Cours 4. Intervalles de confiance

Introduction

L'objectif du statisticien est d'estimer un paramètre ou le comportement d'une certaine population.

Exemples de paramètres que l' on veut connaître : *le pourcentage d'électeurs qui vont voter pour un candidat, la taille ou le poids moyen d'une population, la durée de vie d'un appareil (la population est ici l'ensemble des appareils), le pourcentage de personnes ayant un abonnement ADSL, etc...*

Exemples de comportement : *on veut trouver la loi par exemple des appels téléphoniques, du trafic sur Internet, de la moyenne des notes d'une promotion. On peut aussi vouloir tester une hypothèse, par exemple est ce qu'un régime ou un médicament est efficace.*

Si l'on veut une réponse exacte il faut examiner toute la population (en organisant un recensement ou un vote par exemple). Notez que cet examen est souvent infaisable, la population étant trop importante. Le statisticien va donc tirer au hasard un échantillon de la population (dit échantillon aléatoire) et, à partir des observations faites, donner une estimation de la valeur recherchée.

Exemple : *pour connaître les intentions de vote des électeurs on va faire un sondage sur une petite partie des électeurs. Supposons qu'on interviewe 1 000 personnes et que 520 déclarent voter pour un candidat A, cela ne prouve pas que le jour du vote final A va recueillir 52% des suffrages. Aussi le statisticien ne va pas dire (comme on peut l'entendre dans les médias) «A va recueillir 52% des voix», mais pourra dire «un intervalle de confiance à 95% du pourcentage de gens votant pour A est 48% - 56% (ou 52% $\pm$ 4%)». Ceci veut dire qu'il y a une probabilité 0.95 qu'entre 48% et 56% des électeurs votent pour A.*

La démarche du statisticien est une démarche inductive : il induit une valeur à partir d'une observation. Pour déterminer les intervalles de confiance on va utiliser les probabilités en prenant une démarche déductive. On va déterminer le comportement d'un échantillon aléatoire d'une variable aléatoire modélisant le paramètre étudié.

Une définition intuitive d'un échantillon aléatoire de n éléments consiste à tirer au hasard un sous ensemble de n éléments dans la population considérée. Tirer au hasard veut dire que chaque individu a la même probabilité d'être tiré. Ceci suppose qu'on connaît la population mère du paramètre que l'on veut estimer. *Ce n'est pas toujours le cas*. Enfin notez que dans les sondages, souvent les réponses sont fausses ou biaisées par la question posée.

A. Échantillon Aléatoire

On considère une variable aléatoire X, d'espérance $E(X)$, notée E, et d'écart type $\sigma(X)$, noté σ.

Un échantillon aléatoire à n éléments est un ensemble de n variables aléatoires $X_1, X_2, \ldots, X_n$, 2 à 2 indépendantes et suivant la même loi que X (en particulier $E(X_i) = E$ et $\sigma(X_i) = \sigma$).

Moyenne d'un échantillon aléatoire

Notre but est d'estimer la valeur d'un paramètre, ici l'espérance d'une variable aléatoire. Nous prendrons comme estimateur (voir plus loin) d'une espérance la variable aléatoire qui représente la moyenne des observations faites (appelée moyenne empirique). Par exemple pour estimer la moyenne des notes d'une promotion, on prendra la moyenne observée de n copies tirées au hasard.

Définition : On appelle **moyenne d'un échantillon aléatoire à n éléments** la variable aléatoire m_n définie par

$$m_n = \frac{X_1 + \cdots + X_n}{n}$$

Propriétés :

$$E(m_n) = E \quad \text{et} \quad \sigma(m_n) = \frac{\sigma}{\sqrt{n}}$$

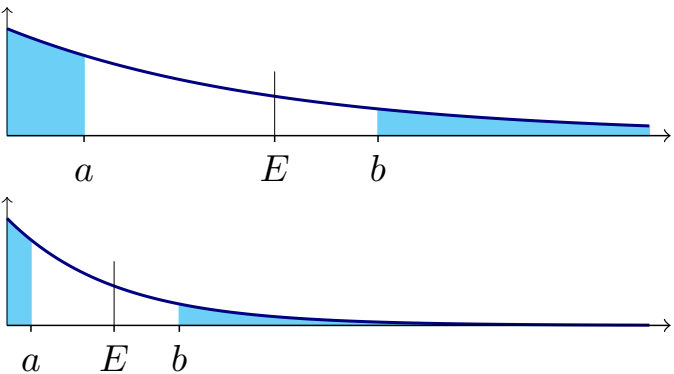
Remarque : si l'écart type dépend de n et décroît quand n augmente, l'espérance de m_n est constante. Un cas important est celui où X suit une loi normale $\mathcal{N}(E, \sigma)$, alors m_n suit une loi normale $\mathcal{N}(E, \frac{\sigma}{\sqrt{n}})$.

