Statistiques

Estimation ponctuelle

Julian Tugaut

Télécom Saint-Étienne

Sommaire

Plan

Introduction - 1

Dans ce chapitre, il s'agit d'estimer certaines caractéristiques statistiques de la loi (moyenne, variance, fonction de répartition) au travers d'une série d'observations.

Introduction - 2

Le problème de l'estimation peut s'énoncer de la façon suivante : disposant d'observations $x[1], x[2], \cdots, x[n]$ d'une variable aléatoire X, obtenues à partir d'un échantillonnage aléatoire (n-échantillon de la variable aléatoire X), quelle loi théorique inconnue peut-on retenir comme loi de X ? (loi parente)

Introduction - 3

Si ce choix devait être fait parmi l'ensemble de toutes les lois de probabilité existantes, on conçoit que le problème serait difficilement résolu sans un échantillon de taille très grande. Mais, dans ce chapitre, nous allons voir que l'on peut le résoudre dès lors que l'on restreint le choix de la loi parente à une famille de lois de probabilité parfaitement déterminées par la donnée d'un ou plusieurs paramètres.

Introduction - 4

Le problème devient alors de choisir une valeur de chaque paramètre (estimation ponctuelle).

Introduction - 4

Le problème devient alors de choisir une valeur de chaque paramètre (estimation ponctuelle).
L'estimation consiste à donner des valeurs approximatives aux paramètres d'une population à l'aide d'un échantillon de n observations issues de cette population. On peut se tromper sur la valeur exacte, mais on donne la "meilleure valeur" possible que l'on peut supposer.

Exemple - 1

Exemple

Avant de choisir un véhicule automobile, on se fixe un critère de choix basé sur le nombre X de pannes par an que l'on est susceptible d'avoir avec un modèle donné. Ayant la possibilité de faire une étude statistique chez un concessionnaire donné, on prélève au hasard n dossiers de véhicules et l'on note $x[1], x[2], \cdots, x[n]$ le nombre de pannes subies la première année de mise en circulation de ces véhicules.

Exemple - 2

La loi de Poisson est bien adaptée à la modélisation du nombre de pannes. Conséquemment, le choix de la loi parente est fait parmi la famille de lois $\{\mathcal{P}(\lambda)\,;\,\lambda > 0\}$.

Exemple - 2

La loi de Poisson est bien adaptée à la modélisation du nombre de pannes. Conséquemment, le choix de la loi parente est fait parmi la famille de lois $\{\mathcal{P}(\lambda)\,;\,\lambda > 0\}$.
L'unique paramètre déterminant la loi est ici λ. Or, on sait que λ est l'espérance de la loi :

$$\lambda = \mathbb{E}[X]\,.$$

Exemple - 2

La loi de Poisson est bien adaptée à la modélisation du nombre de pannes. Conséquemment, le choix de la loi parente est fait parmi la famille de lois $\{\mathcal{P}(\lambda)\,;\,\lambda > 0\}$.
L'unique paramètre déterminant la loi est ici λ. Or, on sait que λ est l'espérance de la loi :

$$\lambda = \mathbb{E}[X]\,.$$

On estime donc la valeur de ce paramètre par la moyenne des valeurs observées sur l'échantillon :

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x[i]\,.$$

Plan

Vocabulaire - 1

On commence par donner les définitions basiques en statistiques inférentielles. Dans tout le reste du cours, Ω désigne une population de taille N. On suppose que N est grand si bien qu'un caractère C peut être vu comme une variable aléatoire sur l'univers Ω.

Vocabulaire - 2

Définition : Échantillon

On appelle échantillon un sous-ensemble de la population Ω.

Vocabulaire - 2

Définition : Échantillon

On appelle échantillon un sous-ensemble de la population Ω.

Définition : Taille de l'échantillon

Un échantillon de taille n est une liste de n individus $\{\omega_1, \cdots, \omega_n\}$ extraits de la population Ω.

Vocabulaire - 2

Définition : Échantillon

On appelle échantillon un sous-ensemble de la population Ω.

Définition : Taille de l'échantillon

Un échantillon de taille n est une liste de n individus $\{\omega_1, \cdots, \omega_n\}$ extraits de la population Ω.

On dit que Ω est la population mère de l'échantillon.

Vocabulaire - 3

Exemple

On considère une population constituée de cinquante étudiants. On a ainsi $\Omega = \{\omega_1, \cdots, \omega_{50}\}$ et $N = 50$. On prend un échantillon de six étudiants, $\{\omega_4, \omega_8, \omega_{15}, \omega_{16}, \omega_{23}, \omega_{42}\}$ et $n = 6$.

Vocabulaire - 4

Définition : Taux d'échantillonnage

Le rapport de l'effectif n de l'échantillon sur l'effectif N de la population mère dans laquelle il a été prélevé est appelé taux d'échantillonnage : $t = \frac{n}{N}$.

Vocabulaire - 4

Définition : Taux d'échantillonnage

Le rapport de l'effectif n de l'échantillon sur l'effectif N de la population mère dans laquelle il a été prélevé est appelé taux d'échantillonnage : $t = \frac{n}{N}$.

Exemple

Dans l'Exemple plus haut, le taux d'échantillonnage est égal à $\frac{6}{50} = 0.12 = 12\%$.

