STATISTIQUE COMPUTATIONNELLE – 6

INTERVALLE DE CONFIANCE

L'estimation ponctuelle fournit la valeur la plus probable du paramètre sous test, en se fondant sur un ensemble d'observations $X_1 = x_1, \ldots, X_n = x_n$ effectuées en utilisant un échantillon de taille n. Néanmoins on ne peut pas affirmer que cette valeur coincide toujours avec la vraie valeur. Plutôt que de chercher à savoir si la valeur estimée est proche ou éloignée de la vraie valeur (ce qui aurait exiger d'avoir une métrique ad hoc dans l'espace des paramètres), on cherche pour un paramètre inconnu Θ à déterminer, à partir d'un estimateur choisi $\widehat{\Theta}$, un intervalle dans lequel il est probable de supposer que se trouve la vraie valeur avec une probabilité de se tromper fixée par l'utilisateur.

6.1

Définition de l'intervalle de confiance

Soit $\theta \in \Theta$ avec $\Theta \subseteq \mathbb{R}^m$ le domaine de variation du paramètre de la population dont on cherche à évaluer l'intervalle de confiance.

Considérons une famille $\mathfrak{I}(\Theta) = \{I_x(\theta) \mid x \in \mathcal{B}_{\mathbb{R}}\}$ d'intervalles de Θ qui ont la propriété suivante :

$$\forall \theta \in \Theta \ : \ P_\theta[x \mid \theta \in I_x(\theta)] = 1 - \alpha \ ; \ \alpha \in [0,1]$$

où α est la probabilité que l'intervalle ne contient pas la vraie valeur du paramètre θ.

La détermination des bornes de l'intervalle de confiance dépend du partage de α en α_1 et α_2. Nous pouvons avoir deux cas de figure :

– Recherche d'un intervalle bilatéral $[a,b]$, correspondant à $\alpha_1 \neq 0$, $\alpha_2 \neq 0$. Dans le cas où la loi est symétrique (e.g. loi normale) nous avons $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$.

– Recherche d'un intervalle unilatéral. Deux cas :
  – $[a,+\infty[$ associé à $\alpha_1 = \alpha$ et $\alpha_2 = 0$. Cet intervalle est à utiliser dans le cas de recherche d'un intervalle du paramètre θ de la forme $\theta > a$ (durée de vie, résistance à la rupture, etc.).
  – $]-\infty,b]$ associé à $\alpha_1 = 0$ et $\alpha_2 = \alpha$. Cet intervalle est à utiliser dans le cas de recherche d'un intervalle du paramètre θ de la forme $\theta < b$ (nombre de pièces défectueuses, temps d'attente, etc.).

6.2

Construction des intervalles de confiance

Afin de construire un intervalle de confiance, on détermine un estimation $u(x,\theta)$ qui, pour chaque $x \in \mathcal{B}_{\mathbb{R}}$, est une fonction monotone de $\theta \in \Theta$. On détermine ensuite deux valeurs u_1, u_2 telles que :

$$P_\theta [u_1 \leq u(x,\theta) \leq u_2] = 1 - \alpha \ ; \ \alpha \in [0,1]$$

et, comme u est monotone, nous avons :

$$P_\theta [\theta_1(x) \leq \theta \leq \theta_2(x)] = 1 - \alpha \ ; \ \alpha \in]0,1[$$

$[\theta_1(x), \theta_2(x)]$ est l'intervalle de confiance.

La règle de décision est fonction du risque α. En effet, on cherche à calculer une valeur z_0 telle que $p(\theta \in W) = \alpha$, avec W le complémentaire de l'intervalle de confiance : $W =]-\infty, \theta_1(x)[\cup]\theta_2(x), +\infty[$. Pour calculer cette valeur on utilise la table de loi correspondante au problème.

Il va de soi que pour calculer l'intervalle de confiance d'un estimateur, il faut connaître sa loi de distribution.

Nous présentons dans la suite quelques exemples des fonctions monotones qui peuvent être utilisées comme des estimateurs pour le calcul des intervalles de confiance. On prendra toujours un échantillon $X_1 = x_1, \ldots, X_n = x_n$ de taille n.

