14 Théorèmes de convergence en probabilité Statistiques inférentielles

14.1 Estimation

Définition 14.1

Soit X une variable aléatoire admettant une espérance μ et une variance σ^2.

Pour tout entier $n \in \mathbb{N}^*$, on appelle n-échantillon de X

On appelle alors *estimateur* d'un paramètre θ de X

Le but d'un estimateur est de pouvoir, à partir d'une suite d'observations empiriques d'un phénomène aléatoire, calculer une approximation du paramètre recherché.

Exemples (à connaître) : Soit X une variable aléatoire admettant une espérance μ et une variance σ^2.

Les deux estimateurs suivants sont à connaître :

- la *moyenne empirique*, notée M_n ou $\overline{X_n}$, définie par

$$M_n =$$

  est un estimateur de

- la *variance empirique*, notée S_n^2, définie par

$$S_n^2 =$$

  est un estimateur de

†. On parlera souvent de suite de variables *i.i.d.*, pour « indépendantes et identiquement distribuées ».

14.2 Convergence en loi

Les suites de variables aléatoires peuvent converger sous différents sens. On étudie ici le cas de la convergence *en loi*.

Définition 14.2

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires, et soit X une variable aléatoire.

Pour tout n, on note F_n la fonction de répartition de X_n, et on note F la fonction de répartition de X.

On dit alors que la suite $(X_n)_{n\in\mathbb{N}}$ converge en loi vers X si pour tout réel t où F est continue, la suite $(F_n(t))_{n\in\mathbb{N}}$ converge vers $F(t)$.

On note alors $X_n \xrightarrow{\mathcal{L}} X$.

Exemple : Soient $U_n \hookrightarrow \mathcal{U}([0,1])$ mutuellement indépendantes. On pose

$$X_n = n(1 - \max(U_1, \ldots, U_n)).$$

On commence par chercher la fonction de répartition de X_n : soit $x \in \mathbb{R}$.

$$\begin{aligned} P(X_n \leqslant x) &= \\ &= \\ &= \\ &= \end{aligned}$$

Or on sait que $(1 - \frac{x}{n})^n \xrightarrow[n\to\infty]{}$, et donc, pour tous n et x

$$F_{X_n}(x) \xrightarrow[n\to\infty]{}$$

On reconnaît alors la fonction de répartition d'une loi

On peut donc dire, en posant $X \hookrightarrow$, que

$$X_n \xrightarrow{\mathcal{L}} X.$$

Proposition 14.3

Soient (X_n) et X des variables aléatoires à valeurs dans $\mathbb{Z}$. Alors

$X_n \xrightarrow{\mathcal{L}} X \quad \Leftrightarrow$

Proposition 14.4 – Approximation de loi binomiale par la loi de Poisson

Soit (X_n) une suite de variables aléatoires de lois respectives $\mathcal{B}(n, p_n)$, avec $np_n \to \lambda > 0$.

Alors $X_n \xrightarrow{\mathcal{L}} X$ où $X \hookrightarrow$.

Démonstration. Les lois binomiales et de Poisson sont à valeurs entières, et on applique donc le résultat précédent.

On a donc

$$P(X_n = k) =$$

On note que la condition $np_n \to \lambda$ est équivalente à dire que $p_n \sim$, et donc

$$\left(\frac{p_n}{1-p_n}\right)^k \sim$$

De plus, on a

$$(1-p_n)^n = e^{n\ln(1-p_n)},$$

avec $n\ln(1-p_n) \sim$. Donc $(1-p_n)^n \to$. On peut aussi montrer facilement que $\binom{n}{k} \sim$.

On a donc

$$P(X_n = k) \sim$$

et on reconnaît la loi d'une variable de Poisson de paramètre λ. □

Dans la pratique, pour n assez grand et certaines conditions, on utilise le résultat d'approximation suivant :

Corollaire 14.5

Si $p \leqslant 0,1$, $n \geqslant 30$ et $np \leqslant 15$, on peut approximer la loi binomiale $\mathcal{B}(n,p)$ par

14.3 Théorèmes limites

Commençons par rappeler que si X est une variable aléatoire admettant une espérance μ et une variance $\sigma^2 \neq 0$, alors on note X^* la variable aléatoire $\frac{X-\mu}{\sigma}$. Cette variable aléatoire est alors centrée réduite.

