STATISTIQUE INFÉRENTIELLE

ESTIMATION PONCTUELLE

Benchikh Tawfik

Faculté de Médecine, UDL, SBA
1ère année Médecine

14 Février 2024

PLAN DE COURS

1. INTRODUCTION
2. EXPOSÉ DU PROBLÈME DE L'ESTIMATION
3. ESTIMATION PONCTUELLE
   - Définitions
   - Principales qualités d'un estimateur
     - Estimateur sans biais
     - Estimateur convergent
   - Précision d'un estimateur
4. QUELQUES MÉTHODES D'ESTIMATION

Démarche statistique: Rappel

- Les démarche statistique consiste:
  - à prélever un échantillon **représentatif** de la population (échantillon aléatoire) par des techniques appropriées. Les différentes méthodes utilisées pour obtenir un tel échantillon relèvent de la théorie de l'échantillonnage ;
  - à étudier les principales caractéristiques d'un échantillon (par rapport à une caractéristique: une variable statistique: X), issu d'une population dont on connaît la loi de probabilité;
  - à savoir réaliser des **échantillons de variables aléatoires** pour vérifier des conclusions (Statistique inférentielle).

Types d'inférence statistique

- Ces lois de probabilités sont généralement:
  - Totalement Inconnues: nous ne connaissons rien de la loi: problème de **statistique inférentielle non-paramétrique** (hors programme)
  - Partiellement inconnues: nous connaissons la famille à laquelle appartient la loi (sa forme) mais pas ses ou un certain nombre de ses paramètres (Exemple: X obéit à une loi normale, mais on ne connaît ni son espérance ni sa variance): problème de **statistique inférentielle paramétrique**

Inférence statistique: objectif

- **Déduction :** prédire, à partir d'une population connue ou supposée connue, les caractéristiques des échantillons qui en seront prélevés.
- **Induction (inférence):** prédire les caractéristiques d'une population inconnue à partir des statistiques déterminées dans un échantillon représentatif de cette population, c'est-à-dire, extrapolation des observations réalisées dans un échantillon à l'ensemble de la population

LES 3 GRANDES LIGNES DE L'INFÉRENCE **statistique paramétrique**

- Les statistiques vont permettre:
  - d'estimer un paramètre inconnu: **Estimation ponctuelle**. Par exemple, approcher , à partir de l'échantillon; l'espérance $\mathbb{E}(X)$ de la variable X: Taux de glycémie,...)
  - de donner une zone dans laquelle un paramètre, a de grande chance de se trouver: **Estimation par intervalle de confiance**.
  - de prendre des décisions: **Tests d'hypothèses**. Par exemple, décider si, au vu de l'échantillon, l'affirmation $\mathbb{E}(X) = 0.9$ est plausible.)
- La pertinence de ces méthodes repose en premier lieu sur la représentativité de l'échantillon.

ESTIMATION: DÉFINITION

- Un aspect important de l'inférence statistique consiste à obtenir des estimations fiables des caractéristiques d'une population à partir d'un échantillon extrait de cette population.
- Estimer un paramètre, c'est donner une valeur approchée de ce paramètre, à partir des résultats obtenus sur un échantillon aléatoire extrait de la population.

Paramètres à estimer

- Estimation c'est un problème de décision concernant des paramètres tels que:
  - l'espérance mathématique notée m ou μ (pour une variable numérique),
  - la variance σ^2 ou l'écart-type notée σ (pour une variable numérique),
  - la proportion p (pour une variable dénombrable).
- Comme un échantillon ne peut donner qu'une information partielle sur la population, les estimations ainsi obtenues seront inévitablement entachées d'erreurs que l'on doit minimiser autant que possible.

Exemple

- Un type de problème relevant de la théorie de l'estimation consiste en l'estimation des paramètres de la loi suivie par une variable aléatoire X , loi dont on connaît la forme.
- Ainsi, par exemple, le nombre d'accidents dans un atelier pendant une semaine suit probablement une loi de Poisson. Cette loi dépend d'un paramètre λ dont on ne connaît pas la valeur.
- Pour donner une valeur à ce paramètre, on note le nombre d'accidents survenus pendant n semaines, c'est l'échantillon aléatoire de taille n.
- Le nombre moyen d'accidents est une estimation ponctuelle du paramètre λ (propriété de la loi de Poisson).

Position du problème

- Sous une forme générale, le problème à résoudre peut se formuler ainsi:
  - Soit X une caractéristique d'un phénomène dont les réalisations dépendent du "hasard". La variable X est donc une variable aléatoire dont les caractéristiques (espérance, variance, ...) sont inconnues.
  - On observe les réalisations d'un échantillon aléatoire issu de la population étudiée.
  - Cette réalisation doit permettre d'induire des valeurs ou estimations des paramètres de la loi suivie par la variable X.

