EISTI

Ecole Internationale des Sciences du Traitement de l'Information

Statistique inférentielle

PREMIERE PARTIE : L'ESTIMATION

« Les faits sont têtus. Il est plus facile de s'arranger avec les statistiques »
Mark Twain, 1835-1910, écrivain et humoriste américain

Table des matières

1. Propriétés élémentaires d'un estimateur 2
   1.1. Estimateur sans biais 2
   1.2. Estimateur convergent 2
   1.3. Loi asymptotique d'un estimateur 3
2. Estimateurs usuels 3
   2.1. La fréquence empirique 4
   2.2. La moyenne 5
   2.3. La variance empirique 6
3. Estimation par intervalle de confiance 7
   3.1. Généralités 7
   3.2. Intervalle de confiance pour une proportion 8
   3.3. Intervalle de confiance pour une moyenne 9
   3.4. Intervalle de confiance pour une variance 10
   3.5. Taille d'échantillon pour une précision donnée 11

Rappels sur les convergences en probabilité 13

Un des objectifs des statistiques est de reconstituer d'après des expériences et/ou des observations le modèle probabiliste d'une situation aléatoire.

Une première étape consiste à trouver parmi les lois usuelles (binomiale, Poisson, exponentielle,...) celle qui correspond le mieux au phénomène étudié. Cette étape est relativement aisée puisque chaque loi est associée à une situation en particulier (Poisson pour un phénomène de comptage, exponentielle pour un temps d'attente,...).

Une fois la loi identifiée, il reste à donner une valeur aux paramètres de cette loi. En effet, si on étudie le nombre de pièces fabriquées par heure sur une machine, on sait que l'échantillon suit une loi de Poisson de paramètre θ. Comment déterminer θ à partir des observations effectuées sur cette machine ? L'objectif de l'estimation est de répondre à cette question.

Considérons les hypothèses de l'échantillonnage, c'est-à-dire que X_1, ..., X_n sont des variables aléatoires i.i.d.

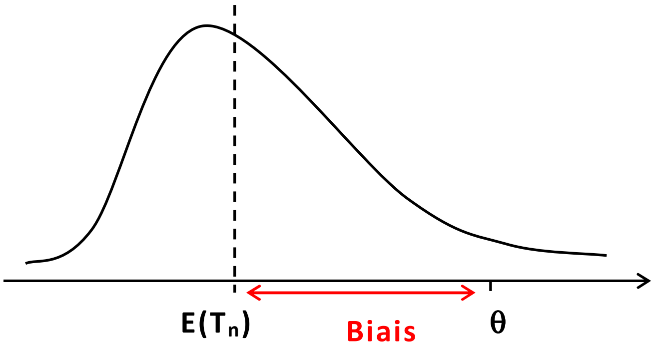
1. Proprietes elementaires d'un estimateur

Soit θ le paramètre à estimer et T_n un estimateur de θ, *i.e* une variable aléatoire fonction de $X_1,..., X_n$.
Avant de déterminer l'estimateur T_n, il est indispensable d'établir les propriétés essentielles auxquelles il doit répondre.

1.1. Estimateur sans biais

L'écart entre l'estimateur et le paramètre peut se décomposer en deux morceaux,

$$T_n-\theta=\underbrace{T_n-E(T_n)}_{\text{fluctuation}}+\underbrace{E(T_n)-\theta}_{\text{erreur}}$$



Le premier morceau correspond à la fluctuation de T_n autour de sa moyenne et le deuxième à une erreur systématique de l'estimateur. L'idéal est de rendre nul le dernier terme.
On dit qu'un estimateur T_n est *sans biais* si et seulement si

$$E(T_n)=\theta.$$

L'écart $b=E(T_n)-\theta$ est appelé le *biais* de l'estimateur.

Exemple

Dans le cas où l'échantillon suit une loi de Poisson, le paramètre θ est usuellement estimé par la moyenne. Cet estimateur est sans biais car

$$E(T_n)=E(\overline{X})=E(X_i)=\theta.$$

1.2. Estimateur convergent

Par ailleurs, il est souhaitable d'améliorer son estimation lorsque la taille de l'échantillon augmente,

$$T_n \underset{n\to+\infty}{\to} \theta.$$

Loi des grands nombres

Si l'estimateur le permet, on peut appliquer la loi des grands nombres pour obtenir une convergence en probabilité ou presque sûre.