On notera toujours $M_n = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ la moyenne empirique, où les variables X_i sont indépendantes et de même loi que X.

On a alors

$$\mathbb{E}(M_n) = \qquad \text{et} \quad \mathbb{V}(M_n) =$$

La variable centrée réduite associée à M_n est donc M_n^* définie par :

$$M_n^* =$$

On rappelle alors l'inégalité de Bienaymé-Tchebychev :

Théorème 14.6 – Inégalité de Bienaymé-Tchebytchev

Soit X une variable aléatoire admettant une variance. Alors

$$\forall \varepsilon > 0,$$

Exemple : On souhaite estimer la valeur de $\frac{\pi}{4}$ à $0,1$ près, avec un risque d'erreur inférieur à 1%. Pour cela, on tire au hasard deux valeurs x et y uniformément dans $[0,1]$. On note que la probabilité que le point de coordonnées (x,y) soit dans le quart de cercle de centre 0 et de rayon 1 est $\frac{\pi}{4}$.

On itère l'expérience : on note X_n la variable aléatoire qui vaut 1 si le n-ième point tiré est dans le cercle, 0 sinon. On note de plus M_n la moyenne empirique des X_n.

On cherche donc n tel que

En notant que pour tout $x \in [0,1]$, $x(1-x) \leqslant \frac{1}{4}$, l'inégalité de Bienaymé-Tchebytchev nous donne

Il suffit donc de prendre

On se fixe alors dans la suite une variable aléatoire X admettant une espérance μ et une variance $\sigma^2 \neq 0$.

On en déduit alors directement la loi faible des grands nombres :

Théorème 14.7 – Loi faible des grands nombres

Soit (X_n) une suite de variables aléatoires indépendantes de même loi que X. Alors

$$\forall \varepsilon > 0,$$

On peut alors raffiner le théorème précédent pour obtenir le célèbre théorème central limite.

Théorème 14.8 – central limite

Soit (X_n) une suite de variables aléatoires indépendantes toutes de même loi que X.

Alors la suite (M_n^*) converge en loi vers une loi normale centrée réduite.

On a donc pour tous réels $a < b$,

Nota : Le résultat reste vrai en faisant tendre a ou b vers $\pm\infty$. Par exemple,

Ce théorème nous dit que, en un certain sens, on peut approximer la quantité M_n par une loi normale de paramètres μ et $\frac{\sigma^2}{n}$.

Un inconvénient de ce théorème est qu'il est nécessaire de connaître l'écart-type du phénomène observé. Dans le cas contraire, on peut le remplacer par l'écart-type empirique.

Avec les mêmes hypothèses, posons

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$
$$S_n^2 = \frac{1}{n}\sum_{k=1}^{n} (X_k - M_n)^2$$

On a alors

Théorème 14.9 – central limite

Dans le contexte précédent, la suite $\left(\frac{M_n-\mu}{\frac{S_n}{\sqrt{n}}}\right)$ converge en loi vers une loi normale centrée réduite.

On a donc

Dans le cas particulier de la loi binomiale, on obtient le théorème suivant :

Théorème 14.10 – Moivre-Laplace

Soit $p \in]0,1[$, et soit (X_n) une suite de variables indépendantes telles que $X_n \hookrightarrow \mathcal{B}(n,p)$.

Alors pour tous réels $a < b$,

$$\mathbb{P}\left(a < \frac{X_n - np}{\sqrt{np(1-p)}} < b\right) \xrightarrow[n\to\infty]{}$$

Démonstration. On peut écrire X_n comme la somme de n variables de Bernoulli identiques et indépendantes, de paramètre p : $X_n = \sum\limits_{k=1}^{n} B_k$.

En notant $\overline{B_n} = \frac{1}{n}\sum\limits_{k=1}^{n} B_k$, on a donc $\overline{B_n}^* = \frac{\overline{B_n}-p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{X_n - np}{\sqrt{np(1-p)}}$.

