République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

RAHIM Bilal

Titre :

Echantillonnage : Théorie et Application

Membres du Comité d'Examen :

Dr.	**BENBRIKA Gozlane**	UMKB	Président
Dr.	**OUANOUGHI Yasmina**	UMKB	Encadreur
Dr.	**ROUBI Afaf**	UMKB	Examinateur

Juin 2018

Dédicace

Je dédie ce modeste travail :

À mon père.

À ma mère et mes frères.

À toute ma famille.

À mes chers amis.

Et à mes collègues de département de mathématiques.

"*Rahim Bilal*"

REMERCIEMENTS

Avant tout, je remercie le dieu tout puissant de m'avoir accordé volonté et patience pour accomplir ce travail.

Mes remerciements les plus sincères vont en particulier, à ***Dr. OUANOUGHI Yasmina*** *pour avoir m'encadrer et diriger, et pour l'effort fournit, ces conseils prodigués et sa patience dans le suivie de ce travail.*

Mes remerciements vont également au membres de jury : ***Dr. ROUBI Afaf*** et ***Dr. BENBRIKA Gozlane*** pour *avoir accepter de juger ce mémoire.*

Un grand merci aux enseignants(es) et personnel du département de mathématique.

A toutes et tous qui ont contribué de près ou de loin à la réalisation de ce travail :

Merci

Table des matières

Table des figures

Introduction

L'étude statistique joue un rôle essentiel dans la recherche, elle peut être utilisée pour la collecte, l'analyse et l'interprétation des données, elle aide aussi les chercheurs à obtenir une excellente conclusion et un bon raisonnement statistique et à obtenir des résultats précis. L'un des mécanismes les plus importants pour les études statistiques est le processus d'échantillonnage.

L'échantillonnage est la sélection d'une partie dans un tout qui produit une série d'échantillon à étudier. Le processus d'échantillonnage dépend de quatre étapes(préciser les objectifs de recherche, identification de la population d'origine à partir de laquelle on sélectionne l'échantillon, ils sont de deux types (population de taille finie ou infinie), Dans cette Mémoire, nous allons étudier le premier type, détermination des caractéristiques de la population, sélectionner la taille de l'échantillon).

L'échantillonnage est devenu une méthode efficace et inévitable dans toutes les études liées à la vie, l'une des utilisations les plus importantes (recherche dans les laboratoires de chimie pour déduire les phénomènes ou les transformations de la matière, histoire, prendre des échantillons de sang pour détecter le groupe sanguin et certaines maladies, bourse et marchés, toutes les études liées aux expériences scientifiques, ...)

L'échantillonnage vise à réduire le temps et économiser l'effort et l'argent. L'étude des échantillons nous permet d'obtenir des résultats précis avec les mêmes caractéristiques que la population d'origine et donc les résultats peuvent être généralisés à la population dans son ensemble.

L'organisation de ce mémoire est la suivante.

Dans le premier chapitre, nous avons mentionné quelques définitions et concepts de base, puis nous avons parlé des méthodes d'échantillonnage et de la différence entre elles et nous avons présenté les deux théorèmes fondamentaux de la statistique asymptotique (théorème central limite et lois des grands nombres).

Dans le deuxième chapitre, nous allons concentrer notre étude sur les distributions des caractéristiques de l'échantillon et propriétés de l'échantillon aléatoire, et les plus importantes (la moyenne empirique, la variance empirique, la fonction de répartition empirique) et nous enfermons chacun d'eux avec les comportements asymptotique. Enfin, nous donnons un aperçu de l'échantillon issus d'une variable normale.

Dans le troisième chapitre, nous présenterons quelques applications des résultats théoriques du deuxième chapitre, à l'aide du logiciel d'analyse statistique R.

Chapitre 1

Notion d'échantillonnage

L'échantillonnage est un moyen de sélectionner un sous-ensemble d'unités dans une population aux fins de la collecte de l'information sur ces unités pour formuler des inférences sur l'ensemble de la population. Dans ce chapitre, nous commençons par résumer quelques définitions et concepts de base, puis nous passons à un examen des principales méthodes d'échantillonnage et des lois fondamentales des statistiques asymptotique et des modes de la convergence .

1.1 Définitions et caractéristiques de base

1.1.1 Définitions sur la population et échantillon

Définition 1.1.1 (Population de taille finie) *Soit E un ensemble, que nous appellerons population mère, contenant un nombre fini N d'éléments. Le statisticien s'intéresse plus particulièrement à une propriété X de la population (l'âge, le prix), appelée propriété statistique.*

Définition 1.1.2 (L'échantillon) *Un échantillon ξ_n est un sous-ensemble de n individus extraits d'une population E composée de N élément. Nous écrivons :*

$$E = \{e_i,\ i \in \{1, ..., N\}\} \quad et \quad \xi_n = \{e_{i_k},\ i_k \in U\}.$$

$U = \{i_1, ..., i_n\}$ *les indices des unités de l'échantillon.*

Définition 1.1.3 *Unité statistique (individu) sont les éléments de la population statistique et les unité de l'échantillon sont les éléments de l'échantillon.*

Définition 1.1.4 *Le taux de sondage, dans le cas de population finie, est le rapport entre la taille de l'échantillon et la taille de la population* $\frac{n}{N}$.

Exemple 1.1.1 *Dans université de Biskra il y a 30 milles étudiants et étudiantes, on doit savoir le niveau moyen d'étudiants pour cela on tire avec hasard seulement 1000 étudiants et étudiantes.*
Nous écrivons : La population est 30 milles étudiants et étudiantes, l'échantillon est 1000 étudiants et étudiantes, et la taux de sondage $\frac{n}{N} = \frac{1000}{30000} = \frac{1}{30}$.

Définition 1.1.5 (Fonction caractéristique) *Soit X une variable aléatoire réelle. On appelle fonction caractéristique de X la fonction de la variable réelle t définie par :*

$$\varphi_X(t) = E\left[e^{(itX)}\right] = \int_D e^{(itX)} \lambda_X(dx).$$

Si X est discrète de loi $\lambda_X = \sum_k P_k \delta_{xk}$ *alors :* $\varphi_X(t) = \sum_k P_k e^{(itx_k)}$.
Si X est absolument continue $\lambda_X(dx) = f_X(x)dx$ *alors :* $\varphi_X(t) = \int_{\mathbb{R}} e^{(itx)} f_X(x)dx$.

1.1.2 Statistique

Définition 1.1.6 *Une statistique T est une variable aléatoire en fonction mesurable de* $X_1, X_2, ..., X_n$.

$$T = h(X_1, X_2, ..., X_n).$$

Exemple 1.1.2 *Les statistiques les plus coramment utilisées*

La moyenne empirique : $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$.

Le moment empirique d'ordre k : $M_n^k := \frac{1}{n}\sum_{i=1}^{n} X_i^k$.

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

La fonction de distribution empirique : $F_n\ (x) := \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}$.

1.1.3 Estimateur sans biais

Définition 1.1.7 *En mathématiques, un estimateur est une statistique permettant d'évaluer un paramètre inconnu relatif à une loi de probabilité (comme son espérance ou sa variance).*

Définition 1.1.8 *Nous appelons biais de l'estimateur* T *de* θ *la fonction* $b_\theta(T) = E_\theta[T] - \theta$.

Nous disons que l'estimateur T *est sans biais si* $b_\theta(T) = E_\theta[T] - \theta = 0$.

Nous disons que l'estimateur T *est un estimateur asymptotiquement sans biais si*

$$\lim_{n\to\infty} b_\theta(T) = 0.$$

Exemple 1.1.3 *1. La moyenne empirique* $T = \bar{X}_n$ *est un estimateur sans biais de la moyenne théorique* $\theta = E(X)$, *Parce que* $b_\theta(\bar{X}_n) = E_\theta[\bar{X}_n] - \theta = 0$.

2. Si le paramètre à estimer est la variance théorique, i.e. $\theta = Var(X)$, *l'estimateur est la variance empirique* $T = S_n^2$. *Alors* $b_\theta\left(S_n^2\right) = E_\theta\left(S_n^2\right) - \theta = \frac{n-1}{n}\theta - \theta = -\frac{1}{n}\theta \neq 0$.

Donc S_n^2 *est un estimateur biaisé de la variance théorique. Mais asymptotiquement est un estimateur sans biais de la variance théorique car* $\lim_{n\longrightarrow\infty} b_\theta\left(S_n^2\right) = \lim_{n\longrightarrow\infty}\left(-\frac{1}{n}\theta\right) = 0$.

1.2 Méthode d'échantillonnage

Il existe deux méthodes d'échantillonnage, le tirage d'un échantillon d'une manière aléatoire ou non aléatoire

1.2.1 Méthode non aléatoires

L'échantillonnage non aléatoire (non probabiliste) repose sur un choix au manière non-aléatoire d'unité dans la population, et le principe de choix des unités n'est soumis à aucune loi.

Echantillonnage systématique

Il arrive parfois que les données de la population soient classées ou rangées dans un fichier, dans ce cas, il est plus simple d'obtenir un échantillon en prenant une progression arithmétique ayant comme base (b) un nombre aléatoire et avec une raison (L) choisie de manière à pouvoir prendre tous les éléments de la population, les éléments choisis seront :

$$x_i = b + (i-1)L \quad i = 1, 2, ..., n.$$

La raison L est calculée en prenant le rapport $\frac{N}{n}$ et b sera un nombre aléatoire entre 1 et L.

Exemple 1.2.1 *On a une population finie composée de 100 individu* $E = \{1, 2, ..., 100\}$. *Nous voulons obtenir un échantillon de 20 individu. On pose* $N = 100, n = 20$ *et* $L = \frac{100}{20} = 5$. *La base b sera un nombre aléatoire entre* 1 *et* 5, *si ce nombre* 4 *l'échantillon comprendra les éléments ayant les numéros suivants :*

$$x_i = 4 + (i-1)5 \quad i = 1, 2, ..., 20, \quad donc \quad \xi_n = \{4, 9, 14, 19, ..., 99\}$$

Echantillonnage unité type

La méthode des unités types consiste à partager la population en groupes homogènes et différentes, pour ensuite choisir dans chacun de ces groupes une unité statistique représentative de ce groupe, que l'on nomme l'unité type, et qui se situe dans la moyenne du groupe. On sous entend que cette unité type aura la même « réaction » sur la variable étudiée que la moyenne des unités du groupe qu'elle représente.

Echantillonnage par quotas

L'échantillonnage par quotas est une méthode d'échantillonnage non aléatoire. Elle est basée sur la répartition connue de la population pour un certain nombre de caractères (sexe, âge, catégorie socioprofessionnelle...). L'échantillon est construit en respectant la distribution de la population, il est choisi de façon à constituer une image aussi fidèle que possible de la population totale.

