Théorie de l'estimation

Cours de Bio-Statistiques

Medouer Nawel
n.meddouer@univ-batna2.dz

Table des matières

1 Introduction : C'est quoi la théorie d'estimation ?

L'objectif de ce cour est de répondre à la problématique suivante :

Comment à partir des informations sur l'échantillon (moyenne, écart-type, la proportion) peut-on prévoir celles d'une population ?

1.1 Problématique

Variable aléatoire d'étude, une population mère caractérisée par deux paramètres exactes (moyenne ou écart-type) ou encore une fréquence d'un caractère dans cette population, un échantillon représentatif issu de cette population. La distribution de X dans la population est normale ou quelconque.

L'objectif de la théorie de l'estimation est de répondre à la problématique suivante : comment à partir des informations (moyenne m ou proportion) calculées sur un échantillon estimer celles d'une population ?

Nous distinguerons deux cas :

— On estime la moyenne d'une distribution définie sur la population

— On estime la proportion

Vocabulaire

- La moyenne théorique dans la population, c'est la valeur attendue μ
- $\overline{X}$ est la variable aléatoire qui est l'estimateur de μ
- $\overline{x}$ est la moyenne empirique
- m est la moyenne de l'échantillon observée
- La loi de $\overline{X}$ est la loi de probabilité de la moyenne empirique, sa distribution théorique, la distribution d'échantillonnage des moyennes

2 Estimation et Estimations ponctuelles

L'estimation ponctuelle consiste à donner une valeur unique du paramètre, soit Φ un estimateur de θ.

$$\textbf{BIAIS} = E(\Phi) - \theta$$

2.1 Estimation sans biais

Si Biais $= 0$ (c'est-à-dire $E(\Phi) = \theta$), alors l'estimateur est dit **sans biais** ou **non biaisé**.

Exemple : (Résultat de théorie d'échantillonnage) $E(\overline{X}) = \mu \rightarrow \overline{X}$ est donc un estimateur sans biais de μ.

m la moyenne de l'échantillon observée est une estimation ponctuelle de μ.

Alors

- Pour estimer la moyenne inconnue d'une population, on prélève un échantillon et on calcule la moyenne de cet échantillon
- Cette moyenne d'échantillon est une estimation ponctuelle de la moyenne μ

Attention ! ! L'écart-type de l'échantillon n'est pas un estimateur sans biais de σ

Exemple

⇒ *Pour estimer l'écart-type d'une population σ, on prélève un échantillon et on calcule l'écart-type corrigé*

$$S = \sqrt{\frac{n}{n-1}}\tilde{S}$$

de cet échantillon.

Cet écart-type corrigé de l'échantillon est une estimation ponctuelle de l'écart-type σ.

Remarque

- L'écart-type de l'échantillon : $\tilde{S} = \sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2}$
- L'écart-type corrigé : $S = \sqrt{\frac{1}{n-1}\sum(x_i - \overline{x})^2}$

Exemple

- Pour estimer la fréquence inconnue d'un caractère dans une population, on prélève un échantillon et on calcule la fréquence d'apparition de ce caractère dans l'échantillon
- Cette fréquence d'apparition est une estimation ponctuelle de la fréquence p

3 Estimation par intervalle de confiance de la moyenne :

L'estimation par intervalle de confiance I consiste à construire un intervalle à l'intérieur duquel le paramètre se trouve avec une probabilité donnée.

$$\mu \in I = \left[m - \bar{Z}_\alpha \frac{\sigma}{\sqrt{n}} \, ; \, m + \bar{Z}_\alpha \frac{\sigma}{\sqrt{n}} \right]$$

☐ m le centre de l'intervalle de confiance.
☐ l'erreur = demi-longueur de l'intervalle = $\bar{Z}_\alpha \frac{\sigma}{\sqrt{n}}$.
☐ $Pr(\mu \in I) = 1 - \alpha$

Remarque

— $z_{\alpha/2}$ est calculé à partir de la table 2 (table de l'écart réduit)
— Cette formule est une conséquence du théorème de la limite centrale

3.1 Population infinie avec ou sans remise - Population finie avec remise

Cas 1 : X suit la loi normale et σ connu
Intervalle de confiance :

$$\left[m - \bar{Z}_\alpha \frac{\sigma}{\sqrt{n}} \, ; \, m + \bar{Z}_\alpha \frac{\sigma}{\sqrt{n}} \right]$$

Cas 2 : X suit la loi normale et σ inconnu, on l'estime par l'estimation ponctuelle $S = \sqrt{\frac{n}{n-1}} \tilde{S}$

1. Pour $n \geq 30$, on cherche dans la table Normale
   Intervalle de confiance :

$$\left[m - \bar{Z}_\alpha \frac{S}{\sqrt{n}} \, ; \, m + \bar{Z}_\alpha \frac{S}{\sqrt{n}} \right]$$

2. Pour $n < 30$, on cherche dans la table de Student
   Intervalle de confiance :

$$\left[m - \bar{t}_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}, \; m + \bar{t}_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}} \right]$$

