Faculté des sciences exactes et sciences de la nature et de la vie
Département : Sciences exactes et sciences de la nature et de la vie de
Module : Biostatistiques

Chapitre II : Théorie d'estimation

1. Introduction :

L'inférence statistique traite principalement de deux types de problèmes : l'**estimation de paramètres** (espérance, variance, probabilité de succès) et **les tests d'hypothèses ;** elle ne conduit jamais à une conclusion stricte mais elle attache toujours **une probabilité à cette conclusion**.

Un phénomène biologique sera entièrement déterminé si l'on connaît la loi de probabilité suivie par la variable aléatoire donnée dans la population. On a alors deux cas de figure :

1) soit la **loi de probabilité suivie par *X* est connue *a priori*** et on vérifie ***a posteriori*** que les observations faites à partir d'un échantillon sont en accord avec elle.

2) soit la **loi de probabilité suivie par *X* est inconnue** mais suggérée par la description de l'échantillon (nature de la variable, forme de la distribution des fréquences, valeurs des paramètres descriptifs). Dans ce cas, il est nécessaire d'**estimer** les paramètres de la loi de probabilité à partir des paramètres établis sur l'échantillon.

2. L'estimation:

L'estimation est l'ensemble des méthodes utilisées pour évaluer un **paramètre (θ)** d'une population à l'aide d'un **estimateur ($\widehat{\theta}$)** pris dans un échantillon extrait de cette population.

3. L'objectif de l'estimation :

L'estimation a pour objectif de déterminer les valeurs inconnues des paramètres de la population (p, μ, σ^2) ou (proportion, moyenne, variance) à partir des données de l'échantillon (f, x, s^2).

Les notations des estimations des paramètres les plus couramment utilisées :

Paramètre	Valeur théorique	Estimation sur échantillon
Moyenne	μ	m
Variance	σ^2	S^2
Fréquence	P	P0

Exemples :

a. quelle est la fréquence de survenue de tel type de cancer chez les souris ?

b. quelle est la vraie valeur de la glycémie de ce patient ?

4. Estimateur :

On dispose d'observations indépendantes des phénomènes, c.à.d. de variables aléatoires X_1, X_2, …, X_n indépendantes et de même loi (celle du phénomène). On parle d'un échantillon. On définit à partir de l'échantillon une nouvelle variable aléatoire notée $\hat{\theta}$ dont les valeurs seront proches de celle de la grandeur θ à estimer. Cette nouvelle variable aléatoire $\hat{\theta}$ sera appelée estimateur de θ.

Il peut y avoir plusieurs estimateurs pour une même grandeur, certains meilleurs que d'autres.

5. Estimation ponctuelle et par intervalle :

L'**estimation** d'un paramètre quelconque θ est **ponctuelle** si l'on associe **une seule valeur** à l'estimateur $\hat{\theta}$ à partir des données observables sur un échantillon aléatoire.

L'**estimation par intervalle** associe à un échantillon aléatoire, **un intervalle** [$\hat{\theta}1, \hat{\theta}2$] qui recouvre θ avec une certaine probabilité.

5.1. Estimation ponctuelle :

On cherche à estimer une valeur inconnue liée a un certain phénomène aléatoire, en général, la moyenne ou la variance ou encore l'écart-type de la loi du phénomène.

- **Estimation de la moyenne et de la variance :**

Etant donne un échantillon X_1, X_2, …, X_n d'un caractère X inconnu, on admet que le meilleur estimateur de la moyenne μ du caractère X est :

$\overline{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$

Le meilleur estimateur de la variance σ^2 = Var(X) du caractère X est la variance empirique corrigée

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

- **Estimation de la fréquence :**

On note *p* la fréquence des individus de la **population** possédant le caractère A. La valeur de ce paramètre étant inconnu, on cherche à estimer la fréquence *p* à partir des données observables sur un échantillon.

A chaque échantillon non exhaustif de taille *n*, on associe l'entier *k*, nombre d'individus possédant le caractère A.