1.2.2 Méthodes aléatoires

L'échantillonnage aléatoire (ou probabiliste) repose sur un choix au hasard d'unité dans la population. La probabilité de sélection d'un unité de la population est connue.

Echantillonnage aléatoire simple

Le principe de cette méthode est la sélection aléatoire des unités de la population, tous les unités de la population ont la même probabilité. Ce choix peut se faire avec ou sans remise.

Tirage avec remise Un individu peut être choisi plusieurs fois. La population reste le même après chaque tirage. Le processus de tirage des individus de la population est indépendant l'un de l'autre. Dans ce cas, il y a N^n échantillon possible.

Exemple 1.2.2 *On a une population finie composée de* 3 *éléments* $P = \{2, 3, 4\}$*, alors le nombre des échantillons possible à être prélevées est une liste* 2 *éléments pris parmi* 3 *éléments c-à-d,* $N^n = 3^2 = 9$*. Les échantillons sont les suivants :*

$$E = \left\{ \begin{array}{l} (2,2); (2,3); (2,4) \\ (3,2); (3,3); (3,4) \\ (4,2); (4,3); (4,4) \end{array} \right\}.$$

Tirage sans remise Un individu peut être choisi au plus une fois. Une unité est déduite de la population chaque tirage. Le processus de tirage des individus de la population devenir non indépendant l'un de l'autre. Dans ce cas, il y a $C_N^n = \frac{N!}{n!(N-n)!}$ échantillon possible.

Exemple 1.2.3 *Le même exemple précédent, parce que le tirage est sans remise donc le nombre des échantillons possible à être prélevées est une combinaison de 2 éléments pris parmi 3 éléments c-à-d :* $C_3^2 = \frac{3!}{2!(3-2)!} = \frac{6}{2} = 3$*. Les échantillons sont les suivants :*

$$E = \{(2,3); (2,4); (3,4)\}.$$

Echantillonnage par grappes

La population est divisée en G grappes, pas forcément de même taille. L'échantillonnage consiste à choisir g grappes selon un plan aléatoire simple sans remise. Le nombre d'échantillons possibles est C_G^g .

1.3 Modèle d'échantillonnage

Définition 1.3.1 *Soit une expérience aléatoire définie par la v.a.* X *telle que :*

$$X : (\Omega, B, P) \longrightarrow (\mathcal{L}, a, P)$$

On appelle modèle d'échantillonnage de taille n l'espace produit $(\mathcal{L}, a, P_\theta)^n$ *égal à* $(\mathcal{L}^n, a_n, P_\theta^n)$ *associé à n expériences aléatoires indépendantes* $(\mathcal{L}, a, P^X)$, *où*

$\mathcal{L}^n$: *le produit des espaces des valeurs de* X.

a_n : *est la tribu produit des événements de* $\mathcal{L}^n$.

P_θ^n : *la loi jointe des expériences.*

On notera X_i *la v.a. de même loi que* X, *associée à la* $i^{éme}$ *expérience et* x_i *sa réalisation. A l'éspace* $(\mathcal{L}, a, P^X)^n$ *correspondra donc une séquence de v.a.* $(X_1, X_2, ..., X_n)$ *indépendantes et identiquement distribuées (i.i.d.) de même loi* P^X.

Remarque 1.3.1 *Le fait que les modèles d'échantillonnage soient composés d'expériences indépendantes les rend très simples à manipuler. Ainsi si* X *est une variable discrète :*

$$\begin{aligned} P^{(X_1,X_2,...,X_n)}(x_1, ..., x_n) &= P(X_1 = x_1, ..., X_n = x_n) \\ &= \prod_{i=1}^{n} P^{X_i}(x_i) \\ &= \prod_{i=1}^{n} P^{X}(x_i). \end{aligned}$$

De même, si X *est continue de densité* f : $f(x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i)$.

Voir Tassi, Philippe [15]

1.4 Modes de convergence

On considère une suite des variables aléatoires ($X_1, ..., X_n$) notée $(X_n)_{n\epsilon\mathbb{N}}$. Nous rappelons ici quatre modes de convergence de $(X_n)_{n\epsilon\mathbb{N}}$ vers X.

Convergence en probabilité

1. L'inégalité de Markov et Bienaymé-Tchebychev

Théorème 1.4.1 (L'inégalité de Markov) *Si* X *est une v.a. de signe quelconque, l'inégalité de Markov en l'appliquant à* $|X|^k$, *pour tout* $k \geq 0$ *tel que* $E|X|^k$ *existe : pour tout*

$\lambda > 0$

$$P(|X|^k \geq \lambda) \leq \frac{E\,|X|^k}{\lambda}.$$

On introduit alors un nombre $\varepsilon > 0$ tel que $\varepsilon^k = \lambda$ et on en déduit pour tout $\varepsilon > 0$:

$$P(|X| \geq \varepsilon) \leq \frac{E\,|X|^k}{\varepsilon^k}. \tag{1.1}$$

Théorème 1.4.2 (L'inégalité de Bienaymé-Tchebychev) *On obtient l'inégalité de Bienaymé-Tchebychev on appliquant l'inégalité de Markov sous derniére forme 1.1, à la v.a. $X - E(X)$ pour $k = 2$, donc pour une variable dont la variance existe, soit pour tout $\varepsilon > 0$ fixé :*

$$\begin{aligned} P(|X - E(X)| \geq \varepsilon) &\leq \frac{E(X - E(X))^2}{\varepsilon^2} \\ P(|X - E(X)| \geq \varepsilon) &\leq \frac{Var(X)}{\varepsilon^2}. \end{aligned} \tag{1.2}$$

2.Convergence en probabilité

Soient $(X_n)_{n\epsilon\mathbb{N}}$ et X des variables aléatoires. On dit que $(X_n)_{n\epsilon\mathbb{N}}$ converge en probabilité vers X et on notera $X_n \xrightarrow{P} X$ si, pour tout $\varepsilon > 0$:

$$P\left(|X_n - X| \geq \varepsilon\right) \longrightarrow 0 \text{ quand } n \longrightarrow \infty.$$

Convergence en loi

Soient $(X_n)_{n\epsilon\mathbb{N}}$ et X des variables aléatoires de fonction de répartition (F_{X_n}), et F_X. Une suite de v.a. $(X_n)_{n\epsilon\mathbb{N}}$ converge en loi vers X si

$$\lim_{n\longrightarrow\infty} F_{X_n}(x) = \ F_X\ (x)$$

en tout point x où F_X est continue. On écrit alors $X_n \xrightarrow{Loi} X$.

Théorème 1.4.3 *Soit $(X_n)_{n\epsilon\mathbb{N}}$ une suite de variables aléatoires réelles. $(X_n)_{n\epsilon\mathbb{N}}$ converge*

en loi vers une variable aléatoire réelle X si et seulement si $(\varphi_{X_n})_{n\epsilon\mathbb{N}}$ converge simplement vers φ_X et on notera $\left\{\forall t \in \mathbb{R} : \varphi_{X_n}(t) \underset{n\to\infty}{\longrightarrow} \varphi_X(t)\right\} \Longleftrightarrow \{X_n \xrightarrow{Loi} X\}$.

Convergence presque sûre

Soient $(X_n)_{n\epsilon\mathbb{N}}$ et X des variables aléatoires. On dit que $(X_n)_n$ converge presque sûrement vers X et on notera $X_n \xrightarrow{p.s.} X$ si pour tout $\varepsilon > 0$:

$$P\left(\sup_{n\geq m}\{|X_n - X| < \varepsilon\}\right) \longrightarrow 1 \text{ quand } n \longrightarrow \infty.$$

Convergence en moyenne quadratique

Soient $(X_n)_{n\epsilon\mathbb{N}}$ et X des variables aléatoires. On dit que $(X_n)_n$ converge en moyenne d'ordre p vers X et on notera $X_n \xrightarrow{L^p} X$ si pour tout $0 < p < \infty$:

$$E\,|X_n - X|^p \longrightarrow 0 \text{ quand } n \longrightarrow \infty.$$

Dans le cas particulier $p = 2$ la convergence L^2 (s'appelle converge en moyenne quadratique) on écrit alors : $X_n \xrightarrow{L^2} X$ si

$$E((X_n - X)^2) \longrightarrow 0 \text{ quand } n \longrightarrow \infty.$$

Relation entre les types de convergences

La convergence presque sûre implique la convergence en probabilité.

La convergence en probabilité implique la convergence en loi.

La convergence en moyenne d'ordre p implique la convergence en probabilité.

Les implications réciproques sont en général fausses.

$$(X_n \xrightarrow{L^p} X) \Longrightarrow (X_n \xrightarrow{P} X) \Longrightarrow (X_n \xrightarrow{Loi} X)$$
$$\Uparrow$$
$$(X_n \xrightarrow{P.S} X).$$

1.5 Lois fondamentales d'échantillonnage

On peut définir deux théorèmes fondamentaux les plus utilisé dans l'étude de la probabilité et de la statistique : lois des grands nombres et théorème limite central

1.5.1 Lois des Grands Nombres

Elles sont de deux types : lois faibles mettant en jeu la convergence en probabilité et lois fortes relatives à la convergence presque sûre. Nous considérons ici des suites de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ de même loi.