Convergence en moyenne quadratique

L'erreur quadratique moyenne (ou risque quadratique) permet de mesurer la précision de l'estimateur.

$$R_\theta(T_n)=E\left[(T_n-\theta)^2\right]=var(T_n)+b^2.$$

Il faut la rendre la plus petite possible, c'est pourquoi on souhaite avoir un estimateur qui converge en moyenne quadratique vers θ, c'est-à-dire tel que

$$R_\theta(T_n) \underset{n\to+\infty}{\to} 0.$$

Avoir une erreur quadratique moyenne faible réduit la variabilité de l'estimateur et assure ainsi que chaque observation est proche du biais.

On remarque que dans le cas d'un estimateur sans biais, cela revient à faire tendre la variance vers 0.

Si T_1 et T_2 sont deux estimateurs de θ, on dit que T_1 est *meilleur* de T_2 si

$$R_\theta(T_1) \leq R_\theta(T_2),$$

c'est à dire s'il est plus précis que T_2. Dans le cas où les estimateurs sont sans biais, le meilleur est celui de plus petite variance.
Le cas idéal est un estimateur sans biais de plus petite variance. Un estimateur biaisé peut cependant être intéressant si son erreur quadratique moyenne est plus petite que la variance d'un estimateur sans biais.

Exemple
Reprenons le cas où l'échantillon suit une loi de Poisson de paramètre θ estimé par la moyenne. Cet estimateur converge en probabilité et presque sûrement car d'après la loi des grands nombres

$$\overline{X} = \frac{1}{n}\sum_1^n X_i \underset{n \to +\infty}{\to} E(X_i) = \theta.$$

Le risque quadratique de l'estimateur est égale à sa variance car il est sans biais,

$$R_\theta(\overline{X}) = \mathrm{var}(\overline{X}) = \frac{\mathrm{var}(X_i)}{n} = \frac{\theta}{n} \underset{n \to +\infty}{\to} 0.$$

Considérons maintenant un deuxième estimateur $T_n=X_1$. Alors T_n est sans biais et $\overline{X}$ est meilleur que T_n. En effet,

$$E(T_n) = E(X_1) = \theta \text{ et } R_\theta(T_n) = \mathrm{var}(X_1) = \theta \geq \frac{\theta}{n}.$$

1.3. Loi asymptotique d'un estimateur

Afin de calculer des probabilités sur l'estimateur, il est nécessaire de connaître sa loi.
Si l'estimateur le permet, on peut appliquer le théorème de la limite centrale pour obtenir une convergence en loi vers une loi normale.
Exemple
Reprenons le cas où l'échantillon suit une loi de Poisson de paramètre θ et considérons l'estimateur $T_n = \overline{X}$. Alors T_n est un estimateur sans biais de θ car $E(T_n)=E(X_i)=\theta$. D'après le théorème de la limite centrale nous avons

$$T_n \underset{n \to +\infty}{\overset{\mathcal{L}}{\to}} N(\theta, \frac{\theta}{n}).$$

D'où par exemple,

$$P(|T_n - \theta| \leq \varepsilon) = 2F(\varepsilon\sqrt{\frac{n}{\theta}}) - 1$$

2. ESTIMATEURS USUELS

L'objectif est d'étudier les propriétés des estimateurs usuels (moyenne, variance empirique, fréquence empirique). Dans un dernier paragraphe, nous verrons comment déterminer un estimateur lorsque les estimateurs usuels ne peuvent pas convenir.

2.1. La fréquence empirique

Soit un évènement A tel que P(A)=p. Sur un échantillon, considérons les variables aléatoires $X_1,...,X_n$ i.i.d telles que

$$X_i = \begin{cases} 1 & \text{si A est réalisé} \\ 0 & \text{sinon} \end{cases}.$$

Alors la fréquence empirique de l'évènement A est définie par

$$F_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{K_n}{n},$$

où K_n est la variable aléatoire qui comptabilise le nombre de fois où l'évènement A se réalise. On remarque que K_n suit une loi binomiale b(n,p).

