Chapitre II

Statistique inférentielle

1 Introduction

Considérons les deux situations suivantes :

Situation 1

Une société s'approvisionne en pièces brutes qui, conformément aux conditions fixées par le fournisseur, doivent avoir une masse moyenne de 780 grammes. Au moment où 500 pièces sont réceptionnées, on en prélève au hasard un échantillon de 36 pièces, dont on mesure la masse. On obtient les résultats suivants :

Masse des pièces (en grammes)	Nombre de pièces
[745; 755[	2
[755; 765[	6
[765; 775[	10
[775; 785[	11
[785; 795[	5
[795; 805[	2

À combien peut-on estimer la moyenne et l'écart-type des masses pour la population constituée des 500 pièces à l'aide des résultats obtenus sur cet échantillon ?

Situation 2

Dans un hôpital important, on prélève au hasard un échantillon de 100 personnes parmi la population des malades et on mesure la pression artérielle diostalique (P.A.D.) de chacune de ces 100 personnes. On obtient les résultats suivants :

P.A.D. (en mm de Hg)	Effectif
[4; 6[	4
[6; 8[	20
[8; 10[	41
[10; 12[	23
[12 ; 14[	12

À combien peut-on estimer la proportion de personnes dont la P.A.D. est strictement inférieure à 8 parmi la population constituée de l'ensemble des malades de l'hôpital ?

Nature du problème

Dans les deux cas, nous cherchons des informations sur une population d'effectif relativement important à partir de l'étude d'un échantillon de quelques dizaines d'unités: dans la situation 2, il s'agit d'une proportion et, dans la situation 1, d'une moyenne et d'un écart-type.
Ce type de situation se rencontre fréquemment dans le monde industriel car, le plus souvent, il n'est pas possible d'étudier la population entière: cela prendrait trop de temps, reviendrait trop cher ou serait aberrant comme, par exemple, dans le cas d'un contrôle de qualité entraînant la destruction des pièces (durée de vie d'une ampoule).
Nous allons apporter à ce problème très important deux types de réponses. Nous proposerons tout d'abord un nombre comme moyenne, proportion ou écart-type de la population : c'est l'*estimation ponctuelle*, séduisante par sa simplicité mais ne donnant pas toujours un résultat utilisable de façon satisfaisante. Aussi, dans une seconde partie, serons nous amenés à introduire la notion d'*intervalle de confiance* associé à un coefficient de confiance.

2 Statistiques

2.1 Définitions

Dans la suite, on considère le cas d'un échantillonnage aléatoire simple, c'est-à-dire que l'on extrait de la population un échantillon de taille n par des tirages aléatoires, équiprobables et indépendants (tirages avec remise ou tirage sans remise dans une population de grande taille).

Soit X la V.A. qui représente le caractère quantitatif que l'on souhaite étudier sur l'ensemble de la population. On note $\mathrm{I\!E}(X) = \mu$ et $\mathrm{Var}(X) = \sigma^2$.

Soit X_k la V.A. qui représente le résultat aléatoire du k-ième tirage. X_k suit la même loi que X. On note x_k sa réalisation.

Définition 12 Le n–uplet $(X_1, \ldots, X_n)$ de V.A. indépendantes et de même loi (celle de X) est appelé n-échantillon ou échantillon de taille n de X.

La réalisation $(x_1, \ldots, x_n)$ de l'échantillon $(X_1, \ldots, X_n)$ est l'ensemble des *valeurs observées*.

On appelle *statistique* sur un échantillon $(X_1, \ldots, X_n)$ une V.A. fonction des X_k :
$Y = f(X_1, \ldots, X_n)$.

Après réalisation, la V.A. Y (statistique) prend la valeur $f(x_1, \ldots, x_n)$.

2.2 Statistiques classiques

2.2.1 Moyenne empirique

Définition 13 On appelle *moyenne empirique* de l'échantillon $(X_1, \ldots, X_n)$ de X la statistique :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sa réalisation $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (qui est la moyenne de l'échantillon) est appelée *moyenne observée*.