La **fréquence observée** du nombre de succès observé dans un échantillon de taille *n* constitue le meilleur estimateur de *p* : $\hat{P} = \frac{K}{n}$

Exemple :

On a prélevé au hasard, dans une population de lapin, 100 individus. Sur ces 100 lapins, 20 sont atteints par la myxomatose. Le pourcentage de lapins atteints par la myxomatose dans la population est donc : $\hat{P} = \frac{K}{n} = \frac{20}{100}$ = **0,2** soit 20% de lapins atteins dans la population

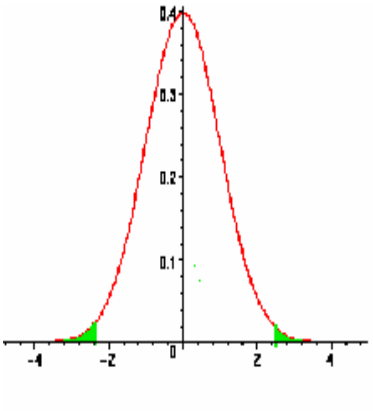
5.2. Estimation par intervalle de confiance :

- **Principe :**

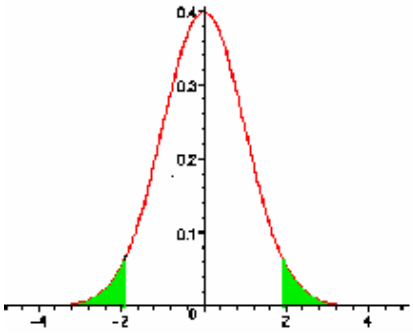
Un estimateur permet de calculer une valeur sur un échantillon qui devrait être proche du paramètre θ sans pour autant savoir si cette valeur est totalement fiable. C'est pourquoi on a introduit la notion d'intervalle de confiance : c'est un intervalle dans lequel se trouve θ avec une probabilité grande 1 - α (où α est un risque qu'on se fixe, en général, petit)

La probabilité 1-α est appelée **niveau de confiance** et α **le risque** (de 1ère espèce), c.à.d. la probabilité que l'intervalle proposé (qu'on notera IC) ne contienne pas la valeur à estimer.

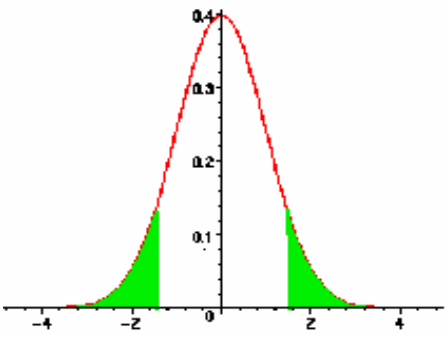
Un intervalle de confiance indique **la précision d'une estimation** car pour un risque α donné, l'intervalle est d'autant plus grand que la précision est faible comme l'indiquent les graphes ci-dessous. Pour chaque graphe, **l'aire hachurée en vert** correspond au coefficient de risque α. Ainsi de part et d'autre de la distribution, la valeur de l'aire hachurée vaut $\frac{\alpha}{2}$.



α = **0,01 : 99 chances sur 100** que la valeur du paramètre recherché se trouve dans l'intervalle de confiance mais la **précision** autour de la valeur prédite est **faible**



α **= 0,05 : 95 chances sur 100** que la valeur du paramètre recherché se trouve dans l'intervalle de confiance et la **précision** autour de la valeur prédite est **correcte**.



α **= 0,10 : 90 chances sur 100** que la valeur du paramètre recherché se trouve dans l'intervalle de confiance mais la **précision** autour de la valeur prédite est **élevée.**

5.2.1. Intervalle de confiance d'une moyenne :

En fonction de la nature de la variable aléatoire continue *X*, de la taille de l'échantillon *n* et de la connaissance que nous avons sur le paramètre σ^2, l'établissement de l'intervalle de confiance autour de μ sera différent.

Quelque soit la valeur de *n*, si $X \rightarrow N$ (μ , σ) et σ^2 est connue :

L'intervalle de confiance de la moyenne μ pour un coefficient de risque α est donné par :

$$\bar{X} - \varepsilon_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + \varepsilon_\alpha \frac{\sigma}{\sqrt{n}}$$

Remarque : La valeur de ε_α est donnée par la **table de l'écart-réduit** pour une valeur α donnée.

Coefficient de risque	Ecart-réduit
α = 0,01	ε_α = 2.576
α = 0,05	ε_α = 1.960
α = 0,10	ε_α = 1.645

- **Exemple :**

Pour des masses comprises entre 50g et 200g, une balance donne une pesée avec une variance de 0,0015. Les résultats des trois pesées d'un même corps sont : 64,32 ; 64,27 ; 64 ,39.

