Chapitre 3. Estimation par intervalle

Pierre Duchesne

February 2, 2017

Considérons le caractère étudié X: l'apport hebdomadaire de capitaux dans une entreprise. Ce caractère est une variable aléatoire.

Il est décidé de modéliser ce phénomène par une variable normale, ainsi $X \sim \mathcal{N}(\mu, \sigma^2)$.

Le paramètre d'intérêt dans un premier temps est μ, l'apport hebdomadaire moyen.

Supposons que l'on ait prélevé un échantillon en centaine de milliers de dollars canadien:

$$\mathcal{E} : 1.37, 0.92, 0.67, 1.16;$$

Ainsi, la réalisation $\mathcal{E} : X_1, X_2, X_3, X_4$ permet de construire une **estimation ponctuelle**.
On a vu que la meilleure estimation ponctuelle de μ est:

$$\hat{\mu} = \bar{x} = \frac{1}{4}\sum_{i=1}^{4} = 1.03,$$

qui impliquerait ici que l'entreprise enregistre en moyenne 103 000$ par semaine.

L'estimation ponctuelle ne fournit qu'un nombre sans mention aucune de la précision de cette estimation.

L'estimation par intervalle en tient compte.

Elle se présente sous la forme: *Il y a 95 chances sur 100 que* μ *soit compris entre 0.99 et 1.07*.

Dans les médias, 95 chances sur 100 est souvent rapporté comme **19 fois sur 20**.

La longueur de l'intervalle, ici 1.07-0.99 = 0.08, est à la fois fonction de la précision de l'estimation ponctuelle ainsi que du degré de confiance qu'on accorde à l'énoncé.
Parfois, on préfèrera rapporter un intervalle sous la forme:

$$1.03 \pm 0.04,$$

afin de faire ressortir **l'estimation ponctuelle** et la **mesure de précision.**

Le degré de confiance dans l'intervalle sera appelé le **niveau de confiance**.

Dans une perspective fréquentiste (par opposition à une perspective bayésienne), un niveau de confiance de 95%, disons, s'interprète ainsi:

Si on prélève 100 échantillons de taille $n = 4$, on doit s'attendre à ce que 95 d'entre eux contiennent la vraie valeur de μ. Autrement dit, 95 d'entre eux auront une valeur de $\bar{x}$ telle que l'écart $|\bar{x} - \mu| \leq 0.04$.

On note que μ n'est pas aléatoire, c'est un paramètre. Il n'y a pas de sens à affirmer que la probabilité que μ soit dans l'intervalle est 95%. C'est plutôt l'intervalle qui est aléatoire, qui contiendra μ avec probabilité 95%.

Construction d'un intervalle de confiance: principle général

Soit X un caractère étudié dont la loi dépend d'un paramètre inconnu θ.

Soit $\mathcal{E}$: $X_1, \ldots, X_n$ un échantillon aléatoire de taille n et soit $x_1, \ldots, x_n$ une réalisation.

On définira un **pivot** pour θ une variable aléatoire à la fois fonction de $\mathcal{E}$ et de θ qui présente les caractéristiques suivantes:

(i) Sa loi soit entièrement connue;

(ii) Elle deviendrait une statistique si la valeur de θ était connue.

Exemple: Supposons que $X \sim \mathcal{N}(\theta, 1)$.
La variable aléatoire $\bar{X}$ est telle que $\bar{X} \sim \mathcal{N}(\theta, n^{-1})$ et ce **n'est pas** un pivot.
Par contre, $n^{1/2}(\bar{X} - \theta) \sim \mathcal{N}(0, 1)$ est fonction de l'échantillon aléatoire, fonction de θ, et de loi entièrement connue. Ainsi c'est bel et bien un pivot.
Nous aurons avantage à définir un pivot à partir d'un estimateur de θ. Plus l'estimateur sera **bon**, plus le pivot correspondant sera considéré **bon**.