Exemple 19 Pour l'échantillon étudié dans la situation 1, la moyenne observée des masses (en grammes) des 36 pièces est (en supposant que les observations sont au centre de chaque classe):

$$\bar{x} = \frac{2\times 750 + 6\times 760 + 10\times 770 + 11\times 780 + 5\times 790 + 2\times 800}{36} \simeq 774,72$$

En l'absence d'informations supplémentaires, on décide de prendre cette valeur comme estimation de la moyenne inconnue μ des masses pour la population constituée des 500 pièces réceptionnées.

Le théorème de la limite centrale permet d'établir le résultat suivant :

Proposition 1 Si la taille *n* de l'échantillon est grande (en pratique *n* > 30), $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Remarque 12 Si $n \leq 30$, mais si $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, on a encore $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

2.2.2 Variance empirique

Définition 14 On appelle *variance empirique* de l'échantillon (X_1, …, X_n) de X la statistique :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Sa réalisation $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ (qui est la variance de l'échantillon) est appelée *variance observée*.

Exemple 20 Pour l'échantillon étudié dans la situation 1, la variance observée des masses des 36 pièces est :

$$\sigma_{36}^2 = \frac{2\times\left(750-\bar{x}\right)^2 + 6\times\left(760-\bar{x}\right)^2 + 10\times\left(770-\bar{x}\right)^2 + 11\times\left(780-\bar{x}\right)^2 + 5\times\left(790-\bar{x}\right)^2 + 2\times\left(800-\bar{x}\right)^2}{36}$$

$$\simeq 157,06$$

Par analogie avec la moyenne, nous sommes tentés de choisir la variance σ_{36}^2 d'un échantillon prélevé au hasard comme estimation ponctuelle de la variance inconnue σ^2 d'une population. Mais en procédant ainsi, nous risquons de sous-estimer la variance de la population, et cela d'autant plus nettement que l'effectif de l'échantillon est petit. Aussi est-on conduit à corriger cette première estimation peu satisfaisante en utilisant le nombre $\frac{36}{35}\sigma_{36}^2$.

D'une manière générale, nous verrons que l'on peut choisir comme estimation ponctuelle de la variance inconnue σ^2 d'une population le nombre :

$$s_{n-1}^2 = \frac{n}{n-1}\sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 .$$

2.2.3 Proportion

Soit une population comportant une modalité *M*. Soit *p* la proportion d'individus de la population possédant la modalité *M*.

On extrait de la population un échantillon de taille *n*.

Soit *R* la V.A. qui représente le nombre d'individus dans l'échantillon possédant la modalité *M*.

Définition 15 On appelle *fréquence empirique* la statistique :

$$F = \frac{R}{n}$$

Sa réalisation $f = \frac{\text{nombre d'individus possédant la modalité } M}{n}$ (qui est la proportion d'individus de l'échantillon possédant la modalité *M*) est appelée *fréquence observée*.

Exemple 21 Pour l'échantillon étudié dans la situation 2, la fréquence observée des personnes dont la P.A.D. est strictement inférieure à 8 est :

$$f = \frac{24}{100} = 0,24$$

En l'absence d'informations supplémentaires, on décide de prendre cette valeur comme estimation, pour la population constituée de l'ensemble des malades de l'hôpital, de la proportion inconnue *p* de personnes dont la P.A.D. est strictement inférieure à 8.

Remarque 13 $R \sim \mathcal{B}(n; p)$

Proposition 2 Si la taille *n* de l'échantillon est grande (en pratique $n > 30$), et si p et $1 - p$ ne sont pas trop petits $\left(p \in [0,1; 0,9]\right)$, alors $F \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$.

Remarque 14 On prend également comme conditions sur *n* et *p* : $n\,p > 5$ et $n\,(1 - p) > 5$.

3 Estimateurs

3.1 Généralités

Exemple 22 Une usine fabrique quelques millions de vis et on veut mesurer le diamètre moyen, appelons-le *d*. Si nous avions l'ensemble des valeurs, il nous suffirait de calculer la moyenne empirique pour trouver la valeur cherchée. Or nous supposerons ne posséder qu'une partie de ces valeurs. Nous allons donc devoir estimer le paramètre *d*.
La première estimation à laquelle nous pensons est la moyenne arithmétique, appelons-la m_1.
Nous pouvons donner une autre estimation de cette moyenne: la moyenne géométrique, appelons-la m_2 $\left(m_2 = \sqrt[n]{\prod_{i=1}^{n} x_i}\right)$. Si nous connaissions la valeur de *d*, il serait facile de voir laquelle des deux valeurs m_1 et m_2 estime le mieux *d*.

