11 CHAP. 11 Théorèmes limites et statistique inférentielle

Les objectifs : Initier à la démarche de la statistique inférentielle sur quelques cas simples, en présentant le problème de l'estimation par intervalle et du test de conformité. Il ne doit en aucun cas faire l'objet d'un développement théorique. On pourra illustrer par des exemples pris dans la vie courante et dans les autres disciplines.

1 Vocabulaire de l'échantillonnage et de l'estimation

Définition

Définition 1.1 :

*X étant une variable aléatoire d'espérance μ et de variance σ^2, un n-échantillon de X est un n-uplet $(X_1, X_2, \cdots, X_n)$ de variables aléatoires **indépendantes** et de **même loi** que X.*

✎ ***Remarque :*** Dans la pratique, il est fréquent que l'on soit dans une situation où l'on désire estimer des paramètres concernant une population d'individus, alors qu'il est impossible d'accéder à cette population dans son intégralité.
Il suffit de penser aux estimations « sortie des urnes » un jour d'élection, à l'échantillonnage pour estimer un caractère donné (diamètre, altération d'un sable, diamètre, age moyen d'un arbre sur une parcelle donnée, etc.), plus généralement à tous résultats expérimentaux issus de n répétitions indépendantes d'une expérience bâtie pour évaluer une grandeur liée à un modèle théorique (pH, température, volume, etc.)

Définition

Définition 1.2 : Estimateur

Soit $(X_1, \cdots, X_n)$ un n-échantillon d'une variable aléatoire X. On appelle estimateur d'un paramètre θ de X (au programme, $\theta = \mu$ ou σ^2) toute suite de variables aléatoires $(T_n)_{n\geq 1}$, fonction de $(X_1, \cdots, X_n)$ qui donne de l'information sur θ.

Remarque

Remarque 1.1.

On dira que la valeur de T_n obtenue à partir d'un échantillon observé de n individus est l'estimateur du paramètre

.

Exemple

Exemple 1.1

Soit $(X_1, \cdots, X_n)$ un n-échantillon d'une variable aléatoire X.

① *La **moyenne empirique** $M_n = \dfrac{X_1 + \cdots + X_n}{n} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ est un estimateur de $\mathbb{E}(X) = \mu$.*

② *Si X est une variable aléatoire centrée, alors $T_n = \dfrac{X_1^2 + X_2^2 + \cdots + X_n^2}{n}$ est un estimateur de $\mathbb{E}(X^2) = \sigma^2$.*

③ *La **variance empirique** $S_n^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - M_n)^2 = \dfrac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - M_n^2$ est un estimateur de σ^2.*

Définition

Définition 1.3 : Erreur d'estimation et biais

On appelle `erreur d'estimation` *la différence entre l'estimateur et la valeur du paramètre, et le* `biais` *l'espérance de l'erreur d'estimation.*
En d'autres termes, le biais est égale à $\mathbb{E}(T_n - \theta) = \mathbb{E}(T_n) - \theta$ *et l'estimateur sera dit* `sans biais` *si* $\mathbb{E}(T_n) = \theta$

Exemple

Exemple 1.2

Si $(X_1, \cdots, X_n)$ *est un n-échantillon d'une variable aléatoire* X *qui admet une espérance* μ *et une variance* σ^2*, alors* M_n *est un estimateur sans biais de* μ *et la variance de l'erreur d'estimation* $M_n - \mu$ *tend vers* 0

Exemple

Exemple 1.3

Si $(X_1, \cdots, X_n)$ *un n-échantillon d'une variable aléatoire* X *qui admet une espérance* μ *et une variance* $\mathbb{V}(X) = \sigma^2$*, alors* S_n^2 *est un estimateur biaisé de* σ^2

Exemple

Exemple 1.4

Soit $(X_1, \cdots, X_n)$ *un n-échantillon d'une variable aléatoire* $X \hookrightarrow \mathcal{U}_{]0,\theta[}$.
Alors la suite (T_n) *où* $T_n = \dfrac{n+1}{n} max(X_1, \cdots, X_n)$ *est un estimateur sans biais de* θ.

2 Théorèmes limites

Propriété

prop.2.1. Théorème de Bienaymé-Tchebychev

Soit X *une variable aléatoire admettant une espérance* μ *et une variance* $\mathbb{V}(X) = \sigma^2$*. Alors :*

$$\boxed{\forall \varepsilon > 0,\ \mathbb{P}\left(|X - \mathbb{E}(X)| \geq \varepsilon\right) \leq \frac{\mathbb{V}(X)}{\varepsilon^2}}$$

Remarque

Remarque 2.1.