But du jeu - 1

On cherche à décrire un caractère qualitatif ou quantitatif (qu'il soit discret ou continu) C dans une population mère Ω à travers l'étude des résultats obtenus sur un échantillon de taille n.

But du jeu - 1

On cherche à décrire un caractère qualitatif ou quantitatif (qu'il soit discret ou continu) C dans une population mère Ω à travers l'étude des résultats obtenus sur un échantillon de taille n.

Remarque

Le taux d'échantillonnage doit répondre à deux critères : il faut qu'il soit suffisamment élevé pour rendre compte de la population mère et il faut qu'il soit suffisamment petit pour être simple à étudier.

But du jeu - 2

Exemple

Étant donné une population d'étudiants, on peut s'intéresser à un caractère quantitatif discret comme la note à un partiel (0, 5, 15, 20...).

But du jeu - 2

Exemple

Étant donné une population d'étudiants, on peut s'intéresser à un caractère quantitatif discret comme la note à un partiel (0, 5, 15, 20...).

Exemple

Étant donné une population d'étudiants, on peut s'intéresser à un caractère quantitatif continu comme la taille.

But du jeu - 3

Exemple

Étant donné une population d'étudiants, on peut s'intéresser à un caractère qualitatif comme la matière préférée ("probabilités", "statistiques", "signaux et systèmes discrets", "traitement des signaux déterministes", "signaux aléatoires", "la sieste"...).

Échantillon

Définition : n-échantillon de valeurs de X

Soit C un caractère quantitatif défini sur une population mère Ω. C est la réalisation d'une variable aléatoire X définie sur Ω par $X(\omega_i) = x[i]$. On appelle n-échantillon de valeurs de X la liste des valeurs $(x_1, \cdots, x_n)$ observées prises par X sur un échantillon $\{\omega_1, \cdots, \omega_n\}$ de la population Ω.

Échantillon

Définition : n-échantillon de valeurs de X

Soit C un caractère quantitatif défini sur une population mère Ω. C est la réalisation d'une variable aléatoire X définie sur Ω par $X(\omega_i) = x[i]$. On appelle n-échantillon de valeurs de X la liste des valeurs $(x_1, \cdots, x_n)$ observées prises par X sur un échantillon $\{\omega_1, \cdots, \omega_n\}$ de la population Ω.

Remarque

Les coordonnées peuvent être considérées comme les valeurs des réalisations d'un vecteur aléatoire $(X_1, \cdots, X_n)$ appelé n-échantillon de X où les variables aléatoires réelles X_i sont indépendantes et identiquement distribuées (de même loi).

Plan

Statistique - 1

Définition

On appelle statistique toute variable aléatoire qui s'écrit à l'aide des variables aléatoires $X_1, \cdots, X_n$.

Statistique - 1

Définition

On appelle statistique toute variable aléatoire qui s'écrit à l'aide des variables aléatoires $X_1, \cdots, X_n$.

Remarque

Une statistique est donc une variable aléatoire qui est mesurable par rapport à la tribu engendrée par les n variables aléatoires $X_1, \cdots, X_n$.

Statistique - 2

Remarque

En d'autres termes, une statistique est une variable aléatoire de la forme $S := \varphi(X_1, \cdots, X_n)$ où φ est une application mesurable de $\mathbb{R}^n$ dans $\mathbb{R}$.

Statistique - 2

Remarque

En d'autres termes, une statistique est une variable aléatoire de la forme $S := \varphi(X_1, \cdots, X_n)$ où φ est une application mesurable de $\mathbb{R}^n$ dans $\mathbb{R}$.

Cette notion est intuitive. En effet, on part du principe que l'on ne dispose d'aucune autre information que celles obtenues par les observations de l'échantillon.

Statistique - 3

Exemple

La variable aléatoire

$$\overline{X_n} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

est une statistique.

Statistique - 3

Exemple

La variable aléatoire

$$\overline{X_n} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

est une statistique.

Exemple

La variable aléatoire

$$S_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2$$

est aussi une statistique.

Statistique - 3

Exemple

La variable aléatoire

$$\overline{X_n} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

est une statistique.

Exemple

La variable aléatoire

$$S_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2$$

est aussi une statistique.

Contre-exemple

La variable aléatoire X_{n+1} n'est pas une statistique.

Fluctuations

Si on extrait plusieurs échantillons de taille n fixée, les résultats que l'on va pouvoir déduire sont variables car ils dépendent de l'échantillon considéré. On parle de "fluctuations d'échantillonnage".

Plan

Moyenne d'échantillon - 1

Définition

On définit la variable aléatoire $\overline{X_n}$, appelée moyenne d'échantillon par :

$$\overline{X_n} := \frac{1}{n} \sum_{i=1}^{n} X_i \, .$$

Moyenne d'échantillon - 2

Comme $\overline{X_n}$ est une variable aléatoire réelle, il peut être judicieux d'étudier ses caractéristiques (espérance et variance).

Moyenne d'échantillon - 2

Comme $\overline{X_n}$ est une variable aléatoire réelle, il peut être judicieux d'étudier ses caractéristiques (espérance et variance).