Seulement, nous ne connaissons pas la valeur de cette moyenne *d* (sinon nous n'aurions pas à l'estimer !).
Il nous faut donc définir au moins un estimateur et, si l'on en possède plusieurs, déterminer certaines de leurs propriétés afin de les comparer et savoir lequel est le meilleur.

Soient *X* une V.A.R. dont la loi dépend d'un paramètre inconnu θ, $(X_1, \dots, X_n)$ un *n*-échantillon de *X* et $(x_1, \dots, x_n)$ sa réalisation.

Il s'agit d'estimer le paramètre θ.

Définition 16 Un *estimateur* de θ est une statistique $T = f(X_1, \dots, X_n)$, et sa réalisation est notée $t = f(x_1, \dots, x_n)$.

Remarque 15 Un paramètre admet une infinité d'estimateurs. Certains sont évidemment "farfelus", d'autres semblent cohérents. Par exemple, le paramètre λ d'une loi de Poisson admet la moyenne empirique et la variance empirique comme estimateurs possibles.

3.2 Biais et convergence

Définition 17 Si *T* est un estimateur du paramètre θ, la V.A. $T - \theta$ est appelée *erreur d'estimation*.

En écrivant $T - \theta = T - \mathrm{IE}(T) + \mathrm{IE}(T) - \theta$, on fait apparaître le terme $T - \mathrm{IE}(T)$ qui traduit la fluctuation de *T* autour de son espérance, et le terme $\mathrm{IE}(T) - \theta = \mathrm{B}(T)$ appelé *biais de l'estimateur.*

Définition 18 Un estimateur *T* de θ est dit *sans biais* si :

$$\mathrm{IE}(T) = \theta \ (\text{ ou } \mathrm{B}(T) = 0)$$

sinon, on dit qu'il est *biaisé*.

Exemple 23 La moyenne empirique $\overline{X}$ est un estimateur sans biais du paramètre λ d'une loi de Poisson. La variance empirique S_n^2 est un estimateur biaisé du même paramètre.

Définition 19 Un estimateur *T* de θ est dit *asymptotiquement sans biais* si :

$$\mathrm{IE}(T) \underset{n \to +\infty}{\to} \theta,$$

Remarque 16 Un estimateur sans biais est également asymptotiquement sans biais.

Définition 20 Si *T* est un estimateur de θ asymptotiquement sans biais et si :

$$\mathrm{Var}(T) \underset{n \to +\infty}{\to} 0,$$

alors *T* est dit *convergent*.

Définition 21 Soient *T* et *T'* deux estimateurs sans biais de θ. *T* est dit *plus efficace* que *T'* si :

$$\mathrm{Var}(T) \leq \mathrm{Var}(T').$$

Un estimateur sans biais de variance minimale est appelé *estimateur efficace.*

3.3 Estimation ponctuelle de paramètres usuels.

3.3.1 Estimation ponctuelle de l'espérance

Proposition 3 Soit *X* une V.A. dont on veut estimer l'espérance $\mu = \mathrm{IE}(X)$ à partir d'un *n*-échantillon (X_1,…, X_n). La moyenne empirique :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

est un estimateur efficace de μ.

3.3.2 Estimation ponctuelle de la variance

Proposition 4 Soit *X* une V.A. suivant une loi normale $\mathcal{N}(\mu, \sigma^2)$ dont on veut estimer la variance σ^2 à partir d'un *n*-échantillon (X_1,…, X_n). La statistique :

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

est un estimateur sans biais et convergent de σ^2.

3.3.3 Estimation ponctuelle d'une proportion

Proposition 5 Soit *p* la proportion d'individus d'une population possédant une modalité *M*.
On extrait de la population un échantillon de taille *n*.
Soit *R* la V.A. qui représente le nombre d'individus dans l'échantillon possédant la modalité *M*.
La fréquence empirique :

$$F = \frac{R}{n}$$

est un estimateur efficace de *p*.