6.3

Intervalles de confiance pour les moyennes

Nous donnons ci-après quelques exemples de construction d'intervalle de confiance pour la moyenne d'une suite des variables aléatoires qui suivent la loi normale. La démarche reste la même pour les autres lois sous l'hypothèse que $n \geq 30$.

Puisqu'on veut établir un estimateur de la moyenne on aura $\Theta = \mu$ et on prendra comme estimateur $\widehat{\Theta} = \overline{X}$ qui est, rappelons-le, une v.a.

Loi normale $\mathcal{N}(\mu,\sigma^2)$ avec variance connue

Considérons l'échantillon aléatoire $X_1,\ldots,X_n$ issu d'une population d'espérance mathématique μ et dont la variance σ^2 est connue. On cherche un intervalle dans lequel se trouve la moyenne μ avec probabilité $1-\alpha$, où $\alpha \in]0,1[$ risque de se tromper, que l'on appelle *niveau* ou ***seuil***.

On suppose que les v.a. X_k suivent la loi normale $\mathcal{N}(\mu,\sigma^2)$. On a donc pour la v.a. $\overline{X}$, l'espérance $E(\overline{X}) = \mu$ et la variance $V(\overline{X}) = \frac{\sigma^2}{n}$. Donc la v.a. centrée, réduite issue de $\overline{X}$ est $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.

On pose donc $\theta = \mu$ et on construit l'estimateur

$$U(X_1,\ldots,X_n;\theta) = \sqrt{n}\,\frac{\overline{X}-\theta}{\sigma} \sim \mathcal{N}(0,1)$$

qui est une fonction monotone par rapport à θ.

On cherche deux nombres a et b , avec $a < b$ et tels que :

$$P_\mu\left[a \le \sqrt{n}\,\frac{\overline{X}-\mu}{\sigma} \le b\right] = 1-\alpha \qquad \Rightarrow$$

$$P_\mu\left[\overline{X} - b\,\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + a\,\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha \qquad \Rightarrow$$

$$P_\mu\left\{\mu \in \left[\overline{X} - b\,\frac{\sigma}{\sqrt{n}},\ \overline{X} + a\,\frac{\sigma}{\sqrt{n}}\right]\right\} = 1-\alpha$$

avec $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$. La longueur de l'intervalle est $\frac{b-a}{\sqrt{n}}\sigma$. On considère un intervalle symétrique autour de 0, et on pose $b = c$ et $a = -c$. Pour que la longueur $2c$ soit minimale il faut que c est le $\alpha/2$ quantile de $\mathcal{N}(0,1)$ qui sera par la suite noté $z_{\alpha/2}$. Si donc on considère une réalisation de l'échantillon $x_1,\ldots,x_n$ avec $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} X_j$ la moyenne empirique, on a

$$P_\Theta\left\{\mu \in \left[\overline{x} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right]\right\} = 1-\alpha$$

La quantité $z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}$ représente la marge d'erreur de l'estimation $\widehat{\theta} = \overline{x}$, c-à-d. $|\overline{x}-\mu| \le z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}$ avec probabilité $1-\alpha$.

La marge d'erreur est utile quand on veut calculer la taille minimale n de l'échantillon afin de ne pas dépasser une marge d'erreur R_Θ donnée. En effet, on a la relation

$$2\,z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} = 2R_\Theta$$

d'où on obtient la taille minimale de l'échantillon

$$n = \left[\frac{z_{\alpha/2}\cdot\sigma}{R_\Theta}\right]^2$$

Loi normale $\mathcal{N}(\mu,\sigma^2)$ avec variance inconnue

La variance est inconnue. Si la taille de l'échantillon n est supérieure ou égale à 30, on procède comme précédemment quand la variance était connue et on remplace dans les formules la quantité σ^2 par son estimation $s_X^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2$ (ou par l'estimation corrigée $s_X^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2$).

