1 Convergence en loi

Définition 1

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles définies sur le même espace probabilisé et X une variable aléatoire définie sur le même espace. Pour tout $n \in \mathbb{N}$, on note F_n (respectivement F) la fonction de répartition de X_n (respectivement X). On dit que $(X_n)_{n\in\mathbb{N}}$ **converge en loi** vers X, et on note $X_n \xrightarrow{\mathcal{L}} X$, si, pour tout réel x **point de continuité** de F, $\lim\limits_{n\to+\infty} F_n(x) = F(x)$.

Proposition 2 (*Caractérisation de la convergence en loi pour les variables à valeurs entières*)

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles à valeurs entières (i.e. $\forall n \in \mathbb{N},\ X_n(\Omega) \subset \mathbb{N}$).
Soit X une variable aléatoire à valeurs entières définie sur le même espace probabilisé.
La suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ converge en loi vers X, si, et seulement si :

$$\forall x \in \mathbb{N},\ \lim_{n\to+\infty} \mathbb{P}(X_n = x) = \mathbb{P}(X = x).$$

Démonstration.

Théorème 3

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires telles que, pour tout $n \in \mathbb{N}^*$, $X_n \hookrightarrow \mathcal{B}(n, p_n)$.
Si $\lim\limits_{n\to+\infty} np_n = \lambda \in \mathbb{R}_+^*$, la suite $(X_n)_{n\in\mathbb{N}^*}$ converge en loi vers une variable aléatoire X suivant la loi de $\mathcal{P}(\lambda)$:

$$\forall k \in \mathbb{N},\ \lim_{n\to+\infty} \mathbb{P}(X_n = k) = \frac{\lambda^k}{k!}e^{-\lambda} \quad \Big(= \mathbb{P}(X = k) \text{ où } X \hookrightarrow \mathcal{P}(\lambda)\Big).$$

Démonstration.

Remarque 4 (Approximation d'une loi binomiale par une loi de Poisson)

Le résultat précédent montre que, pour n suffisamment grand et p suffisamment petit, on peut approcher la loi binomiale $\mathcal{B}(n,p)$ par la loi de Poisson $\mathcal{P}(\lambda)$ où $\lambda = np$. *En pratique, on considère que cette approximation est satisfaisante dès que $\underline{n \geqslant 30 \text{ et } p \leqslant 0,1}$ (ces conditions ne sont pas à connaître).*

2 Théorèmes limites

2.1 Inégalités de concentration

Théorème 5 (*Inégalité de Markov*)

Soit X une variable aléatoire réelle **positive** admettant une espérance. Alors :

$$\forall a > 0,\ \mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}(X)}{a}.$$

Théorème 6 (*Inégalité de Bienaymé-Tchebytchev*)

Soit X une variable aléatoire réelle admettant un moment d'ordre 2. Alors :

$$\forall \varepsilon > 0,\ \mathbb{P}(|X - \mathbb{E}(X)| \geqslant \varepsilon) \leqslant \frac{\mathbb{V}(X)}{\varepsilon^2}.$$

Remarque 7

Ce résultat rappelle que la variance permet de quantifier la dispersion d'une variable aléatoire autour de sa moyenne : la probabilité que X prenne des valeurs éloignées de $\mathbb{E}(X)$ d'au moins ε est d'autant plus faible que $\mathbb{V}(X)$ l'est.

Exemple 8

Soient r un réel strictement plus grand que 1 et X une variable aléatoire suivant la loi géométrique $\mathcal{G}\left(\frac{1}{r}\right)$.

1. Montrer que $\mathbb{P}(X \geqslant r^2) \leqslant \frac{1}{r}$.
2. Montrer que $\mathbb{P}(|X - r| \geqslant r) \leqslant 1 - \frac{1}{r}$. En déduire que $\mathbb{P}(X \geqslant 2r) \leqslant 1 - \frac{1}{r}$.

2.2 Loi faible des grands nombres

Définition 9

On appelle **moyenne empirique** du n-uplet de variables aléatoires $(X_1, \dots, X_n)$ la variable $M_n = \dfrac{1}{n} \sum\limits_{k=1}^{n} X_k$.