3.3.4 Estimation du paramètre λ d'un processus de Poisson homogène

Comme nous avons pu le voir dans le cours de probabilités-fiabilité, que ce soit pour le processus de Poisson homogène ou bien son processus de comptage associé (N_t) la connaissance du paramètre λ est indispensable.
Nous allons donner deux méthodes pour l'estimer, utilisant deux plans d'essais :

- Plans d'essais de type I : on observe le système sur une période *t* et on note le nombre de défaillances N_t. L'estimateur $\hat{\lambda} = \frac{N_t}{t}$ est un estimateur sans biais et convergent de λ.

- Plans d'essais de type II : on observe le système pendant un nombre *n* de défaillances et on note le temps des *n* défaillances. Soit t_n le temps de la dernière défaillance, l'estimateur $\hat{\lambda} = \frac{n}{t_n}$ est un estimateur sans biais et convergent de λ.

4 Estimation de la fiabilité

Soit T une V.A.R. qui modélise la durée de bon fonctionnement d'un élément.
Soient $t_1, t_2, ..., t_n$ n observations de la VAR T.

Exemple 24 Pour estimer la fonction de répartition de T en t on divise le nombre de fois où t_i est inférieur à t par n :

$$\widehat{IP(T<t)} = \frac{1}{n}\sum_{i=1}^{n} 1_{]-\infty;t[}(t_i)$$

Pour l'exemple précédent, nous avons supposé disposer de n observations de la V.A. T.
Dans la pratique, ce ne sera pas toujours le cas. Au départ, on observera toujours n éléments, mais nous ne pourrons pas toujours obtenir les n temps de défaillances.

Voici les différents types d'essais, utilisant n éléments.

- Essais complets : on observe jusqu'à la défaillance du dernier élément. Les données sont donc $t_1, t_2, ..., t_n$.
- Essais incomplets :

- tronqués : on arrête les essais au bout d'un temps t_0, fixé à l'avance;
- censurés : on arrête les essais à la $k^{ième}$ défaillance, fixée à l'avance;
- interrompus : on retire des éléments non défaillants en cours d'observation (on appellera ces éléments des éléments suspendus).

Pour la suite, on notera :

- n : nombre d'éléments mis en service à $t = 0$,
- $v(t)$: nombre d'éléments en vie à l'instant t,
- $d(t_i; t_{i+1})$: nombre d'éléments défaillants dans l'intervalle $]t_i; t_{i+1}]$

Remarque 17 On a de manière évidente $n = v(t) + d(0, t)$.

4.1 Essais complets

Nous allons voir trois méthodes qui permettent d'approcher le mieux possible la fiabilité, suivant la valeur de n.

4.1.1 Méthode des pourcentages simples

Cette méthode est utilisée lorsque la valeur de n est supérieur à 50. On a les estimations suivantes :

- $\widehat{F}(t) = \dfrac{d(0,t)}{n}$
- $\widehat{R}(t) = 1 - \dfrac{d(0,t)}{n} = \dfrac{v(t)}{n}$
- $\widehat{\lambda}(t;t+\Delta t) = \dfrac{1}{\Delta t}\dfrac{d(t;t+\Delta t)}{v(t)}$

En ordonnant les temps de défaillances (ce qui n'enlève rien à la généralité), on obtient les estimations suivantes :

- $\widehat{F}(t_i) = \dfrac{i}{n}$
- $\widehat{R}(t_i) = \dfrac{n-i}{n}$
- $\widehat{\lambda}(t_i;t_{i+1}) = \dfrac{1}{(t_{i+1}-t_i)(n-i)}$

4.1.2 Méthode des rangs moyens

Cette méthode est utilisée lorsque la valeur de *n* est comprise entre 20 et 50. En ordonnant les temps de défaillances, on a les estimations suivantes :

- $\widehat{F}(t_i) = \dfrac{i}{n+1}$
- $\widehat{R}(t_i) = \dfrac{n+1-i}{n+1}$
- $\widehat{\lambda}(t_i;t_{i+1}) = \dfrac{1}{(t_{i+1}-t_i)(n+1-i)}$

