramètre.

La question qui se pose donc est, étant donné un échantillon à partir duquel on construit plusieurs estimateurs sans biais, le quel parmi tout ces estimateurs qui a la variance minimale. Cette question est difficile mais, il existe un théorème qui montre qu'il est possible de déterminer la variance minimale possible (borne inférieure sur la variance) de tout estimateur sans biais obtenu à partir d'un échantillon donné.

Théorème 1.2.1. Borne de Cramer-Rao *Soit $(X_1, \ldots, X_n)$ un échantillon de v.a issues d'une population de densité $f(x;\theta)$ où θ est le paramètre inconnu, et soit $\hat{\Theta} = h(X_1, \ldots, X_n)$ un estimateur sans biais pour θ. La variance de $\hat{\Theta}$ satisfait l'inégalité suivante*

$$\mathrm{var}(\hat{\Theta}) \geq \left[n\mathbb{E}\left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 \right]^{-1} \tag{1.3}$$

si l'espérance et la dérivée existent. Un résultat analogue avec $p(X;\theta)$ en remplaçant $f(X;\theta)$ est obtenue lorsque X est discrète. Cette inéquation fournit donc une borne inférieure de la variance de n'importe quel estimateur sans biais. On note aussi que cette borne s'exprime généralement en fonction du vrai paramètre θ.

La quantité $n\mathbb{E}\left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2$ s'appelle l'*information de Fisher* contenue dans un échantillon de taille n et se note $\mathcal{I}_n(\theta)$. La borne inférieure de Cramér-Rao alors se définie aussi par :

$$\mathrm{var}(\hat{\Theta}) \geq \frac{1}{\mathcal{I}_n(\theta)} \tag{1.4}$$

et énonce donc que l'inverse de l'information de Fisher, $\mathcal{I}_n(\theta)$, d'un paramètre θ, est une borne inférieure de la variance d'un estimateur sans biais de ce paramètre.

⚠ **Remarque 1.2.1.** *En anglais, la borne inférieure de Cramér-Rao s'appelle* **Cramér-Rao Lower Bound** *abréviée par CRLB.*

⚠ **Remarque 1.2.2. Deuxième forme opérationnelle.** *Si le modèle est régulier, l'espérance* $\left[\mathbb{E}\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2\right]^{-1}$ *dans (1.3) est équivalente à* $-\left[\mathbb{E}\left(\frac{\partial^2 \ln f(X;\theta)}{\partial^2\theta}\right)\right]^{-1}$. *L'inégalité de Cramér-Rao peut également être alors mise sous la deuxième forme opérationnelle :*

$$\operatorname{var}(\hat{\Theta}) \geq -\left[n\mathbb{E}\left(\frac{\partial^2 \ln f(X;\theta)}{\partial^2\theta}\right)\right]^{-1}. \tag{1.5}$$

Cette expression alternative souvent offre des avantages de point de vue calcul.

1.2.1 Cas de plusieurs paramètres

Le résultat donnée par l'inéquation (1.3) peut être facilement étendu au cas de plusieurs paramètres. Soient $(\theta_1, \cdots, \theta_m)$ $(m \leq n)$ les paramètres inconnus du modèle (la densité) $f(x;\theta_1,\cdots,\theta_m)$ que l'on cherche à estimer à partir d'un échantillon de données de taille n. Utilisons la notation vectorielle suivante pour le vecteur paramètre et le vecteur pour les estimateurs :

$$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^T$$

et

$$\hat{\boldsymbol{\Theta}} = (\hat{\Theta}_1, \cdots, \hat{\Theta}_m)^T.$$

De façon similaire à (1.3), on peut montrer que l'inégalité de Cramér-Rao, pour le cas de paramètres multiples, est de la forme

$$\operatorname{cov}(\hat{\boldsymbol{\Theta}}) \geq \frac{\Lambda^{-1}}{n}, \tag{1.6}$$

ou le terme général de la matrice Λ est donné par :

$$\Lambda_{ij} = \Lambda(\theta_i,\theta_j) = \mathbb{E}\left[\left(\frac{\partial \ln f(X;\boldsymbol{\theta})}{\partial\theta_i}\right)\left(\frac{\partial \ln f(X;\boldsymbol{\theta})}{\partial\theta_j}\right)\right],\ i,j = 1,2,\cdots,m. \tag{1.7}$$

On a donc remplacé l'information de Fisher par la matrice Λ qui est la *matrice d'information de Fisher*.

Pour le $j$ème paramètre, l'inégalité (1.7) implique que

$$\operatorname{var}(\hat{\boldsymbol{\Theta}}_j) \geq \frac{1}{n\Lambda_{jj}},\ j = 1,\cdots,m.$$

⚠ **Remarque 1.2.3.** *La CRLB peut être transformée sous une transformation du paramètre. Supposons que, au lieu de θ, on s'intéresse à $\phi = g(\theta)$ qui est une transformation un-à-un et différentiable par rapport à θ ; alors*

$$\text{CRLB pour } \operatorname{var}(\hat{\boldsymbol{\Phi}}) = \left[\frac{\mathrm{d}\, g(\theta)}{\mathrm{d}\,\theta}\right]^2 \times \Big(\text{CRLB pour } \operatorname{var}(\hat{\Theta})\Big) \tag{1.8}$$

où $\hat{\boldsymbol{\Phi}}$ est un estimateur sans biais pour ϕ.

⚠ **Remarque 1.2.4. Efficacité d'un estimateur**. *Étant donné un estimateur sans biais $\hat{\Theta}$ de θ, le rapport de sa CRLB par sa variance est appelé l'***efficacité** *de $\hat{\Theta}$*

$$\begin{aligned} e(\hat{\Theta}) &= \frac{\text{CRLB pour } \text{var}(\hat{\mathbf{\Phi}})}{\text{var}(\hat{\Theta})} \\ &= \cdot \end{aligned} \tag{1.9}$$

L'efficacité d'un estimateur sans biais est ainsi inférieure ou égale à 1. Un estimateur sans biais ayant une efficacité égale à 1 est dit **efficace**.

Pour un estimateur, on souhaite aussi pouvoir, en augmentant la taille de l'échantillon, diminuer l'erreur d'estimation. Si c'est le cas, on dit que l'estimateur est consistant (on dit aussi convergent).

Propriété 1.2.3. Consistance (ou convergence). *Un estimateur $\hat{\Theta}$ est dit* **consistant** *pour θ si,*

$$\lim_{n\to\infty} \mathbb{P}(|\hat{\Theta} - \theta| > \epsilon) = 0 \qquad \forall\, \epsilon > 0. \tag{1.10}$$

On l'interprète comme le fait que la probabilité de s'éloigner de la vraie valeur du paramètre de plus de ϵ tend vers 0 quand la taille de l'échantillon augmente. L'estimateur donc converge vers la valeur à estimer quand la taille de l'échantillon augmente.

Théorème 1.2.2. *Soit $\hat{\Theta}$ un estimateur pour θ sur un échantillon de taille n. Alors, si*

$$\lim_{n\to\infty} \mathbb{E}[\hat{\Theta}] = \theta, \quad \textit{et} \quad \lim_{n\to\infty} \text{var}[\hat{\Theta}] = 0, \tag{1.11}$$

l'estimateur $\hat{\Theta}$ est dit *consistant* pour θ.

Afin de décrire une autre propriété d'un estimateur, qui est la *suffisance*, on introduit d'abord la notion de statistique suffisante.