Théorème : Distribution d'échantillonnage de la moyenne

Quelle que soit la loi de la variable aléatoire X d'espérance μ et de variance σ^2 avec $\sigma > 0$ et quel que soit n, on a

$$\mathbb{E}\left[\overline{X_n}\right] = \mu = \mathbb{E}[X]\,,$$

ainsi que

$$\mathrm{Var}\left[\overline{X_n}\right] = \frac{\sigma^2}{n} = \frac{\mathrm{Var}[X]}{n}\,.$$

Moyenne d'échantillon - 3

Preuve

La linéarité de l'espérance nous donne

$$\begin{aligned}
\mathbb{E}\left[\overline{X_n}\right] &= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X] \\
&= \frac{1}{n} n\mathbb{E}[X] \\
&= \mathbb{E}[X]\,,
\end{aligned}$$

ce qui achève la première partie de la preuve.

Moyenne d'échantillon - 4

On calcule désormais la variance de la moyenne d'échantillon. Par définition de la variance, il vient

$$\begin{aligned}
\mathrm{Var}\left[\overline{X_n}\right] &= \mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\} \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X\right] \\
&= \frac{1}{n^2} n\mathrm{Var}\left[X\right] \\
&= \frac{\mathrm{Var}[X]}{n}\,.
\end{aligned}$$

Moyenne d'échantillon et convergence

Proposition

La variable aléatoire $\overline{X_n}$ converge en probabilité vers $\mu = \mathbb{E}[X]$.

Moyenne d'échantillon et convergence

Proposition

La variable aléatoire $\overline{X_n}$ converge en probabilité vers $\mu = \mathbb{E}[X]$.

Preuve

Par définition, la convergence en probabilité de $\overline{X_n}$ vers μ signifie que pour tout $\epsilon > 0$, on a

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X_n} - \mu\right| > \epsilon\right) = 0\,.$$

On utilise l'inégalité de Bienaymé-Tchebychev :

$$\mathbb{P}\left(\left|\overline{X_n} - \mu\right| > \epsilon\right) \leq \frac{\mathrm{Var}\left[\overline{X_n}\right]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}\frac{1}{n} \longrightarrow 0\,.$$

La preuve est ainsi achevée.

Cas normal

Théorème

Si la variable aléatoire X suit la loi normale $\mathcal{N}\left(\mu, \sigma^2\right)$ avec $\sigma > 0$, alors la moyenne empirique $\overline{X_n}$ suit la loi normale $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Puis, la variable aléatoire $\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}}$ suit la loi normale centrée réduite.

Plan

Variance d'échantillon - 1

Définition : Variance d'échantillon

On définit la variable aléatoire S_n^2, appelée variance d'échantillon par :

$$S_n^2 := \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X_n} \right)^2 .$$

Variance d'échantillon - 1

Définition : Variance d'échantillon

On définit la variable aléatoire S_n^2, appelée variance d'échantillon par :

$$S_n^2 := \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X_n} \right)^2 .$$

Comme S_n^2 est une variable aléatoire réelle, il peut être judicieux d'étudier ses caractéristiques (espérance et variance).

Variance d'échantillon - 1

Définition : Variance d'échantillon

On définit la variable aléatoire S_n^2, appelée variance d'échantillon par :

$$S_n^2 := \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X_n} \right)^2 .$$

Comme S_n^2 est une variable aléatoire réelle, il peut être judicieux d'étudier ses caractéristiques (espérance et variance).

ATTENTION

Ne pas confondre la variance d'échantillon S_n^2 avec la variance de la moyenne d'échantillon $\mathrm{Var}\left[\overline{X_n}\right]$.

Distribution d'échantillonnage de la variance - 1

Théorème

Quelle que soit la loi de la variable aléatoire X d'espérance μ et de variance σ^2 avec $\sigma > 0$ et quel que soit n, on a

$$\mathbb{E}\left[S_n^2\right] = \frac{n-1}{n}\sigma^2 = \frac{n-1}{n}\mathrm{Var}[X]\,.$$

Distribution d'échantillonnage de la variance - 2

Preuve

$$
\begin{aligned}
S_n^2 :=& \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{(X_i - \mu) - \left(\overline{X_n} - \mu\right)\right\}^2 \\
=& \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\overline{X_n} - \mu\right)^2 - \frac{2}{n}\left(\overline{X_n} - \mu\right)\sum_{i=1}^{n}(X_i - \mu) \\
=& \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + \left(\overline{X_n} - \mu\right)^2 - 2\left(\overline{X_n} - \mu\right)^2 \\
=& \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \left(\overline{X_n} - \mu\right)^2 .
\end{aligned}
$$

Distribution d'échantillonnage de la variance - 3

$$\mathbb{E}\left[S_n^2\right] = \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \left(\overline{X_n}-\mu\right)^2\right\}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{(X_i-\mu)^2\right\} - \mathbb{E}\left\{\left(\overline{X_n}-\mu\right)^2\right\}.$$

Le second terme est la variance de $\overline{X_n}$. Conséquemment, il vient

$$\mathbb{E}\left[S_n^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[X_i] - \mathrm{Var}\left[\overline{X_n}\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[X] - \frac{\sigma^2}{n} = \mathrm{Var}[X] - \frac{\sigma^2}{n}$$
$$= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Variance de la variance d'échantillon - 1