4.1.3 Méthode des rangs médians

Cette méthode est utilisée lorsque la valeur de *n* est inférieure à 20. En ordonnant les temps de défaillances, on a les estimations suivantes :

- $\widehat{F}(t_i) = \dfrac{i-0,3}{n+0,4}$
- $\widehat{R}(t_i) = \dfrac{n+0,7-i}{n+0,4}$
- $\widehat{\lambda}(t_i;t_{i+1}) = \dfrac{1}{(t_{i+1}-t_i)(n+0,7-i)}$

4.2 Essais incomplets

Dans le cas d'essais tronqués, on a les observations suivantes :

$$t_1, t_2, t_3, \ldots, t_j, \underbrace{t_0, t_0, \ldots, t_0}_{n-j \text{ fois}} \text{ avec } t_i < t_0 \quad \forall i \in \{1, ..., j\}$$

Dans le cas d'essais censurés, on a les observations suivantes :

$$t_1, t_2, t_3, \ldots, t_k, \underbrace{t_0, t_0, \ldots, t_0}_{n-k \text{ fois}}$$

Dans le cas d'essais interrompus, on a les observations suivantes :

$$t_1, t_2, t_3^*, \ldots, t_j^*, \ldots, t_n$$

où les temps étoilés représentent les temps de suspension.

4.2.1 Méthode de Kaplan-Meier

Soit $d(t_j)$ le nombre d'éléments défaillants à l'instant t_j, on a l'estimation suivante :

$$\widehat{R}(t_i) = \prod_{j \text{ tq } t_j \le t_i} \left[1 - \frac{d(t_j)}{v(t_j - \varepsilon)} \right]$$

où ε est strictement positif tel que : $\forall i,\ t_{i-1} < t_i - \varepsilon$.
$v(t_j - \varepsilon)$ représente le nombre d'éléments en vie juste avant le temps de défaillance t_j.

Proposition 6 S'il n'y a pas d'élément suspendu, cette méthode se ramène à la méthode des pourcentages simples.

4.2.2 Méthode de Johnson

Soit i le rang initial de la défaillance à l'instant t_i.
On va utiliser un « rang corrigé » de la défaillance à l'instant t_i, noté i', pour prendre en compte les temps suspendus.

On note $i'\_$ le rang corrigé de la défaillance à l'instant précédant t_i.

Pour le calcul des temps corrigés, on utilise l'algorithme suivant :

1. Pour $i = 1$, $i'\_ = 0$
2. Calcul d'un incrément : $D_i = \begin{cases} 0 & \text{si l'élément est suspendu} \\ \dfrac{n+1-i'\_}{n+2-i} & \text{sinon} \end{cases}$,
3. Calcul d'un rang corrigé : $i' = i'\_ + D_i$,

On estime ensuite les caractéristiques en utilisant une méthode des essais complets avec les couples (i', t_i).

4.3 Cas particulier de la loi exponentielle

Dans le cas d'une durée de vie de loi exponentielle, on a vu que $\lambda(t) = \lambda$ et que $MTTF = \frac{1}{\lambda}$ (où MTTF = Mean Time To Failure = durée de bon fonctionnement).

De plus, il est facile de voir que $R(t) = e^{-\lambda t}$, ce qui entraîne que $\ln\left(R(t)\right) = -\lambda t$.