- Biais
  La fréquence empirique est un estimateur sans biais de p. En effet ,
  $$E(F_n) = \frac{E(K_n)}{n} = \frac{np}{n} = p.$$
- Convergences
  La fréquence empirique est un estimateur convergent. En effet, d'après la loi des grands nombres, F_n converge en probabilité et presque sûrement vers $E(X_i)$=p. De plus, l'erreur quadratique est égale à la variance car elle est sans biais, d'où
  $$R_\theta(F_n) = var(F_n) = \frac{1}{n^2}var(K_n) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$
  Donc F_n converge aussi en moyenne quadratique vers p.
- Loi asymptotique

D'après le théorème de la limite centrale, on connaît la loi asymptotique de la

fréquence empirique car

$$F_n \underset{n\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} N(p, \frac{p(1-p)}{n})$$

Exemple

L'administration d'une école d'ingénieurs en informatique réalise une enquête afin de déterminer en autre le taux (*p*) de filles parmi ses élèves. Afin d'accélérer le traitement, le dépouillement est partagé entre deux personnes. La première relève 30 filles sur 110 étudiants, et la deuxième, 25 filles sur 100 étudiants. On obtient donc trois estimations de *p* suivant l'échantillon considéré. Pour l'échantillon 1, *p*=30/100=0.27, pour l'échantillon 2, *p*=25/100=0.25, et pour l'échantillon total, *p*=(30+25)/(110+100)=0.26. L'estimation dépend de l'échantillon testé, on peut donc se demander quelle est la meilleure estimation ? C'est à cette question que permet de répondre en partie l'estimation par intervalle de confiance du chapitre 4. Supposons donc que *p*=0.26. Quelle est la probabilité d'avoir plus de 50 filles l'année prochaine si la promotion est de 200 étudiants ? Soit f_n le taux de filles de la future promotion de 200 élèves. La taille de l'échantillon est suffisamment grande pour considérer que f_n suit une loi normale d'espérance 0.26 et d'écart-type 0.03, d'où

$$P(F_n > \frac{50}{200}) = P(F_n > 0.25) = P(Z > -0.33) = 1 - P(Z > 0.33) = 0.63,$$

où Z suit une loi N(0,1).

2.2. La moyenne

Soit $X_1,...,X_n$ une suite de variables aléatoires i.i.d telles que $E(X_i)=\mu$ et $var(X_i)=\sigma^2$. La moyenne est la variable aléatoire définie par

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

- Biais
  La moyenne est un estimateur sans biais de μ. En effet,
  $$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu .$$
- Convergences
  La moyenne est un estimateur convergent. En effet, d'après la loi des grands nombres, $\overline{X}$ converge en probabilité et presque sûrement vers $E(X_i)=\mu$. De plus, l'erreur quadratique est égale à la variance car elle est sans biais, d'où
  $$R_\theta(\overline{X}) = var(\overline{X}) = var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) = \frac{\sigma^2}{n} \text{ car } X_i \text{ indépendantes}$$
  Donc $\overline{X}$ converge aussi en moyenne quadratique vers μ.
- Loi asymptotique

D'après le théorème de la limite centrale, on connaît la loi asymptotique de la

moyenne car

$$\overline{X} \underset{n\to+\infty}{\overset{\mathcal{L}}{\rightarrow}} N(\mu, \frac{\sigma^2}{n})$$

- Echantillon gaussien
  Considérons le cas particulier où les X_i suivent une loi normale $N(\mu,\sigma^2)$. Alors on a le résultat exact que $\overline{X}$ suit une loi normale $N(\mu,\sigma^2/n)$. La loi est connue et n'a donc pas besoin d'être approchée à l'aide du théorème de la limite centrale. Ainsi l'échantillon peut être de petite taille.

Exemple

On prélève 40 pièces dans une production industrielle. Le cahier des charges de la machine stipule que les pièces sont produites avec un diamètre d'espérance 20 mm et un écart-type de 3 mm. Toutes les pièces ayant un diamètre moyen supérieur à 21 ou inférieur à 19 sont inutilisables. Quel est le pourcentage de pièces prélevées utilisables.

Soit $\overline{X}$ le diamètre moyen des 40 pièces. On sait que $E(X_i)=20$ et $var(X_i)=9$ où X_i est le diamètre d'une pièce. L'échantillon est suffisamment grand pour approcher la loi de $\overline{X}$ par une loi normale d'espérance 20 et de variance 0.225. Donc

$$P(19 < \overline{X} < 21) = P(-2.108 < Z < 2.108) = 2F(2.108) - 1 = 0.97$$

où Z suit une loi N(0,1) et F sa fonction de répartition.