Proposition

Quelle que soit la loi de la variable aléatoire X d'espérance μ, de variance σ^2 avec $\sigma > 0$ et de moment centré d'ordre quatre égal à μ_4 (c'est-à-dire tel que $\mathbb{E}\left[(X-\mu)^4\right] = \mu_4$) et quel que soit n, on a

$$\mathrm{Var}\left[S_n^2\right] = \frac{n-1}{n^3}\left((n-1)\mu_4 - (n-3)\sigma^4\right) .$$

Variance de la variance d'échantillon - 1

Proposition

Quelle que soit la loi de la variable aléatoire X d'espérance μ, de variance σ^2 avec $\sigma > 0$ et de moment centré d'ordre quatre égal à μ_4 (c'est-à-dire tel que $\mathbb{E}\left[(X-\mu)^4\right] = \mu_4$) et quel que soit n, on a

$$\mathrm{Var}\left[S_n^2\right] = \frac{n-1}{n^3}\left((n-1)\mu_4 - (n-3)\sigma^4\right).$$

Exercice

Démontrer la Proposition.

Variance de la variance d'échantillon - 2

Remarque

On dispose de l'équivalence $\mathrm{Var}\left[S_n^2\right] \sim \sigma^4 \frac{\beta_2 - 1}{n}$ pour n grand. Conséquemment, l'espérance de S_n^2 tend vers la variance de X et la variance de S_n^2 tend vers 0. On peut ainsi montrer la convergence en probabilité de S_n^2 vers $\sigma^2 = \mathrm{Var}[X]$.

Variance d'échantillon corrigée - 1

On peut noter que la moyenne de la variance d'échantillon n'est pas exactement égale à la variance de X. C'est pourquoi on introduit la variance corrigée (ou modifiée) où l'on ne divise pas par le nombre de termes de la somme, mais par le nombre de termes indépendants, à savoir $n-1$. En effet, les termes sont liés par la relation $\sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)=0$.

Variance d'échantillon corrigée - 1

On peut noter que la moyenne de la variance d'échantillon n'est pas exactement égale à la variance de X. C'est pourquoi on introduit la variance corrigée (ou modifiée) où l'on ne divise pas par le nombre de termes de la somme, mais par le nombre de termes indépendants, à savoir $n-1$. En effet, les termes sont liés par la relation $\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right) = 0$.

Définition

On définit la variable aléatoire $\widetilde{S_n^2}$ pour $n \geq 2$, appelée variance d'échantillon corrigée, par

$$\widetilde{S_n^2} := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2 .$$

Variance d'échantillon corrigée - 2

Proposition

Quelle que soit la loi de la variable aléatoire X d'espérance μ et de variance σ^2 avec $\sigma \in]0; +\infty[$ et quel que soit n, on a

$$\mathbb{E}\left[\widetilde{S_n^2}\right] = \sigma^2 = \mathrm{Var}[X]\,.$$

Variance d'échantillon corrigée - 2

Proposition

Quelle que soit la loi de la variable aléatoire X d'espérance μ et de variance σ^2 avec $\sigma \in]0; +\infty[$ et quel que soit n, on a

$$\mathbb{E}\left[\widetilde{S_n^2}\right] = \sigma^2 = \mathrm{Var}[X]\,.$$

On dira par la suite que $\widetilde{S_n^2}$ est un estimateur sans biais tandis que S_n^2 est un estimateur asymptotiquement sans biais.

Cas normal

Théorème

Si la variable aléatoire X suit la loi normale $\mathcal{N}(\mu, \sigma^2)$ avec $\sigma > 0$, alors la variable aléatoire $(n-1)\frac{\widetilde{S}_n^2}{\sigma^2}$ suit la loi du Khi-deux à $n-1$ degrés de liberté, $\chi^2(n-1)$.

Cas normal

Théorème

Si la variable aléatoire X suit la loi normale $\mathcal{N}(\mu, \sigma^2)$ avec $\sigma > 0$, alors la variable aléatoire $(n-1)\frac{\widetilde{S}_n^2}{\sigma^2}$ suit la loi du Khi-deux à $n-1$ degrés de liberté, $\chi^2(n-1)$.

Remarque

Le Théorème se prouve en utilisant le théorème de Cochran sur les vecteurs gaussiens.

Plan

Proportion d'échantillon - 1

Il arrive que le caractère C à estimer ne soit pas quantitatif mais qualitatif. Soit p la proportion d'individus présentant le caractère étudié dans la population mère Ω.

Proportion d'échantillon - 1

Il arrive que le caractère C à estimer ne soit pas quantitatif mais qualitatif. Soit p la proportion d'individus présentant le caractère étudié dans la population mère Ω.
La proportion d'individus ayant le caractère C obtenue dans un n-échantillon est la valeur observée d'une variable aléatoire F_n, fréquence d'apparition dans un échantillon de taille n, appelée proportion d'échantillon. On parle aussi de fréquence statistique.

Proportion d'échantillon - 1

Il arrive que le caractère C à estimer ne soit pas quantitatif mais qualitatif. Soit p la proportion d'individus présentant le caractère étudié dans la population mère Ω.
La proportion d'individus ayant le caractère C obtenue dans un n-échantillon est la valeur observée d'une variable aléatoire F_n, fréquence d'apparition dans un échantillon de taille n, appelée proportion d'échantillon. On parle aussi de fréquence statistique. On peut lier la fréquence statistique à la fréquence d'une série statistique simple.