Considérons maintenant le cas où $n<30$.On utilise l'estimateur de la variance $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j-\overline{X}\right)^2$. On pose $\theta=\mu$ et on construit la fonction

$$U(X_1,\ldots,X_n;\theta) = \sqrt{n}\frac{\overline{X}-\theta}{S} = U_{n-1}\sim t(n-1)$$

qui est une fonction monotone par rapport à θ et qui suit la loi de Student $t(n-1)$ à $n-1$ degrés de liberté (voir poly du cours, chapitre 3, §15). Par conséquent nous avons :

$$P_\mu[a\le U_{n-1}\le b] = 1-\alpha \qquad \Rightarrow$$

$$P_\mu\left[a\le\sqrt{n}\frac{\overline{X}-\mu}{s}\le b\right] = 1-\alpha \qquad \Rightarrow$$

$$P_\mu\left[\overline{X}-b\frac{s}{\sqrt{n}}\le\mu\le\overline{X}+a\frac{s}{\sqrt{n}}\right] = 1-\alpha$$

La longueur de l'intervalle est $\frac{b-a}{\sqrt{n}}s$. Cette longueur est minimale si $b=c$ et $a=-c$, avec $c=t_{n-1;\alpha/2}$.Si donc on considère une réalisation de l'échantillon $x_1,\ldots,x_n$ avec $\bar{x}=\frac{1}{n}\sum_{j=1}^{n}X_j$ la moyenne empirique, on a

$$P_\mu\left\{\mu\in\left[\bar{x}-t_{n-1;\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x}+t_{n-1;\alpha/2}\frac{s}{\sqrt{n}}\right]\right\} = 1-\alpha$$

Si on veut que l'intervalle ait une longueur égale à $2c$, alors il faut prendre

$$n = \left[\frac{t_{n-1;\alpha/2}\cdot s}{c}\right]^2$$

6.4

Intervalle de confiance pour les variances

Nous donnons ci-après quelques exemples de construction d'intervalle de confiance pour la variance d'une suite des variables aléatoires qui suivent la loi normale. La démarche reste la même pour les autres lois.

Loi normale $\mathcal{N}(\mu,\sigma^2)$ avec moyenne connue

Considérons l'échantillon aléatoire $X_1,\ldots,X_n$ issu d'une population de variance σ^2 et dont la moyenne μ est connue. On cherche un intervalle dans lequel se trouve la variance σ^2 avec probabilité $1-\alpha$, où $\alpha \in]0,1[$ risque de se tromper.

On pose $\theta = \sigma^2$ et on construit l'estimateur

$$U(X_1,\ldots,X_n;\theta) = \frac{1}{\theta}\sum_{j=1}^{n}(X_j-\mu)^2 = \frac{nS_\mu^2}{\sigma^2} \sim \chi_n^2$$

qui est une fonction monotone par rapport à θ et qui suit la loi du χ^2 avec n degrés de liberté (cf. poly du cours, chapitre 3, §14) et où nous avons posé $S_\mu^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j-\mu)^2$.

On cherche donc à calculer deux nombres a et b tels que :

$$P\left[a \le \chi_n^2 \le b\right] = 1-\alpha \qquad \Rightarrow$$

$$P_{\sigma^2}\left[a \le \frac{ns_\mu^2}{\sigma^2} \le b\right] = 1-\alpha \qquad \Rightarrow$$

$$P_{\sigma^2}\left[\frac{ns_\mu^2}{b} \le \sigma^2 \le \frac{ns_\mu^2}{a}\right] = 1-\alpha \qquad \Rightarrow$$

$$P_{\sigma^2}\left\{\sigma^2 \in \left[\frac{ns_\mu^2}{b},\ \frac{ns_\mu^2}{a}\right]\right\} = 1-\alpha$$

La longueur de l'intervalle est $\left(\frac{1}{a}-\frac{1}{b}\right)ns_\mu^2$. Bien qu'il ne s'agit pas d'un choix optimal, pour évaluer a et b on prend

$$a = \chi^2_{n;1-\alpha/2} \quad ; \quad b = \chi^2_{n;\alpha/2}$$

Loi normale $\mathcal{N}(\mu,\sigma^2)$ avec moyenne inconnue

La démarche est la même que dans le cas où la moyenne est connue. On pose $\theta = \sigma^2$ et on construit l'estimateur

$$U(X_1,\ldots,X_n;\theta) = \frac{1}{\theta}(n-1)\cdot S^2 \sim \chi^2_{n-1}$$

qui est une fonction monotone par rapport à θ. Nous avons noté $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j-\overline{X}\right)^2$.

