

COURS N°6 : Estimations

On peut rappeler que les **biostatistiques** ont pour objectif de **prendre en compte la variabilité interindividuelle**, de **résumer et décrire des données** et de **comparer des échantillons**.

Nous avons fait, durant les 5 premiers cours, des **statistiques descriptives**, nous connaissons les caractéristiques des **populations** et nous avons **calculé des probabilités d'événements** pour tenter de **prévoir** certaines choses concernant des **échantillons**.

Nous allons maintenant réaliser des **statistiques inférentielles**, c'est-à-dire qu'on va **prendre en compte des résultats expérimentaux obtenus sur des échantillons pour tenter de généraliser des données sur toute une population**.

C'est cette seconde partie de l'enseignement qui est le plus important et qui permet aux biostatistiques d'avoir une **place importante en médecine**, dans plusieurs domaines :

- En **recherche clinique**, où l'on va **étudier et comparer des groupes de malades pour généraliser des données sur une population**.
- En **recherche diagnostique**, pour **évaluer la performance des tests ou des stratégies diagnostiques**.
- En **recherche thérapeutique**, pour :
 - Pour **évaluer la toxicité et l'efficacité de médicaments**.
 - Pour **comparer les nouveaux traitements par rapport aux précédents**.
 - Pour **prendre en compte la variabilité des réponses aux traitements entre patients pour adapter le traitement**.
- En **recherche pronostique**, pour **évaluer ou prédire l'évolution des maladies sous différentes stratégies thérapeutiques**.

[Je passe les rappels de la moyenne de variables aléatoires et l'intervalle de pari]

I. Théorie de l'estimation

A. Introduction

Pour les intervalles de pari, on connaît les valeurs théoriques de la moyenne et de la variance (ou de l'écart-type).

En **statistiques**, on a la **problématique inverse** : on a un échantillon de n valeurs et on veut en déduire quelque chose au niveau d'une population.

Se posent alors les questions de la **précision** de notre étude et de la **taille** efficiente ou non de notre échantillon.

B. Estimation statistique

L'estimation statistique consiste à **obtenir le maximum d'informations d'un échantillon en vue d'estimer un ou plusieurs paramètres inconnus dans la population.**

On se base alors sur l'échantillon représentatif qui est un **sous-ensemble de la population d'étude constitué par tirage au sort** (de façon aléatoire donc).

Cependant, souvent en médecine, on réalise un échantillonnage « systématique » (on va par exemple prendre tous les patients d'un service etc...).

Soit un n-échantillon de X (X_1, X_2, \dots, X_n) (les X sont **indépendants**) dont on cherche le paramètre θ inconnu.

NB : θ n'est pas une variable aléatoire !

Un estimateur de θ est une **fonction des valeurs de l'échantillon**. On le note T_n .

$$T_n = T(X_1, X_2, \dots, X_n)$$

A chaque fois qu'on tire un au sort un nouvel échantillon, l'application de la fonction T donne un résultat différent : on parle d'estimation de θ notée t_n .

$$t_n = T(x_1, x_2, \dots, x_n)$$

Les fluctuations d'échantillonnage de t_n sont liées aux valeurs x_1, x_2, \dots, x_n observées.

C. Propriétés d'un estimateur

1. Biais d'un estimateur

Le **biais d'un estimateur** est l'**écart entre la vraie valeur et la valeur indiquée**. On parle aussi de **déformation systématique**. Le biais d'un estimateur est donnée par la relation :

$$E(T_n) - \theta$$

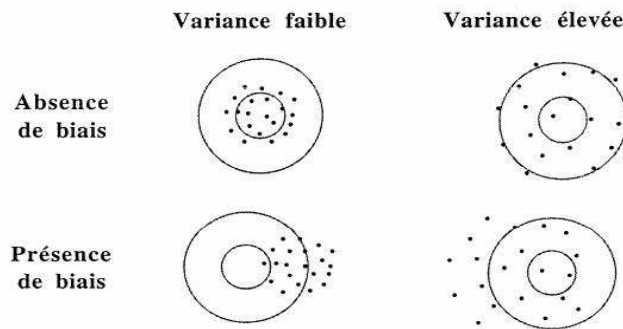
Un estimateur est **sans biais** si $E(T_n) = \theta$: les estimations obtenues ne s'écartent pas de la vraie valeur de façon systématique.