Loi faible des grands nombres

Théorème 1.5.1 *Soit une suite* $(X_n)_{n\in\mathbb{N}}$ *de v.a indépendantes, de même loi, intégrables. On pose* $E(X_1) = \mu$. *Alors*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\longrightarrow\infty}{\longrightarrow} \mu \text{ , en probabilité et en moyenne.}$$

Preuve. X_i $\forall 1 \leq i \leq n$ carré intégrable indépendantes,On pose $E(X_1) = \mu$ et $\sigma^2 = Var(X_1)$, en effet :

$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n}{n}E(X_1) = \mu.$$
$$Var(\bar{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n}{n^2}Var(X_1) = \frac{\sigma^2}{n}.$$

Appliquons l'inégalité de Bienaymé-Tchebychev 1.2 pour $\bar{X}_n$: Pour tout $\varepsilon > 0$

$$P(\left|\bar{X}_n - \mu\right| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2 n} \underset{n\longrightarrow\infty}{\longrightarrow} 0.$$

Maintenant déduire la convergence en probabilité de $\bar{X}_n$ vers μ.

$$E(\left|\bar{X}_n - \mu\right|) \leq \sqrt{E(\bar{X}_n - \mu)^2} = \sqrt{Var(\bar{X}_n)} = \sqrt{\frac{\sigma^2}{n}} \underset{n\longrightarrow\infty}{\longrightarrow} 0.$$

Alors déduire la convergence en moyenne de $\bar{X}_n$ vers μ. ■

Loi forte des grands nombres

Théorème 1.5.2 *Soit une suite* $(X_n)_{n\in\mathbb{N}}$ *de v.a indépendantes, de même loi, intégrables et* $E(X_1) = \mu$ *est finie. Alors*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p.s.} \mu \text{ ,quand } n \longrightarrow \infty.$$

Pour la démonstration de la loi forte des grands nombres voir [9]

1.5.2 Théorème Limite Central

Théorème 1.5.3 *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de variables aléatoires indépendantes, de même loi et de carré intégrable (et non constantes). Notons* $E(X_1) = \mu$, $Var(X_1) = \sigma^2$ *avec* $\sigma > 0$. *Alors*

$$S_n^{\bullet} := \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{Loi} \mathcal{N}(0,1) \text{ quand } n \longrightarrow \infty, \quad \text{où } S_n := \sum_{i=1}^{n} X_i. \tag{1.3}$$

Exemple 1.5.1 *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de variables aléatoires indépendantes, de même loi de Bernoulli de paramétre* $p \in]0,1[$, *avec* $q := 1-p$. *On pose* $E(S_n) = np$ *et* $Var(S_n) = npq$. *Alors*

$$S_n^{\bullet} := \frac{S_n - np}{\sqrt{npq}} = \sqrt{\frac{n}{pq}}\left(\frac{S_n - np}{n}\right) \xrightarrow{Loi} \mathcal{N}(0,1) \text{ quand } n \longrightarrow \infty, \quad \text{où } S_n := \sum_{i=1}^{n} X_i \rightsquigarrow B(n,np).$$

Chapitre 2

Distributions des caractéristiques d'un échantillon

La distribution d'échantillonnage est le concept de base dans l'inférence statistique. Dans ce chapitre, en déduire quelques caractéristiques de la population qui ne sont pas connues selon sur la base d'informations et les caractéristiques spécifiques d'un échantillon sélectionné. Par conséquent, nous pouvons prendre un échantillon $(X_1, ..., X_n)$ de taille n en considérant les caractéristiques d'un échantillon aléatoire, principalement la moyenne et sa variance et leur distributions.

2.1 Statistique $\bar{\mathbf{X}}_n$: (moyenne empirique)

Définition 2.1.1 *La statistique $\bar{X}_n$ ou moyenne empirique de l'échantillon est :*

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

Propriétés 2.1.1 *Soit $(X_1, ..., X_n)$ un échantillon aléatoire d'une v.a. X de moyenne μ*

et de variance σ^2.

1. $E(\bar{X}_n) = \mu \quad et \quad Var(\bar{X}_n) = \dfrac{\sigma^2}{n}$ *(tirage avec remise).*
2. $E(\bar{X}_n) = \mu \quad et \quad Var(\bar{X}_n) = \left(\dfrac{N-n}{N-1}\right)\dfrac{\sigma^2}{n}$ *(tirage sons remise).*

Pour calculer $E(\bar{X}_n)$ *et* $Var(\bar{X}_n)$, *il convient de distinguer le mode de tirage.*

Preuve. 1. Calculer la moyenne et la variance de $\bar{X}_n$ dans le cas d'un échantillon de v.a. i.i.d.(tirage avec remise d'une population finie). L'espérance de $\bar{X}_n$ est égale μ, en effet :

$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n\mu}{n} = \mu.$$

La variance de $\bar{X}_n$ est égale $\frac{\sigma^2}{n}$, en effet :

$$Var(\bar{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

D'après l' indépendance des X_i.

2. Calculer la moyenne et la variance de $\bar{X}_n$ dans le cas (tirage sans remise d'une population finie) les $X_i \ \ \forall 1 \leq i \leq .n$ ne sont pas indépendantes. L'espérance de $\bar{X}_n$ est égale μ.
La variance de $\bar{X}_n$ est égale $\left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$, en effet :

$$\begin{aligned} Var(\bar{X}_n) &= \frac{1}{n^2}\left[\sum_{i=1}^{n} Var(X_i) + \sum_{i=1}^{n}\sum_{i\neq j} cov(X_i, X_j)\right] \\ &= \frac{1}{n^2}\left[nVar(X_i) + n\sum_{i\neq j} cov(X_i, X_j)\right] \\ &= \frac{1}{n^2}\left[n\sigma^2 + n(n-1)cov(X_i, X_j)\right] \\ &= \frac{1}{n}\left[\sigma^2 + (n-1)cov(X_i, X_j)\right] \end{aligned}$$

avec

$$\begin{aligned} cov(X_i, X_j) &= E\left[(X_i - \mu)(X_j - \mu)\right] \\ &= \sum_{l=1}^{N}\sum_{r=1}^{N}(X_l - \mu)(X_r - \mu)P(X_i - \mu, X_j - \mu) \end{aligned}$$

d'autre part

$$\begin{aligned} P(X_i - \mu, X_j - \mu) &= P(X_i = x_l)P(X_j = x_r \diagdown X_i = x_l) \\ &= \frac{1}{N}\frac{1}{N-1}. \end{aligned}$$

D'après la probabilité conditionnelle.

Nous obtenons :

$$cov(X_i, X_j) = \begin{cases} \sum_{l=1}^{N}\sum_{r=1}^{N}(X_l - \mu)(X_r - \mu)\frac{1}{N}\frac{1}{N-1} & \text{si } r \neq l \\ 0 & \text{si } r = l \end{cases}$$

donc

$$cov(X_i, X_j) = \frac{1}{N}\frac{N}{N-1}\sum_{r\neq l}^{N}(X_l - \mu)(X_r - \mu).$$

Comme

$$\left[\sum_{l=1}^{N}(X_l - \mu)\right]^2 = \sum_{r\neq l}^{N}(X_l - \mu)^2 + \sum_{l,r=1}^{N}\sum_{l\neq r}^{N}(X_l - \mu)(X_r - \mu)$$

$$\begin{aligned} N\sum_{l\neq r}^{N}(X_l - \mu)(X_r - \mu) &= \left[\sum_{l=1}^{N}(X_l - \mu)\right]^2 - \sum_{r\neq l}^{N}(X_l - \mu)^2 \\ &= \left[N\mu - N\mu\right]^2 - N\sigma^2 \\ &= 0 - N\sigma^2 \end{aligned}$$

$$\sum_{l\neq r}^{N}(X_l - \mu)(X_r - \mu) = -\sigma^2$$

On obtient

$$cov(X_i, X_j) = \frac{1}{N}\frac{N}{N-1}(-\sigma^2) = \frac{-\sigma^2}{N-1}.$$

Finalement

$$\begin{aligned} Var(\bar{X}_n) &= \frac{1}{n}\left[\sigma^2 + (n-1)cov(X_i, X_j)\right] \\ &= \frac{1}{n}\left[\sigma^2 + (n-1)\frac{-\sigma^2}{N-1}\right] \\ &= \frac{\sigma^2}{n}\left[1 - \frac{(n-1)}{N-1}\right] \\ &= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right). \end{aligned}$$

La fin de la démonstration. ■

Remarque 2.1.1 *1. En terme de variance, X est donc toujours moins dispersé un tirage sans remise que pour un tirage avec remise.*

2. Si $N \longrightarrow \infty$ et $\frac{n}{N} \longrightarrow 0$, $\left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} \longrightarrow \frac{\sigma^2}{n}$ et donc il n'y a pas différence entre les deux modes de tirage, ce qui est intuitif.

Propriétés 2.1.2 *Nous nommons $\mu_k^{(c)} = E\,(X-\mu)^k$ le moment centrée d'ordre k ($k \in \mathbb{N}$) de X, alors le troisiéme et le quatriéme moment de $\bar{X}_n$ sont données par :*

$$\mu_3^{(c)}(\bar{X}_n) = \frac{\mu_3^{(c)}(X)}{n^2}, \qquad et \qquad \mu_4^{(c)}(\bar{X}_n) = \frac{\mu_4^{(c)}(X) + 3(n-1)\sigma^4}{n^3}.$$

Preuve. Le moment centrée d'ordre 3 de $\bar{X}_n$ est égale $\frac{\mu_3^{(c)}}{n^2}$, en effet :

$1^{ére}$ cas X centrée, $\mu = 0$

$$\mu_3^{(c)}(\bar{X}_n) = E(\bar{X}_n)^3 = \frac{1}{n^3}E\left[\sum_{i=1}^{n} X_i\right]^3.$$

Donc

$$\left[\sum_{i=1}^{n} X_i\right]^3 = \sum_{i=1}^{n} X_i^3 + C\sum_{i=1}^{n}\sum_{i\neq j}^{n} X_i X_j^2 \quad (C = \text{constant})$$

donc

$$\begin{aligned}
\mu_3^{(c)}(\bar{X}_n) &= \frac{1}{n^3} E\left[\sum_{i=1}^{n} X_i^3 + C\sum_{i=1}^{n}\sum_{i\neq j}^{n} X_i X_j^2\right] \\
&= \frac{1}{n^3}\left[E\left(\sum_{i=1}^{n} X_i^3\right) + nCE\left(\sum_{i\neq j}^{n} X_i X_j^2\right)\right] \\
&= \frac{1}{n^3}\left[\sum_{i=1}^{n} EX_i{}^3 + nC\sum_{i\neq j}^{n} E\left(X_i X_j^2\right)\right] \\
&= \frac{1}{n^3}\left[n\mu_3^{(c)}(X) + nC\sum_{i\neq j}^{n} EX_i EX_j^2\right]
\end{aligned}$$

comme X centrée, alors $\sum_{i\neq j}^{n} EX_i EX_j^2 = 0$, nous obtenons :

$$\begin{aligned}
\mu_3^{(c)}(\bar{X}_n) &= \frac{1}{n^3}\left[n\mu_3^{(c)}(X)\right] \\
&= \frac{\mu_3^{(c)}(X)}{n^2}.
\end{aligned}$$

$2^{émé}$ cas X quelconque

Supposons que $Y = X - \mu$ centrée, alors $\bar{X}_n = \bar{Y}_n + \mu$, d'où

$$\begin{aligned}
\mu_3^{(c)}(\bar{X}_n) &= E\left[\bar{X}_n - \mu\right]^3 \\
&= E\left[\bar{Y} + \mu - \mu\right]^3 \\
&= E\left[\bar{Y}\right]^3 \\
&= \mu_3^{(c)}(\bar{Y}_n) \\
&= \frac{\mu_3^{(c)}(Y)}{n^2} \\
&= \frac{E\left[X - \mu\right]^3}{n^2} \\
&= \frac{\mu_3^{(c)}(X)}{n^2}.
\end{aligned}$$

Le moment centrée d'ordre 4 de $\bar{X}_n$ est égale $\frac{\mu_4^{(c)}(X)+3(n-1)\sigma^4}{n^3}$, en effet