Définition 1.2.1. *Soit X un vecteur d'observations de taille n avec les X_i $i.i.d.$. Soit θ un paramètre de la loi de probabilité des X_i. Une statistique $S(X)$ est dite exhaustive pour le paramètre θ (on dit aussi suffisante) si la probabilité conditionnelle d'observer X sachant $S(X)$ est indépendante de θ. Cela peut se traduire par la formule suivante :*

$$\mathbb{P}(X = x|S(X) = s, \theta) = \mathbb{P}(X = x|S(X) = s), \tag{1.12}$$

En pratique l'on se sert peu de cette formule pour montrer qu'une statistique est exhaustive et l'on préfère en règle générale utiliser le critère suivant appelé critère de factorisation (parfois aussi appelé critère de Fisher-Neyman) :

Soit $f_\theta(x)$ la densité de probabilité du vecteur aléatoire X. Une statistique S est exhaustive si et seulement s'il existe deux fonctions g et h telles que : $f_\theta(x) = h(x)\, g(\theta, S(x))$,

Propriété 1.2.4. Estimateur suffisant. *Soit $(X_1, X_2, \cdots, X_n)$ un échantillon $i.i.d.$ de X de distribution à paramètre θ. Si $Y = h(X_1, X_2, \cdots, X_n)$ est une statistique telle que, pour toute autre statistique $Z = g(X_1, X_2, \cdots, X_n)$, la distribution conditionnelle de Z, étant donné $Y = y$, ne dépend pas de θ, ç.à.d*

$$\mathbb{P}(Z = z|Y = y, \theta) = \mathbb{P}(Z = z|Y = y)$$

, alors Y est appelée une **statistique exhaustive (suffisante)** *pour θ. Si l'on a également $\mathbb{E}[Y] = \theta$, alors Y est dit un* **estimateur suffisant** *pour θ.*

Autrement dit, la définition de la suffisance dit que, si Y est une statistique suffisante pour θ, toute l'information de l'échantillon concernant θ est contenue dans Y.

Si une statistique suffisante pour un paramètre θ existe, le théorème 1.2.3 suivant, fournit un moyen de la trouver.

Théorème 1.2.3. Critère de factorisation de Fisher-Neyman *Soit $Y = h(X_1, \ldots, X_n)$ une statistique basée sur un échantillon i.i.d de taille n. Alors Y est une statistique exhaustive pour θ si et seulement si la densité jointe des X_i peut être factorisée selon la forme :*

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) = g\left(h(x_1, \ldots, x_n), \theta\right) \phi(x_1, \ldots, x_n). \tag{1.13}$$

Dans le cas discret on a :

$$P(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} P(x_i; \theta) = g\left(h(x_1, \ldots, x_n), \theta\right) \phi(x_1, \ldots, x_n). \tag{1.14}$$

Le résultat ci-dessus peut être étendu au cas de paramètres multiples. Soit $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^T$, $m \leq n$ le vecteur paramètre. Alors $Y_1 = h(X_1, \ldots, X_n), \ldots, Y_r = h(X_1, \ldots, X_n)$, $r \geq m$ est un ensemble de statistique suffisantes pour $\boldsymbol{\theta}$ si et seulement si

$$f(x_1, \ldots, x_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = g\left(\boldsymbol{h}(x_1, \ldots, x_n), \boldsymbol{\theta}\right) \phi(x_1, \ldots, x_n). \tag{1.15}$$

avec $\boldsymbol{h} = (h_1, \ldots, h_n)^T$. L'expression dans le cas discret est similaire.

1.3 Méthodes d'estimation

Il existe plusieurs méthodes d'estimation de paramètres, notamment estimation ponctuelle comme la méthode des moments, la méthode du maximum de vraisemblance, la méthode du maximum a posteriori ou la méthode d'estimation par intervalle. Dans ce cours, nous verrons en particulier les méthodes du maximum de vraisemblance et de maximum a posteriori (estimation ponctuelle) pour leur usage très commun en estimation de modèles probabilistes et aussi la méthode d'estimation par intervalle de confiance.

L'estimation d'un paramètre quelconque θ est *ponctuelle* si l'on associe une seule valeur à l'estimateur à partir des données observées sur un échantillon aléatoire. L'estimation *par intervalle* associe quant à elle à un échantillon aléatoire, un intervalle $[\hat{\theta}_a\,,\, \hat{\theta}_b]$ qui recouvre θ avec une certaine probabilité.

Chapitre 2

Méthode du maximum de vraisemblance

Introduite par le statisticien Fischer en 1922, la méthode du maximum de vraisemblance est devenue la méthode générale la plus importante de l'estimation d'un point de vue théorique. Son plus grand atout réside dans le fait que certaines propriétés très générale associées à cette procédure peuvent être dérivées et, dans le cas de grands échantillons, ce sont des propriétés optimales en fonction des critères d'absence de biais, de variance minimale, de consistance et d'efficacité.

2.1 Définition de la fonction de vraisemblance

Soit $f(x;\theta)$ la fonction de densité de probabilité d'une a.a X où θ est le paramètre (vrai paramètre) à estimer (Nous prenons ici le cas simple d'un seul paramètre). Soit $\boldsymbol{x} = (x_1, \ldots, x_n)$ un échantillon d'observations des variables aléatoires $(X_1, \ldots, X_n)$. La *vraisemblance* du paramètre θ pour l'échantillon $\boldsymbol{x}$ est donnée par la densité jointe de $\boldsymbol{x}$ et se note ainsi :

$$L(\theta; x_1, \ldots, x_n) = L(\theta; \boldsymbol{x}) = f(x_1, \ldots, x_n; \theta). \tag{2.1}$$

Sous l'hypothèse que les individus son $i.i.d.$, on a donc

$$\begin{aligned} L(\theta; x_1, \ldots, x_n) &= f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta) \\ &= \prod_{i=1}^{n} f(x_i;\theta). \end{aligned} \tag{2.2}$$

Dans le cas où les X_i sont des v.a discrètes, on a

$$\begin{aligned} L(\theta; x_1, \ldots, x_n) = L(\theta; \boldsymbol{x}) &= P(x_1;\theta) P(x_2;\theta) \cdots P(x_n;\theta) \\ &= \prod_{i=1}^{n} P(x_i;\theta). \end{aligned} \tag{2.3}$$

⚠ **Remarque 2.1.1.** *On peut aussi rencontrer la notation $L(\theta)$ de la vraisemblance de θ au lieu de $L(\theta; x_1, \ldots, x_n)$.*

On voit que pour des valeurs d'échantillon données, la fonction de vraisemblance est seulement fonction du paramètre θ.

2.2 Maximum de vraisemblance

Définition 2.2.1. Maximum de vraisemblance. *L'estimation de θ par la méthode du maximum de vraisemblance consiste à choisir, comme estimation de θ, la valeur de θ qui maximise la fonction de vraisemblance $L(\theta)$.*

En effet, en choisissant une valeur de θ qui maximise L (ou $\ln L$), cela revient à dire que, parmi les valeurs possible de θ, nous prenons la valeur qui rend le plus probable que possible l'évènement que les les valeurs de l'échantillon observé $(x_1, \ldots, x_n)$ viennent de la population de densité $f(x; \theta)$.

2.2.1 Cas d'un seul paramètre à estimer

Mathématiquement, le maximum de $L(\theta)$ correspond à la valeur de θ pour laquelle la dérivée de L par rapport à θ est nulle : $\frac{\mathrm{d}\, L(\theta)}{\mathrm{d}\, \theta} = 0$ (extremum) et la dérivée seconde par rapport à θ est négative : $\frac{\mathrm{d}^2\, L(\theta)}{\mathrm{d}^2\, \theta} < 0$ pour identifier le maximum parmi les possibles extrema obtenus. Ainsi, l'*estimateur du maximum de vraisemblance* (MV) de θ, souvent noté $\hat{\theta}$, à partir des valeurs de l'échantillon $(x_1, \cdots, x_n)$ peut être déterminé à partir de

$$\frac{\mathrm{d}\, L(\theta; x_1, \ldots, x_n)}{\mathrm{d}\, \theta}|_{\theta=\hat{\theta}} = 0. \tag{2.4}$$

qui permet d'identifier les extrema de $L(\theta)$ (mais ne permet pas de savoir lesquels parmi ces extrema sont des maxima (que nous recherchons) ou bien des minima (qui ne nous intéressent pas). Il faut donc, après que les solutions de l'équation aient été trouvées, sélectionner celles qui correspondent à des maxima. Un authentique vérifie

$$\frac{\mathrm{d}^2\, L(\theta; x_1, \ldots, x_n)}{\mathrm{d}^2\, \theta}|_{\theta=\hat{\theta}} < 0 \tag{2.5}$$

il faut donc sélectionner parmi les solutions de la première équation celles qui vérifient cette deuxième équation.