Notre but est d'avoir un estimateur sans biais pour notre étude.

2. Variance d'un estimateur

La **variance d'un estimateur** correspond à la **dispersion des estimations**. Elle est notée $\text{Var}(T_n)$.
 Quand les estimations sont peu dispersées, la variance est faible.
 Notre but est d'avoir une variance la plus basse possible.

Représentation du biais et de la variance
d'un estimateur



3. Convergence d'un estimateur

Un **estimateur est dit convergent** si $E([T_n - \theta]^2)$ tend vers 0 lorsque n tend vers l'infini.

$E([T_n - \theta]^2)$ est l'**erreur quadratique moyenne** (car elle dépend du biais et de la variance) :

$$E([T_n - \theta]^2) = E(T_n^2) - 2\theta E(T_n) + \theta^2 = E(T_n^2) - E(T_n)^2 + E(T_n)^2 - 2\theta E(T_n) + \theta^2$$

D'après la définition de variance, on a $\text{Var}(T_n) = E(T_n^2) - E(T_n)^2$

On remarque aussi que $E(T_n)^2 - 2\theta E(T_n) + \theta^2 = (E[T_n] - \theta)^2$ or $E(T_n) - \theta$ est le biais de l'estimateur.

Donc :

$$E([T_n - \theta]^2) = \text{Var}(T_n) + (E[T_n] - \theta)^2 = \text{Var}(T_n) + \text{biais}^2$$

D. Estimateur de la moyenne μ

Il est possible d'utiliser **3 estimateurs pour la moyenne μ** :

- $T_n = X_1$
- $T_n = \frac{(X_1, X_2, \dots, X_n)}{(n-1)}$
- $T_n = M_n = \frac{(X_1, X_2, \dots, X_n)}{n}$

1. Estimateur 1 : $T_n = X_1$

Pour cet estimateur, on a :

- $E(T_n) = E(X_1) = \mu$; d'où $E(T_n) - \mu = 0$
L'estimateur est donc **sans biais**.
- $\text{Var}(T_n) = \text{Var}(X_1) = \sigma^2$
- $E([T_n - \mu]^2) = \sigma^2$ (pas de biais)
Cette valeur ne dépend pas de n donc elle **ne tend pas vers 0 quand n augmente**. L'estimateur n'est **pas convergent**.

2. Estimateur 2 : $T_n = \frac{(X_1, X_2, \dots, X_n)}{(n-1)}$

Pour cet estimateur, équivalent à $\frac{M_n \times n}{(n-1)}$, on a :

$$- E(T_n) = \frac{E(M_n) \times n}{(n-1)} = \frac{\mu \times n}{(n-1)}$$

L'estimateur est **biaisé** ($E[T_n]$ est différent de μ). Le biais équivaut à $\frac{\mu \times n}{(n-1)} - \mu = \frac{\mu}{(n-1)}$

$$- \text{Var}(T_n) = \text{Var}\left(\frac{M_n \times n}{(n-1)}\right) = \frac{\sigma^2 \times n}{(n-1)^2}$$

$$- E([T_n - \mu]^2) = \frac{\sigma^2 \times n}{(n-1)^2} + \left(\frac{\mu}{(n-1)}\right)^2$$

L'erreur quadratique moyenne **tend vers 0 quand n tend vers 0** donc l'estimateur est **convergent**.

3. Estimateur 3 : $T_n = M_n = \frac{(X_1, X_2, \dots, X_n)}{n}$

Pour cet estimateur, on a :

- $E(T_n) = E(M_n) = \mu$
L'estimateur n'est pas biaisé.

- $Var(T_n) = Var(M_n) = \frac{\sigma^2}{n}$

- $E([T_n - \mu]^2) = \frac{\sigma^2}{n}$

L'erreur quadratique moyenne tend vers 0 quand n tend vers 0 donc l'estimateur est convergent.