Proposition 10

Soit $(X_n)_{n \in \mathbb{N}^*}$ une suite de variables aléatoires indépendantes, admettant une même espérance notée μ et une même variance notée σ^2.
Pour tout entier $n \in \mathbb{N}^*$, la moyenne empirique $M_n = \dfrac{X_1 + \cdots + X_n}{n}$ admet une espérance et une variance et :

$$\mathbb{E}(M_n) = \mu, \quad \mathbb{V}(M_n) = \frac{\sigma^2}{n} \quad \text{et} \quad \sigma(M_n) = \frac{\sigma}{\sqrt{n}}.$$

Démonstration.

Théorème 11 (*Loi faible des grands nombres*)

Soit $(X_n)_{n \in \mathbb{N}^*}$ une suite de variables aléatoires indépendantes admettant une même espérance μ et une même variance σ^2. Pour tout $n \in \mathbb{N}^*$, la variable aléatoire $M_n = \dfrac{X_1 + \cdots + X_n}{n}$ vérifie :

$$\forall \varepsilon > 0, \ \lim_{n \to +\infty} \mathbb{P}\left(|M_n - \mu| \geqslant \varepsilon\right) = 0.$$

Démonstration.

Remarque 12

- La loi faible des grands nombres assure qu'il devient improbable que la moyenne empirique M_n s'écarte de l'espérance d'une valeur fixée ε à mesure que n prenne de grandes valeurs.
- La loi faible des grands nombres fournit une approximations de loi dont on ignore l'espérance (lorsqu'elle existe !) : il suffit de choisir un grand nombre de réalisations et d'en calculer la moyenne empirique.

Exemple 13

On lance un dé à 8 faces, numéroté de 1 à 8. On considère la variable aléatoire X définie par :

$$X = \begin{cases} 1 \text{ si on obtient un nombre divisible par 3} \\ -1 \text{ si on obtient un nombre divisible par 4} \\ 0 \text{ sinon.} \end{cases}.$$

1. Écrire une fonction `simule_X` simulant la réalisation de la variable X.
2. Écrire un script calculant une approximation de l'espérance de X à l'aide de 1000 simulation.

2.3 Théorème central limite

Corollaire 14

Avec les notation de la proposition 10 et si $\sigma \neq 0$, la variable aléatoire $M_n^* = \dfrac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ est centrée réduite.

Définition 15

Soit $(X_1, \ldots, X_n)$ un n-uplet de variables aléatoires définies sur le même espace probabilisé, de moyenne empirique M_n. On appelle **variance empirique** de $(X_1, \ldots, X_n)$, la variable aléatoire notée S_n^2 définie par :

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

On appelle **écart-type empirique** de $(X_1, \ldots, X_n)$, la variable aléatoire $S_n = \sqrt{S_n^2}$.

Proposition 16 (*Hors-programme*)

Soit $(X_1, \ldots, X_n)$ un n-uplet de variables aléatoires indépendantes, de moyenne empirique M_n et de variance empirique S_n^2. Alors S_n^2 admet une espérance, égale à :

$$\mathbb{E}\left(S_n^2\right) = \frac{n-1}{n} \sigma^2.$$

Démonstration.

Remarque 17 (Correction de la variance empirique)

La variable aléatoire ${S'}_n^2 = \dfrac{1}{n-1}\displaystyle\sum_{k=1}^{n}(X_k - M_n)^2$, appelée **variance empirique corrigée**, a pour espérance σ^2.

Théorème 18 (*Théorème central limite version variance connue*)

Si $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes **identiquement distribuées** (i.e. de même loi), admettant une espérance notée μ et une variance notée σ^2, alors la suite de variables aléatoires $\left(\dfrac{M_n-\mu}{\frac{\sigma}{\sqrt{n}}}\right)_{n\in\mathbb{N}^*}$ converge en loi vers une variable aléatoire Z suivant la loi normale centrée réduite.
En particulier, pour tout $(a,b) \in \overline{\mathbb{R}}^2$ tel que $a < b$, on a :

$$\lim_{n\to+\infty} \mathbb{P}\left(a < \frac{M_n-\mu}{\frac{\sigma}{\sqrt{n}}} < b\right) = \mathbb{P}(a < Z < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,\mathrm{d}t \ = \Phi(b) - \Phi(a).$$

où Φ désigne la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$.

Démonstration admise.