Bien que la plupart des vraisemblances soient différentiables (avec l'importante exception de la distribution uniforme), les solutions de l'équation (2.4) ne s'expriment pas toujours par des formes analytiques. On a souvent recours à des méthodes d'optimisations numériques pour identifier les maxima de la fonction de vraisemblance (par exemple comme en régression Logistique, mélange de densités, modèles de Markov cachés, etc) comme par exemple la montée de gradient, l'algorithme de Newton Raphson, l'algorithme EM, etc.

De manière équivalente, cela revient à annuler la dérivée du logarithme de la fonction de vraisemblance, la fonction de vraisemblance étant en effet positive et le logarithme est monotone et la vraisemblance atteint donc son maximum pour la même valeur que son logarithme. Le logarithme de la fonction de vraisemblance s'appelle *log-vraisemblance*. Manipuler le log de la fonction de vraisemblance au lieu de la vraisemblance elle même vient aussi du fait que, comme cette dernière s'écrit souvent comme produit de densités (de probabilités dans le cas discrets), cela peut résulter en des valeurs très faibles qui peuvent dans certains cas dépasser la précision de calculateurs. Ainsi, traiter des logarithmes revient plutôt à sommer et donc d'éviter des problèmes numériques.

L'estimateur de MV de θ est donc aussi donné par

$$\frac{\mathrm{d} \ln L(\theta; x_1, \ldots, x_n)}{\mathrm{d}\, \theta}|_{\theta=\hat{\theta}} = 0. \tag{2.6}$$

C'est l'équation de vraisemblance.

La solution désirée est une racine de cette equation, ou par équivalence celle en L, (si cette fonction admet des racines). Dans le cas où cette fonction est concave et admet donc une seule racine, l'estimateur du maximum de vraisemblance correspond à cette racine et on parle de *maximum global*. Cependant, la fonction de vraisemblance peut avoir plus d'un maximum (maxima). Dans ce cas, on parle de *maxima locaux* et l'estimateur du maximum de vraisemblance correspond au maximum global (lorsque tous les maxima ont été identifiés, seul le plus grand d'entre-eux doit être retenu).

2.2.2 Cas de plusieurs paramètres à estimer (Vraisemblance multivariée)

Plusieurs densités admettent plus d'un paramètre. Par exemple l'estimation d'une densité normale monodimensionnelle nécessite l'estimation de la moyenne μ et de la variance σ^2. L'extension au cas de plusieurs paramètres est simple. Dans le cas de m paramètres à estimer, la fonction de vraisemblance devient

$$\ln L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n)$$

et les estimateurs de MV de $\theta_j, j = 1, \cdots, m$, sont obtenus en résolvant simultanément le système d'équations de vraisemblance

$$\frac{\mathrm{d} \ln L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n)}{\mathrm{d}\, \theta_j}|_{\theta_j=\hat{\theta}_j} = 0 \quad \text{pour } j = 1, \ldots, m \tag{2.7}$$

et comme pour le cas univarié, mais dans ce cas multivarié c'est plus complexe, il faut en plus que au moins une des dérivées partielles secondes de L soit strictement négative pour au moins un j

$$\frac{\mathrm{d}^2 \ln L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n)}{\mathrm{d}^2\, \theta_j}|_{\theta_j=\hat{\theta}_j} = 0 \quad \text{pour au moins un } j$$

et le déterminant de la matrice des dérivées partielles secondes de L soit strictement positif :

$$\left| \frac{\mathrm{d}^2 \ln L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n)}{\mathrm{d}^2\, \theta_j} \right|_{\theta_j=\hat{\theta}_j} > 0$$

Cette dernière condition est en général difficile à vérifier, même dans les cas simples.

2.3 Propriétés

Soit $\hat{\theta}$ la valeur de l'estimateur du maximum de vraisemblance de θ estimée à partir de l'échantillon $(x_1, \cdots, x_n)$ de la population X de densité $f(x; \theta)$ L'estimation MV $\hat{\theta}$, calculée à partir de $(x_1, \cdots, x_n)$ peut être notée comme

$$\hat{\theta} = h(x_1, \cdots, x_n).$$

L'*estimateur du maximum de vraisemblance* $\hat{\Theta}$ [1] pour θ est donc

$$\hat{\Theta} = h(X_1, \cdots, X_n).$$

Propriété 2.3.1. Consistance. *L'estimateur obtenu par la méthode du maximum de vraisemblance est convergent :* $\lim_{n\to\infty} \mathbb{P}(|\hat{\Theta}_n - \theta| > \epsilon) = 0 \qquad \forall\, \epsilon > 0.$

Propriété 2.3.2. Absence de biais et efficacité asymptotiques. *Soit* $\hat{\Theta}$ *l'estimateur du maximum de vraisemblance (EMV) pour* θ *sous la densité* $f(x;\theta)$ *à partir d'un échantillon de taille* n*. Alors, quand* n *tend vers l'infini on a :*

$$\lim_{n\to\infty} \mathbb{E}[\hat{\Theta}] = \theta \tag{2.8}$$

L'EMV est donc asymptotiquement sans biais. On a également

$$\lim_{n\to\infty} \mathrm{var}[\hat{\Theta}] = \frac{1}{n\mathbb{E}\left[\left(\frac{\partial f(X;\theta)}{\partial\theta}\right)^2\right]} = \frac{1}{\mathcal{I}_n(\theta)} = \mathrm{CRLB}\cdot \tag{2.9}$$

L'EMV est donc asymptotiquement efficace Des résultats analogues sont obtenus lorsque la loi de X est discrète.

Propriété 2.3.3. Normalité asymptotique *En outre, la distribution de* $\hat{\Theta}$ *tend vers une distribution normale lorsque* n *devient grand. L'EMV est donc asymptotiquement normal.*

$$\sqrt{n}\,(\hat{\Theta} - \theta) \xrightarrow[n\to\infty]{\mathrm{loi}} \mathcal{N}\big(0, \mathcal{I}_n(\theta)^{-1}\big). \tag{2.10}$$

Propriété 2.3.4. Invariance. *On peut montrer que, si* $\hat{\Theta}$ *est l'EMV de* θ*, alors l'EMV d'une fonction bijective différentiable de* θ*, soit* $g(\theta)$*, est* $g(\hat{\Theta})$*.*

Cette importante propriété d'invariance implique que, par exemple, si $\hat{\Sigma}$ est l'EMV de l'écart type σ pour une distribution donnée, alors l'EMV de la variance σ^2 est $\hat{\Sigma}^2$.

2.4 Cas gaussien

Soit $(X_1, \ldots, X_n)$ un échantillon de variables aléatoires réelles issues d'une population de densité normale $\mathcal{N}(\mu, \sigma^2)$, alors la moyenne empirique $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur sans biais de l'espérance μ et la variance empirique *corrigée* $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_i)^2$ est un estimateur sans biais de la variance σ^2 et on a

$$\bar{X} \sim \mathcal{N}\big(\mu, \frac{\sigma^2}{n}\big) \tag{2.11}$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{2.12}$$

On peut donc remarquer que $\frac{\sqrt{n}(\bar{X}-\mu)}{S}$ suit une loi de student de paramètre $n-1$. Cela vient du fait que $(\bar{X} - \mu) \sim \mathcal{N}(0, \frac{\sigma^2}{n})$ donc $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim \mathcal{N}(0,1)$ et comme S^2 suit une loi de χ^2, en remplaçant σ par son estimateur S on a alors la loi de student t_{n-1}.