2.3. La variance empirique

Soit $X_1,\ldots,X_n$ une suite de variables aléatoires i.i.d telles que $E(X_i)=\mu$ et $var(X_i)=\sigma^2$. La variance empirique est la variable aléatoire définie par

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

- <u>Biais</u>
  La variance empirique S^2 est un estimateur biaisé de σ^2. En effet,

$$E(S^2)=\frac{1}{n}\sum_{i=1}^{n}E(X_i^2)-E(\overline{X}^2)=\frac{1}{n}\sum_{i=1}^{n}\left(var(X_i)+E(X_i)^2\right)-\left(var(\overline{X})+E(\overline{X})^2\right)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2+\mu^2\right)-\left(\frac{\sigma^2}{n}+\mu^2\right)=\frac{n-1}{n}\sigma^2.$$

  C'est pourquoi, on utilise l'estimateur sans biais de σ^2,

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

- <u>Convergences</u>

La variance empirique est un estimateur convergent. En effet, d'après la loi des grands nombres,

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2\underset{n\to+\infty}{\to}E(X_i^2)=\sigma^2+\mu^2 \text{ et } \overline{X}^2\underset{n\to+\infty}{\to}\mu^2,$$

  donc S^2 converge en probabilité et presque sûrement vers σ^2. En revanche, le calcul de l'erreur quadratique n'est pas possible ici car le moment d'ordre 4 est inconnu.

- <u>Loi asymptotique</u>

Le théorème de la limite centrale ne peut s'appliquer que si on connaît le moment d'ordre 4. Dans le cas contraire, la loi asymptotique de la variance empirique n'est pas connue.

- <u>Echantillon gaussien</u>
  Considérons le cas particulier où les X_i suivent une loi normale $N(\mu,\sigma^2)$. Alors deux nouveaux résultats peuvent être établis.
  - les variables aléatoires

$$\frac{n}{\sigma^2}S^2 \quad\text{et}\quad \frac{n-1}{\sigma^2}S^{*2}$$

  suivent une loi du chi-deux à n-1 degrés de liberté, χ^2_{n-1}. La loi est connue et n'a donc pas besoin d'être approchée à l'aide du théorème de la limite centrale. Ainsi l'échantillon peut être de petite taille.
  - l'estimateur S^{*2} converge en moyenne quadratique vers σ^2 car

$$var\left(\frac{n-1}{\sigma^2}S^{*2}\right)=n-1,$$

  d'après la loi du χ^2. Donc

$$var(S^{*2})=\frac{2\sigma^4}{n-1}$$

  d'où

$$R_\theta(S^{*2}) = var(S^{*2}) \underset{n\to+\infty}{\longrightarrow} 0$$

<u>Exemple</u>
On prélève 40 pièces dans une production industrielle. Le cahier des charges de la machine stipule que le diamètre des pièces suit une loi normale d'espérance 20mm et d'écart-type 2 mm. On considère que la production n'est plus homogène lorsque l'écart-type empirique dépasse de 5% sa valeur théorique. Quelle est la probabilité que la production ne soit plus homogène ?

$$P(S-2>0.05) = P(S^2>4.2) = P(\frac{40}{4}S^2>42) = 0.01.$$

3. ESTIMATION PAR INTERVALLE DE CONFIANCE

Aussi précise soit une estimation, elle ne peut être exacte. C'est toujours une valeur approchée du paramètre et on peut souhaiter disposer d'un outil mathématique pour mesurer la qualité de l'estimation. L'idée est, non pas de fournir une seule valeur pour le paramètre, mais une « fourchette » de valeurs probables. La détermination de cet intervalle s'effectue par la donnée du niveau de confiance, *i.e* la probabilité que la vraie valeur du paramètre appartient à l'intervalle. Ainsi, l'objectif est de déterminer un intervalle [a,b] tel que

$$P(a\le\theta\le b) = 1-\alpha,$$

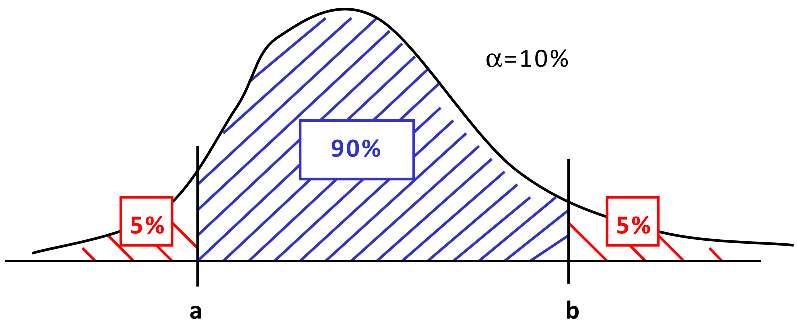
où α détermine le niveau de confiance de l'intervalle et correspond au risque d'erreur. Les risques d'erreur utilisés sont le plus souvent 10%, 5% et 1%.