$1^{ére}$ cas X centrée, $\mu = 0$

$$\begin{aligned}
\mu_4^{(c)}(\bar{X}_n) &= E\left[\bar{X}_n\right]^4 \\
&= \frac{1}{n^4} E\left[\sum_{i=1}^{n} X_i\right]^4 \\
&= \frac{1}{n^4} E\left[\sum_{i=1}^{n} X_i^4 + 3\sum_{i=1}^{n}\sum_{i\neq j}^{n} X_i^2 X_j^2 + 2\sum_{i=1}^{n}\sum_{i\neq j}^{n} X_i X_j^3\right] \\
&= \frac{1}{n^4}\left[\sum_{i=1}^{n} E(X_i^4) + 3n\sum_{i\neq j}^{n} E(X_i^2 X_j^2) + 2n\sum_{i\neq j}^{n} E(X_i X_j^3)\right] \\
&= \frac{1}{n^4}\left[\sum_{i=1}^{n} E(X_i^4) + 3n\sum_{i\neq j}^{n} E(X_i^2)E(X_j^2) + 2n\sum_{i\neq j}^{n} E(X_i)E(X_j^3)\right] \\
&= \frac{1}{n^4}\left[\sum_{i=1}^{n} \mu_4^{(c)}(X) + 3n\sum_{i\neq j}^{n} \sigma^2\sigma^2 + 0\right] \qquad \text{car } X \text{ centrée} \\
&= \frac{1}{n^4}\left[n\mu_4^{(c)}(X) + 3n(n-1)\sigma^4\right] \\
&= \frac{\mu_4^{(c)}(X) + 3(n-1)\sigma^4}{n^3}.
\end{aligned}$$

$2^{émé}$ cas X quelconque, Supposons que $Y = X - \mu$ centrée, alors $\bar{X}_n = \bar{Y}_n + \mu$

$$\begin{aligned}
\mu_4^{(c)}(\bar{X}_n) &= E\left[\bar{X}_n - \mu\right]^4 \\
&= E\left[\bar{Y}_n + \mu - \mu\right]^4 \\
&= E\left[\bar{Y}_n\right]^4 \\
&= \mu_4^{(c)}(\bar{Y}_n) \\
&= \frac{\mu_4^{(c)}(Y) + 3(n-1)\sigma^4}{n^3} \\
&= \frac{E\left[X-\mu\right]^4 + 3(n-1)\sigma^4}{n^3} \\
&= \frac{\mu_4^{(c)}(X) + 3(n-1)\sigma^4}{n^3}.
\end{aligned} \tag{2.1}$$

Fin de la preuve. ■

Définition 2.1.2 (Coefficients d'asymétrie et d'aplatissement) *Si σ est l'écart-type de la population, et μ la moyenne, alors*

1. Le coefficients d'asymétrie de X noté cd, définie par $cd(X) := \frac{\mu_3^{(c)}(X)}{\sigma^3}$.

2. Le coefficients d'aplatissement de X noté ca, définie par $ca(X) := \frac{\mu_4^{(c)}(X)}{\sigma^4}$.

Propriétés 2.1.3 *Si on designe par $cd(X)$ et $ca(X)$ le coefficients d'asymétrie et d'aplatissement de X, alors :*

$$cd(\bar{X}_n) = \frac{cd(X)}{\sqrt{n}} \qquad et \quad ca(\bar{X}_n) = 3 + \frac{ca(X) - 3}{n}$$

le coefficients d'asymétrie et d'aplatissement de $\bar{X}_n$, en effet :

$$cd(\bar{X}_n) = \frac{\mu_3^{(c)}(\bar{X}_n)}{(\sigma_{\bar{X}_n})^3} = \frac{\frac{\mu_3^{(c)}}{n^2}}{(\frac{\sigma}{\sqrt{n}})^3} = \frac{\mu_3^{(c)}}{n^2}\frac{n^{\frac{3}{2}}}{\sigma^3} = \frac{cd(X)}{\sqrt{n}},$$

et

$$ca(\bar{X}_n) = \frac{\mu_4^{(c)}(\bar{X}_n)}{(\sigma_{\bar{X}_n})^4} = \frac{\frac{\mu_4^{(c)}+3(n-1)\sigma^4}{n^3}}{(\frac{\sigma}{\sqrt{n}})^4} = \frac{\mu_4^{(c)} + 3(n-1)\sigma^4}{n\sigma^4} = 3 + \frac{\frac{\mu_4^{(c)}}{\sigma^4} - 3}{n} = 3 + \frac{ca(X) - 3}{n}.$$

Remarque 2.1.2 *On voit que $\lim\limits_{n\longrightarrow\infty} cd(\bar{X}_n) = 0$ et $\lim\limits_{n\longrightarrow\infty} ca(\bar{X}_n) = 3$, ce qui la normalité asymptotique de $\bar{X}_n$.*

2.1.1 Comportement asymptotique

Théorème 2.1.1 (T.L.C) *Dans le cadre d'une expérience renouvelable, on peut idéalement faire appel à l'asymptotique et, en utilisant le T.L.C, on obtient directement sous la condition que X soit de carré intégrable ($E(X^2) < \infty$) :*

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{Loi} \mathcal{N}(0,1) \qquad quand \;\; n \longrightarrow \infty. \tag{2.2}$$

Pour n suffisamment grand ($n \geq 30$ ou 50 en général), on utilise l'approximation normale :

$$\bar{X}_n \rightsquigarrow \mathcal{N}(\mu, \frac{\sigma^2}{n}). \tag{2.3}$$

Théorème 2.1.2 (Loi des grands nombres) *La moyenne empirique converge en probabilité et presque sûrement vers $E(X) = \mu$ quand n tend vers l'infini.*

$$\begin{aligned} &Loi\ faible\ GN \quad \bar{X}_n \xrightarrow{P} \mu \qquad quand\ n \longrightarrow \infty. \\ &Loi\ forte\ GN \quad \bar{X}_n \xrightarrow{p.s.} \mu \qquad quand\ n \longrightarrow \infty. \end{aligned} \tag{2.4}$$

2.2 Statistique S_n^2 :(variance empirique)

Définition 2.2.1 *On appelle variance empirique, la statistique notée S_n^2 définie par :*

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

Propriétés 2.2.1 *Soit X une variable aléatoire d'écart-type σ et de moment centré d'ordre 4, $\mu_4^{(c)} = E(X-\mu)^4$. On a :*

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

De plus, lorsque n tend vers l'infini, $Var(S_n^2) \simeq \frac{\mu_4^{(c)} - \sigma^4}{n}$.

Preuve. 1. L'espérance de S_n^2 est égale à $\frac{n-1}{n}\sigma^2$.

$1^{ére}$ cas X $-\mu$ centrée

La décomposition de S_n^2

$$\begin{aligned} S_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\mu) - \left(\bar{X}_n - \mu\right)\right]^2 \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(X_i-\mu)^2 + n\left(\bar{X}_n-\mu\right)^2 - 2\left(\bar{X}_n-\mu\right)\sum_{i=1}^{n}(X_i-\mu)\right] \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(X_i-\mu)^2 + n\left(\bar{X}_n-\mu\right)^2 - 2n\left(\bar{X}_n-\mu\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \left(\bar{X}_n-\mu\right)^2. \end{aligned} \tag{2.5}$$

Donc la linéarité de l'espérence donne :

$$\begin{aligned} E(S_n^2) &= \frac{1}{n}\sum_{i=1}^{n}E\left(X_i-\mu\right)^2 - E\left(\bar{X}_n-\mu\right)^2 \\ &= Var(X_i) - Var(\bar{X}_n) \\ &= \sigma^2 - \frac{\sigma^2}{n} \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

$2^{émé}$ cas X centrée, $\mu = 0$ et $\mu_2^{(c)} = \sigma^2$.

On peut écrire S_n^2 sous la forme :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i)^2 - (\bar{X}_n)^2. \tag{2.6}$$

Donc la linéarité de l'espérence donne :

$$\begin{aligned}
E(S_n^2) &= \frac{1}{n}\sum_{i=1}^{n} E(X_i)^2 - E(\bar{X}_n)^2 \\
&= \frac{n}{n}E(X_i)^2 - E\left[\frac{1}{n^2}\sum_{i=1}^{n} {X_i}^2 + 2n\sum_{i\neq j}^{n} X_i X_j\right] \\
&= \sigma^2 - \left[\frac{1}{n^2}\sum_{i=1}^{n} E{X_1}^2 + 2n\sum_{i\neq j}^{n} EX_i X_j\right] \\
&= \sigma^2 - \left[\frac{1}{n}\sigma^2 + 0\right] \qquad \text{car } X \text{ centrée} \\
&= \sigma^2 - \frac{1}{n}\sigma^2 \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}$$

2. La variance de S_n^2 est égale à $\dfrac{n-1}{n^3}\left[(n-1)\mu_4^{(c)} - (n-3)\sigma^4\right]$.

$1^{ére}$ cas $X\ -\mu$ centrée : La démonstration voir O.Wintenberger [10]

$2^{émè}$ cas X centrée :

Ainsi $\mu_1^{(c)} = 0$ et $\mu_2^{(c)} = \sigma^2$. En appliquant la forme 2.6 de S_n^2, on trouve :

$$\begin{aligned}
\sum_{i,j}^{n}(X_i - X_j)^2 &= \sum_{i,j}^{n}\left(X_i^2 - 2X_iX_j + X_j^2\right) \\
&= \sum_{j=1}^{n}\sum_{i=1}^{n} X_i^2 - 2\sum_{j=1}^{n}\sum_{i=1}^{n} X_iX_j + \sum_{i=1}^{n}\sum_{j=1}^{n} X_j^2 \\
&= 2n\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i \sum_{j=1}^{n} X_j \\
&= 2n\sum_{i=1}^{n} X_i^2 - 2\left(n\bar{X}_n\right)\left(n\bar{X}_n\right) \\
&= 2n^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right) \\
&= 2n^2 S_n^2.
\end{aligned}$$

On peut donc calculer la variance de S_n^2, en utilisant la relation suivante :

$$\begin{aligned}
Var(S_n^2) &= Cov(S_n^2, S_n^2) \\
&= \frac{1}{(2n^2)^2} \sum_{i,j,k,l}^{n} Cov((X_i - X_j)^2, (X_k - X_l)^2) \\
&= \frac{1}{(2n^2)^2} \left[\begin{array}{c} \sum_{i \neq j}^{n} Cov((X_i - X_j)^2, (X_i - X_j)^2) + \sum_{i \neq j \neq k}^{n} Cov((X_i - X_j)^2, (X_k - X_j)^2) \\ + \sum_{i \neq j \neq k \neq l}^{n} Cov((X_i - X_j)^2, (X_k - X_l)^2) \end{array} \right] \\
&= \frac{1}{(2n^2)^2} \left[\begin{array}{c} 2n(n-1) Cov((X_i - X_j)^2, (X_i - X_j)^2) + 4n(n-1)(n-2) \\ \times Cov((X_i - X_j)^2, (X_k - X_j)^2) + 0 \end{array} \right].
\end{aligned}$$

On remarque que si $i = j$ ou $k = l$, alors on obtient une covariance avec zéro (de la forme $Cov(0, (X_k - X_l)^2)$ ou $Cov((X_i - X_j)^2, 0)$) qui est nulle.