1. $\hat{\Theta}$ représente la variable aléatoire associé à $\hat{\theta}$.

2.5 Estimation par intervalle

Nous allons maintenant voir une autre approche d'estimation des paramètres *l'estimation par intervalle*. L'estimation par intervalle fournit, à partir d'un échantillon d'une population, non seulement des valeurs des paramètres à estimer, mais un intervalle de valeurs centré sur la valeur numérique estimée du paramètre inconnu avec un *niveau de confiance* donné. Ce niveau de confiance représente la probabilité que le vrai paramètre se trouve dans l'intervalle que l'on donne comme estimation. L'intervalle est appelé *intervalle de confiance* le niveau de confiance est aussi appelé *précision* ou *cœfficient de confiance*.

2.6 Intervalle de confiance

Définition 2.6.1. Intervalle de confiance. *Soit $(X_1, \ldots, X_n)$ un échantillon de variables aléatoires issues d'une population de densité $f(x;\theta)$, θ étant le (vecteur) paramètre à estimer. Supposons aussi que $T_1(X_1, \ldots, X_n)$ et $T_2(X_1, \ldots, X_n)$ deux statistiques sur l'échantillon telle que $T_1 < T_2$. L'intervalle $[T_1, T_2]$ est dit intervalle de confiance à $100(1-\alpha)\%$ pour θ si*

$$P[T_1 < \theta < T_2] = 1 - \alpha\cdot \tag{2.13}$$

α représente le *risque* que le vrai paramètre θ ne soit pas dans cet intervalle et $1-\alpha$ s'appelle niveau ou (cœfficient) de confiance. Une estimation par intervalle de confiance sera donc d'autant meilleure que l'intervalle sera petit pour un cœfficient de confiance grand (proche de 1) ou de manière équivalente pour un risque α proche de zéro. Les valeurs généralement prises pour $1-\alpha$ sont 0.90, 0.95, 0.99, and 0.999. Les limites de l'intervalle T_1 et T_2 sont appelés respectivement la *limite inférieure de confiance* et *limite supérieure de confiance*.

2.7 Cas d'estimation d'une gaussienne

2.7.1 Intervalle de confiance pour μ dans $\mathcal{N}(\mu, \sigma^2)$ avec σ connu

Chapitre 3

Méthode des Moindres Carrées

3.1 Méthode des Moindres Carrées

La méthode des moindres carrés consiste à estimer les paramètres d'un modèle en minimisant les écarts quadratiques entre les données observées, d'une part, et leurs valeurs attendues, d'autre part

Très utilisée notamment en régression où l'on cherche à expliquer la variation d'une variable de sortie (expliquée) Y, par la variation d'une variable d'entrée (explicative, covariable) X

Compte tenu de la valeur de X, la meilleure prédiction de Y (en termes d'erreur quadratique) est l'espérance $f(X)$ de Y sachant X.

On dit que Y est une fonction de X plus un bruit (erreur) :

$$Y = f(X) + E \tag{3.1}$$

f est appelée la fonction de régression, et E est un bruit souvent supposé d'espérance nulle.

L'estimateur des MC à des propriétés optimales d'absence de biais, de variance minimale (sous certaines conditions)

3.1.1 Définition des Moindres Carrés

Critères des moindres carrés

Soit le modèle

$$Y_i = f(X_i) + E_i \tag{3.2}$$

La fonction f est à estimer à partir d'un échantillon des couples de covariables X_i et leur réponses Y_i : $((x_1, y_1), ..., (x_n, y_n))$

Cette estimation est effectuée en minimisant la somme des écarts (erreurs) quadratiques

Définition 3.1.1. *Définition :* **Critères des moindres carrés** *L'erreur quadratique est donnée par la somme des carrés des résidus (Residual Sum of Squares (RSS)) :*

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2. \tag{3.3}$$

3.1.2 Moindres Carrés

Erreur quadratique dans le cas d'une fonction paramétrique

Soit $f(x;\theta)$ une fonction de paramètre θ à estimer. La somme des écarts qudratique dans ce cas est donnée par

$$RSS(\theta) = \sum_{i=1}^{n} \left(y_i - f(x_i;\theta)\right)^2 . \tag{3.4}$$

Définition 3.1.2. Définition de l'estimateur des moindres carrés *L'estimation de θ par la méthode des moindres carrés consiste à choisir, comme estimation de θ, la valeur de θ qui minimise la fonction RSS(θ).*

$$\hat{\theta} = \arg\min_{\theta} RSS(\theta) \tag{3.5}$$

En effet, en choisissant une valeur de θ qui minimise RSS(θ), cela revient à dire que, parmi les valeurs possible de θ, nous prenons la valeur qui correspond à une erreur minimale que les réponses y s'écartent de $f(x;\theta)$ pour l'échantillon observé $((x_1, y_1), \ldots, (x_n, y_n))$.

Cas d'un seul paramètre à estimer

Cas d'un seul paramètre θ

Définition 3.1.3. *Estimateur des moindres carrés (MC)L'estimateur de moindres carrés (MC) de θ, noté $\hat{\theta}$, à partir des valeurs de l'échantillon $((x_1, y_1), \cdots, (x_n, y_n))$ peut être déterminé à partir de*

$$\frac{\mathrm{d}\, RSS(\theta)}{\mathrm{d}\,\theta}|_{\theta=\hat{\theta}} = 0. \tag{3.6}$$

qui permet d'identifier les extrema de RSS(θ). Il faut donc, après que les solutions de l'équation aient été trouvées, sélectionner celles qui correspondent à des minima. Un minimum vérifie

$$\frac{\mathrm{d}^2\, RSS(\theta)}{\mathrm{d}^2\,\theta}|_{\theta=\hat{\theta}} > 0 \tag{3.7}$$

il faut donc sélectionner parmi les solutions de la première équation celles qui vérifient cette deuxième équation.

Moindres Carrés

⚠ **Remarque 3.1.1.** *Bien que la plupart des critères d'EQ soient différentiables, la minimisation du critère des MC ne s'effectue pas toujours de façon analytique*

⇒ On a souvent recours à des méthodes d'optimisations numériques (par exemple comme en réseau de neurones, etc)

⇒ la descente de gradient, l'algorithme de Newton Raphson, etc

Dans le cas où la fonction d'erreur est convexe, l'estimateur du Moindres Carrés fournit le minimum global. Cependant, dans beaucoup de problèmes réels, la fonction d'erreur n'est pas convexe et l'on a un minimum local ; atteindre le minimum global n'est pas toujours garanti

Des procédures algorithmiques existent (plusieurs initialisations, etc) et peuvent permettre d'atteindre un "bon" minimum local

Cas de plusieurs paramètres à estimer

Dans le cas d'un paramètre multiple $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$,le critère d'erreur est donné par

$$\text{RSS}(\boldsymbol{\theta}) = \text{RSS}(\theta_1, \ldots, \theta_m)$$

Les estimateurs de MC de $\theta_j, j = 1, \cdots, m$, sont obtenus en résolvant simultanément le système d'équations suivant

$$\frac{\partial \text{RSS}(\boldsymbol{\theta})}{\partial \theta_j}|_{\theta_j = \hat{\theta}_j} = 0 \quad \text{pour } j = 1, \ldots, m \tag{3.8}$$

3.1.3 Propriétés de l'estimateur des moindres carrés

Soit $\hat{\theta}$ la valeur de l'estimateur des Moindres Carrés $\hat{\Theta}$ de θ estimée à partir de l'échantillon $((x_1, y_1), \cdots, (x_1, y_n))$ de taille n

Propriété 3.1.1. *Absence de biais Si l'on suppose que les erreurs (le bruit) sont d'espérance nulle* ($\mathbb{E}[E_i] = 0$)*, l'estimateur des MC est sans bias*

Propriété 3.1.2. *variance minimale Si les erreurs sont d'espérance nulle* ($\mathbb{E}[E_i] = 0$) *et homoscédastiques décorrélées (*$\mathbb{E}[E_i^T E_i] = \sigma^2 \boldsymbol{I}$*) L'EMC est alors à variance minimale*

$\Rightarrow$ efficace et est donc le meilleur estimateur sans biais
Ces propriétés sont valables quelle que soit la distribution des erreurs.
Si en plus on fait l'hypothèse de normalité sur les erreurs ($e_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$) :

Propriété 3.1.3. *Normalité La distribution de* $\hat{\Theta}$ *est normale centrée sur le vrai paramètre* θ

Chapitre 4

Régression linéaire

4.1 Introduction

Les outils développés dans les chapitres 4 et 2 précédents pour l'estimation de paramètres et l'évaluation de la qualité d'un estimateur seront appliqués dans se chapitre à une famille de problèmes très connues en statistique et estimation qui est celle de la régression simple et multiple.