3.1. Généralités

Les bornes a et b de l'intervalle de confiance dépendent du partage du risque $\alpha=\alpha_1+\alpha_2$.
On dit que l'intervalle est *unilatéral* si un des α_i est nul.

- si $\alpha_2=0$, l'intervalle est de la forme $[a,+\infty[$. Il est utilisé dans le cas où $\theta\ge a$ (durée de vie, résistance à la rupture, etc...)
- si $\alpha_1=0$, l'intervalle est de la forme $]-\infty,b]$. Il est utilisé dans le cas où $\theta\le a$ (nombre de pièces défectueuses, temps d'attente, etc...)

Dans le cas contraire, on dit que l'intervalle est *bilatéral*. En général, l'intervalle est symétrique, *i.e* $\alpha_1=\alpha_2$.



3.2. Intervalle de confiance pour une proportion

Soit $X_1,\dots,X_n$ une suite de variables aléatoires i.i.d. et p la probabilité d'un évènement A. On cherche l'intervalle [a,b] tel que

$$P(a \le p \le b) = 1-\alpha .$$

On suppose que n est suffisamment grand pour approcher la loi de l'estimateur F_n par la loi normale (cf. 3.1.). Alors

$P(a \le p \le b) = P[\sqrt{\frac{n}{p(1-p)}}(f_n - b) \le \sqrt{\frac{n}{p(1-p)}}(F_n - p) \le \sqrt{\frac{n}{p(1-p)}}(f_n - a)] = P(b' \le Z \le a')$, où Z suit la loi N(0,1). Si on suppose un risque symétrique et, étant donnée que la loi N(0,1) est symétrique par rapport à Oy, nous pouvons poser a'=-b'=k. D'où

$$P(a \le p \le b) = 1-\alpha \Leftrightarrow P(-k \le Z \le k) = 1-\alpha$$

$$\Leftrightarrow 2F(k) - 1 = 1-\alpha \Leftrightarrow F(k) = 1-\alpha/2,$$

où F est la fonction de répartition de Z. La valeur de k s'obtient par lecture de la table inverse de la loi N(0,1). D'où

$$a = f_n - k\sqrt{\frac{p(1-p)}{n}} \text{ et } b = f_n + k\sqrt{\frac{p(1-p)}{n}} .$$

Ces bornes dépendent de p (inconnu), cependant, on peut montrer que lorsque que n est grand,

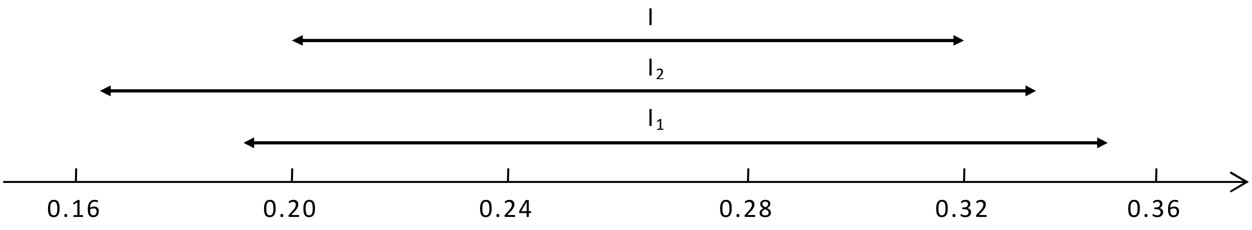
$$k\sqrt{\frac{p(1-p)}{n}} \text{ est équivalent à } k\sqrt{\frac{f_n(1-f_n)}{n}} .$$

Finalement, l'intervalle de confiance symétrique d'une proportion pour un risque α est

$$\left[f_n - k\sqrt{\frac{f_n(1-f_n)}{n}} ; f_n + k\sqrt{\frac{f_n(1-f_n)}{n}} \right].$$

Exemple

Reprenons l'exemple de l'école d'ingénieurs. Pour chacun des échantillons, on peut déterminer un intervalle de confiance avec α=5%. Par lecture de la table de la loi N(0,1), on a k=1.96. On obtient I_1=[0.19 ;0.35] pour l'échantillon 1, I_2=[0.16 ;0.33] pour l'échantillon2 et I=[0.2 ;0.32] pour l'échantillon total.