Commençons par le calcul de $Cov((X_i - X_j)^2, (X_i - X_j)^2)$ lorsque $(k = i, l = j)$ ou $(k = j, l = i)$ avec $i \neq j$.

$$Cov((X_i - X_j)^2, (X_i - X_j)^2) = E\left[(X_i - X_j)^4\right] - \left[E((X_i - X_j)^2)\right]^2.$$

On introduit la moyenne μ dans le calcul de l'espérance.

Utilisez le changement de variable suivant : $Y_i = (X_i - \mu)$, $Y_k = (X_k - \mu)$, et $Y_j = (X_j - \mu)$

$$\begin{aligned}
(X_i - X_j)^4 &= [(X_i - \mu) - (X_j - \mu)]^4 \\
&= Y_i^4 - 4Y_iY_j^3 + 6Y_i^2Y_j^2 - 4Y_i^3Y_j + Y_j^4 \\
E\left[(X_i - X_j)^4\right] &= 2\mu_4^{(c)} + 8\mu_1^{(c)}\mu_3^{(c)} + 6(\mu_2^{(c)})^2 \\
&= 2\mu_4^{(c)} + 6\sigma^4 \quad \text{car } \mu_1^{(c)} = 0 \text{ et } \mu_2^{(c)} = \sigma^2. \\
(X_i - X_j)^2 &= [(X_i - \mu) - (X_j - \mu)]^2 \\
&= Y_i^2 - 2Y_iY_j + Y_j^2 \\
E\left[(X_i - X_j)^2\right] &= 2\mu_2^{(c)} = 2\sigma^2.
\end{aligned}$$

Ainsi, pour $i \neq j$,

$$Cov((X_i - X_j)^2, (X_i - X_j)^2) = 2\mu_4^{(c)} + 2\sigma^4. \tag{2.7}$$

Continuons par le calcul de $Cov((X_i - X_j)^2, (X_k - X_j)^2)$ lorsque ($l = i$ *ou* $l = j$) avec i, j, k différents.

$$\begin{aligned}
&Cov((X_i - X_j)^2, (X_k - X_j)^2) \\
&= E\left[(X_i - X_j)^2 (X_k - X_j)^2\right] - \left[E((X_i - X_j)^2)E((X_k - X_j)^2)\right] \\
&= E\left[(X_i - X_j)^2 (X_k - X_j)^2\right] - (2\sigma^2)^2 \\
&(X_i - X_j)^2 (X_k - X_j)^2 \\
&= [(X_i - \mu) - (X_j - \mu)]^2 [(X_k - \mu) - (X_j - \mu)]^2 \\
&= \left[Y_i^2 - 2Y_iY_j + Y_j^2\right]\left[Y_k^2 - 2Y_kY_j + Y_j^2\right] \\
&= Y_i^2Y_k^2 - 2Y_iY_jY_k^2 + Y_jY_k^2 - 2Y_i^2Y_kY_j + Y_j^2Y_k^2 + 4Y_iY_kY_j^2 \\
&- 2Y_kY_j^3 + Y_i^2Y_j^2 - 2Y_iY_j^3 + Y_j^4.
\end{aligned}$$

$$E\left[(X_i - X_j)^2 (X_k - X_j)^2\right] = EY_i^2 EY_k^2 + EY_i^2 EY_j^2 + EY_j^2 EY_k^2 + EY_j^4 + 0$$
$$= 3(\mu_2^{(c)})^2 + \mu_4^{(c)} = 3\sigma^4 + \mu_4^{(c)}.$$

Ainsi, pour i, j, k différents,

$$Cov((X_i - X_j)^2, (X_k - X_j)^2) = \mu_4^{(c)} - \sigma^4. \tag{2.8}$$

Finalement, d'aprés 2.7 et 2.8 on obtient :

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

C'est le résultat désiré. ■

2.2.1 Variance empirique corrigée (S_n^{*2})

Pour obtenir l'estimateur S_n^{*2}, en multipliant l'estimateur S_n^2 fois $\frac{n}{n-1}$.

Définition 2.2.2 *La variance empirique corrigée en notée S_n^{*2}, définie par*

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

est un estimateur sans biais et convergent de σ^2.

Preuve. La moyenne de la variance empirique corrigée S_n^{*2} est égale à σ^2, en effet :

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

Donc S_n^{*2} est un estimateur sans biais de σ^2. ■

2.2.2 Comportement asymptotique

Théorème 2.2.1 *Soit $X_1, ..., X_n \sim X$ telle que $E(X^2) < \infty$, alors on a*

$$S_n^2 \xrightarrow{p.s.} Var(X) \quad quand\ n \longrightarrow \infty, \tag{2.9}$$

et

$$S_n^{*2} \xrightarrow{p.s.} Var(X) \quad quand\ n \longrightarrow \infty.$$

Preuve. On peut écrire S_n^2 sous la forme 2.6.

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

D'aprés L.G.N, on a

$$\bar{X}_n \xrightarrow{p.s.} E(X) \text{ alors } (\bar{X}_n)^2 \xrightarrow{p.s.} E^2(X) \quad et \quad \overline{X^2} \xrightarrow{p.s.} E(X^2).$$

Donc

$$\overline{X^2} - (\bar{X}_n)^2 \xrightarrow{p.s.} E(X^2) - E^2(X).$$

D'où

$$S_n^2 \xrightarrow{p.s.} Var(X).$$

De même façons S_n^{*2} converge aussi p.s. vers $Var(X)$. ■

Théorème 2.2.2 *Soit $X_1, ..., X_n \sim X$ telle que $E(X^4) < \infty$, alors on a*

$$\sqrt{n}\frac{(S_n^2 - \sigma^2)}{\sqrt{\mu_4^{(c)} - \sigma^4}} \xrightarrow{Loi} \mathcal{N}(0,1), \ quand\ n \longrightarrow \infty. \tag{2.10}$$

Preuve. D'aprés la décomposition 2.5 de $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \left(\bar{X}_n-\mu\right)^2$ on a :

$$\begin{aligned}\sqrt{n}\frac{(S_n^2-\sigma^2)}{\sqrt{\mu_4^{(c)}-\sigma^4}} &= \sqrt{n}\frac{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \left(\bar{X}_n-\mu\right)^2 - \sigma^2\right)}{\sqrt{\mu_4^{(c)}-\sigma^4}} \\ &= \sqrt{n}\frac{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \sigma^2\right)}{\sqrt{\mu_4^{(c)}-\sigma^4}} - \sqrt{n}\frac{\left(\bar{X}_n-\mu\right)^2}{\sqrt{\mu_4^{(c)}-\sigma^4}} \\ &= Z_n - W_n.\end{aligned}$$

D'aprés le théorème central limite, on peut ecrire $Z_n \xrightarrow{Loi} \mathcal{N}(0,1)$.Car

$$E((X_i-\mu)^2) = \mu_2^{(c)} = \sigma^2,$$

et

$$Var((X_i-\mu)^2) = E\left[(X_i-\mu)^4\right] - \left[E\left(X_i-\mu\right)^2\right]^2 = \mu_4^{(c)} - \sigma^4.$$

Reste à prouver que W_n est négligeable. On sait par le T.L.C classique, que

$$\sqrt{n}\left(\bar{X}_n-\mu\right) \xrightarrow{Loi} N(0,\sigma^2).$$

On conclut en utilisant L.G.N et le théorème de Slutsky avec

$$\bar{X}_n \xrightarrow{p.s.} \mu \Rightarrow \bar{X}_n \xrightarrow{P} \mu \;\text{ que } \sqrt{n}\left(\bar{X}_n-\mu\right)^2 \xrightarrow{P} 0.$$

Donc $W_n \xrightarrow{P} 0$, d'où le résultat. ■

Remarque 2.2.1 *La même convergence est vraie pour* S_n^{*2} :

$$\sqrt{n}\left(S_n^{*2}-\sigma^2\right) \xrightarrow{Loi} N(0,\mu_4^{(c)}-\sigma^4) \quad quand\ n \longrightarrow \infty.$$

2.3 Corrélation entre la moyenne empirique et la variance empirique

Théorème 2.3.1 *Si $(X_1, ..., X_n)$ un échantillon d'une v.a. X telque $E(X^3) < \infty$ et $\mu = 0$ (X centrée), alors*

$$Cov(\bar{X}_n, S_n^2) := \frac{n-1}{n^2}\mu_3^{(c)}.$$

Preuve. En utilisant la forme 2.6 de S_n^2, on a

$$\begin{aligned} Cov(\bar{X}_n, S_n^2) &= E(\bar{X}_n S_n^2) \\ &= E\left(\bar{X}_n(\overline{X^2} - (\bar{X}_n)^2)\right) \\ &= E\left(\bar{X}_n\overline{X^2}\right) - \mu_3^{(c)}(\bar{X}_n) \\ &= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} E(X_i)\right)\left(\frac{1}{n}\sum_{j=1}^{n} E(X_j^2)\right)\right] - \frac{\mu_3^{(c)}}{n^2} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E(X_i X_j^2) - \frac{\mu_3^{(c)}}{n^2}, \end{aligned}$$

car

$$E(X_i X_j^2) = \begin{cases} E(X_i^3) = \mu_3^{(c)} & \text{Si } i = j \\ E(X_i)E(X_j^2) = 0 & \text{Si } i \neq j \end{cases}$$

pour $i \neq j$ à cause de l'indépendance, on déduit :

$$\begin{aligned} Cov(\bar{X}_n, S_n^2) &= \frac{1}{n^2}\sum_{i=1}^{n}\mu_3^{(c)} - \frac{\mu_3^{(c)}}{n^2} \\ &= \left(\frac{1}{n} - \frac{1}{n^2}\right)\mu_3^{(c)} \\ &= \frac{n-1}{n^2}\mu_3^{(c)}. \end{aligned}$$

D'où la preuve. ■

Remarque 2.3.1 ***1.*** *Si la loi de X est symétrique, alors ($\mu_3^{(c)} = 0$) et $\bar{X}_n$ et S_n^2 sont non corrélées.*

2. *Si $\lim_{n\longrightarrow\infty} Cov(\bar{X}_n, S_n^2) = 0$, alors $\bar{X}_n$ et S_n^2 sont asymptotiquement non corrélées.*

2.4 Moments empiriques

Définition 2.4.1 *Soient $X_1, ..., X_n \sim X$ et $k \in \mathbb{N}^*$, alors le moment d'ordre k vaut $E(X^k)$ et sont notés μ_k. Le moment centrés d'ordre k vaut $E((X-\mu)^k)$ et sont notés $\mu_k^{(c)}$ où $\mu = \mu_1$ est la vraie moyenne. Ils ont des équivalent empiriques :*

$$M_n^k := \frac{1}{n}\sum_{i=1}^{n} X_i^k \quad et \quad M_n^{k\star} := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^k$$

appelés le moment empirique d'ordre k et le moment empirique centré d'ordre k.