Une situation qui arrive couramment est celle dans laquelle une variable aléatoire Y est fonction d'une ou plusieurs variables indépendantes déterministes $(x_1, \ldots, x_m)$. Par exemple le prix d'un logement (Y) est une fonction de sa localisation (x_1) et de son âge (x_2); la durée de vie d'un composant électronique (Y) peut être liée à la température (x_1), la pression (x_2), etc ; la vitesse d'un automobiliste (Y) en fonction du temps t, etc. Notons que les variables indépendantes est aussi appelées *variables explicatives* car à travers elles on cherche à expliquer les variables Y qui sont dites *expliquées* [1]

L'objectif est donc d'estimer "la relation" entre Y et les variables indépendantes $(x_1, \ldots, x_m)$ étant donnés un échantillon $(Y_i, \ldots, Y_n)$ de la variable Y et les valeurs associées des variables explicatives x_j, $j = 1, \ldots, m$ pour chaque valeurs observée de Y_i, $i = 1 \ldots, n$.

4.2 Le modèle linéaire simple

prenons le cas simple où l'on suppose que Y ne dépend que d'une seule variable explicative x et que cette relation est supposée linéaire. en d'autres termes on a la relation suivante

$$Y = \beta_0 + \beta_1 x + \epsilon, \tag{4.1}$$

avec $(\beta_0, \beta_1) \in \mathbb{R}^2$ sont les deux paramètres de la droite de régression, β_0 appelé *ordonnée à l'origine (intercept)* et β_1 *pente (slope)*. Ce sont les **coefficients de régression**. ϵ est une variable aléatoire représentant un résidu (erreur de mesure). En effet, ce modèle suppose que la variable expliquée que l'on observée résulte du vrai modèle (ici le modèle linéaire représentée par la droite) et un bruit (de mesure par exemple) ou tout autre type d'erreur. Ce bruit est généralement supposé d'espérance nulle et de

1. En informatique et en particulier en machine learning (apprentissage), on trouve aussi l'appellation entrées/sorties pour respectivement variables explicatives et expliquées.

variance σ^2 et décorrélé ($\mathrm{cov}(\epsilon_i, \epsilon_j)_{i \neq j} = 0$). Dans ce cas σ^2 devient également un paramètre du modèle et est donc aussi à estimer. Dans le cadre de ce cours, on va supposer que ce bruit est en plus Gaussien. Il en découle donc qu'il est indépendant (les ϵ_i dont i.i.d).

Les deux paramètres (β_0, β_1) sont inconnus et donc à estimer. Cette estimation sera effectuée à partir d'un échantillon de couples $((x_1, Y_1), \ldots, (x_n, Y_n))$ [2]. Le modèle s'écrit donc sous la forme

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{4.2}$$

où ϵ_i est le bruit associée à la $i$ème variable aléatoire. Étant donnés donc une réalisation (valeur) y_i de chaque Y_i et le résidu associé (réalisation de la variable aléatoire représentant le bruit) que nous notons e_i on obtient alors :

$$y_i = \beta_0 + \beta_1 x_i + e_i, \tag{4.3}$$

Le modèle de régression linéaire est souvent estimé par la méthode des moindres carrés mais il existe aussi de nombreuses autres méthodes pour estimer ce modèle. On peut par exemple estimer le modèle par maximum de vraisemblance ou encore par inférence bayésienne (maximum a posteriori).

4.2.1 Estimation par moindres carrés

La méthode des moindres carrés est une approche d'estimation ponctuelle des paramètres de régression (β_0, β_1). Elle consiste à fournir les estimations $(\hat{\beta}_0, \hat{\beta}_1)$ qui minimisent la somme des écarts quadratiques (somme des carrées résidus) entre les valeurs observées y_i et l'espérance $\beta_0 + \beta_1 x_i$ du modèle de Y_i. D'après l'équation (4.3), l'écart entre la valeur d'une observation et l'espérance du modèle est donné par

$$e_i = y_i - (\beta_0 + \beta_1 x_i).$$

La somme des carrés des résidus est donc donnée par

$$\mathrm{RSS}(\beta_0, \beta_1) = Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2. \tag{4.4}$$

La fonction de deux variables Q est une fonction quadratique et sa minimisation, comme nous allons le voir ci-dessous, peut s'effectuer facilement de façon analytique.

Théorème 4.2.1. *Considérons le modèle de régression simple donné par l'équation (4.1). Soit $((x_1, y_1), \ldots, (x_n, y_n))$ un échantillon de valeurs observées de Y et leurs veleurs associées de x. Alors les estimations des moindres carrées ordinaires (MCO) de β_0 et β_1 sont données par :*

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{4.5}$$

$$\hat{\beta}_1 = \left[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right] \left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{-1}, \tag{4.6}$$

2. Ici nous utilisons la notations (x_i, Y_i) vu que x est déterministe mais cela ne change rien au modèle si X est aléatoire.

où $\bar{x}$ représente la moyenne empirique des x_i et $\bar{y}$ la moyenne moyenne empirique es y_i et sont donnés par

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

et

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \cdot$$

Preuve. Selon les MCO, les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ sont celles qui minimisent la somme des carrées des résidus (4.4) par rapport à $\hat{\beta}_0$ et $\hat{\beta}_1$. Mathématiquement on note

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta}_1) &= \arg\min_{(\beta_0,\beta_1)\in\mathbb{R}^2} Q(\beta_0, \beta_1) \\ &= \arg\min_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 \cdot \end{aligned} \tag{4.7}$$

Minimiser la fonction RSS revient à prendre les paramètres (β_0, β_1) qui annulent sa dérivée première et pour lesquels la dérivée second est positive. Nous aurons donc déterminé les estimations $(\hat{\beta}_0, \hat{\beta}_1)$. Ainsi, nous avons :

$$\begin{aligned}\frac{\partial Q}{\partial \beta_0} &= \frac{\partial \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2}{\partial \beta_0} = -2\sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big) \\ \frac{\partial Q}{\partial \beta_1} &= \frac{\partial \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i\big(y_i - (\beta_0 + \beta_1 x_i)\big) \cdot \end{aligned}$$

Ensuite, en annulant ces dérivés nous avons :

$$\begin{aligned}\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i &= 0 \\ \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= 0 \cdot \end{aligned}$$

ce qui donne

$$\begin{aligned} n\hat{\beta}_0 &= \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

qu'on appelle les *équations normales*. La première équation donne

$$n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

et en divisant par n on obtient la valeur de $\hat{\beta}_0$

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \bar{y} - \hat{\beta}_1 \bar{x} \cdot$$

La seconde équation donne

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \cdot$$

En remplaçant $\hat{\beta}_0$ par sa valeur on obtient

$$\sum_{i=1}^{n} x_i \bar{y} - \hat{\beta}_1 \sum_{i=1}^{n} x_i \bar{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \,,$$

ce qui donne enfin

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y}}{\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \bar{x}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot$$