On remarque que l'intervalle de confiance pour un même risque est d'autant moins large, donc d'autant plus précis, que la taille de l'échantillon est grande.

Remarque

Dans le cas où le risque n'est pas symétrique, la méthodologie est la même, seule l'hypothèse a'=-b' n'est plus valide. Il faut donc faire du cas par cas.

Remarque
Dans le cas où l'échantillon est de petite taille, il est possible d'utiliser la loi exacte de K_n ($=nF_n$), et ainsi les tables de la loi $b(n,p)$.

3.3. Intervalle de confiance pour une moyenne

Soit $X_1,\ldots,X_n$ une suite de variables aléatoires i.i.d telles que $E(X_i)=\mu$ et $var(X_i)=\sigma^2$. On cherche l'intervalle $[a,b]$ tel que

$$P(a\le\mu\le b)=1-\alpha.$$

1er cas
Supposons que n est grand et que l'échantillon est de loi inconnue mais avec σ^2 connue. Alors nous pouvons approcher la loi de $\overline{X}$ par une loi normale $N(\mu,\sigma^2/n)$, d'où

$$P(a\le\mu\le b)=P[\sqrt{n}\frac{\overline{x}-b}{\sigma}\le\sqrt{n}\frac{\overline{X}-\mu}{\sigma}\le\sqrt{n}\frac{\overline{x}-a}{\sigma}]=P(b'\le Z\le a'),$$

où Z suit la loi $N(0,1)$. Si on suppose un risque symétrique et, étant donnée que la loi $N(0,1)$ est symétrique par rapport à Oy, nous pouvons poser $a'=-b'=k$. D'où

$$P(a\le\mu\le b)=1-\alpha \Leftrightarrow P(-k\le Z\le k)=1-\alpha$$
$$\Leftrightarrow 2F(k)-1=1-\alpha \Leftrightarrow F(k)=1-\alpha/2,$$

où F est la fonction de répartition de Z. La valeur de k s'obtient par lecture de la table inverse de la loi $N(0,1)$. D'où

$$a=\overline{x}-k\frac{\sigma}{\sqrt{n}} \text{ et } b=\overline{x}+k\frac{\sigma}{\sqrt{n}}.$$

Finalement, l'intervalle de confiance symétrique d'une moyenne pour un risque α est

$$\left[\overline{x}-k\frac{\sigma}{\sqrt{n}};\overline{x}+k\frac{\sigma}{\sqrt{n}}\right].$$

Exemple
On s'intéresse à la longueur en millimètre des faces de lunettes d'un certain modèle. L'écart-type de cette longueur est spécifié dans le cahier des charges à 0.48 mm. On prélève un échantillon de 64 faces de lunettes et on mesure la moyenne des longueurs $\overline{x}=130.10$ mm. Alors l'intervalle de confiance à 95% de la longueur moyenne des faces est [129.98 ;130.22].

2ème cas
Supposons que l'échantillon soit de loi normale avec σ^2 connue. Alors nous pouvons procéder de la même façon que précédemment car nous savons que $\overline{X}$ suit une loi normale $N(\mu,\sigma^2/n)$. La différence vient du fait que la loi est exacte et donc que l'échantillon peut être de petite taille.

3ème cas
Supposons que l'échantillon soit de loi normale avec σ^2 inconnue. Il faut alors remplacer σ^2 par son estimation S^{*2} et alors nous savons que

$$\sqrt{n}\frac{\overline{X}-\mu}{S^*}$$

suit une loi de Student à n-1 degré de liberté. La procédure est ensuite la même que précédemment puisque la loi de Student est elle aussi symétrique par rapport à Oy.

Exemple

Dans la fabrication de comprimés effervescents, il est prévu que le dosage de bicarbonate de sodium suit une loi normale. On a prélevé un échantillon de 25 comprimés et mesuré une moyenne $\overline{x}=1625$ mg et un écart-type $s^*=10$ mg. Déterminer un intervalle de confiance à 95% pour le dosage moyen des comprimés.