La procédure générale afin de construire un intervalle de confiance est de fixer un **niveau de confiance** $1 - \alpha$. Les valeurs de α les plus courantes sont $\alpha = 1\%, 5\%, 10\%$.
À partir du pivot $T_\theta = t_\theta(X_1, \ldots, X_n)$, on détermine alors deux **bornes aléatoires**, c'est-à-dire deux statistiques:

$$T_1 = t^*(X_1, \ldots, X_n),$$

et

$$T_2 = t^{**}(X_1, \ldots, X_n),$$

satisfaisant la relation

$$P(T_1 < \theta < T_2) = 1 - \alpha.$$

Ainsi, un intervalle de confiance de niveau $1 - \alpha$ pour le paramètre θ, est donné par l'intervalle:

$$t_1 < \theta < t_2,$$

où

$$t_1 = t^*(x_1, \ldots, x_n),$$

et

$$t_2 = t^{**}(x_1, \ldots, x_n).$$

Ainsi t_1 et t_2 sont des réalisations de T_1 et T_2, respectivement.

Intervalle de confiance pour μ et pour σ^2 dans une population normale

Soit X un caractère étudié de loi $\mathcal{N}(\mu, \sigma^2)$. Soit $\mathcal{E}$: $X_1, \ldots, X_n$ un échantillon aléatoire.

On présume disposer d'une réalisation $x_1, \ldots, x_n$.

Intervalle de confiance pour μ quand σ^2 est connu

L'estimateur naturel de μ est $\bar{X}$ qui est de loi $\mathcal{N}(\mu, \frac{\sigma^2}{n})$. Le pivot à considérer est:

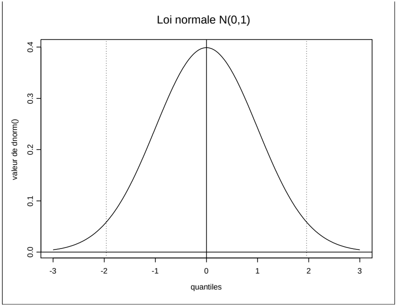
$$T_\mu(X_1, \ldots, X_n) = n^{1/2}\left(\frac{\bar{X} - \mu}{\sigma}\right) \sim \mathcal{N}(0, 1).$$

Considérons α_1 et α_2 tels que $\alpha_1 + \alpha_2 = \alpha$. Posons z_{α_1} et z_{α_2} deux quantiles tels que

$$P(\mathcal{N}(0,1) \leq z_{1-\alpha_i}) = 1 - \alpha_i,$$

c'est-à-dire

$$P(\mathcal{N}(0,1) > z_{1-\alpha_i}) = \alpha_i.$$



Nous aurons alors:

$$P\left(-z_{1-\alpha_1} < n^{1/2}\left(\frac{\bar{X}-\mu}{\sigma}\right) < z_{1-\alpha_2}\right) = 1-\alpha_1-\alpha_2 = 1-\alpha.$$

Or

$$\begin{aligned} n^{1/2}\left(\frac{\bar{X}-\mu}{\sigma}\right) < z_{1-\alpha_2} \quad &\Longleftrightarrow \quad \bar{X}-\mu < z_{1-\alpha_2}\frac{\sigma}{n^{1/2}}, \\ &\Longleftrightarrow \quad \bar{X}-z_{1-\alpha_2}\frac{\sigma}{n^{1/2}} < \mu. \end{aligned}$$

De même,

$$\begin{aligned} n^{1/2}\left(\frac{\bar{X}-\mu}{\sigma}\right) > -z_{1-\alpha_1} \quad &\Longleftrightarrow \quad \bar{X}-\mu > -z_{1-\alpha_1}\frac{\sigma}{n^{1/2}}, \\ &\Longleftrightarrow \quad \bar{X}+z_{1-\alpha_1}\frac{\sigma}{n^{1/2}} > \mu. \end{aligned}$$