Maintenant il faut vérifier si la dérivée partielle seconde de Q par rapport à au moins l'un des paramètres est positive (minimum de la fonction minimisée) et que le déterminant de la matrice des dérivées partielles secondes de Q est strictement positif. Partons de la dérivée partielle dans (4.8), on obtient

$$\begin{aligned} \frac{\partial^2 Q}{\partial^2 \beta_0} &= \frac{-2\partial \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)}{\partial \beta_0} \\ &= -2 \sum_{i=1}^{n} \frac{\partial \beta_0}{\partial \beta_0} = 2n \geq 0 \cdot \end{aligned} \tag{4.8}$$

Maintenant il reste à vérifier que le déterminant de la matrice des dérivées partielles secondes de Q est strictement positif. Le déterminant de cette matrice est donné par

$$\det \begin{pmatrix} \frac{\partial^2 Q}{\partial^2 \beta_0} & \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 Q}{\partial^2 \beta_1} \end{pmatrix} = \det \begin{pmatrix} 2n & 2\sum_i x_i \\ 2\sum_i x_i & 2\sum_i x_i^2 \end{pmatrix} = 4n \sum_i (x_i - \bar{x})^2 > 0 \cdot$$

Notez que ce déterminant est nul si tous les x_i prennent la même valeur. Ainsi, au moins deux valeurs x_i distinctes sont nécessaires pour la détermination des coefficients de régressions $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$.

4.2.2 Formulation vectorielle

maintenant nous reformulons le modèle que nous venons de voir sous forme de vecteurs-matrices. Comme nous allons le voir, les résultats sous la forme matricielle sont obtenus à partir de calculs simples. Cela permettra aussi de généraliser le modèle linéaire simple à des modèles généraux notamment la régression multiple.

Soit $(y_1, \ldots, y_n)$ l'ensemble des valeurs observées de la variable dépendante Y et $(x_1, \ldots, x_n)$ l'ensemble des valeurs observées de la variable explicative x.

⚠ **Remarque 4.2.1.** *L'ensemble des couples* $\big((x_1, y_1), \ldots, (x_n, y_n)\big)$ *s'appelle aussi ensemble d'apprentissage. Car c'est l'ensemble de données à partir duquel on va estimer notre modèle (donc apprendre le modèle) pour pouvoir ensuite prédire la valeur de* Y *pour une nouvelle valeur de* x

Selon le modèle de régression linéaire simple on a

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1, \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2, \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_i + \epsilon_n \cdot \end{aligned} \tag{4.9}$$

Si l'on note par $\mathbf{y} = (y_1, \ldots, y_n)^T$ le vecteur des valeurs d'observations de Y, $\mathbf{e} = (e_1, \ldots, e_n)^T$ le vecteur des valeurs des résidus, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ le vecteur paramètre à estimer, et enfin par $\mathbf{X}$ la *matrice de régression* (appelée aussi matrice de *design* ou de *Vendermonde*

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \tag{4.10}$$

le modèle (4.9) s'écrit donc sous la forme matricielle suivante :

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \cdot \tag{4.11}$$

La somme des écarts quadratiques donnée par l'équation (4.4) est maintenant donnée par :

$$\begin{aligned} \mathrm{RSS}(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}) = \mathbf{e}^T\mathbf{e} &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) && (4.12) \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \cdot && (4.13) \end{aligned}$$

L'estimation $\hat{\boldsymbol{\beta}}$ par moindres carrés de $\boldsymbol{\beta}$ s'obtient en minimisant (4.13) qui est une fonction quadratique en $\boldsymbol{\beta}$. En dérivons (4.13) par rapport à $\boldsymbol{\beta}$ et en annulant cette dérivée on obtient

$$-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \tag{4.14}$$

et donc les équations normales

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \cdot \tag{4.15}$$

En multipliant cette équation par $(\mathbf{X}^T\mathbf{X})^{-1}$ on obtient l'estimation de $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \cdot \tag{4.16}$$

Notez que l'inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ existe si les données comportent au moins deux valeurs distinctes de x_i.

Chapitre 5

Estimation par intervalle

Nous allons maintenant voir une autre approche d'estimation des paramètres *l'estimation par intervalle*. L'estimation par intervalle fournit, à partir d'un échantillon d'une population, non seulement des valeurs des paramètres à estimer, mais un intervalle de valeurs centré sur la valeur numérique estimée du paramètre inconnu avec un *niveau de confiance* donné. Ce niveau de confiance représente la probabilité que le vrai paramètre se trouve dans l'intervalle que l'on donne comme estimation. L'intervalle est appelé *intervalle de confiance* le niveau de confiance est aussi appelé *précision* ou *cœfficient de confiance*.

5.1 Intervalle de confiance

Définition 5.1.1. Intervalle de confiance. *Soit $(X_1, \ldots, X_n)$ un échantillon de variables aléatoires issues d'une population de densité $f(x;\theta)$, θ étant le (vecteur) paramètre à estimer. Supposons aussi que $T_1(X_1, \ldots, X_n)$ et $T_2(X_1, \ldots, X_n)$ deux statistiques sur l'échantillon telle que $T_1 < T_2$. L'intervalle $[T_1, T_2]$ est dit intervalle de confiance à $100(1-\alpha)\%$ pour θ si*

$$P[T_1 < \theta < T_2] = 1 - \alpha\cdot \tag{5.1}$$

α représente le *risque* que le vrai paramètre θ ne soit pas dans cet intervalle et $1-\alpha$ s'appelle niveau ou (cœfficient) de confiance. Une estimation par intervalle de confiance sera donc d'autant meilleure que l'intervalle sera petit pour un cœfficient de confiance grand (proche de 1) ou de manière équivalente pour un risque α proche de zéro. Les valeurs généralement prises pour $1-\alpha$ sont 0.90, 0.95, 0.99, and 0.999. Les limites de l'intervalle T_1 et T_2 sont appelés respectivement la *limite inférieure de confiance* et *limite supérieure de confiance*.

⚠Remarques

— L'intervalle de confiance est fonction de l'estimation du paramètre θ

— L'intervalle de confiance est également fonction de α. A taille d'échantillon n fixée, lorsqu'on augmente le niveau de confiance $1-\alpha$, la largeur de l'Intervalle de Confiance (IC)

— Pour un niveau de confiance $1-\alpha$ fixé, lorsqu'on augmente la taille de l'échantillon n, la largeur de l'IC diminue.

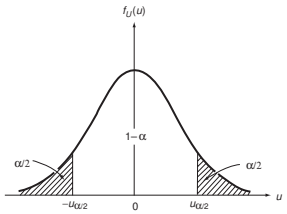


FIGURE 5.1 – Loi normale centrée réduite et IC à $[100(1-\alpha)]\%$

5.1.1 Calcul d'un intervalle de confiance

Soit a et b les bornes d'un intervalle de confiance $\text{IC}_{1-\alpha}(\theta)$ pour le paramètre θ on a On a :

$$\mathbb{P}(a < \theta < b) = 1 - \alpha \Longrightarrow \mathbb{P}(\theta < a) + \mathbb{P}(\theta > b) = \alpha \tag{5.2}$$

En posant $\alpha = \alpha_1 + \alpha_2$, il existe donc une infinité de choix possibles pour α_1 et α_2, et donc de choix pour a et b et donc de l'IC. Pour l'instant, nous ne considérons que le cas d'un intervalle de confiance bilatéral symétrique, où on a les mêmes risques $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$

Notons que, connaissant la loi de l'estimateur, il est possible de donner un intervalle de confiance. Ici nous considérons les intervalles de confiance les plus classiques.