On veut connaître le poids moyen de ce corps dans la population avec un coefficient de confiance de 99%.

Avec $\bar{X}$ = 64,33g et ε_α = 2,576 alors $\varepsilon_\alpha \frac{\sigma}{\sqrt{n}}$= 2,576 x $\frac{0,039}{1,732}$ 0,039 1,732 = **0,058** et donc μ = $\bar{X} \pm \varepsilon_\alpha \frac{\sigma}{\sqrt{n}}$ **= 64,33g** ± **0,058**

D'où le poids moyen de ce corps est compris dans l'intervalle [64,27 ; 64,39] avec une probabilité de 0,99.

***Si n ≤ 30* et $X \rightarrow N$ (μ,σ) et σ^2 est inconnue :**

L'intervalle de confiance de l'espérance μ pour un coefficient de risque α est donné par :

$$\bar{X} - t_\alpha \frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{X} + t_\alpha \frac{\hat{\sigma}}{\sqrt{n}}$$

***Si n > 30* et $X \rightarrow N$ (μ,σ) et σ^2 est inconnue :**

L'intervalle de confiance de l'espérance μ pour un coefficient de risque α est donné par :

$$\bar{X} - \varepsilon_\alpha \frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{X} + \varepsilon_\alpha \frac{\hat{\sigma}}{\sqrt{n}}$$

- **Exemples :**

1) Dans un échantillon de **20 étudiants** de même classe d'âge et de même sexe, la taille moyenne observée est de 1,73m et l'écart-type de 10 cm. La taille moyenne de l'ensemble des individus est donc

Avec $\bar{x}$ = 1,73m ; $\hat{\sigma}^2 = \frac{n}{n-1}s^2 = \frac{20}{19}$ x 0.01= 0,011 et t_α = 2,086

D'où $t_\alpha \frac{\hat{\sigma}}{\sqrt{n}}$ = 2,086 x $\sqrt{\frac{0.011}{20}}$ = 0.049 ainsi μ = $\overline{X} \pm t_\alpha \frac{\hat{\sigma}}{\sqrt{n}}$ = **1,73m** ± **0,049**

La taille moyenne des étudiants dans la population est comprise dans l'intervalle [1,68 ; 1,78] avec une probabilité de 0,95.

2) Dans un échantillon de **100 étudiants**, la taille moyenne de la population est :

$\bar{x}$ = 1,73m ; $\hat{\sigma}^2 = \frac{n}{n-1}s^2 = \frac{100}{99}$ x 0.01 = 0.01 et ε_α = 1.960

D'où $\varepsilon_\alpha \frac{\hat{\sigma}}{\sqrt{n}}$ = 1,960 x $\sqrt{\frac{0.010}{100}}$ = 0.02 ainsi μ= $\overline{X} \pm \varepsilon_\alpha \frac{\hat{\sigma}}{\sqrt{n}}$ = **1,73m** ± **0,02**

La **taille moyenne** des étudiants dans la population est comprise dans l'intervalle **[1,71 ; 1,75]** avec une probabilité de 0,95.

Ainsi lorsque **la taille** de l'échantillon **augmente** pour un même coefficient de confiance (1-α), l'estimation autour de μ est **plus précise**.

5.2.2. Intervalle de confiance d'une proportion :

L'intervalle de confiance de la fréquence *p* pour un coefficient de risque α est donné par :

$$\frac{k}{n} - \varepsilon_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \frac{k}{n} - \varepsilon_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Vraie seulement si ***n* est grand et *np, nq* > 5**

- **Exemple :**

Un laboratoire d'agronomie a effectué une étude sur le maintien du pouvoir germinatif des graines de *Papivorus subquaticus* après une conservation de 3 ans. Sur un lot de 80 graines, 47 ont germé. Ainsi la probabilité de germination des grPaines de *Papivorus subquaticus* après trois ans de conservation avec un coefficient de confiance de 95% est donc :

Avec : $\hat{p} = \frac{k}{n} = \frac{47}{80} = 0.588$, $\hat{p} = \frac{n-k}{n} = \frac{33}{80} = 0.412$ et $\varepsilon_\alpha = 1.96$

Alors : $\varepsilon_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \times \sqrt{\frac{0.588.0.412}{80}} = 0.108$ d'où $p = 0{,}588 \pm 0{,}108$ ainsi la probabilité de germination est comprise dans l'intervalle [0,480 et 0,696] avec une probabilité de 0,95.