Position du problème

- Ces estimations peuvent revêtir deux formes:
  - soit une valeur unique, l'estimation ponctuelle, ou valeur la plus probable que prendra le paramètre,
  - soit un ensemble de valeurs appartenant à un intervalle, l'estimation par intervalle de confiance. Un intervalle de confiance doit avoir de "grandes chances" de contenir la vraie valeur du paramètre, il est toujours associé à un risque d'erreur α.

DÉFINITION D'UN ESTIMATEUR PONCTUELLE

- Le but de la théorie de l'estimation ponctuelle est de choisir, parmi toutes les statistiques possibles, le meilleur estimateur, c'est-à-dire celui qui donnera une estimation ponctuelle la plus proche possible du paramètre θ et ceci, quel que soit l'échantillon.

Définition

- Un **estimateur** est une statistique qui a des propriétés bien définies.
- Une suite T_n de statistiques, $T_n = \varphi(\underline{X}_n)$, est appelée estimateur du paramètre θ si T_n tend vers θ quand n tend vers l'infini, la convergence étant une convergence en probabilité, presque sûre ou en moyenne quadratique (critères).
- Une fois l'échantillon effectivement réalisé, l'estimateur prend une valeur numérique, appelée estimation ponctuelle du paramètre θ par la statistique T.
- On note, en général, $\hat{\theta}(\underline{X})$ ou $\hat{\theta}_n$, ou plus simplement $\hat{\theta}$, l'estimateur du paramètre θ.

Définition d'un estimarteur: Exemple

- Soit $\overline{X}$ la statistique:

$$\overline{X} = T(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i \, ,$$

  c'est-à-dire la fonction moyenne arithmétique des n observations d'un échantillon.
- Cette statistique peut être considérée comme un estimateur, a priori raisonnable, de l'espérance mathématique $\mathbb{E}(X) = m$ (d'après la Loi faible des grands nombre).

Définition d'un estimateur: Exemple

- En effet:
  - cette statistique prend en compte toutes les observations,
  - cette statistique possède les propriétés suivantes:

$$\mathbb{E}(T) = \mathbb{E}(X) = m \qquad Var(T) = \frac{Var(X)}{n},$$

  - si la loi de la variable aléatoire X est connue, on peut en déduire celle de $\overline{X}$.
- Par la suite, on montrera que les deux premières propriétés de cet estimateur sont, en fait, de "bonnes propriétés" pour un estimateur.

Estimateur sans biais: définition

- Un estimateur est **sans biais** si $\mathbb{E}(T) = \theta$.
- Un estimateur est **biaise** si $\mathbb{E}(T) \neq \theta$.
- Un estimateur est asymptotiquement sans biais si $\mathbb{E}(T) \to \theta$, quand la taille n de l'échantillon tend vers l'infini.

ESTIMATEUR SANS BIAIS: REMARQUE

- Un estimateur biaisé donne des estimations qui peuvent s'écarter systématiquement de la valeur à estimer; il est donc moins satisfaisant qu'un estimateur sans biais.
- Cependant, l'absence de biais n'est pas une garantie absolue de "bon estimateur". Il faut aussi tenir compte de sa variance.

Estimateur sans biais: exemple 1

- **Estimateur de l'espérance mathématique $m = \mu$:**
  - La statistique $\overline{X}$ déjà étudiée est un estimateur sans biais pour l'espérance mathématique $\mathbb{E}(X)$.
  - En effet, $\mathbb{E}(\overline{X}) = \mathbb{E}(X)$.

ESTIMATEUR SANS BIAIS: EXEMPLE 2

- **Estimateur de la variance:**
  - La statistique $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ est un estimateur biaisé pour la variance:
$$\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2 = \left(1 - \frac{1}{n}\right)\sigma^2 ,$$
le biais est égal à $\frac{\sigma^2}{n}$.
  - En revanche, la statistique S^{*2} définie par $S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ est un estimateur sans biais pour la variance. En effet:
$$\mathbb{E}(S^{*2}) = \frac{n}{n-1}\mathbb{E}(S^2) = \sigma^2 .$$

ESTIMATEUR CONVERGENT OU BON ESTIMATEUR

- Un estimateur est convergent (en probabilité) si sa distribution tend à se concentrer autour de la valeur inconnue du paramètre.
- Un estimateur sans biais, dont la variance tend vers 0 quand n tend vers l'infini, est donc convergent.
- Mais il n'est pas nécessairement unique comme le montrent les exemples ci-dessous.