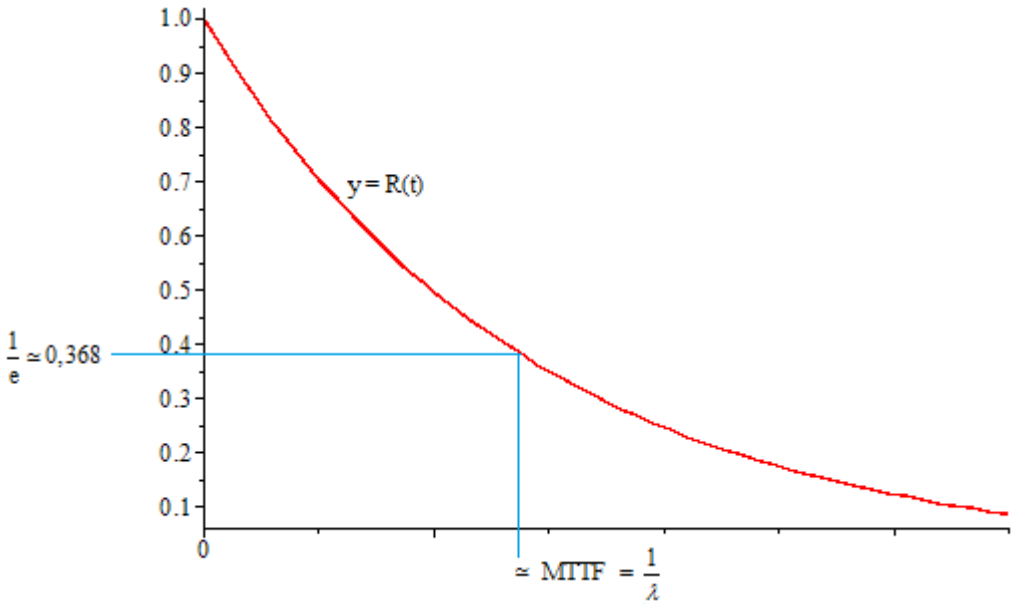
4.3.1 Méthode graphique

On trace :

- soit sur du papier semi-log la droite de régression, qui doit passer par le point (0 ; 1)
- soit la courbe d'équation y = R(t) obtenue par interpolation des points $M_i\left(t_i ; \widehat{\mathrm{R}}\left(t_i\right)\right)$

On sait que $\mathrm{R}\left(\frac{1}{\lambda}\right) \simeq 36{,}8\%$ car $\frac{1}{\mathrm{e}} \simeq 0{,}368$.

Il reste à lire sur l'axe du temps l'abscisse qui correspond à l'ordonnée 0,368. On obtient alors une estimation du MTTF.



4.3.2 Méthode numérique

On calcule la valeur de la pente de régression pour la série $\left(t_i ; \ln\left(\widehat{R}\left(t_i\right)\right)\right)$.

Cette valeur correspond à une estimation de $-\lambda$.

5 Estimation par intervalle de confiance

Dans la situation 1, en choisissant un nouvel échantillon de 36 pièces, on obtiendrait une nouvelle moyenne pour les masses de ces 36 pièces. De même, dans la situation 2, un nouvel échantillon de 100 personnes donnerait une nouvelle proportion de malades possédant la même propriété.
Ainsi, les estimations ponctuelles proposées ci-dessus de la moyenne d'une population et d'un pourcentage d'éléments de la population dépendent très directement de l'échantillon prélevé au hasard. Dans de nombreux cas, l'importance attribuée au hasard dans le choix des éléments d'un échantillon, et donc dans le résultat des estimations ponctuelles, est grande. Cela conduit à s'interroger avant d'utiliser ces estimations pour prendre des décisions dont les conséquences économiques, financières, sociales, …, peuvent être très grandes : refus éventuel d'une livraison, choix d'une stratégie commerciale, fixation d'un minimum de ressources pour l'obtention d'une aide, …
Aussi est-on amené à chercher un nouveau type d'estimation de la moyenne d'une population ou d'un pourcentage d'éléments d'une population sous forme d'intervalle, en utilisant le calcul des probabilités.
Il s'agit en fait de situer le paramètre inconnu θ par un intervalle qui a une forte probabilité (généralement 95% ou 99%) de le contenir.

5.1 Cas général

Définition 22 On appelle intervalle de confiance (noté I.C.) d'un paramètre inconnu θ au *niveau de sécurité* (ou de *confiance*) γ fixé, tout intervalle $[T_1, T_2]$ tel que :

$$\mathrm{IP}\left(T_1 < \theta < T_2\right) = \gamma$$

En général, on prend $\gamma = 0{,}95$ ou $0{,}99$.

Remarque 18 On construit les intervalles de confiance à partir de la loi de probabilité d'un bon estimateur $\hat{\theta}$ de θ. Compte tenu de l'infinité des intervalles possibles, on construira des intervalles de confiance symétriques, c'est à dire tels que : $\mathrm{IP}\left(T_1 < \theta\right) = \mathrm{IP}\left(\theta < T_2\right) = \dfrac{1-\gamma}{2}$

Pour la suite, nous nous contenterons de construire des intervalles de confiance pour la moyenne et pour la proportion. La technique se généralise à un paramètre inconnu quelconque (par exemple, la variance d'une loi gaussienne).