Remarque 19

Chacune des inégalités strictes peut être remplacée par une inégalité large dans le théorème précédent.

Remarque 20

Le théorème central limite permet d'approcher, lorsque n est suffisamment grand, la loi de $\dfrac{M_n-\mu}{\frac{\sigma}{\sqrt{n}}}$ par la loi normale centrée réduite $\mathcal{N}(0,1)$, donc la loi de M_n par la loi normale $\mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$.

Corollaire 21 (*Théorème de Moivre-Laplace*)

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes identiquement distribuées, suivant la loi de Bernoulli de paramètre $p \in]0,1[$. On note, pour tout $n \in \mathbb{N}^*$, la variable aléatoire $T_n = X_1 + \cdots + X_n$.
Alors la suite $\left(\dfrac{T_n - np}{\sqrt{np(1-p)}}\right)_{n\in\mathbb{N}^*}$ converge en loi vers une variable aléatoire T suivant la loi normale centrée réduite.
En particulier, pour tout $(a,b) \in \overline{\mathbb{R}}^2$ tel que $a < b$, on a :

$$\lim_{n\to+\infty} \mathbb{P}\left(a < \frac{T_n - np}{\sqrt{np(1-p)}} < b\right) = \mathbb{P}(a < T < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,\mathrm{d}t = \Phi(b) - \Phi(a).$$

Démonstration.

Remarque 22

Le théorème de Moivre-Laplace permet d'approcher la loi binomiale $\mathcal{B}(n,p)$ par la loi normale $\mathcal{N}(np, np(1-p))$ lorsque n est suffisamment grand. *En pratique, cette approximation est considérée valide dès $n \geqslant 30$, $np \geqslant 5$ et $np(1-p) \geqslant 5$ (ces conditions ne sont pas à connaître).*

Théorème 23 (*Théorème central limite version variance inconnue*)

Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes et identiquement distribuées (i.e. de même loi), admettant une espérance notée μ et une variance notée σ^2.
Alors, les suites de variables aléatoires $\left(\dfrac{M_n-\mu}{\frac{S_n}{\sqrt{n}}}\right)_{n\in\mathbb{N}^*}$ et $\left(\dfrac{M_n-\mu}{\frac{S'_n}{\sqrt{n}}}\right)_{n\in\mathbb{N}^*}$ convergent en loi vers une variable aléatoire suivant la loi normale centrée réduite, où S_n et S'_n désignent respectivement l'écart-type empirique et l'écart-type empirique corrigé de $(X_1,\dots,X_n)$.

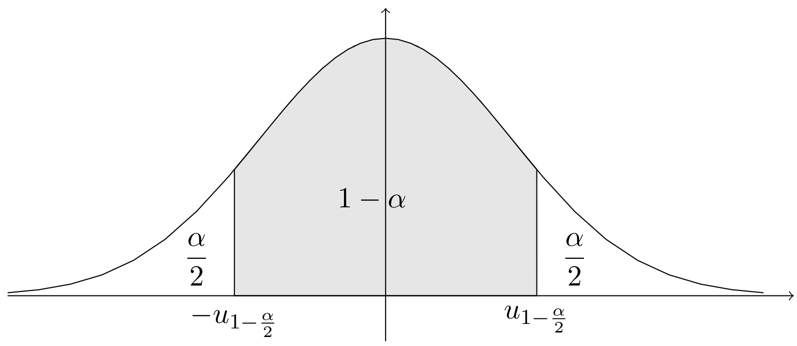
Exemple 24

Lors de la construction d'un collège accueillant 500 élèves, il est prévu la construction d'une cantine comprenant deux salles, chacune disposant de N places. On fait l'hypothèse que les élèvent qui mange à la cantine choisissent au hasard et de façon équiprobable l'une des deux salles, indépendamment les uns des autres.
On note X la variable aléatoire correspondant au nombre d'élèves choisissant la première salle.
Montrer qu'il faut et il suffit que $N \geqslant 279$ pour que la probabilité que chaque élève trouve une place dans la salle qu'il a choisie soit supérieure à 0,99. *On donne :* $\Phi(2,58) \approx 0,995$.