ESTIMATEUR CONVERGENT: EXEMPLE 3

- **Comparaison d'estimateurs de l'espérance mathématique**
  - Comme estimateur de l'espérance mathématique m, on peut choisir la statistique T_1:

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \quad \mathbb{E}(T_1) = m \quad Var(T_1) = \frac{\sigma^2}{n}\,.$$

    T_1 est un estimateur sans biais dont la variance tend vers 0 quand n tend vers l'infini, il est donc convergent.
  - On peut aussi choisir la statistique T_2, moyenne arithmétique des observations de rang impair. On obtient si on suppose n pair:

$$T_2 = \frac{2}{n}\sum_{i=1}^{p-1} X_{2i+1}\,, \quad n = 2p \quad \mathbb{E}(T_2) = m \quad Var(T_2) = \frac{2\sigma^2}{n}$$

    L'estimateur T_2 est sans biais et sa variance tend vers 0 quand n tend vers l'infini, il est donc convergent.

ESTIMATEUR CONVERGENT: EXEMPLE 3

- **Conclusion:** les deux estimateurs T_1 et T_2 ont les mêmes propriétés. Il est évident cependant que T_1 est "meilleur" que T_2. En effet:
  - il tient compte de toute l'information apportée par l'échantillon,
  - sa variance est la plus petite: $Var(T_1) < Var(T_2)$.

ESTIMATEUR SANS BIAIS: EXEMPLE 4

- **Comparaison d'estimateurs de la variance:**
  - On suppose connue l'espérance mathématique $\mathbb{E}(X) = m$ de la loi de probabilité de la variable aléatoire X et on cherche alors le meilleur estimateur de la variance.
  - Comparons les deux estimateurs sans biais T et S^{*2}:

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

  - Un calcul rapide donne (μ_4 est le moment centré d'ordre 4):

$$Var(T) = \frac{1}{n}(\mu_4 - \sigma^4) \qquad Var(S^{*2}) = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4) .$$

  - D'où $Var(T) < Var(S^{*2})$.
  - **Conclusion:** si l'espérance mathématique m est connue, l'estimateur T est un estimateur de la variance "meilleur" que la statistique S^{*2}

Exemple: estimation du paramètre p d'une loi binomiale

- On veut étudier une caractéristique X d'un phénomène médical, par exemple la proportion p des individus d'une population P d'effectif N, présentant un certain caractère: avoir une une telle maladie, avoir une allergie pour un médicament, etc.
- Pour obtenir la valeur exacte de p, il suffirait d'interroger les N individus de la population ce qui, en général, est impossible; on interroge donc les individus d'un échantillon aléatoire, de taille n, représentatif de la population. On obtient un ensemble de n valeurs x_i telles que:
  - $x_i = 1$ si l'individu i présente le caractère X,
  - $x_i = 0$ sinon.

Exemple

- Soit k_n le nombre de réponses favorables.
- Soit $f_n = \frac{k}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}_n$: la proportion des individus ayant le caractère X dans l'échantillon.
- Cette proportion f_n converge vers p quand n tend vers l'infini (loi des grands nombres).

$$\mathbb{E}(k_n) = n\,p \Rightarrow \mathbb{E}(f_n) = p\,,$$

et

$$Var(k_n) = n\,p\,(1-p) \Rightarrow Var(f_n) = \frac{p\,(1-p)}{n}$$

-
- La fréquence f_n est un estimateur sans biais du paramètre p. De plus, sa variance tend vers 0 quand n tend vers l'infini, il est donc convergent.
- Par un choix adéquat de n, on peut essayer de minimiser l'erreur due à cette approximation.

PRÉCISION D'UN ESTIMATEUR

- La précision d'un estimateur est mesurée par l'erreur quadratique moyenne:

$$\mathbb{E}\left[(T-\theta)^2\right] .$$

- La propriété la plus désirable pour un estimateur est d'avoir une faible Erreur Quadratique Moyenne (ce qui n'exige pas forcément d'être sans biais).

Précision d'un estimateur

- En écrivant comme précédemment $T - \mathbb{E}(T) + \mathbb{E}(T) - \theta$, on obtient:

$$\mathbb{E}\left[(T-\theta)^2\right] = Var(T) + \left[\mathbb{E}(T) - \theta\right]^2 .$$

- Pour rendre l'erreur quadratique moyenne la plus petite possible, il faut que:
  - $\mathbb{E}(T) = \theta$, donc choisir un estimateur sans biais,
  - $Var(T)$ soit petite.