Cas 3 : X suit la loi quelconque : il faut que $n \geq 30$

1. Pour σ connu, on cherche dans la table Normale
   Intervalle de confiance :

$$\left[m - \bar{Z}_\alpha \frac{\sigma}{\sqrt{n}} \, ; \, m + \bar{Z}_\alpha \frac{\sigma}{\sqrt{n}} \right]$$

2. Pour σ inconnu, on cherche dans la table Normale
   Intervalle de confiance :

$$\left[m - \bar{Z}_\alpha \frac{S}{\sqrt{n}} \, ; \, m + \bar{Z}_\alpha \frac{S}{\sqrt{n}} \right]$$

Remarque : Parmi les cinq intervalles de confiance, la lecture est faite dans la table de Student (table 3) dans un seul cas : il s'agit du cas $n < 30$, σ inconnu, et X suit la loi normale.

Attention ! ! Population finie tirage sans remise : on ajoute le facteur d'exhaustivité : $\sqrt{\frac{N-n}{N-1}}$

3.2 Population finie et tirage exhaustif (sans remise)

Cas 1 : X suit la loi normale et σ connu
Intervalle de confiance :

$$\left[m - \bar{Z}_\alpha \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \,;\, m + \bar{Z}_\alpha \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}\right]$$

Cas 2 : X suit la loi normale et σ inconnu, on l'estime par l'estimation ponctuelle $S = \sqrt{\frac{n}{n-1}}\tilde{S}$

1. Pour $n \geq 30$, on cherche dans la table Normale
   Intervalle de confiance :
$$\left[m - \bar{Z}_\alpha \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \,;\, m + \bar{Z}_\alpha \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}\right]$$
2. Pour $n < 30$, on cherche dans la table de Student
   Intervalle de confiance :
$$\left[m - \bar{t}_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}},\; m + \bar{t}_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}\right]$$

Cas 3 : X suit la loi quelconque : il faut que $n \geq 30$

1. Pour σ connu, on cherche dans la table Normale
   Intervalle de confiance :
$$\left[m - \bar{Z}_\alpha \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \,;\, m + \bar{Z}_\alpha \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}\right]$$
2. Pour σ inconnu, on cherche dans la table Normale
   Intervalle de confiance :
$$\left[m - \bar{Z}_\alpha \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \,;\, m + \bar{Z}_\alpha \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}\right]$$

Exemple : Sur 20 patients a été mesuré le taux de fer sérique (distribution Normale) exprimé en :
83.0 ; 98.0 ; 183.3 ; 119.6 ; 78.5 ; 162.6 ; 155.7 ; 147.3 ; 100.1 ; 139.2 ; 172.1 ; 102.0 ; 162.8 ; 113.8 ; 157.4 ; 128.5 ; 136.2 ; 129.3 ; 131.6 ; 157.3

1. Les estimations ponctuelles non biaisées de la moyenne et la variance du taux de fer sérique à partir de cet échantillon
2. L'estimation de la moyenne au risque de α par intervalle de confiance

Solution : Sur 20 patients a été mesuré le taux de fer sérique (distribution normale), en $\mu g/dL$:

$83.0, 98.0, 183.3, 119.6, 78.5, 162.6, 155.7, 147.3, 100.1, 139.2, 172.1, 102.0, 162.8, 113.8, 157.4, 128.5, 136.2,$

1. **Estimations ponctuelles biaisées** :
   En introduisant les x_i dans la calculatrice, on obtient :
   $$m = 145{,}435 \quad \text{et} \quad S = 28{,}72 \quad \text{(écart-type corrigé)}$$
2. **Intervalle de confiance à 95% de la moyenne :**
   Puisque σ inconnu et $n = 20 < 30$, on utilise la loi de Student :
   $$\bar{t}_\alpha = t_{0.05;\,19} \approx 2{,}093$$

$$\left[m - \bar{t}_\alpha \cdot \frac{S}{\sqrt{n}}\,;\, m + \bar{t}_\alpha \cdot \frac{S}{\sqrt{n}}\right] = \left[145{,}435 - 2{,}093 \cdot \frac{28{,}72}{\sqrt{20}}\,;\, 145{,}435 + 2{,}093 \cdot \frac{28{,}72}{\sqrt{20}}\right]$$

$$\frac{S}{\sqrt{n}} \approx 6{,}42 \Rightarrow IC = [145{,}435 - 13{,}43\,;\, 145{,}435 + 13{,}43] = [132{,}01\,;\, 158{,}87]$$

Conclusion : la moyenne du taux de fer sérique est estimée à $145{,}4\,\mu g/dL$, avec un intervalle de confiance à 95% compris entre 132,01 et $158{,}87\,\mu g/dL$.