Proportion d'échantillon - 2

Définition

On définit la variable aléatoire F_n, que l'on appelle proportion d'échantillon ou fréquence statistique par

$$F_n := \frac{K_n}{n},$$

où K_n est la variable aléatoire qui compte le nombre d'apparitions du caractère considéré dans un échantillon de taille n.

Loi de la proportion d'échantillon

La variable aléatoire K_n suit la loi binomiale $\mathcal{B}(n, p)$. On en déduit donc

$$\mathbb{E}[K_n] = np\,,$$

ainsi que

$$\mathrm{Var}[K_n] = np(1-p)\,.$$

On en déduit le théorème suivant.

Loi de la proportion d'échantillon

La variable aléatoire K_n suit la loi binomiale $\mathcal{B}(n, p)$. On en déduit donc

$$\mathbb{E}[K_n] = np\,,$$

ainsi que

$$\mathrm{Var}[K_n] = np(1-p)\,.$$

On en déduit le théorème suivant.

Théorème

Pour tout entier strictement positif n, on a

$$\mathbb{E}[F_n] = p\,,$$

ainsi que

$$\mathrm{Var}[F_n] = \frac{p(1-p)}{n}\,.$$

Loi de la proportion d'échantillon

La variable aléatoire K_n suit la loi binomiale $\mathcal{B}(n, p)$. On en déduit donc

$$\mathbb{E}[K_n] = np\,,$$

ainsi que

$$\mathrm{Var}[K_n] = np(1-p)\,.$$

On en déduit le théorème suivant.

Théorème

Pour tout entier strictement positif n, on a

$$\mathbb{E}[F_n] = p\,,$$

ainsi que

$$\mathrm{Var}[F_n] = \frac{p(1-p)}{n}\,.$$

Exercice

Démontrer le théorème.

Plan

Cadre des estimateurs

Dans la suite, nous nous intéressons à l'estimation d'un paramètre $\theta^* \in \Theta \subset \mathbb{R}^d$ d'une loi parente, à partir d'un n-échantillon. Ce paramètre θ^* peut être la moyenne μ, la variance σ^2, une proportion p ou même le couple (μ, σ^2).

Premières définitions

Définition

Un estimateur $\widehat{\theta_n}$ est une statistique permettant d'évaluer un paramètre **inconnu** θ^* relatif à la loi de probabilité parente.

Premières définitions

Définition

Un estimateur $\widehat{\theta_n}$ est une statistique permettant d'évaluer un paramètre **inconnu** θ^* relatif à la loi de probabilité parente.

Le paramètre θ^* étant inconnu, la fonction φ telle que $\widehat{\theta_n} := \varphi(X_1, \cdots, X_n)$ ne doit pas dépendre de θ^*.

Premières définitions

Définition

Un estimateur $\widehat{\theta_n}$ est une statistique permettant d'évaluer un paramètre **inconnu** θ^* relatif à la loi de probabilité parente.

Le paramètre θ^* étant inconnu, la fonction φ telle que $\widehat{\theta_n} := \varphi(X_1, \cdots, X_n)$ ne doit pas dépendre de θ^*.

Définition

On parle d'estimation de θ^* associée à cet estimateur la valeur observée lors de l'expérience, c'est-à-dire la valeur prise par la fonction au point observé $(x[1], \cdots, x[n])$.

Fil rouge - 1

Exemple

Une usine d'embouteillage possède une machine qui remplit des bouteilles de un litre. Cette machine n'est pas très précise et le volume de boisson versé dans une bouteille est toujours supérieur à un litre. On supposera par la suite que le surplus sur une journée suit une loi uniforme entre 0 et θ, θ représentant le nombre maximal de litres que peut verser en trop la machine en une journée.

Fil rouge - 2

On aimerait avoir une estimation de ce nombre. Pour cela, un technicien regarde la machine pendant n jours et note pour chaque $i \in \{1, \cdots, n\}$, le nombre de litres X_i que la machine a versé en trop durant la journée numéro i.

Fil rouge - 2

On aimerait avoir une estimation de ce nombre. Pour cela, un technicien regarde la machine pendant n jours et note pour chaque $i \in \{1, \cdots, n\}$, le nombre de litres X_i que la machine a versé en trop durant la journée numéro i.
On rappelle que si X suit une loi uniforme sur $[0; \theta]$, alors :
$\mathbb{E}[X] = \frac{\theta}{2}$ et $\mathrm{Var}[X] = \frac{\theta^2}{12}$.

Fil rouge - 3

On propose deux stratégies pour estimer θ. La première consiste à considérer l'estimateur

$$\widehat{\theta_1}(n) := 2\overline{X_n}\,.$$

La deuxième consiste à définir $\widehat{\theta_2}(n)$ comme étant la plus grande des variables aléatoires parmi $X_1, \cdots, X_n$.

Fil rouge - 3

On propose deux stratégies pour estimer θ. La première consiste à considérer l'estimateur

$$\widehat{\theta_1}(n) := 2\overline{X_n}\,.$$

La deuxième consiste à définir $\widehat{\theta_2}(n)$ comme étant la plus grande des variables aléatoires parmi $X_1, \cdots, X_n$.
Quel est le meilleur estimateur ?