3 Test de conformité de la moyenne

Remarque 25 (Rappel)

Pour tout $x \in]0,1[$, on appelle **quantile** d'ordre x de la loi $\mathcal{N}(0,1)$ et on note u_x l'unique réel tel que $\Phi(u_x) = x$, où Φ désigne la fonction de répartition de la loi $\mathcal{N}(0,1)$.
Si $X \hookrightarrow \mathcal{N}(0,1)$ alors, pour tout $\alpha \in]0,1[$, $\mathbb{P}\left(-u_{1-\frac{\alpha}{2}} \leqslant X \leqslant u_{1-\frac{\alpha}{2}}\right) = 1-\alpha$.



Proposition-Définition 26

Soit $X_1, \ldots, X_n$ des variables indépendantes identiquement distribuées, d'espérance μ et d'écart-type σ^2.
Alors, pour tout $\alpha \in]0,1[$, on a :

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

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

$$\lim_{n\to+\infty} \mathbb{P}\left(\mu \in \left[M_n - u_{1-\frac{\alpha}{2}}\frac{S'_n}{\sqrt{n}}, M_n + u_{1-\frac{\alpha}{2}}\frac{S'_n}{\sqrt{n}}\right]\right) = 1-\alpha.$$

où $u_{1-\frac{\alpha}{2}}$ désigne le quantile d'ordre $1-\dfrac{\alpha}{2}$ de la loi normale $\mathcal{N}(0,1)$, i.e. l'unique réel u tel que $\Phi(u) = 1-\dfrac{\alpha}{2}$.
On appelle **intervalles (asymptotiques) de confiance** au niveau $1-\alpha$ les intervalles :

$$\left[M_n - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, M_n + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right],\ \left[M_n - u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}, M_n + u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right] \text{ et } \left[M_n - u_{1-\frac{\alpha}{2}}\frac{S'_n}{\sqrt{n}}, M_n + u_{1-\frac{\alpha}{2}}\frac{S'_n}{\sqrt{n}}\right].$$

Démonstration.

Remarque 27

Cette propriété n'est pas à connaître, il faut la retrouver systématiquement d'après le programme.

Remarque 28

On considère que l'approximation est satisfaisante dès que $n \geqslant 30$.

On présente ci-dessous le contexte du test de conformité à la moyenne :

- On considère une population dans laquelle les individus possèdent un certain caractère X et dont l'espérance μ est inconnue. Une hypothèse est formulée sur la valeur de ce paramètre $\mu = \mu_0$.
- On étudie un échantillon de la population et, à partir des résultats obtenus, rejeter ou non l'hypothèse formulée.
- L'hypothèse selon laquelle on fixe a priori un paramètre de la population à une valeur particulière s'appelle l'**hypothèse nulle** et est notée $(H_0) : \mu = \mu_0$. On définit l'**hypothèse alternative** $(H_1) : \mu \neq \mu_0$.
- Un **test statistique** est une démarche qui a pour but de fournir une règle permettant, à partir des résultats obtenus sur l'échantillon, de faire un choix entre ces deux hypothèses. C'est l'hypothèse H_0 qui est soumise au test et toute la démarche du test s'effectue en supposant cette hypothèse validée.
- On peut alors établir des règles de décision qui nous permettent d'accepter ou de rejeter l'hypothèse H_0. Toutefois, il est statistiquement impossible de prendre, à coup sûr, la bonne décision. Le risque, consenti à l'avance et que nous notons α, de rejeter l'hypothèse H_0 alors qu'elle est vraie, s'appelle le **risque de première espèce** ; ainsi :

  $$P_{[H_0 \text{ est vrai}]}(\text{rejeter } H_0) = \alpha$$

  On choisit en général $\alpha = 0,05$, voire $\alpha = 0,01$.
- Il existe un autre risque, à savoir celui d'accepter l'hypothèse H_0 alors qu'elle est fausse. Ce risque est appelé le **risque de seconde espèce** β. Le risque de première espèce α est fixé au départ; celui de deuxième espèce reste à calculer (quand c'est possible).

Méthode 29. Test de conformité à la moyenne

(i) On choisit la valeur μ_0 définissant l'hypothèse nulle $(H_0) : \mu = \mu_0$

(ii) Le réel α étant fixé, on détermine le quantile $u_{1-\frac{\alpha}{2}}$ d'ordre $1 - \dfrac{\alpha}{2}$.