Précision d'un estimateur: Résumé

- Parmi les estimateurs sans biais, on choisira donc celui qui a la variance la plus petite, cette propriété traduit l'*efficacité* de l'estimateur.
- Ainsi, si l'espérance mathematique m est connue, on choisira comme estimateur de la variance σ^2, la statistique T et non la statistique S^{*2}.
- **Remarque:** Il est possible de trouver des estimateurs biaisés plus précis que le meilleur estimateur sans biais.
- **Problème:** le calcul de la variance d'un estimateur et donc l'existence et la définition d'un estimateur de variance minimale nécessite généralement la connaissance de la loi de probabilité jointe de l'échantillon aléatoire.

RÉSUMÉ: ESTIMATEUR CONVERGENT ET SANS BIAIS

- D'une moyenne $\mathbb{E}(X)$: la statistique $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
  - c'est un estimateur sans biais: $\mathbb{E}(\overline{X}) = m$,
  - son risque quadratique est asymptotiquent nul: $\mathbb{E}\left[(\overline{X} - m)^2\right] = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0$.
  - c'est un estimateur convergent $\overline{X} \xrightarrow{\Pr} m$, grasse au loi faible des grands nombres.
- d'une variance: la statistique $S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, c'est un estimateur sans biais et convergent.
- d'une proportion p: la fréquence $f_n = \frac{k_n}{n}$ (k_n nombre de réalisations de l'événement étudié au cours de n épreuves). C'est un estimateur sans biais convergent et son risque quadratique est asymptotiquement nul.

RÉSUMÉ: ESTIMATEUR CONVERGENT ET SANS BIAIS

- **Remarque:** S^* n'est pas un estimateur sans biais de l'écart-type σ, en effet:

$$\mathbb{E}(\sqrt{S^{*2}}) \neq \sqrt{\mathbb{E}(S^{*2})}\,.$$

MÉTHODE DES MOMENTS: DÉFINITION

- Soit $(X_1, X_2, \ldots, X_n)$ un échantillon aléatoire de taille n dans la population de distribution de probabilité $f(x, \theta)$, où θ est un vecteur de paramètres inconnus de dimension p et que l'on cherche à estimer.
- Soient

$$\overline{X^r} = \frac{1}{n}\sum_{i=1}^{n} X_i^r \quad \text{et} \quad m_r(\theta) = \mathbb{E}_\theta(X^r)\,,$$

les moments empiriques et théoriques d'ordre r de l'échantillon et de la loi de X respectivement (lorsque $r = 1$, on a la moyenne empirique et l'espérance). Les moments théoriques dépendent de θ.

Méthode des moments: définition

- L'estimateur des moments de θ est le vecteur Θ_n qui vérifie le système d'équations en θ: $\overline{X^r} = m_r(\theta)$
- Si ce système peut être résolu, $\Theta_n = k(\overline{X^1}, \ldots, \overline{X^p})$ est un estimateur des moments de θ.
- Quand la réalisation d'un échantillon est donné $(x_1, \ldots, x_n)$, la valeur de l'estimateur est

$$\hat{\theta}_n = k(\overline{x^1}, \ldots, \overline{x^p}) \, .$$

EM: REMARQUES

1. La méthode des moments fournit des estimateurs convergents
2. La méthode des moments fournit des estimateurs peu précis lorsque n est modéré.

EXERCICE 1

1. Supposons que l'on ait noté chez 11 individus normaux, pris au hasard, la valeur du rythme cardiaque. On a : 64, 80, 72, 88, 78, 88, 78, 88, 88, 72, 60.
2. On veut estimer la moyenne et la variance du rythme cardiaque chez les individus normaux (population cible).
3. On a alors :
   - Estimation de la moyenne de la population (= moyenne de l'échantillon): $\overline{X} = 77.82$. C'est un estimateur sans biais et convergent.
   - Variance de l'échantillon : $S^2 = 90.51$ et l'estimation de la variance de la population (estimateur sans biais et convergent): $S'^2 = \frac{n}{n-1} S^2 = 99.56$.
   - Estimation de l'écart-type de la population : $S' = \sqrt{99.56} = 9.98$

EXERCICE 2

On considère l'échantillon statistique $(1, 0, 2, 1, 1, 0, 1, 0, 0)$.

1. Calculer sa moyenne et sa variance empiriques.
2. En supposant que les données de cet échantillon sont des réalisations d'une variable de loi inconnue, donner une estimation non biaisée de l'espérance et de la variance de cette loi.
3. On choisit de modéliser les valeurs de cet échantillon par une loi binomiale $\mathcal{B}(2, p)$. Utiliser la moyenne empirique pour proposer une estimation ponctuelle pour p.
4. Avec le même modèle, utiliser la variance empirique pour proposer une autre estimation de p.
5. On choisit de modéliser les valeurs de cet échantillon par une loi de Poisson $\mathcal{P}(\lambda)$, Quelle estimation ponctuelle proposez-vous pour λ ?