Le théorème central limite appliqué à la suite (B_n) permet d'affirmer que la suite $(\overline{B_n}^*)$ converge en loi vers une loi normale centrée réduite. $\square$

En pratique, on approche la loi binomiale $\mathcal{B}(n,p)$ par une loi normale $\mathcal{N}(np, \sqrt{npq}^2)$ si $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.

14.4 Test de conformité de la moyenne

En statistique, un *test d'hypothèse* est une méthode qui nous permettra de choisir s'il est raisonnable d'accepter une hypothèse H_0.

L'hypothèse H_0, appelée *hypothèse nulle* est alors faite ; tous les résultats qui suivent supposent donc que cette hypothèse est vraie. L'hypothèse contraire à H_0 est appelée *hypothèse alternative*, notée H_1.

Le test sera alors une méthode qui nous permettra d'accepter ou de rejeter H_0, en fonction d'observations empiriques. Il est cependant statistiquement impossible d'être sûr d'avoir eu raison dans notre conclusion. On distingue alors deux types d'erreurs :

- le risque de première espèce, qui est le risque de rejeter l'hypothèse H_0 sachant qu'elle est vraie ; ce risque est choisi à l'avance, souvent 5% ou 1%
- le risque de seconde espèce, qui est le risque d'accepter l'hypothèse sachant qu'elle est fausse ; il est plus difficile à calculer, et son étude dépasse le cadre de ce cours.

On peut alors utiliser le test suivant pour la valeur de la moyenne :

Proposition 14.11

Soit X une variable aléatoire admettant une espérance μ et une variance $\sigma^2 \neq 0$. Soit $\mu_0 \in \mathbb{R}$.

On suppose $H_0 : \mu = \mu_0$, et on définit l'hypothèse alternative $H_1 : \mu \neq \mu_0$.

Alors

où $u_{1-\frac{\alpha}{2}}$ est l'unique réel tel que $\Phi\left(u_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$.

En pratique, pour faire un test d'hypothèse de conformité de la moyenne :

-
-

EXEMPLE : Un médicament doit contenir $\mu_0 = 2,5$g de substance active.

En prenant 100 comprimés au hasard, on trouve une moyenne de $M_n = 2,6$g de substance active, avec une variance $S_n^2 = 0.16$.

On utilise un test de conformité à la moyenne, avec un risque de 5%.

On choisit alors $u_{1-\frac{\alpha}{2}} = 1.96$, et on a alors

On peut donc notre hypothèse.

14.5 Exercices

Exercice 1

Soit (X_n) une suite de variables aléatoires identiques et indépendantes, admettant une espérance μ et une variance σ^2.

1. Montrer que $\mathbb{E}(S_n^2) \neq \sigma^2$.

2. On appelle alors *estimateur corrigé de la variance* la suite

$$S_n'^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M_n)^2.$$

   Montrer que $\mathbb{E}(S_n'^2) = \sigma^2$.

Exercice 2

Soit f la fonction définie sur $\mathbb{R}$ par $f(x) = e^{-x-e^{-x}}$.

1. Vérifier que f est une fonction de densité.

On dit qu'une variable admettant f pour densité suit la loi de Gumbel.

2. Soit (X_n) une suite de variables aléatoires indépendantes, suivant toutes la loi exponentielle de paramètre 1.

   On pose pour tout n : $M_n = \max(X_1, \ldots, X_n)$.

   Calculer la fonction de répartition F_n de M_n.

3. Montrer que la suite $(M_n - \ln(n))$ converge en loi vers une loi de Gumbel.

Exercice 3

Soit $\lambda > 0$. Pour tout entier $n \geqslant \lambda$, on définit une suite $(X_{n,i})_{i \in \mathbb{N}^*}$ de variables aléatoires indépendantes suivant toutes une loi de Bernoulli de paramètre $p_n = \frac{\lambda}{n}$.

On définit alors pour tout n :

$$N_n = \frac{1}{n} \min\{i \in \mathbb{N}^* \mid X_{n,i} = 1\}.$$

1. Déterminer la loi de la variable nN_n, et en déduire la fonction de répartition de N_n.

2. Montrer que la suite (N_n) converge en loi vers une variable exponentielle de paramètre λ.

Exercice 4

Une compagnie aérienne, disposant de 300 places dans un avion, souhaite optimiser au mieux son remplissage. On estime qu'une personne ayant réservé oublie de se présenter à l'aéroport avec une probabilité de 10%. Les choix des passagers de venir ou non sont supposés indépendants.