Pour μ on utilise l'estimateur $\overline{X} = \frac{1}{n}\sum_{i=1}^{n}x_i$. Ainsi la fonction pour $\theta = \sigma^2$ devient $S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j-\overline{X}\right)^2$. La longueur de l'intervalle est maintenant $\left(\frac{1}{a}-\frac{1}{b}\right)(n-1)\,s_X^2$ et on prend

pour les bornes de l'intervalle

$$a = \chi^2_{n-1;1-\alpha/2} \quad ; \quad b = \chi^2_{n-1;\alpha/2}$$

6.5
Le cas particulier des proportions

Nous avons déjà vu le cas où on cherche à estimer la proportion ou le nombre d'éléments dans une population qui présentent une caractéristique particulière. Posons q la vraie proportion dans la population et $\widehat{q}$ la fréquence observée dans un échantillon de taille n. L'estimateur $\widehat{Q}$ est une v.a. et si $n\widehat{q} \geq 5$ et $n(1-\widehat{q}) \geq 5$, alors $\widehat{Q} \sim \mathcal{N}\left(q, \frac{q(1-q)}{n}\right)$. La v.a. standarisée correspondante à Q et qui suit la loi normale centrée réduite est donnée par

$$Z_Q = \frac{\widehat{Q} - q}{\sqrt{\frac{q(1-q)}{n}}}$$

Le problème qui se pose ici est qu'on ne connaît pas q. Dans ce cas, nous pouvons

– soit remplacer q par la proportion $\widehat{q}$ observée dans l'échatillon et nous avons donc pour l'intervalle de confiance

$$P_Q\left\{q \in \left[\widehat{q} - z_{\alpha/2}\sqrt{\frac{\widehat{q}(1-\widehat{q})}{n}},\ \widehat{q} + z_{\alpha/2}\sqrt{\frac{\widehat{q}(1-\widehat{q})}{n}}\right]\right\} = 1-\alpha$$

– soit on prend comme valeur de q celle qui maximise la variance (c-à-d. on calcule l'intervalle de confiance le plus grand possible). Dans ce cas on a $\sigma^2 = \frac{q(1-q)}{n}$ d'où on coclut qu'il faut maximiser la quantité $q(1-q)$. On a donc $\frac{d}{dq}[q(1-q)] = 0 \Rightarrow q = \frac{1}{2}$ et par conséquent l'intervalle de confiance devient

$$P_Q\left\{q \in \left[\widehat{q} - z_{\alpha/2}\frac{0.5}{\sqrt{n}},\ \widehat{q} + z_{\alpha/2}\frac{0.5}{\sqrt{n}}\right]\right\} = 1-\alpha$$

Il y a une troisième méthode qui utilise des abaques et que nous n'examinerons pas ici.

6.6
Exercices

EXERCICE 6.1 *On mesure la force de compression d'un ciment en moulant de petits cylindres et en mesurant la pression X (exprimée en kg/cm^2) à partir de laquelle ils se cassent.*

On suppose que $X \sim \mathcal{N}(\mu, \sigma^2)$.

(1) On se propose à faire des estimations ponctuelles.

(a) Pouvez-vous proposer des estimateurs pour les paramètres inconnus ?

(b) Pouvez-vous calculer d'autres estimateurs pour les paramètres inconnus ?

(c) Étudier leurs propriétés.

(2) On se propose à faire des estimations par des I.D.C.

(a) Calculer un I.D.C. pour les paramètres inconnus au niveau $\alpha = 0,1$.