L'inégalité de Bienaymé-Tchebychev illustre le fait que $\mathbb{V}(X)$ *mesure la dispersion de* X *autour de son espérance. En effet, la probabilité que* X *s'écarte de* $\mathbb{E}(X)$ *de plus de* ε *est d'autant plus faible que* $\mathbb{V}(X)$ *est faible.*

Propriété

prop.2.2. Loi faible des grands nombres

Soit $(X_n)_{n\in\mathbb{N}^*}$ *une suite de variables aléatoires réelles* ***deux à deux indépendantes*** *et* ***de même loi****, admettant* ***une même espérance*** μ *et* ***un même écart-type*** σ*.*

Si $M_n = \dfrac{X_1 + \cdots + X_n}{n} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ *désigne la moyenne empirique, alors :*

$$\forall \varepsilon > 0,\ \boxed{\mathbb{P}\left(|M_n - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}} \text{ ou encore } \boxed{\lim_{n\to\infty} \mathbb{P}\left(|M_n - \mu| \geq \varepsilon\right) = 0}$$

Remarque

Remarque 2.1.

Le loi faible des grands nombres peut prendre une autre forme si on passe par le complémentaire.
Sous les mêmes hypothèses, $\boxed{\lim_{n\to\infty} \mathbb{P}\left(|M_n - \mu| < \varepsilon\right) = 1}$

Propriété

prop.2.3. Le cas particulier du théorème de Bernoulli

Soit $(X_n)_{n\in\mathbb{N}^*}$ *une suite de variables aléatoires réelles* ***deux à deux indépendantes*** *et* ***de même loi de Bernoulli de paramètre*** p*. Alors* M_n *désigne la fréquence observée du succès et :*

$$\forall \varepsilon > 0,\ \boxed{\mathbb{P}\left(|M_n - p| \geq \varepsilon\right) \leq \frac{p(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}}$$

Exemple

Exemple 2.1

Une pièce de monnaie est lancée 1000 *fois et* 480 *piles sont obtenus. Si on note* p *la probabilité d'obtenir pile, déterminer l'intervalle* $]a, b[$ *dans lequel* p *a une probabilité au moins égale à* 0.9 *de se trouver.*

Propriété

prop.2.4. Théorème central limite (Première forme)

Soit $(X_n)_{n\in\mathbb{N}^*}$ *une suite de variables aléatoires réelles* ***indépendantes*** *et* ***de même loi****, admettant* ***une même espérance*** μ *et* ***un même écart-type*** σ *non nul.*

Si $M_n = \dfrac{X_1 + \cdots + X_n}{n} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ *désigne la moyenne empirique, alors :*

$$\forall a, b \in \overline{\mathbb{R}}, a < b,\ \lim_{n\to\infty} \mathbb{P}(a < \frac{M_n - \mu}{\sigma/\sqrt{n}} < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt$$

ou encore

$$\lim_{n\to\infty} \mathbb{P}(a < M_n^* < b) = \phi(b) - \phi(a)$$

où ϕ *désigne la fonction de répartition de la loi normale centrée réduite.*

Remarque

Remarque 2.2.

Cette loi peut être vue comme le moyen d'obtenir une approximation asymptotique de la loi de l'erreur d'estimation réduite $\varepsilon(M_n) = \dfrac{M_n - \mu}{\sigma/\sqrt{n}}$

✎ ***Illustration numérique :*** Vérifier la convergence précédente en simulant des tirages répétés et indépendants d'une loi uniforme ou d'une loi exponentielle.

Exemple

Exemple 2.2

Une montre fait une erreur d'au plus une demi minute par jour selon une loi uniforme. Quelle est la probabilité que l'erreur commise au bout d'une année soit inférieure à 15 *minutes ?*

Propriété

prop.2.5. Théorème de Moivre-Laplace

Soit X *une variable aléatoire qui suit une loi binomiale de paramètres* n *et* p*. Si* n *est suffisamment grand et si* p *n'est ni trop proche de* 0*, ni trop proche de* 1 *(dans la pratique, on vérifiera* $n \geq 30$*,* $np \geq 10$ *et* $nq \geq 10$*), alors :*

$$\forall (a,b) \in \bar{\mathbb{R}}^2,\ a < b,\ \mathbb{P}(a < X^* < b) = \mathbb{P}\left(a < \frac{X - np}{\sqrt{npq}} < b\right) \approx \phi(b) - \phi(a)$$

ou encore, sous les mêmes conditions :

$$\text{Si } X \hookrightarrow \mathcal{B}(n,p),\ \text{alors } \forall x \in \mathbb{R},\ \mathbb{P}(X \leq x) \approx \phi_{np,\sqrt{npq}}(x)$$

où $\phi_{np,\sqrt{npq}}$ *est une densité de* $\mathcal{N}(np, \sqrt{npq})$*.*

Exemple

Exemple 2.3

Une pièce de monnaie amène « pile » avec la probabilité 0.4*. On la lance* 100 *fois.*