Un deuxième cas important est celui où n est assez grand (en pratique $n > 20$) ; en effet dans ce cas le théorème central limite implique qu'on peut approximer m_n par la même loi normale $\mathcal{N}(E, \frac{\sigma}{\sqrt{n}})$.

B. Intervalle de confiance

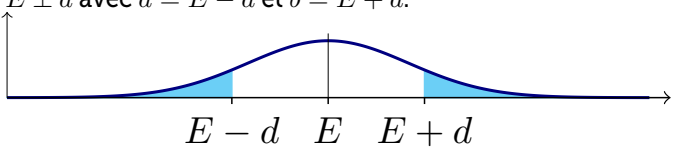
On appelle **intervalle de confiance** à $(1 - \alpha)$ d'une variable aléatoire X, l'intervalle $[a, b]$ centré en E tel que :

$$\mathbb{P}(X < a) = \frac{\alpha}{2}$$
$$\mathbb{P}(a \leq X \leq b) = 1 - \alpha$$
$$\mathbb{P}(b < X) = \frac{\alpha}{2}$$



Intervalles de confiance à $(1-\alpha)$ $[a,b]$ sur loi exponentielle $\mathcal{E}(\lambda)$
en haut : $\alpha = 0.5, \lambda = 0.4$ et en bas : $\alpha = 0.6, \lambda = 1$

Pour une loi symétrique, l'intervalle de confiance est centré autour de l'espérance. C'est par exemple le cas pour la loi normale. On note alors cet intervalle sous la forme $E \pm d$ avec $a = E - d$ et $b = E + d$.



Dans le cas particulier de la loi normale centrée réduite $\mathcal{N}(0,1)$, on note par $Z_{\frac{\alpha}{2}}$ la valeur de d.
$Z_{\frac{\alpha}{2}}$ est la valeur telle que

$$\mathbb{P}(-Z_{\frac{\alpha}{2}} \leq Z \leq Z_{\frac{\alpha}{2}}) = 1 - \alpha$$

ou vu la symétrie de Z :

$$\mathbb{P}(Z \leq -Z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

Exemple canonique : *pour* $\alpha = 0.05$, $Z_{\frac{\alpha}{2}} = 1.96$ *(lu dans la table pour la valeur* $\frac{\alpha}{2} = 2.5\%$*).*

Corollaire : L'intervalle de confiance à $(1-\alpha)$ d'une variable aléatoire X qui suit loi normale $\mathcal{N}(E,\sigma)$ est : $E \pm Z_{\frac{\alpha}{2}}\sigma$.

Corollaire : Pour la loi de m_n, $\mathcal{N}(E, \frac{\sigma}{\sqrt{n}})$, l'intervalle de confiance est : $E \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

Utilisation de l'intervalle de confiance (idée) : *Supposons qu'on veuille tester l'hypothèse H_0 : «l'espérance d'une variable aléatoire X vaut E». On va utiliser un intervalle de confiance à $(1-\alpha)$ de la variable aléatoire m_n. D'après la définition, avec une probabilité $(1-\alpha)$, la valeur observée de la moyenne d'un échantillon aléatoire à n éléments doit se trouver dans l'intervalle de confiance donné par le corollaire ci-dessus.*
Le statisticien va faire une observation de la moyenne des valeurs des éléments tirés au hasard. Si cette valeur dite moyenne observée et notée m_{obs} appartient à l'intervalle, il acceptera la valeur E avec une confiance $(1-\alpha)$ dans l'hypothèse H_0 (ou dans la personne qui formule l'hypothèse). Sinon, si m_{obs} n'appartient pas à l'intervalle de confiance, il rejettera la valeur E avec un risque α de se tromper.

On appelle alors cet intervalle de confiance **un intervalle à priori** (on a un a priori sur la valeur à estimer). En pratique E est inconnue. On définit alors un intervalle de confiance à posteriori de la manière suivante $m_{obs} \pm d$ où d représente la borne de l'intervalle de confiance. Cet intervalle représente l'ensemble des valeurs possibles au vu de l'observation faite.