5.1.2 Loi normale : Intervalle de confiance sur μ

Loi normale : Intervalle de confiance pour μ avec σ connu

On a vu que $\bar{X}$ est le meilleur estimateur de μ et que

$$U = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \quad \sim \quad \mathcal{N}(0,1) \tag{5.3}$$

En prenant des risques symétriques ($\frac{\alpha}{2}$), pour un risque fixé, on peut donc lire dans les tables de probabilités de la loi normale centrée réduite, le **quantile** $\mathbb{P}(U \leq \frac{\alpha}{2})$

Remarque : Comme le risque est symétrique ici, on a donc

$$\mathbb{P}(U \geq \frac{\alpha}{2}) = \alpha - \mathbb{P}(U \leq \frac{\alpha}{2}) = \frac{\alpha}{2} \tag{5.4}$$

La notion de quantile est définie de la façon suivante :

Définition 5.1.2. *pour une variable aléatoire continue X, le quantile q_α d'ordre α de la loi de X est telle que*

$$\mathbb{P}(U \leq q_\alpha) = \alpha \tag{5.5}$$

Remarque : la notation généralement utilisée pour les quantile est : u_α pour la loi normale, t_α pour la loi de Student à n degrés de liberté, χ^n_α pour la loi χ^2_n, etc

Le risque étant symétrique, d'après (5.4) on a

$$\mathbb{P}(-u_{\frac{\alpha}{2}} \leq U \leq u_{\frac{\alpha}{2}}) = 1 - \alpha \tag{5.6}$$

et d'après (5.3) on obtient

$$\mathbb{P}\left(\bar{X} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{5.7}$$

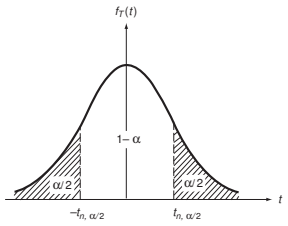


FIGURE 5.2 – Loi de Student à $n-1$ degrés de liberté et IC à $[100(1-\alpha)]\%$

d'où l'intervalle de confiance sur μ :

$$\mathrm{IC}_{1-\alpha}(\mu) = \left[\bar{X} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\,,\ \bar{X} + u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \tag{5.8}$$

exemple pour $\alpha = 0.05$

$$\mathrm{IC}_{0.95}(\mu) = \left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}\,,\ \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] \tag{5.9}$$

Loi normale : Intervalle de confiance pour μ avec σ inconnu

Pour calculer l'IC sur μ on a vu que la statistique à utiliser est $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$. Or, comme la variance σ^2 est inconnue, on utilise a sa place son meilleur estimateur : la variance empirique corrigée $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_i)^2$

La statistique à utiliser est donc

$$T = \frac{\sqrt{n}(\bar{X}-\mu)}{S} \tag{5.10}$$

On sait que $Z = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

la statistique (5.10) pour calculer l'IC s'écrit dont

$$T = \frac{U}{\sqrt{\frac{Z}{n-1}}} \sim \text{ Loi de Student à n-1 degrés de liberté} \tag{5.11}$$

Soit $t_{n-1,\frac{\alpha}{2}} = \mathbb{P}(T \leq \frac{\alpha}{2})$ le quantile d'ordre $\alpha/2$ de la loi de Student à $n-1$ degrés de liberté.

L'intervalle de confiance est donc donné par

$$\mathbb{P}(-t_{n-1,\frac{\alpha}{2}} \leq T \leq t_{n-1,\frac{\alpha}{2}}) = 1-\alpha \tag{5.12}$$

On obtient donc l'intervalle de confiance comme précédemment

$$\mathrm{IC}_{1-\alpha}(\mu) = \left[\bar{X} - t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\,,\ \bar{X} + t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right] \tag{5.13}$$

où $t_{n-1,\frac{\alpha}{2}}$ est le quantile d'ordre $\alpha/2$ de la loi de Student à $n-1$ degrés de liberté

⚠ **Remarque 5.1.1.** *Si la loi de X n'est pas normale, on sait d'après le théorème central limite que lorsque la taille d'échantillon est grande, $\bar{X}$ suit une loi normale, et donc les résultats précédents sont applicables.*

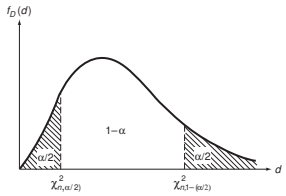


FIGURE 5.3 – Loi de χ^2 à n degrés de liberté et IC à $[100(1-\alpha)]\%$

5.1.3 Loi normale : Intervalle de confiance sur σ^2

Loi normale : Intervalle de confiance pour σ lorsque μ connu

On sait que la variance observée $V^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ constitue le meilleur estimateur de σ^2 lorsque μ est connue.

D'autre part :

$$D = \frac{n}{\sigma^2}V^2 \sim \chi^2_n \tag{5.14}$$

Soit $\chi^2_{n,\frac{\alpha}{2}} = \mathbb{P}(D \leq \frac{\alpha}{2})$ le quantile d'ordre $\frac{\alpha}{2}$ de la loi de χ^2 à n degrés de liberté.

l'IC$_{1-\alpha}(\sigma^2)$ est donc donné par

$$\mathbb{P}\left(\chi^2_{n,\frac{\alpha}{2}} \leq D = \frac{n}{\sigma^2}V^2 \leq \chi^2_{n,1-\frac{\alpha}{2}}\right) = 1-\alpha$$

donc

$$\mathbb{P}\left(\frac{nV^2}{\chi^2_{n,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{nV^2}{\chi^2_{n,\frac{\alpha}{2}}}\right) = 1-\alpha$$

finalement on obtient

$$\text{IC}_{1-\alpha}(\sigma^2) = \left[\frac{nV^2}{\chi^2_{n,1-\frac{\alpha}{2}}}, \leq \frac{nV^2}{\chi^2_{n,\frac{\alpha}{2}}}\right] \tag{5.15}$$

Loi normale : Intervalle de confiance pour σ lorsque μ inconnu

Lorsque μ est inconnue, on sait que que la variance empirique corrigée $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ constitue le meilleur estimateur de σ^2

On sait également que :

$$D = \frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1} \tag{5.16}$$

l'IC$_{1-\alpha}(\sigma^2)$ est donc donné par

$$\mathbb{P}\left(\chi^2_{n-1,\frac{\alpha}{2}} \leq D = \frac{n-1}{\sigma^2}S^2 \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha$$

donc

$$\text{IC}_{1-\alpha}(\sigma^2) = \left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right] \tag{5.17}$$

Remarque : Ces intervalles de confiance sur la variance ne sont valables que pour une loi normale. Contrairement au cas de la moyenne, ces résultats ne peuvent être étendus aux cas d'autres lois

Chapitre 6

Tests d'hypothèses

Dans cette partie nous allons étudier le *test statistique d'hypothèse*. Un *test statistique* est un procédé qui permet de décider entre deux ou plusieurs hypothèses sur une population selon les résultats obtenus à partir d'un échantillon de cette population. Généralement, on teste une hypothèse sur laquelle on se demande si les données observées fournissent une évidence suffisante pour la rejeter, sinon elle est retenue. Cette hypothèse s'appelle l'*hypothèse nulle* et se note H_0. Par exemple, si le test concerne la valeur d'un paramètre θ, cette hypothèse nulle peut s'écrire

$$H_0 : \theta \in \Theta_0 \tag{6.1}$$

où Θ_0 est l'ensemble de valeurs supposée du paramètre θ selon H_0. Toute autre hypothèse qui diffère de l'hypothèse nulle s'appelle l'*hypothèse alternative* (ou contre-hypothèse) qui se note H_1. L'hypothèse nulle est donc testée contre l'hypothèse alternative.

Une hypothèse est dite *simple* si elle ne contient qu'un seul élément, ce qui est généralement le cas pour $H_0 : \theta = \theta_0$; sinon elle est composite. L'hypothèse alternative est généralement composite

$$H_1 : \theta \in \Theta_1 \tag{6.2}$$

avec Θ_1 un sous ensemble de l'ensemble des paramètres disjoint de θ_0. H_1 se ramène so uvent aux trois cas suivants

1. $H_1 : \theta < \theta_0$,
2. $H_1 : \theta > \theta_0$,
3. $H_1 : \theta \neq \theta_0$.