① *Quelle est la probabilité d'obtenir au plus cinquante « pile » ?*

② *Quelle est la probabilité que le nombre de « pile » soit exactement de* 50 *?*

Propriété

prop.2.6. Théorème central limite (Seconde forme)

Soit $(X_n)_{n\in\mathbb{N}^*}$ *une suite de variables aléatoires réelles* ***indépendantes*** *et* ***de même loi****, admettant* ***une même espérance*** μ *et* ***une même variance*** *inconnue.*
Alors, S_n *désignant l'écart-type empirique :*

$$\forall a,b \in \overline{\mathbb{R}}, a < b,\ \lim_{n\to\infty} \mathbb{P}\left(a < \frac{M_n - \mu}{S_n/\sqrt{n}} < b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt$$

3 Applications statistiques

On rappelle que si ϕ désigne la fonction de répartition de la loi normale centrée réduite, alors ϕ est une bijection de $\mathbb{R}$ sur $]0,1[$ dont la réciproque ϕ^{-1} est appelée la fonction des quantiles.
Sous Python, après avoir importé le module adéquat, à savoir : `from scipy.stats import norm`, on a : `norm.cdf(x)` $= \phi(x)$ et `norm.ppf(y)` $= \phi^{-1}(y)$.

Dans la suite, on notera $u_{1-\frac{\alpha}{2}} = \phi^{-1}(1-\frac{\alpha}{2})$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$.
On a notamment, si $X \hookrightarrow \mathcal{N}(0,1)$: $\mathbb{P}\left(-u_{1-\frac{\alpha}{2}} < X < u_{1-\frac{\alpha}{2}}\right) = 2\phi(u_{1-\frac{\alpha}{2}}) - 1 = 1-\alpha$

Propriété

prop.3.1. Intervalle de confiance

Soit $(X_1, \cdots, X_n)$ un n-échantillon d'une variable aléatoire X admettant une espérance μ et une variance σ^2.
Si M_n et S_n désignent respectivement la moyenne et l'écart-type empirique de l'échantillon, alors :

$$\lim_{n\to\infty} \mathbb{P}\left(\left[M_n - u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}} < \mu < M_n + u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right]\right) = 1-\alpha$$

L'intervalle ci-dessus est appelé intervalle de confiance au niveau de confiance $1-\alpha$.

Remarque

Remarque 3.1.

Le plus souvent, on prend $\alpha = 0.05 = 5\%$, c'est-à-dire un niveau de confiance de 95%.

Dans ce cas, $\phi^{-1}(1-\frac{\alpha}{2}) = \phi^{-1}(0.975) =$ `norm.ppf(0.975)` $= 1.96$ et

$$\mu \in \left[M_n - 1.96\frac{S_n}{\sqrt{n}}, M_n + 1.96\frac{S_n}{\sqrt{n}}\right] \text{ avec un niveau de confiance de } 0.95$$

Si le niveau de confiance est de 0.9, $u_{1-\frac{\alpha}{2}} =$ ..... et l'intervalle de confiance est IC= ... ...

Interprétation : On dira que « on a confiance à 95% que l'intervalle de confiance contienne la valeur μ » ou encore « la probabilité qu'un intervalle construit de cette manière contienne μ est de 95% ».
On veillera par exemple à **ne pas écrire** : $\mathbb{P}(M_n - E_n < \mu < M_n + E_n) = 0.95$.

Remarque

Remarque 3.2. Test de conformité de la moyenne

Soit un n-échantillon de moyenne μ_0 d'une variable aléatoire X.

Si on souhaite tester l'hypothèse (H_0) : « $\mu = \mu_0$ », alors on utilise que $\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{M_n - \mu_0}{S_n/\sqrt{n}}\right| > u_{1-\frac{\alpha}{2}}\right) = \alpha$.

Dès lors, on décide de rejeter l'hypothèse (H_0) avec un risque α de se tromper si :

$$\frac{M_n - \mu_0}{S_n/\sqrt{n}} \text{ n'est pas dans l'intervalle } [-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}]$$