C. Estimation de probabilité

On veut estimer une certaine probabilité p. On considère donc une population mère où les individus ont une probabilité p d'avoir un certain comportement ou propriété (exemple voter pour un candidat, fumer, être satisfait, tomber en panne, ...). Cela veut dire que si on tire au hasard un élément de la population mère, il a la probabilité p d'avoir le comportement. La loi pour le tirage d'un individu est donc une loi binomiale d'espérance p et de variance $p(1-p)$.

Pour estimer p on va utiliser la variable aléatoire f_n qui donne la fréquence (ou la proportion, le pourcentage) d'individus d'un échantillon aléatoire à n éléments ayant le comportement ou la propriété recherchée.

Rappel : Si on dénote par S_n la variable aléatoire qui compte le nombre total d'individus ayant une propriété alors Sn suit une loi binomiale $\mathcal{B}(n,p)$.

On en déduit donc le comportement de $f_n = \frac{S_n}{n}$ qui suit une loi dite loi binomiale en proportion à valeur dans $\left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\right\}$.

$$E(f_n) = p \qquad V(f_n) = \frac{p(1-p)}{n}$$

Remarque : Notez l'analogie avec la loi de m_n si on prend $E = p$ et $\sigma^2 = p(1-p)$.

Si n est assez grand on peut approximer f_n par une loi normale $\mathcal{N}\left(\mu = p,\ V = \frac{p(1-p)}{n}\right)$.

En fait, une bonne approximation est obtenue si la valeur de $np(1-p)$ est assez grande (en pratique supérieur à 30).

Corollaire : Pour n assez grand, un intervalle de confiance à $(1-\alpha)$ à priori de la variable aléatoire f_n est :

$$p \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

D. Facteur correctif

On a supposé que l'on avait un échantillon aléatoire à n éléments ce qui suppose un tirage **avec** remise. Dans la pratique on fait souvent des tirages **sans** remise (en particulier pour les sondages où on n'interroge pas 2 fois la même personne). Si on effectue un tirage sans remise il faut appliquer la règle suivante.

Règle pour un tirage sans remise : Si on effectue un tirage sans remise d'un échantillon aléatoire à n éléments dans une population mère de N éléments, il faut multiplier l'écart type de la loi considérée par $\sqrt{\frac{N-n}{N-1}}$.

Exemple : *Pour un tirage sans remise m_n suit une loi normale $\mathcal{N}(\mu = E, V = \frac{(N-n)\sigma^2}{(N-1)n})$.*

Remarque : Notez que si $N = 1$ le facteur correctif vaut 1 vu que l'on ne tire qu'un élément. Si $N = n$, on obtient un écart type corrigé de 0 ce qui est normal vu qu'on a tiré toute la population et qu'on connaît donc la valeur du paramètre avec certitude.
La valeur du facteur correctif peut-être trouvée en calculant la variance de l'ensemble des tirages possibles sans remise. L'ensemble des évènement élémentaire sera alors de cardinal $\binom{N}{n}$ (au lieu de N^n pour le tirage avec remise).

Remarque : *Dans de nombreux cas, la valeur du facteur correctif est si proche de 1 que le facteur est négligeable. Par exemple, pour un sondage politique en France, on a $N = 43\,000\,000$ d'électeurs, et $n = 1000$ la taille moyenne d'un échantillon. Alors*

$$\sqrt{\frac{N-n}{N-1}} = 1 - (1.1616 \times 10^{-5})$$

Ainsi, si l'on ne tient pas compte du facteur correctif, l'erreur n'est que d'un cent-millième, elle est donc négligeable.

Conclusion

De nombreux facteurs ne sont pas considérés par les calculs statistiques : la population qui accepte de répondre à un sondage n'est pas forcément représentative de la population totale, les réponses données ne sont pas forcément sincères et enfin, la question elle-même peut être mal comprise, ou comprise différement par les personnes interrogées.
Les résultats obtenus sont ainsi le plus souvent modifiés, avant d'être publiés, en utilisant des méthodes empiriques (issues de la sociologie) se basant sur l'historique connu de l'opinion que l'on cherche à mesurer. C'est pourquoi un sondage peut complétement se tromper lorsqu'on in-

terroge un échantillon d'une population sur une question jamais posée auparavant ou sur une situation nouvelle.

Enfin, l'impact des sondages sur la population est important, et plusieurs études ont constaté l'influence des résultats d'un sondage publiés sur la population. Si les sondages sont un thermomètre de l'opinion, ils sont aussi un faiseur d'opinion, par des phénomènes d'amplification médiatique : ce n'est pas tant le résultat d'un sondage qui influence l'opinion que l'interprétation qui en est faite et la manière dont il est présenté à la population.