L'échantillon est gaussien avec σ^2 inconnu, donc

$$Z=\sqrt{n}\frac{\overline{X}-\mu}{S^*}$$

suit une loi de Student à 24 degrés de liberté. Donc si on suppose un risque symétrique, la valeur de k s'obtient par lecture de la table inverse de la loi de Student tel que $P(Z\geq k)=0.025$. D'où

$$a=\overline{x}-2.064\frac{s^*}{\sqrt{n}} \text{ et } b=\overline{x}+2.064\frac{s^*}{\sqrt{n}}.$$

Finalement, l'intervalle de confiance symétrique à 95% est

$$[1620.9;1629.1].$$

Remarque

Dans le cas où le risque n'est pas symétrique, la méthodologie est la même, seule l'hypothèse a'=-b' n'est plus valide. Il faut donc faire du cas par cas.

3.4. Intervalle de confiance pour une variance

Soit $X_1,\ldots,X_n$ une suite de variables aléatoires i.i.d telles que $E(X_i)=\mu$ et $var(X_i)=\sigma^2$. On cherche l'intervalle [a,b] tel que

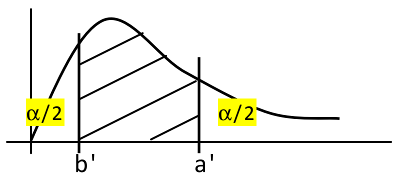
$$P(a\leq\sigma^2\leq b)=1-\alpha.$$

Etant donné que la loi asymptotique de la variance empirique n'est pas connue, l'intervalle de confiance ne peut être déterminé que dans le cas d'un échantillon gaussien. Supposons donc que les X_i suivent une loi $N(\mu,\sigma^2)$ avec μ inconnu, alors $Z=nS^2/\sigma^2$ ou $Z=(n-1)S^{*2}/\sigma^2$ suit une loi du chi-deux à n-1 degrés de liberté (cf.3.3.), d'où

$$P(a\leq\sigma^2\leq b)=P[\frac{ns^2}{b}\leq\frac{nS^2}{\sigma^2}\leq\frac{ns^2}{a}]=P(b'\leq Z\leq a'),$$

Si on suppose un risque symétrique alors

$$P(a\leq\sigma^2\leq b)=1-\alpha \Leftrightarrow \begin{cases}P(Z\geq a')=\alpha/2\\P(Z\geq b')=1-\alpha/2\end{cases}$$



Les valeurs de a' et b' s'obtiennent par lecture de la table de la loi du chi-deux. D'où

$$a=\frac{ns^2}{a'} \text{ et } b=\frac{ns^2}{b'}.$$

Finalement, l'intervalle de confiance symétrique d'une variance pour un risque α est

$$\left[\frac{ns^2}{a'};\frac{ns^2}{b'}\right].$$

On remplace n par n-1 si on utilise l'estimateur S^{*2} à la place de S^2.

Exemple

Reprenons le cas des comprimés effervescents (§ 4.3), alors l'intervalle de confiance à 95 % pour l'écart-type est défini par la loi de Student à 24 d.d.l.,

$$\begin{cases} P(Z\geq a')=0.025 \\ P(Z\geq b')=0.975 \end{cases} \Rightarrow \begin{cases} a'=39.36 \\ b'=12.40 \end{cases} \Rightarrow \begin{cases} a=\dfrac{24\times10^2}{39.36}=60.97 \\ b=\dfrac{24\times10^2}{12.40}=193.55 \end{cases}.$$

Finalement, l'intervalle de confiance symétrique à 95% de l'écart-type est

$$[7.81;13.91].$$

Tableau récapitulatif de la loi des estimateurs dans le cas gaussien

Estimation moyenne $\theta=\mu$	σ^2 connu	$\sqrt{n}\dfrac{\overline{X}-\theta}{\sigma}\sim N(0,1)$
	σ^2 inconnu	$\sqrt{n}\dfrac{\overline{X}-\theta}{S^*}\sim t_{n-1}$
Estimation variance $\theta=\sigma^2$	μ connu	$\dfrac{1}{\theta}\sum_{i=1}^{n}(X_i-\mu)^2\sim\chi^2_{n-1}$
	μ inconnu	$\dfrac{n-1}{\theta}S^{*2}\sim\chi^2_{n-1}$

3.5. Taille d'échantillon pour une précision donnée

Le problème est inverse, c'est-à-dire que l'on veut une précision donnée ε sur l'estimation et fixer la taille de l'échantillon en conséquence. La taille n est donc l'inconnue du problème qu'il faut déterminer telle que l'amplitude de l'intervalle de confiance ne dépasse pas ε.