Remarque 2.4.1

1. *Moyenne et variance sont des cas particuliers car $\mu = \mu_1$ et $\sigma^2 = \mu_2^{(c)}$, d'où les versions empiriques $M_n^1 = \bar{X}_n \quad et \quad M_n^{2\star} = S_n^2$.*
2. *Le moment centré d'ordre 1 ($\mu_1^{(c)}$) vaut toujours 0.*

Proposition 2.4.1 *Si $(X_1, ..., X_n)$ est un échantillon d'un v.a. X admet un moment μ_k d'ordre k, alors l'espérance et la variance de M_n^k est :*

$$E\left(M_n^k\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^k) = E(X^k) = \mu_k,$$

et

$$Var(M_n^k) = \frac{1}{n}\left[E(X^{2k}) - E^2(X^k)\right] = \frac{1}{n}\left(\mu_{2k} - \mu_k^2\right).$$

Par contre $E(M_n^{k\star}) \neq \mu_k$ et on peut corriger le moment empirique centré (cas $k = 2$ où $M_n^{2\star} = S_n^2$).

2.4.1 Loi asymptotique du moment empirique

Une application du T.L.C nous donne la loi asymptotique des moments (centrés ou non) :

Proposition 2.4.2 *Si $E(X^{2k}) < \infty$, i.e. μ_{2k} existe, alors*

$$\sqrt{n}\frac{\left(M_n^k - \mu_k\right)}{\sqrt{\mu_{2k} - \mu_k^2}} \xrightarrow{Loi} \mathcal{N}(0,1) \qquad quand\ n \longrightarrow \infty.$$
$$\sqrt{n}\frac{\left(M_n^{k\star} - \mu_k^{(c)}\right)}{\sqrt{\mu_{2k}^{(c)} - (\mu_k^{(c)})^2}} \xrightarrow{Loi} \mathcal{N}(0,1) \qquad quand\ n \longrightarrow \infty.$$

2.5 Fonction de distribution empirique

Définition 2.5.1 *La fonction de distribution empirique F_n basée sur l'échantillon $X_1, ..., X_n$ est définie par :*

$$F_n(x) := \frac{nombre\ d'observation \leq x\ dans\ l'échantillon}{n} := \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}, \quad x \in \mathbb{R}.$$

2.5.1 Loi de $F_n(x)$ avec $x \in \mathbb{R}$

Pour tout $x \in \mathbb{R}$ fixé, posons $Y_i = 1_{\{X_i \leq x\}}$. Il est facile de voir que $F_n(x) = \bar{Y}_n$ et que $Y = 1_{\{X \leq x\}}$ est une variable aléatoire valant soit 0 soit 1, donc $Y \sim B(p)$ avec $p = P(Y = 1) = E(Y)$, d'où

$$p = E(Y = 1_{\{X \leq x\}}) = P(X \leq x) = F(x)$$

On déduit, d'après l'étude de la loi de la moyenne empirique dans le cas Bernoulli que

$$nF_n(x) \rightsquigarrow B(n, F(x))$$

De plus,

$$E(F_n(x)) = E(\bar{Y}_n) = E(Y) = F(x).$$
$$Var(F_n(x)) = Var(\bar{Y}_n) = \frac{Var(Y)}{n} = \frac{F(x)(1-F(x))}{n}.$$

2.5.2 Loi asymptotique de $\mathbf{F}_n(\mathbf{x})$

En appliquant le T.L.C aux $Y_i = 1_{\{X_i \leq x\}}$ et la loi des grands nombres sur $F_n(x)$ on trouve les théorèmes suivants :

Théorème 2.5.1 *Soit $F = F_X$ la fonction de répartition de X, alors $\forall x \in \mathbb{R}$*

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1-F(x))}} \xrightarrow{Loi} \mathcal{N}(0,1), \quad \textit{quand } n \to \infty. \tag{2.11}$$

Théorème 2.5.2 *Pour tout $x \in \mathbb{R}$*

$$F_n(x) \xrightarrow{p.s.} F(x), \quad \textit{quand } n \to \infty. \tag{2.12}$$

Preuve. D'aprés la loi des grands nombres, on a

$$F_n(x) \xrightarrow{p.s.} E\left(\frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}\right) = E\left(1_{\{X_1 \leq x\}}\right)$$

avec $E\left(1_{\{X_1 \leq x\}}\right) = F(x)$, d'où le résultat. ■

Dans le cas des statistique non paramétrique utiliser la théorème de **Glivenko-Cantelli**.

Théorème 2.5.3 (Glivenko-Cantelli) *La fonction de répartition empirique F_n converge uniformément vers la fonction de répartition F, ou bien, de manière équivalente :*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{p.s.} 0, \ \textit{quand } n \to \infty.$$

La démonstration théorème voir[16]

2.6 Echantillons gaussiens

On considère que la variable X suit une loi normale de moyenne $\mu \in \mathbb{R}$ et d'écart-type $\sigma > 0$, $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$. Soit $(X_1, ..., X_n)$ un échantillon aléatoire de X.

2.6.1 Etude de la moyenne et la variance de l'échantillon

Loi de la statistique $\bar{X}_n$

La statistique $\bar{X}_n$ est une combinaison linéaire de n variables aléatoires gaussiennes indépendantes. C'est donc une variable gaussienne : $\bar{X}_n \rightsquigarrow \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Propriétés 2.6.1 *Si* $\bar{X}_n \rightsquigarrow \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ *et* $X_i - \bar{X}_n \rightsquigarrow \mathcal{N}\left(0, \sigma\sqrt{\frac{n-1}{n}}\right)$, *alors*

$$Cov(\bar{X}_n, X_i - \bar{X}_n) = 0.$$

Théorème 2.6.1 $\bar{X}_n$ *et* $X_i - \bar{X}_n$, $\forall i = 1, ..., n$, *sont indépendantes.*

Loi de la statistique $\mathbf{S}_n^2$

Définition 2.6.1 (Loi du Chi-deux) *Soit* $Z_1, ..., Z_\nu$ *une suite de variables aléatoires i.i.d de loi* $\mathcal{N}(0, 1)$. *Alors la v.a.* $\sum_{i=1}^{\nu} Z_i^2$ *suit une loi appelée* ***loi du Chi-deux****,* χ^2, *à* ν *degrés de liberté :*

$$\chi^2(\nu) = \sum_{i=1}^{\nu} Z_i^2 = \sum_{i=1}^{\nu} \left(\frac{X_i - \mu}{\sigma}\right)^2.$$

En utilisant la décomposition 2.5 de S_n^2 et en divisant par σ^2 on obtient :

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{nS_n^2}{\sigma^2} + \left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}\right)^2.$$

Le premier membre est une somme de carrés de variables normales centrées réduites indépendantes donc une variable du Chi-deux $\chi^2(n)$.

Le second membre est la somme de deux formes quadratiques sur ces variables, l'une de rang $n-1$, l'autre de rang 1. Le deuxième terme est le carré d'une variable normale centrée réduite donc une variable du Chi-deux à un degré de liberté.

On en déduit les deux théorème suivants :

Théorème 2.6.2 *La loi de la variable* $\frac{nS_n^2}{\sigma^2}$ *est une loi du Chi-deux à* $n-1$ *degrés de liberté :*

$$\frac{nS_n^2}{\sigma^2} \rightsquigarrow \chi^2(n-1).$$

Théorème 2.6.3 *Les statistiques* $\bar{X}_n$ *et* S_n^2 *sont indépendantes.*

Voir Veysseyre, Renée [17]

2.6.2 Théorème de Fisher

Théorème 2.6.4 (Théorème de Fisher) *Soit* $(X_1, ..., X_n)$ *des variables aléatoires indépendantes et de même loi* $\mathcal{N}(0,1)$, *les variables*

$$\sqrt{n}\bar{X}_n \quad et \quad \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 = nS_n^2 = (n-1)S_n^{*2}$$

sont indépendantes suivent respectivement $\mathcal{N}(0,1)$ *et* $\chi^2(n-1)$.

La démonstration de théorème. Voir [15].

Corollaire 2.6.1 *soit un échantillon* $(X_1, ..., X_n)$ *i.i.d extrait d'une loi* $\mathcal{N}(\mu, \sigma^2)$. *Alors* $\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ *et* $(n-1)\frac{S_n^{*2}}{\sigma^2}$ *sont indépendantes suivent respectivement* $\mathcal{N}(0,1)$ *et* $\chi^2(n-1)$.

La démonstration. Voir [15]

Conséquences du théorème Fisher

1. Calcul de $Var(S_n^2)$ et $Var(S_n^{*2})$

On a $(n-1)\frac{S_n^{*2}}{\sigma^2} \rightsquigarrow \chi^2(n-1)$ et $\frac{nS_n^2}{\sigma^2} \rightsquigarrow \chi^2(n-1)$. Alors

$$Var((n-1)\frac{S_n^{*2}}{\sigma^2}) = 2(n-1) = \frac{(n-1)^2}{\sigma^4}Var(S_n^{*2}).$$
$$Var(S_n^{*2}) = \frac{2\sigma^4}{n-1}.$$
$$Var(\frac{nS_n^2}{\sigma^2}) = 2(n-1) = \frac{n^2}{\sigma^4}Var(S_n^2).$$
$$Var(S_n^2) = \frac{2(n-1)\sigma^4}{n^2}.$$

2. Loi de Student

Définition 2.6.2 *Soit Z et Q deux v.a. indépendantes telles que $Z \rightsquigarrow \mathcal{N}(0,1)$ et $Q \rightsquigarrow \chi^2(\nu)$. Alors la v.a.*

$$T = \frac{Z}{\sqrt{\frac{Q}{\nu}}}$$

suit une loi appelée ***loi de Student à*** *ν degrés de liberté, notée $\mathcal{T}(\nu)$.*

Puisque $\frac{\bar{X}_n-\mu}{\sigma}\sqrt{n} \rightsquigarrow \mathcal{N}(0,1)$ et $\frac{nS_n^2}{\sigma^2} \rightsquigarrow \chi^2(n-1)$ on aura :

$$T = \frac{\frac{\bar{X}_n-\mu}{\sigma}\sqrt{n}}{\sqrt{\frac{nS_n^2}{(n-1)\sigma^2}}} = \frac{\bar{X}_n-\mu}{S_n}\sqrt{(n-1)}.$$

où T est une variable de Student à $n-1$ degrés de liberté. Ce résultat est extrêmement utile car il ne dépend pas de σ et servira donc chaque fois que σ est inconnu, on écrit :

$$\frac{\bar{X}_n-\mu}{S_n}\sqrt{(n-1)} \rightsquigarrow \mathcal{T}(n-1).$$

Chapitre 3

Application sous R

La simulation est une méthode de mesure et d'étude consistant à remplacer un phénomène, un système par un modèle plus simple mais ayant un comportement analogue. La simulation est aussi une étape essentielle dans la plupart des études. Dans ce chapitre, nous allons présenter quelques graphiques qui reflètent la simulation numérique des résultats des deux chapitres précédents, à l'aide du logiciel d'analyse statistique (R).