Biais - 1

Biais - 1

Lorsqu'on évalue les qualités d'un appareil de mesure, on considère essentiellement la justesse (évaluée *a contrario* par l'erreur systématique du réglage).

Biais - 1

Lorsqu'on évalue les qualités d'un appareil de mesure, on considère essentiellement la justesse (évaluée *a contrario* par l'erreur systématique du réglage).
Un estimateur nous fournit, comme un appareil de mesure, une valeur numérique : d'une manière analogue, nous évaluons les qualités d'un estimateur par sa justesse.

Biais - 1

Lorsqu'on évalue les qualités d'un appareil de mesure, on considère essentiellement la justesse (évaluée *a contrario* par l'erreur systématique du réglage).
Un estimateur nous fournit, comme un appareil de mesure, une valeur numérique : d'une manière analogue, nous évaluons les qualités d'un estimateur par sa justesse.
Pour ce faire, on regardera l'espérance de l'estimateur.

Biais - 2

Remarque

La loi de probabilité d'un estimateur $\widehat{\theta_n}$ dépend de la valeur du paramètre θ^*. C'est pourquoi, dans la suite de ce cours, on note $\mathbb{E}_{\theta^*}[\widehat{\theta_n}]$ (respectivement $\mathrm{Var}_{\theta^*}[\widehat{\theta_n}]$) l'espérance (respectivement la variance) de la statistique $\widehat{\theta_n}$.

Biais - 3

Définition

[Biais d'un estimateur] On appelle biais de $\widehat{\theta_n}$ pour θ^* la valeur

$$b_{\theta^*}(\widehat{\theta_n}) := \mathbb{E}_{\theta^*}[\widehat{\theta_n}] - \theta^* .$$

Biais - 3

Définition

[Biais d'un estimateur] On appelle biais de $\widehat{\theta_n}$ pour θ^* la valeur

$$b_{\theta^*}(\widehat{\theta_n}) := \mathbb{E}_{\theta^*}[\widehat{\theta_n}] - \theta^* .$$

Définition

Un estimateur est dit sans biais si $b_{\theta^*}(\widehat{\theta_n}) = 0$ c'est-à-dire si

$$\mathbb{E}_{\theta^*}[\widehat{\theta_n}] = \theta^* .$$

Biais - 3

Définition

[Biais d'un estimateur] On appelle biais de $\widehat{\theta_n}$ pour θ^* la valeur

$$b_{\theta^*}(\widehat{\theta_n}) := \mathbb{E}_{\theta^*}[\widehat{\theta_n}] - \theta^* .$$

Définition

Un estimateur est dit sans biais si $b_{\theta^*}(\widehat{\theta_n}) = 0$ c'est-à-dire si

$$\mathbb{E}_{\theta^*}[\widehat{\theta_n}] = \theta^* .$$

On dit qu'un estimateur est sans biais si en moyenne, on ne fait pas d'erreur systémique.

Biais - 4

Exemple

Si le paramètre à estimer est l'espérance de la loi $\theta^* = \mathbb{E}[X]$, l'estimateur naturel est la moyenne d'échantillon :

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \, .$$

Alors, on a $\mathbb{E}_{\theta^*}[\overline{X}_n] = \mathbb{E}[X] = \theta^*$. Il s'ensuit que $\overline{X}_n$ est un estimateur sans biais de l'espérance.

Biais - 5

Pour certains estimateurs, la propriété précédente ne peut pas être strictement vérifiée, le biais diminuant seulement quand la taille n de l'échantillon augmente. Ceci correspond à la définition suivante.

Biais - 5

Pour certains estimateurs, la propriété précédente ne peut pas être strictement vérifiée, le biais diminuant seulement quand la taille n de l'échantillon augmente. Ceci correspond à la définition suivante.

Définition

On dit que $\widehat{\theta_n}$ est un estimateur asymptotiquement sans biais de θ^* si pour tout $\theta \in \Theta$, on a $\lim_{n\to\infty} b_\theta[\widehat{\theta_n}] = 0$ c'est-à-dire si

$$\lim_{n\to\infty} \mathbb{E}_\theta[\widehat{\theta_n}] = \theta\,.$$

Biais - 6

Exemple

Si le paramètre à estimer est la variance de la loi, $\theta^* = \mathrm{Var}[X]$, l'estimateur naturel est la variance d'échantillon :

$$S_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 .$$

Or, $\mathbb{E}_{\theta^*}[S_n^2] = \frac{n-1}{n}\mathrm{Var}[X] = \frac{n-1}{n}\theta^*$. Il s'ensuit que S_n^2 n'est pas un estimateur sans biais de la variance. Mais, c'est un estimateur asymptotiquement sans biais.

Biais - 7

D'autre part, un autre estimateur de la variance est la variance d'échantillon corrigée

$$\widetilde{S_n^2} := \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n \right)^2 .$$

Puis, l'on rappelle que $\mathbb{E}_{\theta^*} \left[\widetilde{S_n^2} \right] = \mathrm{Var}[X] = \theta^*$. Il s'ensuit que la variance d'échantillon corrigée, $\widetilde{S_n^2}$, est un estimateur sans biais de la variance.