Exemple : La tension artérielle systolique (en mmHg) a été mesurée chez 85 patients âgés de plus de 65 ans dans le cadre d'une étude gériatrique. Les mesures obtenues sont telles que :

$$\sum x_i = 11\,220 \qquad \sum x_i^2 = 1\,484\,300$$

1. Donner une estimation ponctuelle de la tension artérielle systolique moyenne pour la population des patients âgés de plus de 65 ans.
2. Donner une estimation ponctuelle de la variance de la tension artérielle systolique et de son écart-type pour cette population.
3. Estimer par intervalle de confiance, la tension artérielle systolique moyenne pour la population étudiée.

Solution Une étude mesure la tension artérielle systolique (en mmHg) chez 85 patients de plus de 65 ans. On observe les résultats suivants :

$$\sum x_i = 11\,220 \qquad \sum x_i^2 = 1\,484\,300$$

— Effectif de l'échantillon : $n = 85$
— Moyenne :
$$m = \frac{\sum x_i}{n} = \frac{11\,220}{85} = 132$$

— Variance :

$$S^2 = \frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right) = \frac{1}{84}\left(1\,484\,300 - \frac{(11\,220)^2}{85}\right) = \frac{3260}{84} \approx 38{,}81$$

— Écart-type :

$$S = \sqrt{38{,}81} \approx 6{,}23$$

Intervalle de confiance à 95% pour la moyenne de la tension artérielle :

Puisque $n \geq 30$, on utilise la loi normale centrée réduite. Le quantile de confiance est :

$$\bar{Z}_\alpha = 1{,}96 \quad \text{(pour 95\%)}$$

L'intervalle de confiance s'écrit :

$$\left[m - \bar{Z}_\alpha \frac{S}{\sqrt{n}}\,;\, m + \bar{Z}_\alpha \frac{S}{\sqrt{n}}\right] = \left[132 - 1{,}96 \cdot \frac{6{,}23}{\sqrt{85}}\,;\, 132 + 1{,}96 \cdot \frac{6{,}23}{\sqrt{85}}\right]$$

$$\frac{S}{\sqrt{n}} = \frac{6{,}23}{9{,}22} \approx 0{,}6756$$

$$\Rightarrow IC = [132 - 1{,}324\,;\, 132 + 1{,}324] = [130{,}68\,;\, 133{,}32]$$

Conclusion : avec un niveau de confiance de 95%, la tension artérielle systolique moyenne des patients âgés est comprise entre **130,68 mmHg** et **133,32 mmHg**.

Exercice d'entrainement (arrondir à 0.01 près) La maladie d'Alzheimer est une maladie neurodégénérative touchant principalement les personnes âgées. On s'intéresse à la moyenne d'âge d'apparition des premiers symptômes de la maladie.

On sélectionne un échantillon aléatoire de 200 malades, et on relève pour l'apparition des premiers symptômes une moyenne observée de 75 ans et un écart-type $\tilde{S} = 5$ ans. On acceptera un risque de 5%.

QCM 1 : Choisir la bonne réponse

(A) L'estimation ponctuelle de la moyenne théorique est de 75.
(B) L'estimateur sans biais de la moyenne théorique est la moyenne observée.
(C) L'estimation non biaisée de la moyenne théorique est de 75.19.
(D) L'estimateur sans biais de la moyenne théorique est de 75.
(E) Aucune réponse juste.

QCM 2 : Choisir la bonne réponse (les bonnes réponses)

(A) L'estimation non biaisée de l'écart-type théorique c'est l'écart-type observé.
(B) L'estimateur sans biais de l'écart-type théorique est de 5.
(C) L'estimation ponctuelle de l'écart-type théorique est de 5.
(D) L'estimateur sans biais de l'écart-type théorique est de 5.01.
(E) Aucune réponse juste.

QCM 3 : Concernant l'intervalle de confiance

(A) L'intervalle calculé est centré sur la moyenne d'âge d'apparition des symptômes de la maladie d'Alzheimer dans la population.

(B) L'intervalle calculé est centré sur la moyenne d'âge d'apparition des symptômes de la maladie d'Alzheimer des 200 malades.

(C) Il y a 95% de chances que la moyenne d'âge d'apparition des premiers symptômes des 200 malades soit comprise dans l'intervalle de confiance.

(D) Il y a 95% de chances que la moyenne d'âge d'apparition des premiers symptômes dans la population soit comprise dans l'intervalle de confiance.

(E) Aucune réponse juste.

QCM 4 : Choisir la bonne réponse (les bonnes réponses)

(A) L'intervalle calculé est de [74.31 — 75.69].

(B) L'intervalle calculé est de [74 — 75].

(C) Conditions non satisfaites pour calculer l'intervalle de confiance.

(D) L'intervalle calculé est de [74.71 — 75.86].

(E) Aucune réponse juste.

4 Estimation par intervalle de confiance de la proportion

Un intervalle de confiance pour la proportion ne se calcule que pour les grands échantillons $n \geq 30$ et sous cette condition :

$$n \cdot \hat{p} > 5 \quad n \cdot (\hat{p} - 1) > 5$$

$$I = \left[\hat{p} - \bar{z}_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + \bar{z}_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

tels que :

Centre de l'intervalle : $\hat{p}$

Erreur : $\bar{z}_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$Pr(p \in I) = 1 - \alpha$