3.1 Moyenne empirique

Nous avons étudiée le comportement asymptotique de la moyenne empirique (voir 2.3 et 2.4) :

1 Loi forte des grands nombres : Pour des échantillons de la loi uniforme sur $[0, 10]$ de tailles variant jusqu'à $n = 1000$, calculons les moyennes empiriques successives et traçons la moyenne empirique en fonction de la taille de l'échantillon. La figure 3.1 représente la convergence de la moyenne empirique vers la valeur $\mu = 5$ (la moyenne théorique).

Programme sous R :

```
n=1000 ; d=1 :n ; m=numeric(n)
for(i in 1 :n){
x=runif(i,0,10) ; m[i]=mean(x)}
```

plot(d,m,xlab="Taille d'échantillon",ylab="Moyenne empirique",main="Loi des grands nombres ",col='4')

abline(h=5,lwd=3,col="red") #Tracer un droite horizontale rouge

text(300,7,expression(mu==5),col="red")

legend(500,8,legend=c("Moyenne empirique", "Moyenne théorique"),lwd=c(1,2), lty=c(3,1), col=c('4', 'red'))

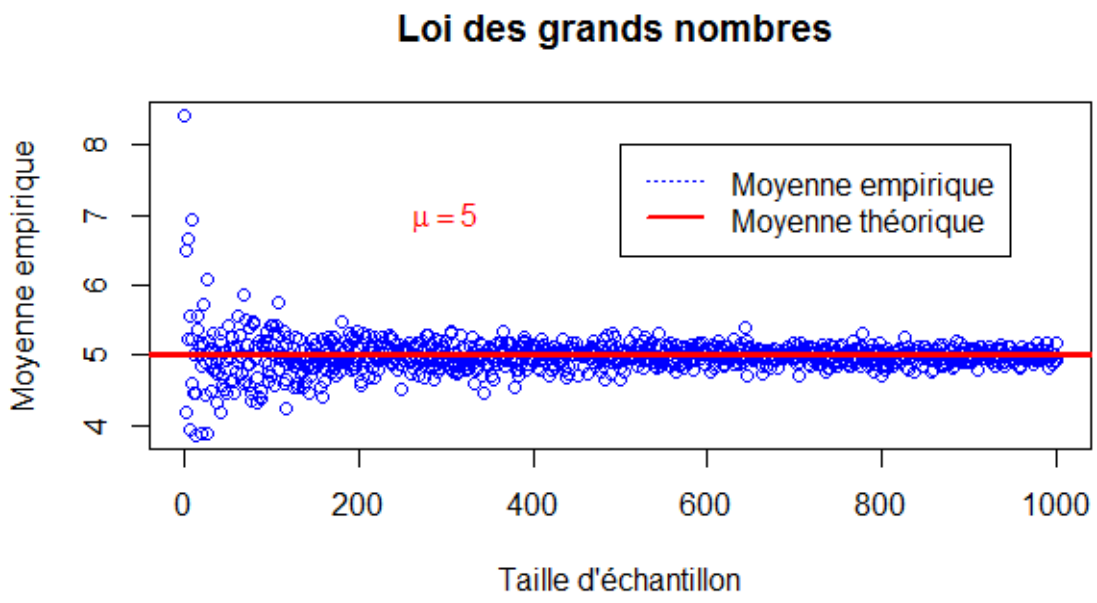

FIG. 3.1 – L.G.N. la convergence p.s. de la moyenne empirique.

Nous remarquons dans la graphe 3.1 la moyenne empirique converge bien vers la valeur $\mu = 5$ (la moyenne théorique) représentée par une droite horizontale rouge.

2. Théorème de la limite centrale : Soit une distribution uniforme sur $[0, 10]$ avec moyenne $\mu = 5$ et écart-type $\sigma = \frac{10}{\sqrt{12}}$. On tire un échantillon de grandeur $n = 1000$. La figure 3.2 représente la distribution uniforme dans la première graphe, nous avons comparer l'histogramme des moyennes (la distribution de la moyenne de l'échantillon) par la densité de la loi $\mathcal{N}(5, \frac{100}{12000})$, dans la deuxième graphe.

Programme sous R :

```
par(mfrow=c(1,3))
Nbsimul<-1000 ; taille<-900
t=runif(taille,0,10)
hist(t, freq = FALSE, ylim = c(0, 0.2),col='blue',main="Distribution uniforme")
m<-5 ; sigma<-sqrt(100/12)
Result<-NULL ; M=numeric(taille)
for(i in 1 :Nbsimul){
tirage<-runif(taille,0,10) # Générer échantillon d'une loi uniforme
Result[i]<-mean(tirage) # Calculer la moyenne de chaque échantillon
M[i]=sqrt(taille)*((Result[i]-5)/sqrt(100/12))}
hist(Result,freq=FALSE,col='blue', ylim=c(0,dnorm(m,m,sigma/sqrt(taille))), main = "L'ap-
proximation normale",xlab="Moyennes observées",ylab="Fréquences observées")
curve(dnorm(x,m,sigma/sqrt(taille)),add=T,col='red')
qqnorm(M,col=3)
qqline(M,lwd='2')
A=c(-3,3) ;B=c(-3,3)
lines(A,B,col='2',lwd='2')
```

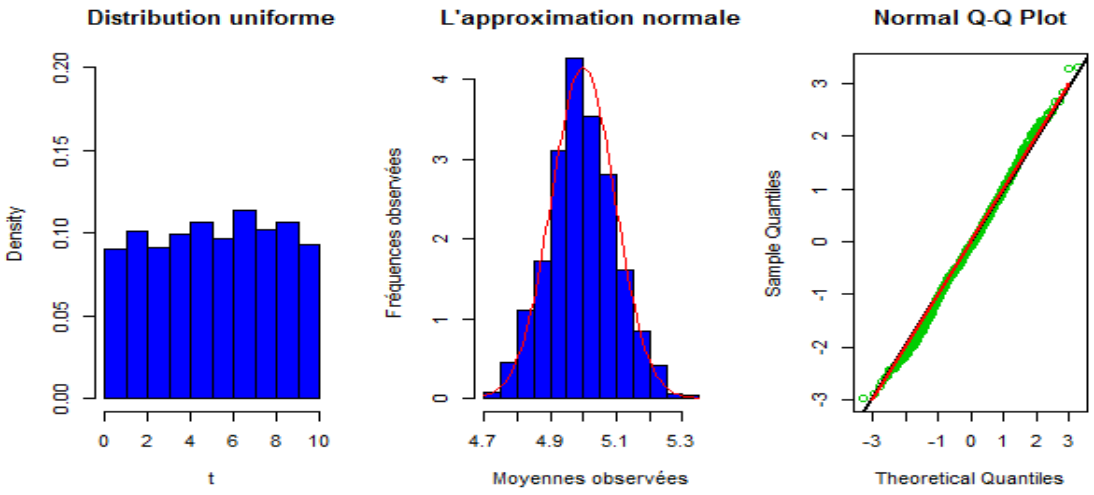

FIG. 3.2 – T.C.L.. la convergence en loi de la moyenne empirique.

Nous remarquons dans la deuxième graphe 3.2 la distribution de la moyenne de l'échantillon approchera une distribution normale $\mathcal{N}(5, \frac{100}{12000})$ lorsque n tend vers l'infini. On a déjà une bonne approximation ($\bar{X}_n \rightsquigarrow \mathcal{N}(\mu, \frac{\sigma^2}{n})$) lorsque $n \geq 30$. Dans la troisième graphe, nous utilisons l'instruction ($qqnorm$) pour une bonne comparaison entre la distribution de $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$ et la distribution normale $\mathcal{N}(0,1)$, nous remarquons dans ce graphe la distribution d'échantillon des moyennes s'approche d'une distribution normale $\mathcal{N}(0,1)$ quand n est grand.

3.2 Variance empirique

Nous avons étudiée le comportement asymptotique de la variance empirique (voir 2.9 et 2.10) :

1 Loi forte des grands nombres : Pour des échantillons de la loi exponentielle de paramètre 2 de tailles variant jusqu'à $n = 1000$, calculons les variances empiriques successives et traçons la variance empirique en fonction de la taille de l'échantillon. La figure 3.3 représente la convergence de la variance empirique vers la valeur $\sigma^2 = \frac{1}{4}$ (la variance théorique).