(b) Évaluer un intervalle de confiance dissymétrique pour ces paramètres avec $\alpha_1 = 0,02$*, et* $\alpha_2 = 0,03$*, (*α_1 *à gauche,* α_2 *à droite).*

(3) On suppose que $\sigma^2 = 0,69$.

(a) Évaluer un intervalle de confiance pour la moyenne de X *au niveau* $\alpha = 0,1$.

(b) Comparer avec le résultat précédent. Commenter.

(4) Les cylindres doivent passer dans des anneaux fabriqués par une autre entreprise. On suppose que la v.a. Y *mesure le diamètre des cylindres et suit une loi normale, la v.a.* Z *mesure le diamètre des anneaux et suit une loi normale. On suppose que les écart-types de* Y *et* Z *sont égaux. Dans quel intervalle doit se trouver la différence de deux moyennes pour que les articles puissent s'emboîter avec probabilité* 95% *?*

Application numérique.-

- *Valeurs observées dans l'échantillon de la variable aléatoire* X *:*
  $19,6; 19,9; 20,4; 19,8; 20,5; 21; 18,5; 19,7; 18,4; 19,4$
- *Valeurs observées dans l'échantillon de la variable aléatoire* Y *:*
  $5,3; 4,8; 4,9; 4,9; 5,2; 5,1; 5,2; 4,7$
- *Valeurs de la variable aléatoire* Z *:*
  $5,1; 4,9; 4,7; 4,9; 5,3; 5,0; 5,1; 4,8$

EXERCICE 6.2 *On veut déterminer le poids d'un objet à l'aide d'une balance à deux plateaux. Le poids marqué à l'équilibre est un v.a.* X *qui, compte tenu de l'imprécision, suit une loi* $\mathcal{N}(\mu, \sigma^2)$*, les paramètres* μ *et* σ *étant inconnus. On considère que* σ *caractérise la précision de la balance. On a réalisé* 25 *pesées du même objet et on a calculé*

$$\sum_{i=1}^{25} (x_i - \bar{x})^2 = 280$$

.

(1) Comparer les intervalles de confiance pour σ^2 *avec* $\alpha = 0,05$ *:*

(a) bilatéral symétrique ;

(b) bilatéral dissymétrique , $\alpha_1 = 0,03$ *et* $\alpha_2 = 0,02$.

(c) unilatéral à droite, de la forme $]a, +\infty[$.

(2) Quelle aurait dû être la taille de l'échantillon, dans le cas d'un intervalle bilatéral symétrique, pour que l'intervalle soit de longueur inférieur à 15 *?*

EXERCICE 6.3 *On mesure quatre fois une distance. La moyenne de ces quatre mesures est 1,215m. Supposons que les mesures suivent une loi normale avec écart-type 0,01 m.*

(1) Quel est, l'intervalle de confiance pour cette distance avec un risque de $\alpha = 5\%$ *?*

(2) Si on veut, avec la même confiance, avoir un ***I.D.C.*** *de 0,015 m, calculer le nombre de mesures qu'il faut effectuer.*

EXERCICE 6.4 *Considérons une v.a.* $X \sim \mathcal{B}(p)$*. On cherche à calculer de façon approchée un intervalle de confiance pour* p *au niveau* α*, sur la base de* n *observations* $X_1, \ldots, X_n$*.*
Application numérique.- *On jette 40 fois une pièce et on obtient 24 fois face. Calculer, de façon approchée, l'intervalle de confiance pour la fréquence des faces, avec confiance 95%.*

EXERCICE 6.5 *L'écart-type des durées de vie d'un échantillon de 200 ampoules électriques est de 100 heures. Déterminer l'intervalle de confiance de l'écart-type, au niveau 0,05,.*

EXERCICE 6.6 *Considérons une v.a.* $X \sim \mathcal{B}(p)$*. On cherche à calculer de façon approchée un intervalle de confiance pour* p *à niveau* α*, sur la base de* n *observations* $X_1, \ldots, X_n$*.*
Application numérique.- *On jette 40 fois une pièce et on obtient 24 fois face. Calculer, de façon approchée, l'intervalle de confiance pour la fréquence des faces, avec confiance 95%.*