Pour un entier n représentant le nombre de réservations faites, on note S_n le nombre de personnes qui se présentent à l'aéroport.

1. Calculer $\mathbb{E}(S_n)$ et $\mathbb{V}(S_n)$.
2. Déterminer le nombre maximal de réservations que peut proposer la compagnie pour que le risque de ne pas pouvoir embarquer tout le monde soit inférieur à 5%.

Exercice 5

Montrer que

$$\lim_{n\to\infty} e^{-n}\sum_{k=0}^{n}\frac{n^k}{k!}=\frac{1}{2}.$$

On pourra utiliser des variables aléatoires X_i indépendantes, suivant toutes une loi de Poisson $\mathcal{P}(1)$.

Exercice 6

On pourra utiliser pour les programmes `Python` *la fonction* `linalg.matrix_rank()` *du module* `numpy`*, qui permet de déterminer le rang d'une matrice, comme le montre l'exemple suivant :*

```
import numpy as np
A = np. array ( [ [1 ,2 ,1] , [2 ,3 ,2], [3 ,5 ,3]] )
print ( np. linalg . matrix_rank (A) )

```

La dernière ligne affiche le rang de la matrice $\begin{pmatrix}1&2&1\\2&3&2\\3&5&3\end{pmatrix}$*, c'est-à-dire* 2.

On pourra aussi utiliser la fonction `randint()` *du module* `random`*. Pour a et b deux entiers* `randint(a,b)` *retourne un entier équiprobablement entre a et b (a et b étant inclus).*

On considère la matrice :

$$A=\begin{pmatrix}3&1&1\\1&0&1\\-3&0&-1\end{pmatrix}.$$

1. a) Écrire une fonction Python prenant en arguments deux vecteurs de taille 3 et renvoyant un booléen (`True` ou `False`) indiquant s'ils sont colinéaires. *(On pourra représenter les vecteurs par des listes).*

   b) Écrire une fonction Python `vecteurs_propres(u)` prenant en argument un vecteur de taille 3 et renvoyant un booléen (`True` ou `False`) indiquant s'il est un vecteur propre de A.

2. a) Vérifier que $-1, 1, 2$ sont valeurs propres de A et préciser pour chacune un vecteur propre associé.

   b) La matrice A est-elle diagonalisable ?

3. Soient $X_1, \cdots, X_n$, n variables aléatoires indépendantes suivant la loi de Bernoulli de paramètre $p \in]0;1[$. On note : $M_n = \frac{1}{n}\sum\limits_{k=1}^{n} X_k$ et $M_n^* = \frac{M_n - p}{\sqrt{\frac{p(1-p)}{n}}}$.

   a) Donner, pour $\alpha \in \mathbb{R}_+^*$, l'approximation de la probabilité $P([-\alpha < M_n^* < \alpha])$ donnée par le théorème central limite.

   b) En déduire que $\mathbb{P}\left(p \in \left[M_n - \frac{1}{\sqrt{n}}, M_n + \frac{1}{\sqrt{n}}\right]\right) \geqslant 0.95$.

   On pourra admettre que, $\forall x \in [0;1]$, $x(1-x) \leq \frac{1}{4}$ et si Φ désigne la fonction de répartition d'une variable suivant une loi normale centrée réduite, alors $\Phi(1;96) \approx 0,975$.

4. On note N_V le nombre de vecteurs propres de A dont les cœfficients sont des entiers de $[\![-5, 5]\!]$.

   a) Expliquer comment le programme suivant permet d'estimer la valeur de N_V :

```python
def simul ():
    u = [ randint ( -5 ,5) for k in range (3) ]
    return vecteurs_propres (u)
n = 10000       # Valeur de n a definir.
nb = 0
for k in range (n):
    if simul ():
        nb += 1
print ( round (nb/n *11**3)) # round (x) = l'entier le plus proche de x.

```

   b) Comment choisir n pour que l'on soit sûr à 95% de la valeur affichée ?

   c) Commenter le résultat obtenu.