5.2 I.C. pour une moyenne

Soit X une V.A. suivant une loi gaussienne, de variance σ^2 et d'espérance μ que l'on veut estimer.
On rappelle que pour un n-échantillon $(X_1, X_2, \ldots, X_n)$ de la population, un estimateur efficace de la moyenne est :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

5.2.1 Cas où σ^2 est connue

Si σ^2 est connue, on sait que : $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ et donc que $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0\ ;\ 1)$.

Si U suit une loi $\mathcal{N}(0\ ;1)$, alors pour tout $\alpha \in]0;1[$, il existe un réel positif, noté $u_{1-\alpha/2}$, tel que :

$$\mathrm{IP}\left(-u_{1-\alpha/2} < U < u_{1-\alpha/2}\right) = \gamma = 1-\alpha \Leftrightarrow \mathrm{IP}\left(U < u_{1-\alpha/2}\right) = 1-\frac{\alpha}{2}$$

On a donc : $\mathrm{IP}\left(-u_{1-\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le u_{1-\alpha/2}\right) = \mathrm{IP}\left(\overline{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \gamma = 1-\alpha$

Théorème 3 L'I.C. au niveau de confiance $\gamma = 1-\alpha$ de la moyenne μ d'une population gaussienne $\mathcal{N}(\mu, \sigma^2)$ où σ^2 est connue est :

$$I_\alpha = \left[\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} ; \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (moyenne observée)

qui s'interprète par : « avec la probabilité $1-\alpha$, $\mu \in I_\alpha$ »
On trouve grâce à la table de la loi normale centrée réduite que :

- si $\alpha = 5\%$ (la confiance étant de 95%), alors $u_{1-\alpha/2} = 1{,}96$
- si $\alpha = 1\%$ (la confiance étant de 99%), alors $u_{1-\alpha/2} = 2{,}57$

Illustration numérique : $\sigma^2 = 4,\ n = 25,\ \bar{x} = 8$ donne :

$$I_{5\%} = \left[8 - 1{,}96\frac{2}{5} ; 8 + 1{,}96\frac{2}{5}\right] = [7{,}22;8{,}78]$$

$$I_{1\%} = \left[8 - 2{,}57\frac{2}{5} ; 8 + 2{,}57\frac{2}{5}\right] = [6{,}97;9{,}03]$$

Remarques 19

1. $\alpha_1 > \alpha_2 \Rightarrow I_{\alpha_1} \subset I_{\alpha_2}$: plus la confiance exigée est grande, plus l'amplitude de l'intervalle de confiance est grand.

2. Si l'on veut réduire l'amplitude de l'intervalle de confiance I_α (à un niveau de confiance fixé) dans un rapport *k*, il faut multiplier la taille de l'échantillon par k^2.

Illustration numérique : $\sigma^2 = 4,\ \alpha = 5\%,\ \bar{x} = 8$ donne :

$n = 25 \quad \rightarrow \quad I_{5\%} = [7{,}22;8{,}78]$

$n = 100 \quad \rightarrow \quad I_{5\%} = [7{,}61;8{,}39]$

Pratiquement, plus la confiance choisie est forte, plus l'échantillon devra être important.

5.2.2 Cas où σ^2 est inconnue

Si σ^2 est inconnue, on sait que $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ est un estimateur sans biais et convergent de σ^2.

On montre de plus que : $\frac{\overline{X} - \mu}{S_{n-1}/\sqrt{n}} \sim T_{n-1}$ où T_{n-1} suit une loi de Student à n–1 degrés de liberté.