M_n est donc un « bon » estimateur de la moyenne (en tout cas le meilleur des 3).
La moyenne expérimentale m est une estimation de μ et donc une réalisation de M_n .

E. Estimateur de la variance

A partir d'un n-échantillon, on prend une variable aléatoire V_n en tant qu'estimateur de la variance σ^2 , définie par :

$$V_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - M_n)^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n \times M_n^2 \right)$$

Une réalisation de V_n sur un échantillon est s^2 , variance « expérimentale » ou bien estimation de la variance :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n \times m^2 \right)$$

[à connaître pour estimer des variances ! Cette formule n'est pas dans le formulaire.]

Biais de l'estimateur :

$$E(V_n) = \frac{1}{n-1} \left(\sum_{i=1}^n E[X_i^2] - n \times E[M_n^2] \right)$$

On sait que $Var(X) = E(X^2) - E(X)^2$ donc :

$$E(X_i^2) = E(X^2) = Var(X) + E(X)^2 = \sigma^2 + \mu^2 \text{ donc } \sum_{i=1}^n E[X_i^2] = n(\sigma^2 + \mu^2) = n\sigma^2 + n\mu^2$$

De la même façon :

$$E(M_n^2) = \text{Var}(M_n) + E(M_n)^2 = \frac{\sigma^2}{n} + \mu^2$$

Donc

$$E(V_n) = \frac{1}{n-1} \left(\sum_{i=1}^n E[X_i^2] - n \times E[M_n^2] \right) = \frac{1}{n-1} (n\sigma^2 + n\mu^2 - n[\frac{\sigma^2}{n} + \mu^2]) = \frac{1}{n-1} (n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2)$$

$$E(V_n) = \frac{1}{n-1} (\sigma^2 [n-1]) = \frac{\sigma^2 (n-1)}{(n-1)} = \sigma^2$$

$$\boxed{E(V_n) = \sigma^2}$$

$$E(V_n) - \sigma^2 = 0$$

L'estimateur n'est pas biaisé.

La variance expérimentale s^2 est une estimation de σ^2 . C'est aussi une réalisation de V.

F. Estimateur d'une proportion

Soit π la moyenne d'une variable de Bernoulli. L'estimateur de π est donné par :

$$\boxed{T_n = P_n = \frac{(\text{nombre de } X_i = 1)}{n} = \frac{(X_1, X_2, \dots, X_n)}{n}}$$

C'est un estimateur non biaisé et convergent (il a donc les propriétés de M_n).

La proportion observée p est une estimation de π . C'est aussi une réalisation de P.

II. Intervalle de confiance

Soit un échantillon X_1, X_2, \dots, X_n d'une loi ayant un paramètre θ inconnu.

On pourrait donner une estimation ponctuelle de θ à partir d'un estimateur. Mais cela n'est **pas très satisfaisant** puisque l'estimation dépend de l'échantillon et qu'il existe de nombreuses fluctuations d'échantillonnage.

On peut aussi donner un intervalle de valeurs possibles (ou une « fourchette ») de θ pour donner une **précision** à la valeur estimée.

A. Définition

L'intervalle de confiance de niveau $1 - \alpha$, noté $IC_{1-\alpha}$, est un intervalle qui a pour probabilité $1 - \alpha$ de contenir la vraie valeur θ .

(Si on calcule $IC_{1-\alpha}$ sur un nombre infini d'échantillons, alors $\theta \in IC_{1-\alpha}$ dans une proportion de $1 - \alpha$ cas.)

B. Intervalle de confiance pour la moyenne

Soit X une variable aléatoire continue définie par $E(X) = \mu$ et $\text{Var}(X) = \sigma^2$. On considère qu'on a un grand échantillon et que $n \geq 30$ (on peut appliquer le TLC).

On sait que :

$$P(|\mu - M| \leq \varepsilon_\alpha \sqrt{\frac{\sigma^2}{n}}) = 1 - \alpha$$

On obtient l'estimation par intervalle de μ , ou l'intervalle de confiance de μ , en considérant que μ est l'inconnue que σ^2 est estimé par s^2 (le calcul est impossible sinon).