(iii) On effectue un nombre $n \geqslant 30$ de mesures, dont on calcule la moyenne M_n et l'écart-type S_n empiriques.

(iv) On rejette l'hypothèse nulle (H_0) si (de manière équivalente) :

- la valeur observée $\dfrac{M_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$ n'appartient pas à l'intervalle $[-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}]$,
- ou la valeur μ_0 n'appartient pas à l'intervalle de confiance $\left[M_n - u_{1-\frac{\alpha}{2}} \dfrac{S_n}{\sqrt{n}}, M_n + u_{1-\frac{\alpha}{2}} \dfrac{S_n}{\sqrt{n}}\right]$.

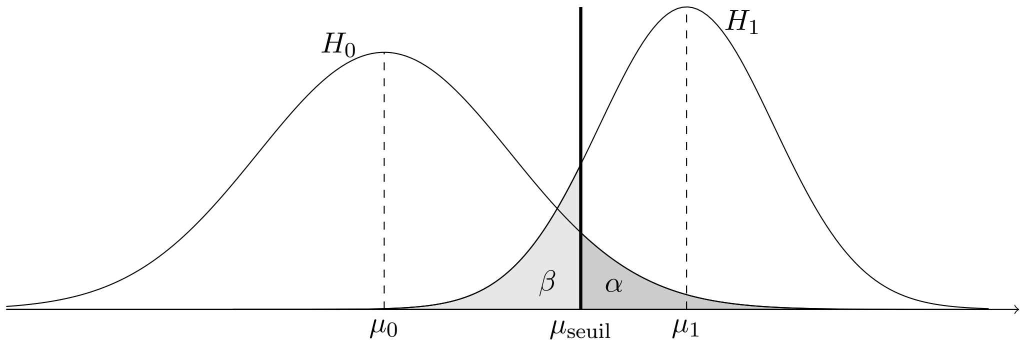
Le risque de rejeter (H_0) alors que cette hypothèse est vraie est asymptotiquement égal à α.

Remarque 30

Le tableau ci-dessous rassemble ce qu'il faut retenir des risques de première et seconde espèce.

	Hypothèse H_0 vrai	Hypothèse H_1 vraie
Hypothèse H_0 accepté	bonne décision $(1-\alpha)$	mauvaise décision (risque β)
Hypothèse H_1 accepté	mauvaise décision (risque α)	bonne décision $(1-\beta)$

On représente ci-dessous un exemple de distribution des lois (ici normales) de la statistique de test selon que l'hypothèse H_0 ou H_1 est vraie, ainsi que le risque β et la valeur μ_{seuil} permettant, une fois le risque α fixé, de décider d'accepter ou de rejeter H_0.



Remarque 31

On présente ci-dessous le tableau des valeurs usuelles de niveau de confiance et quantile correspondant.

niveau de confiance $(1-\alpha)$	90%	95%	99%
valeur de $\Phi\left(u_{1-\frac{\alpha}{2}}\right)$	0,95	0,975	0,995
quantile $u_{1-\frac{\alpha}{2}}$	1.65	1.96	2.58

Exemple 32

Dans un centre avicole, des études antérieures ont montré que la masse d'un œuf choisi au hasard peut être considérée comme la réalisation d'une variable aléatoire normale X, de moyenne m et de variance σ^2. On admet que les masses des œufs sont indépendantes les unes des autres. On prend un échantillon de n $=36$ œufs que l'on pèse. Les mesures (en grammes) sont présentées ci-dessous :

50,34	52,62	53,79	54,99	55,82	57,67	51,41	53,13	53,89	55,04	55,91	57,99
51,51	53,28	54,63	55,12	55,95	58,10	52,07	53,30	54,76	55,24	57,05	59,30
52,22	53,32	54,78	55,28	57,18	60,58	52,38	53,39	54,93	55,56	57,31	63,15

1. Déterminons un intervalle de confiance au niveau 95% puis 99% de la masse moyenne m d'un œuf.

2. Testons si la moyenne de cette variable est égale à 56g.

Exemple 33

Dans une population de renards sauvages d'effectif inconnu, on capture 154 individus dont 14 sont atteints de la rage. Déterminer un intervalle de confiance à 95% de la fréquence de renards atteints de la rage dans la population totale.