Convergence - 1

Définition

[Estimateur convergent] $\widehat{\theta_n}$ est un estimateur convergent si la suite de variables aléatoires $(\widehat{\theta_n})_n$ converge en probabilité vers θ^*.

Convergence - 1

Définition

[Estimateur convergent] $\widehat{\theta_n}$ est un estimateur convergent si la suite de variables aléatoires $(\widehat{\theta_n})_n$ converge en probabilité vers θ^*.

En anglais, on dit "consistent".

Convergence - 2

En d'autres termes, un estimateur est convergent si pour tout $\epsilon > 0$, la suite numérique de terme général

$$\mathbb{P}\left(\left|\widehat{\theta_n} - \theta^*\right| > \epsilon\right)$$

converge vers 0.

Convergence - 3

Remarque

La distribution d'un estimateur convergent tend à se concentrer autour de la valeur θ^* du paramètre à estimer quand la taille de l'échantillon augmente.

Convergence - 3

Remarque

La distribution d'un estimateur convergent tend à se concentrer autour de la valeur θ^* du paramètre à estimer quand la taille de l'échantillon augmente.

Démontrer la convergence en probabilité n'est pas facile. On peut toutefois montrer qu'un estimateur est convergent en utilisant le théorème suivant.

Convergence - 4

Théorème

Un estimateur asymptotiquement sans biais dont la variance tend vers zéro est convergent.

Convergence - 4

Théorème

Un estimateur asymptotiquement sans biais dont la variance tend vers zéro est convergent.

Preuve

Fixons $\epsilon > 0$ arbitrairement petit. Montrons que

$$\mathbb{P}\left(\left|\widehat{\theta_n} - \theta^*\right| > \epsilon\right)$$

tend vers 0 quand n tend vers l'infini.

Convergence - 5

L'inégalité triangulaire nous donne :

$$\left|\widehat{\theta_n} - \mathbb{E}_{\theta^*}[\widehat{\theta_n}]\right| \geq \left|\widehat{\theta_n} - \theta^*\right| - \left|\mathbb{E}_{\theta^*}[\widehat{\theta_n}] - \theta^*\right| .$$

Conséquemment, si $\left|\widehat{\theta_n} - \theta^*\right| > \epsilon$, alors on a $\left|\widehat{\theta_n} - \mathbb{E}_{\theta^*}[\widehat{\theta_n}]\right| > \epsilon - \left|\mathbb{E}_{\theta^*}[\widehat{\theta_n}] - \theta^*\right|$. Il s'ensuit l'inégalité

$$\mathbb{P}\left(\left|\widehat{\theta_n} - \theta^*\right| > \epsilon\right) \leq \mathbb{P}\left(\left|\widehat{\theta_n} - \mathbb{E}_{\theta^*}[\widehat{\theta_n}]\right| > \epsilon - \left|\mathbb{E}_{\theta^*}[\widehat{\theta_n}] - \theta^*\right|\right) .$$

Or, l'estimateur est asymptotiquement sans biais donc la quantité $\left|\mathbb{E}_{\theta^*}[\widehat{\theta_n}] - \theta^*\right|$ est plus petite que $\frac{\epsilon}{2}$ si n est assez grand. Donc, pour n assez grand, il vient

$$\mathbb{P}\left(\left|\widehat{\theta_n} - \theta^*\right| > \epsilon\right) \leq \mathbb{P}\left(\left|\widehat{\theta_n} - \mathbb{E}_{\theta^*}[\widehat{\theta_n}]\right| > \frac{\epsilon}{2}\right) .$$

On utilise ensuite l'inégalité de Bienaymé-Tchebychev :

$$\mathbb{P}\left(\left|\widehat{\theta_n} - \theta^*\right| > \epsilon\right) \leq \frac{4}{\epsilon^2}\mathbb{E}_{\theta^*}\left\{\left(\widehat{\theta_n} - \mathbb{E}_{\theta^*}[\widehat{\theta_n}]\right)^2\right\} = \frac{4}{\epsilon^2}\mathrm{Var}_{\theta^*}[\widehat{\theta_n}] ,$$

ce qui converge vers 0 quand n tend vers l'infini. La preuve est ainsi achevée.

Exemple

Si le paramètre à estimer est l'espérance de la loi, $\theta = \mathbb{E}[X]$, ce paramètre est estimé sans biais par $\overline{X}_n$. Or, on sait que l'on a

$$\mathrm{Var}_\theta[\overline{X}_n] = \frac{\mathrm{Var}[X]}{n} \longrightarrow 0\,.$$

Il s'ensuit que la moyenne d'échantillon est un estimateur convergent de l'espérance.

Exemple

Si le paramètre à estimer est l'espérance de la loi, $\theta = \mathbb{E}[X]$, ce paramètre est estimé sans biais par $\overline{X}_n$. Or, on sait que l'on a

$$\mathrm{Var}_\theta[\overline{X}_n] = \frac{\mathrm{Var}[X]}{n} \longrightarrow 0\,.$$

Il s'ensuit que la moyenne d'échantillon est un estimateur convergent de l'espérance.

De la même manière, on sait, sous certaines hypothèses, que la variance de S_n^2 tend vers 0 quand n tend vers l'infini. Par conséquent, S_n^2 (ainsi que $\widetilde{S_n^2}$) est un estimateur convergent de la variance.