- Dans le cas de la fréquence empirique, on a l'équation

$$\left(f_n+k\sqrt{\frac{f_n(1-f_n)}{n}}\right)-\left(f_n-k\sqrt{\frac{f_n(1-f_n)}{n}}\right)=\varepsilon$$

$$\Leftrightarrow\ n\geq\left(\frac{2k}{\varepsilon}\right)^2 f_n(1-f_n)$$

- Dans le cas de la moyenne (1^{er} ou $2^{ème}$ cas), on a l'équation

$$\left(\overline{x}+k\frac{\sigma}{\sqrt{n}}\right)-\left(\overline{x}-k\frac{\sigma}{\sqrt{n}}\right)=\varepsilon$$

$$\Leftrightarrow \ n \geq \left(\frac{2k\sigma}{\varepsilon}\right)^2$$

Les autres cas sont plus compliqués puisque la loi de l'estimateur dépend elle même de n.

<u>Exemple</u>
Un laborantin utilise un thermomètre sophistiqué mais dont les mesures sont incertaines. On sait que l'écart-type de l'appareil est de 1.1 degré Celsius. Combien doit-il effecteur de mesures pour connaître avec un coefficient de confiance de 95%, la température moyenne à 0.5 degré près ?
Soit $\overline{X}$ la moyenne des mesures. On suppose *a priori* n suffisamment grand pour considérer que $\overline{X}$ suit une loi normale $N(\mu,\sigma^2/n)$. Alors l'intervalle de confiance à 95% est donné par

$$\left[\overline{x}-1.96\frac{1.1}{\sqrt{n}}\,;\overline{x}+1.96\frac{1.1}{\sqrt{n}}\right]$$

d'où

$$n \geq \left(\frac{2\times 1.96\times 1.1}{0.5}\right)^2 = 74.37$$

Il doit effectuer au moins 75 mesures pour avoir la précision souhaitée. On constate que l'hypothèse sur la taille de l'échantillon est validée *a posteriori*.

RAPPELS SUR LES CONVERGENCES EN PROBABILITE

Convergence en loi

Notons F_1, ..., F_n les fonctions de répartition des variables X_1, ..., X_n. On dit que la suite *converge en loi* vers X, noté $X_n \underset{n\to+\infty}{\overset{\mathcal{L}}{\to}} X$, si et seulement si

$$\forall t \in D,\ \lim_{n\to+\infty} F_n(t) = F(t)$$

où F est la fonction de répartition de X et D le domaine de continuité de F.

> Théorème central-limite
>
> Supposons que $(X_n)_{n\in N}$ est une suite de variables aléatoires indépendantes de même loi (variables aléatoires i.i.d.) telles que $E(X_i)=\mu$ et $var(X_i)=\sigma^2$, alors le théorème central-limite permet de dire que
>
> $$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \underset{n\to+\infty}{\overset{\mathcal{L}}{\to}} N(0,1)$$

Ainsi, si l'échantillon est suffisamment grand, on peut approcher la loi de la moyenne $\overline{X}_n$ par une loi normale, $N(\mu, \frac{\sigma^2}{n})$, et ce, quelle que soit la loi initiale de l'échantillon.

Convergence en probabilité

On dit que la suite *converge en probabilité* vers la variable aléatoire X, noté $X_n \underset{n\to+\infty}{\overset{p}{\to}} X$, si et seulement si

$$\forall \varepsilon>0,\ \lim_{n\to+\infty} P(|X_n - X| < \varepsilon) = 1.$$

> Loi (faible) des grands nombres
>
> Supposons $X_1,\ldots,X_n$ indépendantes telles que $E(X_i)=\mu$ et $var(X_i)=\sigma^2$ existent, alors
>
> $$\frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to+\infty}{\overset{p}{\to}} \mu.$$

Convergence en moyenne quadratique

Supposons $X_1,\ldots,X_n$ telles que $E(X_i)$ et $var(X_i)$ existent. On dit que la suite *converge en moyenne quadratique vers X*, noté $X_n \underset{n\to+\infty}{\overset{m.q}{\to}} X$, si et seulement si

$$\lim_{n\to+\infty} E\left[|X_n - X|^2\right] = 0.$$

La convergence en moyenne quadratique permet de connaître la vitesse de convergence de la suite.