Dans les deux premiers cas, le test est dit *unilatéral* et dans le dernier le test est dit *bilatéral*.

6.1 Région de rejet d'un test

Soit X une variable aléatoire et $\mathcal{X}$ l'ensemble de ses valeurs. Le test s'effectue en trouvant un sous-ensemble $R \subseteq \mathcal{X}$ appelé la *région de rejet*. Ainsi, si $X \in R$,

l'hypothèse nulle (H_0) est rejetée, sinon, elle est retenue. Cette région se définit sou la forme suivante

$$R = \{x : T(x) > s\} \tag{6.3}$$

où T est une *statistique de test* et s un *seuil*. le problème en test d'hypothèse est donc de trouver une statistique de test convenable et une valeur convenable pour le seuil de rejet s.

6.2 Erreurs associées à un test

Bien sur, en effectuant un test d'hypothèses, on peut se tromper en rejetant l'hypothèse nulle ou en l'acceptant. Il existe donc deux types d'erreur : l'erreur de première espèce (dite aussi erreur de type I) et l'*erreur de deuxième espèce* (dite aussi erreur de type II). L'erreur de première espèce correspond au cas où l'on rejette H_0 (décider H_1) alors que celle-ci est vraie. L'erreur de deuxième espèce correspond quant à elle au cas où l'on rejette H_1 (décider H_0) alors que celle-ci est vraie. Les décisions possibles sont récapitulées par le tableau suivant. Pour chacune des deux erreurs, on associe un une

Décision\ Vérité	H_0	H_1
H_0	décision correcte	erreur de deuxième espèce
H_1	erreur de première espèce	décision correcte

TABLE 6.1 – Récapitulatif des décisions en test d'hypothèse

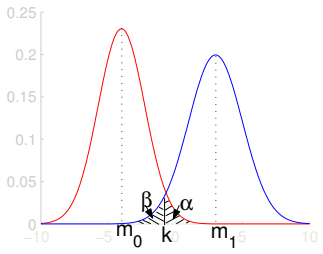
probabilité (*un risque*). Le risque de première espèce est notée α. Il représente le risque de rejeter H_0 à tort. Des valeurs de ce risque sont $1\%, 5\%, 10\%$ qui correspondent aux niveaux de confiance $99\%, 95\%$ et 90%. Le *niveau de confiance du test* est donc $1-\alpha$ qui correspond à retenir H_0 à raison. Le risque de deuxième espèce représente quant à lui le risque de retenir H_0 à tort. Il est noté β.

La *puissance d'un test* est la probabilité de rejeter l'hypothèse nulle à raison. La puissance du test est donc le complément de l'erreur de deuxième espèce et est donc égale à $1-\beta$.

On peut résumer cela par le tableau suivant :

Décision\ Vérité	H_0	H_1
H_0	niveau de confiance $1-\alpha$	risque β
H_1	risque α	puissance de test $1-\beta$

TABLE 6.2 – Récapitulatif sur les risques associés à un test d'hypothèses



6.3 Statistiques de test

Le choix de la statistique de test et de la région de rejet s'effectue de façon à maximiser la puissance du test $1-\beta$ pour un risque de première espèce α fixé.

6.3.1 Test du rapport de vraisemblance

Si l'on se place dans le cadre d'un test entre deux hypothèses simples

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1,$$

Le théorème de Neyman et Pearson montre que le *test du rapport de vraisemblance* est le test le plus puissant avec un risque α. Selon ce test, la région critique (de rejet) optimale est définie par

$$R = \{x : \frac{L(\theta_1; x)}{L(\theta_0; x)} > s_\alpha\}$$

avec $L(\theta_k; x)$ étant la vraisemblance de θ_k pour x. Le seuil de rejet s_α, qui dépende de α, est déterminé par $\alpha = \mathbb{P}_{\theta_0}(X \in R)$.

Test du rapport de vraisemblance (ou Test de Neyman-Pearson) : Exemple

Soit un échantillon d'observations $i.i.d.$ $(x_1, \ldots, x_n)$ où $X_i \sim \mathcal{N}(x_i; \mu, \sigma^2)$
supposons que la variance σ^2 est connue et que l'espérance μ est inconnue
Considérons le test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu = \mu_1,$$

avec $\mu < \mu_1$

On la vraisemblance de μ pour l'échantillon $\mathbf{x} = (x_1, \ldots, x_n)$ est

$$\begin{aligned} L(\mu; \mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\,\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] \\ &= \frac{1}{(\sigma^2\sqrt{2\,\pi})^n} \exp\left[-\sum_{i=1}^{n} \frac{1}{2\sigma}\left(x_i - \mu\right)^2\right] \end{aligned} \tag{6.4}$$

le rapport de vraisemblance est don donné par

$$\frac{L(\mu_1; \mathbf{x})}{L(\mu_0; \mathbf{x})} = \exp\left[\frac{1}{2\sigma^2} 2\left(\mu_1 - \mu_0\right) \sum_{i=1}^{n} x_i - \frac{n}{2\sigma^2}\left(\mu_1^2 - \mu_0^2\right)\right] \tag{6.5}$$

Donc, en prenant le logarithme, $\frac{L(\mu_1;\mathbf{x})}{L(\mu_0;\mathbf{x})} > s_\alpha$ est équivalent à

$$\bar{x} > \log(s_\alpha)\frac{\sigma^2}{n(\mu_1 - \mu_0)} + \frac{(\mu_1 + \mu_0)}{2} = \text{cste}$$

On a vu que cette $cste$ qui dépende α est déterminé par $\alpha = \mathbb{P}_{\mu_0}(X \in R)$ qui vaut dans ce cas $\alpha = \mathbb{P}_{\mu_0}(\bar{x} > \text{cste})$

La région de rejet du test est donc donnée par

$$R = \{\mathbf{x} : \bar{x} > \mu_0 + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\} \tag{6.6}$$

6.3.2 Test de Wald

Soit θ un paramètre et soit $\hat{\Theta}$ un estimateur de ce paramètre et soit $\hat{\sigma}$ l'écart type de cet estimateur $\hat{\Theta}$. Considérons le test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_1,$$

Supposons que $\hat{\Theta}$ est asymptotiquement normal : $\frac{\sqrt{n}(\widehat{\Theta}-\theta_0)}{\sqrt{\text{var}(\hat{\Theta})}} \xrightarrow[n\to\infty]{loi} \mathcal{N}\big(0,1\big)$. Dans le *test de Wald*, la statistique de test est donnée par

$$\frac{\widehat{\Theta} - \theta_0}{\sqrt{\text{var}(\hat{\Theta})}} \tag{6.7}$$

où $\sqrt{\text{var}(\hat{\Theta})}$ représente l'écart type de l'estimateur. Le test de Wald consiste à comparer cette statistique à la loi normale centrée réduite. Il consiste alors à rejeter H_0 si

$$\left| \frac{(\widehat{\Theta} - \theta_0)}{\sqrt{\text{var}(\hat{\Theta})}} \right| > u_{\alpha/2}$$

où $u_{\alpha/2}$ est le quantile d'ordre $\alpha/2$ de la loi normale centrée réduite.

Exemple : cas d'un grand échantillon gaussien

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0,$$

lorsque σ est connue

la statistique de test sous H_0 dans ce cas est donnée par

$$U = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{6.8}$$

On sait que $U \sim \mathcal{N}(0,1)$

On rejette donc H_0 si

$$\left| \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right| > u_{\frac{\alpha}{2}}$$

où $u_{\frac{\alpha}{2}}$ est le quantile d'ordre $\alpha/2$ de la loi normale centrée réduite

ou par équivalence : rejeter H_0 si

$$\left|\bar{X} - \mu_0\right| > u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$