Convergence - 8

Définition

Un estimateur $\widehat{\theta_n}$ est dit absolument correct si $\widehat{\theta_n}$ est sans biais et si sa variance tend vers 0 quand n tend vers l'infini.

Convergence - 8

Définition

Un estimateur $\widehat{\theta_n}$ est dit absolument correct si $\widehat{\theta_n}$ est sans biais et si sa variance tend vers 0 quand n tend vers l'infini.

Exemple

La moyenne d'échantillon est un estimateur absolument correct.

Normalité asymptotique - 1

En vu des intervalles de confiance, il est parfois crucial que la loi de l'estimateur puisse être approchée par une loi normale quand n est grand.

Normalité asymptotique - 2

Définition

Un estimateur réel $\widehat{\theta_n}$ d'un paramètre réel θ^* est dit asymptotiquement normal s'il existe deux suites de réels $(m_n(\theta^*))_{n\in\mathbb{N}^*}$ et $(v_n(\theta^*))_{n\in\mathbb{N}^*}$ telles que $v_n(\theta^*) > 0$ pour tout $n \geq 1$ et

$$\frac{\widehat{\theta_n} - m_n(\theta^*)}{v_n(\theta^*)} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)\,.$$

Normalité asymptotique - 3

Remarque

En d'autres termes, pour tout $u \in \mathbb{R}$, on a

$$\mathbb{P}_{\theta^*}\left(\frac{\widehat{\theta_n} - m_n(\theta^*)}{v_n(\theta^*)} \leq u\right) \longrightarrow \Phi(u)\,,$$

quand n tend vers l'infini. Ici, $\Phi(u)$ est la fonction de répartition de la loi normale centrée réduite.

Normalité asymptotique - 4

Normalité asymptotique - 4

Pour établir ce genre de résultats, on s'adosse la plupart du temps au théorème central de la limite.

Normalité asymptotique - 4

Pour établir ce genre de résultats, on s'adosse la plupart du temps au théorème central de la limite.

Exemple

La proportion d'échantillon est asymptotiquement normale. Il en est de même avec la moyenne d'échantillon. Enfin, si le moment d'ordre quatre de la variable aléatoire X est fini, la variance d'échantillon est également asymptotiquement normale.

Efficacité - 1

La qualité d'un estimateur se mesure aussi par l'erreur quadratique moyenne.

Efficacité - 1

La qualité d'un estimateur se mesure aussi par l'erreur quadratique moyenne.

Définition

L'erreur quadratique moyenne d'un estimateur $\widehat{\theta_n}$ est définie comme étant $\mathbb{E}_{\theta^*}\left[\left(\widehat{\theta_n} - \theta^*\right)^2\right]$.

Efficacité - 2

Proposition

Soit $\widehat{\theta_n}$ un estimateur du paramètre θ^* à étudier. On a alors

$$\mathbb{E}_{\theta^*}\left[\left(\widehat{\theta_n} - \theta^*\right)^2\right] = \mathrm{Var}_{\theta^*}[\widehat{\theta_n}] + \left(\mathbb{E}_{\theta^*}[\widehat{\theta_n}] - \theta^*\right)^2 .$$

Efficacité - 2

Proposition

Soit $\widehat{\theta_n}$ un estimateur du paramètre θ^* à étudier. On a alors

$$\mathbb{E}_{\theta^*}\left[\left(\widehat{\theta_n} - \theta^*\right)^2\right] = \mathrm{Var}_{\theta^*}[\widehat{\theta_n}] + \left(\mathbb{E}_{\theta^*}[\widehat{\theta_n}] - \theta^*\right)^2 .$$

Exercice

Démontrer la proposition.

Efficacité - 3

Remarque

Entre deux estimateurs sans biais, le meilleur est celui dont la variance est minimale. On parle d'efficacité.

Efficacité - 3

Remarque

Entre deux estimateurs sans biais, le meilleur est celui dont la variance est minimale. On parle d'efficacité.

Remarque

Le critère d'erreur quadratique moyenne n'est pas parfait mais il est préféré à d'autres critères qui semblent plus naturels comme l'erreur absolue moyenne $\mathbb{E}_{\theta^*}\left[\left|\widehat{\theta_n} - \theta^*\right|\right]$ car il s'exprime en fonction de notions simples comme le biais et la variance et est relativement facile à manipuler analytiquement.

Robustesse

Robustesse

La dernière propriété que l'on aime voir satisfaite par un estimateur est la robustesse. Cela signifie que l'ajout d'une nouvelle donnée ne perturbe pas complètement l'estimateur. On pense ici à des données extrêmales et qui peuvent être aberrantes.

Robustesse

La dernière propriété que l'on aime voir satisfaite par un estimateur est la robustesse. Cela signifie que l'ajout d'une nouvelle donnée ne perturbe pas complètement l'estimateur. On pense ici à des données extrêmales et qui peuvent être aberrantes.
Ainsi, la moyenne d'échantillon n'est pas toujours robuste. Il en est de même avec l'estimateur du maximum de vraisemblance que nous verrons par la suite. Par conséquent, de nombreux statisticiens travaillent à fournir une version plus robuste de ce dernier.