Si T_{n-1} suit une loi de Student à n–1 degrés de liberté, alors pour tout $\alpha \in]0;1[$, il existe un réel positif, noté $t_{n-1,1-\alpha/2}$, tel que :

$$\mathrm{IP}\left(-t_{n-1,1-\alpha/2} < T_{n-1} < t_{n-1,1-\alpha/2}\right) = \gamma = 1-\alpha \Leftrightarrow \mathrm{IP}\left(T_{n-1} < t_{n-1,1-\alpha/2}\right) = 1 - \frac{\alpha}{2}$$

D'où : $\mathrm{IP}\left(-t_{n-1,1-\alpha/2} \le \frac{\overline{X} - \mu}{S_{n-1}/\sqrt{n}} \le t_{n-1,1-\alpha/2}\right) = \mathrm{IP}\left(\overline{X} - t_{n-1,1-\alpha/2}\frac{S_{n-1}}{\sqrt{n}} \le \mu \le \overline{X} + t_{n-1,1-\alpha/2}\frac{S_{n-1}}{\sqrt{n}}\right) = \gamma$

Théorème 4 L'I.C. au niveau de confiance $\gamma = 1-\alpha$ de la moyenne μ d'une population gaussienne $\mathcal{N}\left(\mu, \sigma^2\right)$ où σ^2 est inconnue est :

$$I_\alpha = \left[\bar{x} - t_{n-1,1-\alpha/2}\frac{s_{n-1}}{\sqrt{n}} ; \bar{x} + t_{n-1,1-\alpha/2}\frac{s_{n-1}}{\sqrt{n}}\right],$$

où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (moyenne observée) et $s_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$ (*déviation standard)*

5.2.3 Cas où *n* est « grand » (*n* > 30)

Si *n* est « grand », il n'est pas nécessaire que *X* soit gaussienne : le théorème de la limite centrale donne $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ pour $n \to \infty$

Lorsque σ^2 est inconnue (ce qui est presque toujours les cas !), on utilise l'approximation :

$$\frac{\overline{X} - \mu}{S_n/\sqrt{n}} \xrightarrow{L} \mathcal{N}(0\ ;\ 1)$$

où S_n^2 est la variance empirique. On a alors :

Théorème 5 Lorsque *n* est « grand », l'I.C. au niveau de confiance $\gamma = 1-\alpha$ de la moyenne μ est :

$$I_\alpha = \left[\bar{x} - u_{1-\alpha/2}\frac{\sigma_n}{\sqrt{n}} ; \bar{x} + u_{1-\alpha/2}\frac{\sigma_n}{\sqrt{n}}\right]$$

où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (moyenne observée), et $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ (variance observée)

5.3 I.C. pour une proportion

Soit *p* la proportion d'individus d'une population possédant une modalité *M* ($p \in [0;1]$).
On extrait de la population un échantillon de taille *n*.

Soit *R* la V.A. qui représente le nombre d'individus dans l'échantillon possédant la modalité *M*.
on rappelle que $F = \frac{R}{n}$ est un estimateur efficace de *p*.
On sait de plus que $R \sim B(n; p)$

Utilisons la convergence en loi de la loi binomiale vers la loi normale quand $n \to +\infty$

$$\frac{R - np}{\sqrt{np(1-p)}} \xrightarrow{L} \mathcal{N}(0\ ;\ 1)$$

Soit $u_{1-\alpha/2}$ la valeur déterminée dans la table de la loi normale centrée réduite suivie par la V.A. U telle que : $\mathrm{IP}\left(-u_{1-\alpha/2} < U < u_{1-\alpha/2}\right) = \gamma = 1-\alpha \Leftrightarrow \mathrm{IP}\left(U < u_{1-\alpha/2}\right) = 1 - \frac{\alpha}{2}$. Alors :

$$\mathrm{IP}\left(-u_{1-\alpha/2} < \frac{R - np}{\sqrt{np(1-p)}} < u_{1-\alpha/2} \right) = 1-\alpha$$

Lorsque *n* est « grand », si f est la fréquence observée on considère que $p(1-p) \simeq f\left(1-f\right)$ et on obtient :

Théorème 6 L'I.C. au niveau de confiance $\gamma = 1-\alpha$ de la proportion *p* est :

$$\mathrm{I}_\alpha = \left[f - u_{1-\alpha/2}\sqrt{\frac{f\left(1-f\right)}{n}} ; f + u_{1-\alpha/2}\sqrt{\frac{f\left(1-f\right)}{n}} \right]$$