Les manipulations ont isolé pour μ. De

$$P\left(-z_{1-\alpha_1} < n^{1/2}\left(\frac{\bar{X}-\mu}{\sigma}\right) < z_{1-\alpha_2}\right) = 1-\alpha,$$

nous avons obtenu:

$$P\left(\bar{X} - z_{1-\alpha_2}\frac{\sigma}{n^{1/2}} < \mu < \bar{X} + z_{1-\alpha_1}\frac{\sigma}{n^{1/2}}\right) = 1-\alpha.$$

Ainsi l'intervalle de confiance de niveau $1-\alpha$ est:

$$\bar{X} - z_{1-\alpha_2}\frac{\sigma}{n^{1/2}} < \mu < \bar{X} + z_{1-\alpha_1}\frac{\sigma}{n^{1/2}}.$$

Si nous posons $\alpha_1 = \alpha_2 = \alpha/2$, alors l'intervalle de confiance est symétrique en ce sens que $\bar{X}$ est le point milieu de l'intervalle. Autrement dit, l'estimation ponctuelle de μ est le point milieu de l'intervalle de confiance.
Autre écriture: $\mu \in \bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{n^{1/2}}$, au niveau de confiance $1-\alpha$.

Intervalle de confiance pour μ et pour σ^2 dans une population pas nécessairement normale

On vient de voir que l'intervalle $\mu \in \bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{n^{1/2}}$, au niveau de confiance $1 - \alpha$.

L'intervalle de confiance est exactement de niveau de confiance $1 - \alpha$, sous l'hypothèse que le caractère étudié X est de loi normale. En repassant la démarche, on voit que l'on a utilisé la normalité de la moyenne échantillonnale $\bar{X}$.

Si X est un caractère étudié qui n'est pas nécessairement de distribution normale, mais qui est tel que $E(X) = \mu$ et $\text{var}(X) = \sigma^2$, alors le théorème limite central peut être invoqué afin de construire un intervalle de confiance dont le niveau de confiance sera approximativement égal à $1 - \alpha$.

Rappel: Théorème Limite Central

Si on dispose d'un échantillon aléatoire, dont le caractère X est tel que $E(X) = \mu$ et $\mathrm{var}(X) = \sigma^2$, le Théorème Limite Central mentionne que:

$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

On constate que la démarche est la même, mais alors l'intervalle de confiance $\mu \in \bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{n^{1/2}}$, est **approximativement** de niveau de confiance $1 - \alpha$. L'approximation s'avère valable dès que $n \geq 30$ dans ce contexte.

Influence de la taille de l'échantillon sur la longueur de l'intervalle

Reprenons l'intervalle pour μ, qui est $\mu \in \bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{n^{1/2}}$, avec un niveau de confiance $1 - \alpha$.
La longueur de l'intervalle est donc:

$$2 z_{1-\alpha/2} \frac{\sigma}{n^{1/2}}.$$

Ainsi, si n augmente, la longueur de l'intervalle diminue: *Plus nous disposons d'information, plus n sera donc grand, et plus l'inférence sera précise pour estimer μ, et donc l'intervalle de confiance devient de plus en plus court (précis) à mesure que n augmente.*

Influence du niveau de confiance sur la longueur de l'intervalle

La longueur de l'intervalle est:

$$2z_{1-\alpha/2}\frac{\sigma}{n^{1/2}}.$$

Plus le niveau de confiance $1-\alpha$ augmente, plus la longueur de l'intervalle augmente: *Plus nous désirons un niveau de confiance élevé que l'intervalle contiendra μ, plus il faudra que l'intervalle de confiance soit large. Pourquoi ne pas choisir un niveau de confiance de 100%? Dans un tel cas, la longueur de l'intervalle serait infinie!*

Les niveaux de confiance courants:

$1-\alpha$	90%	95%	99%
$z_{1-\alpha/2}$	1.644854	1.959964	2.575829