Sur un échantillon de n valeurs, on observe la moyenne m et la variance s^2 , toutes deux expérimentales, pour calculer l'intervalle de confiance de μ de niveau $1 - \alpha$, ou de risque α (en considérant que $n \geq 30$) :

$$IC_{1-\alpha} = \left[m \pm \varepsilon_\alpha \sqrt{\frac{s^2}{n}} \right]$$

Remarque : souvent, $\alpha = 5\%$ et donc $\varepsilon_\alpha = 1,96 \approx 2$, d'où :

$$IC_{95} = \left[m \pm 2 \sqrt{\frac{s^2}{n}} \right]$$

C. Intervalle de confiance pour une proportion

Soit X une variable de Bernoulli de paramètre π . On admet qu'on est sur un grand échantillon et que $n\pi \geq 5$ **ET** $n(1-\pi) \geq 5$

On sait que :

$$P(|\pi - M| \leq \varepsilon_\alpha \sqrt{\frac{\pi(1-\pi)}{n}}) = 1 - \alpha$$

On obtient l'estimation par intervalle de π , ou l'intervalle de confiance de π , en considérant que π est l'inconnue que π est estimé par p (le calcul est impossible sinon).

Sur un échantillon de n valeurs, on observe la proportion p expérimentale, pour calculer l'intervalle de confiance de π de niveau $1 - \alpha$, ou de risque α (en considérant que $n\pi \geq 5$ **ET** $n(1-\pi) \geq 5$):

$$IC_{1-\alpha} = \left[p \pm \varepsilon_\alpha \sqrt{\frac{p(1-p)}{n}} \right]$$

Attention : Après votre calcul, il faut vérifier les conditions de validité aux bornes de l'intervalle de confiance. Avec vos deux valeurs de π limites π_1 et π_2 , il faut s'assurer que les quatre termes $n\pi_1$; $n(1-\pi_1)$; $n\pi_2$; $n(1-\pi_2)$ sont tous supérieurs ou égaux à 5.

Remarque : souvent, $\alpha = 5\%$ et donc $\varepsilon_\alpha = 1,96 \approx 2$, d'où :

$$IC_{95} = \left[p \pm 2 \sqrt{\frac{p(1-p)}{n}} \right]$$

[formulaire page 2]

D. Largeur, précision et nombre de sujets nécessaires pour un intervalle de confiance

La largeur d'un intervalle $[a ; b]$ vaut $b - a$.

- Pour l'intervalle de confiance d'une moyenne, la largeur l est donnée par :

$$l = 2 \varepsilon_\alpha \frac{s}{\sqrt{n}}$$

Elle dépend de s et n .

- Pour l'**intervalle de confiance d'une proportion**, la **largeur l** est donnée par :

$$l = 2 \varepsilon_{\alpha} \sqrt{\frac{p(1-p)}{n}}$$

Elle dépend de **p** et de **n**.

(Pour diviser par 2 la largeur, il faut multiplier n par 4).

Pour rappel, la **précision i** est la **demi-largeur**.

Nombre de sujets nécessaires pour avoir une précision i donnée :

- Pour l'**intervalle de confiance d'une moyenne**, on a :

$$n \geq \frac{\varepsilon_{\alpha}^2 \times s^2}{i^2}$$

- Il faut se donner une **valeur à priori pour s²**.
- **n** augmente avec s².

- Pour l'**intervalle de confiance d'une proportion**, on a :

$$n \geq \frac{\varepsilon_{\alpha}^2 \times p(1-p)}{i^2}$$

- Il faut se donner une **valeur à priori pour p**.
- **n** est maximum pour **p = 0,5**.

*[Ces formules sont à connaître ou à retrouver, elles ne sont pas dans le formulaire. Il ne faut pas les confondre avec le nombre de sujets nécessaires pour les tests que nous allons voir plus tard !
Je passe sur la conclusion, tout est dans le cours.]*

Ce document, ainsi que tous les cours P1, sont disponibles gratuitement sur
<http://coursplbichat-larib.weebly.com>