Programme sous R :

```
n=1000 ; d=1 :n ; S=numeric(n)
for(i in 1 :n){
x=rexp(i,2) ;S[i]=var(x)}
plot(d,S,xlab="Taille d'échantillon",ylab="Variance empirique",main="Loi des grands nombres ",col='4')
abline(h=1/4,lwd=3,col="red")
text(300,0.5,expression(sigma^2==1/4),col="red")
legend(550,0.4, yjust=0,c("Variance empirique", "Variance théorique"),lwd=c(1,2), lty=c(3,1), col=c('4', 'red'))
```

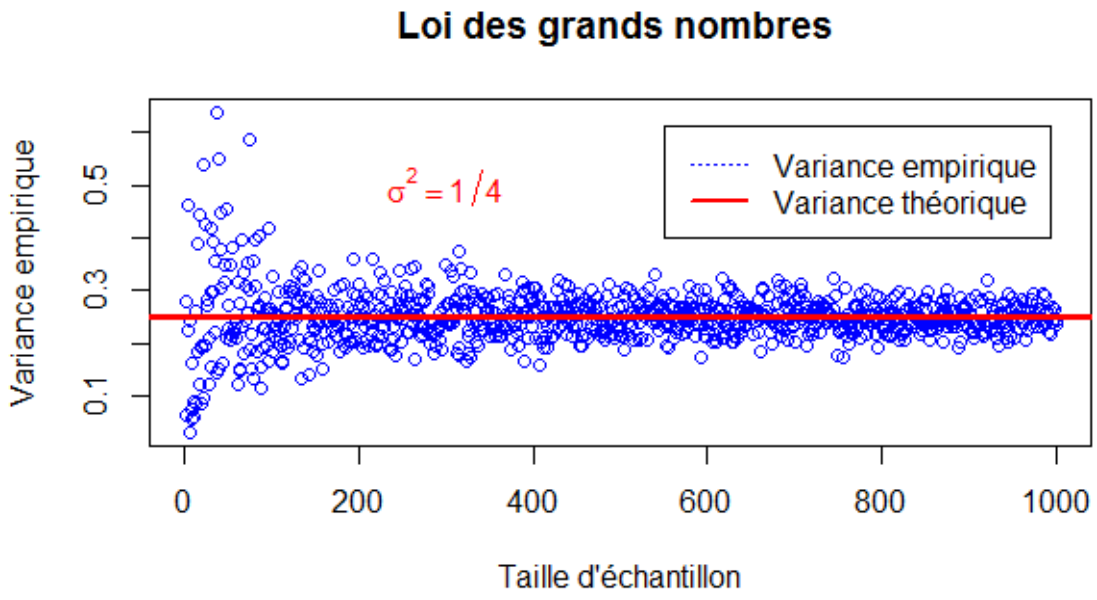

FIG. 3.3 – L.G.N. la convergence p.s. de la variance empirique.

Nous remarquons dans la graphe 3.3 la variance empirique converge bien vers la valeur $\sigma^2 = \frac{1}{4}$ (la variance théorique) représentée par une droite horizontale rouge.

2. Théorème de la limite centrale : Soit une distribution exponentielle de paramètre 2 avec moyenne $\mu = \frac{1}{2}$ et écart-type $\sigma = \frac{1}{4}$. On tire un échantillon de grandeur $n = 1000$. La figure 3.4 représente la distribution exponentielle dans la première graphe, nous avons comparer l'histogramme des variances (la distribution de la variance de l'échantillon) par la densité de la loi $\mathcal{N}\left(\sigma^2, \sqrt{\frac{\mu_4^{(c)} - \sigma^4}{n}}\right)$, dans la deuxième graphe.

Programme sous R :

```
par(mfrow=c(1,3))
Nbsimul<-1000 ; taille<-900
t=rexp(taille,2)
hist(t, freq = FALSE, ylim = c(0, 0.2),col='blue',main="Distribution exponentielle ")
Eloi<-1/2 ;Vloi<-1/4
VS<-((9/16)-(1/16))/taille
```

```
S=numeric(taille) ;sq=numeric(taille)
for(i in 1 :Nbsimul){
tirage<-rexp(taille,2)
S[i]<-var( tirage)
sq[i]=sqrt(taille)*((S[i]-(1/4))/sqrt(0.5))}
hist(S,freq=FALSE,col='blue', ylim=c(0,dnorm(Vloi,Vloi,sqrt(VS))), main = "L'approximation normale ",xlab="Varaince observées",ylab="Fréquences observées")
curve(dnorm(x,Vloi,sqrt(VS)),add=T,col='red') #
qqnorm(sq,col=3)
qqline(sq,lwd='2')
a=c(-3,3) ;b=c(-3,3)
lines(a,b,col='2',lwd='2')
```

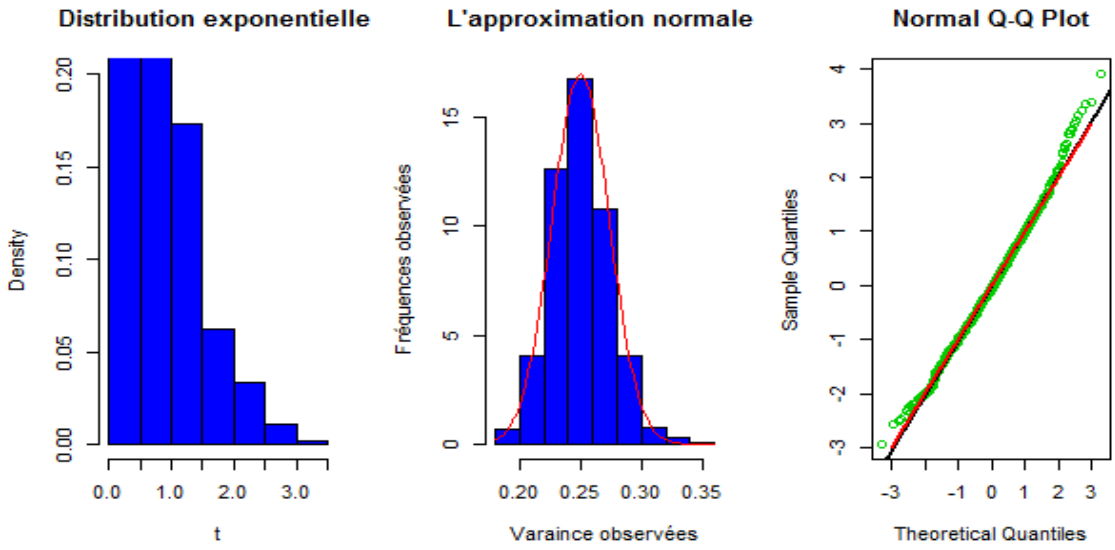

FIG. 3.4 – T.C.L.. la convergence en loi de la variance empirique.

Nous remarquons dans la deuxième graphe 3.4 la distribution de la variance de l'échantillon approchera une distribution normale $\mathcal{N}\left(\sigma^2, \sqrt{\frac{\mu_4^{(c)}-\sigma^4}{n}}\right)$ lorsque n tend vers l'infini. Dans la troisième graphe, nous utilisons l'instruction ($qqnorm$) pour une bonne comparaison entre la distribution de $\sqrt{n}\frac{(S_n^2-\sigma^2)}{\sqrt{\mu_4^{(c)}-\sigma^4}}$ et la distribution normale $\mathcal{N}(0,1)$, nous remarquons dans

ce graphe la distribution d'échantillonnage des variances s'approche d'une distribution normale $\mathcal{N}(0,1)$ quand n est grand.

3.3 Fonction de distribution empirique

Nous avons étudiée le comportement asymptotique de la fonction de distribution empirique (voir 2.11 et 2.12)

1.Convergence p.s. de $\mathbf{F}_n$ vers F : Pour un échantillon de la loi exponentielle de paramètre 1 et de taille $n = 200$, Dans la figure 3.5, nous allons tracer la distribution empirique par taille d'échantillon Puis nous étudions la convergence de la distribution empirique vers la distribution théorique.

Programme sous R :

```
n=200 ; y=numeric(n) ; z=numeric(n) ; X=rexp(n)
Fn=function(x){
for(i in 1 :n){
if (X[i]<=x) y[i]=1 else y[i]=0 }
mean(y)}
for(i in 1 :n){
z[i]=Fn(X[i])}
plot(X,z,ylab="Fonction de distribution",main="La convergence p.s de Fn vers F",col=4,lwd=2)
x = seq(min(X),max(X),0.1)
lines(x,pexp(x),col=2,lwd=2)
legend(3,0.4,legend=c("Distribution empirique Fn", "Distribution théorique F"),lwd=c(1,2),
lty=c(3,1), col=c('4', 'red'))
```

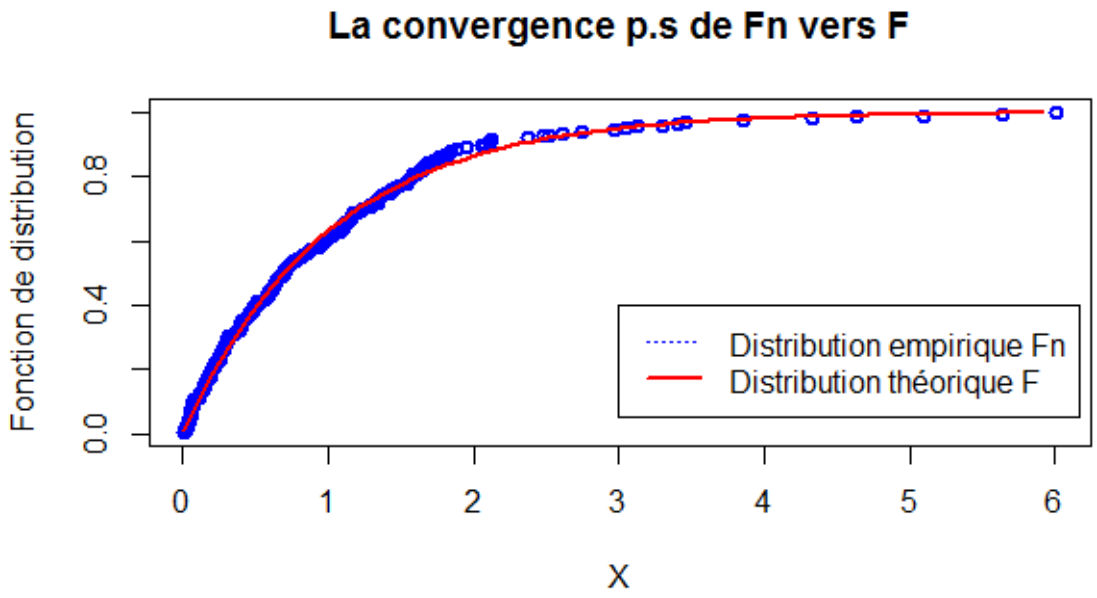

FIG. 3.5 – La convergence de la distribution empirique vers la distribution théorique.

Nous remarquons dans la figure 3.5, que les deux distributions sont trés proche. De cela, nous concluons une bonne convergence de la distribution empirique vers la distribution théorique.

2. Théorème de la limite centrale : Pour un échantillon de la loi exponentielle de paramètre 1 et de taille $n = 200$, Dans la figure 3.6, nous avons comparé la distribution de $\frac{\sqrt{n}(F_n(x)-F(x))}{\sqrt{F(x)(1-F(x))}}$ (pour $t = 1$) avec la distribution normale $\mathcal{N}(0, 1)$.

Programme sous R :

```
n=200 ; a=numeric(n)
for (i in 30 :n){
x=rexp(i)
Fn=ecdf(x)
t=1 ; F=pexp(t)
sd=sqrt(Fn(t)*(1-Fn(t))/i)
a[i]=(Fn(t)-F)/sd}
a=a[30 :n]
```

qqnorm(a,col=3, xlab="Quantile empirique",ylab="Quantile théorique")

qqline(a,col=2,lwd=2)

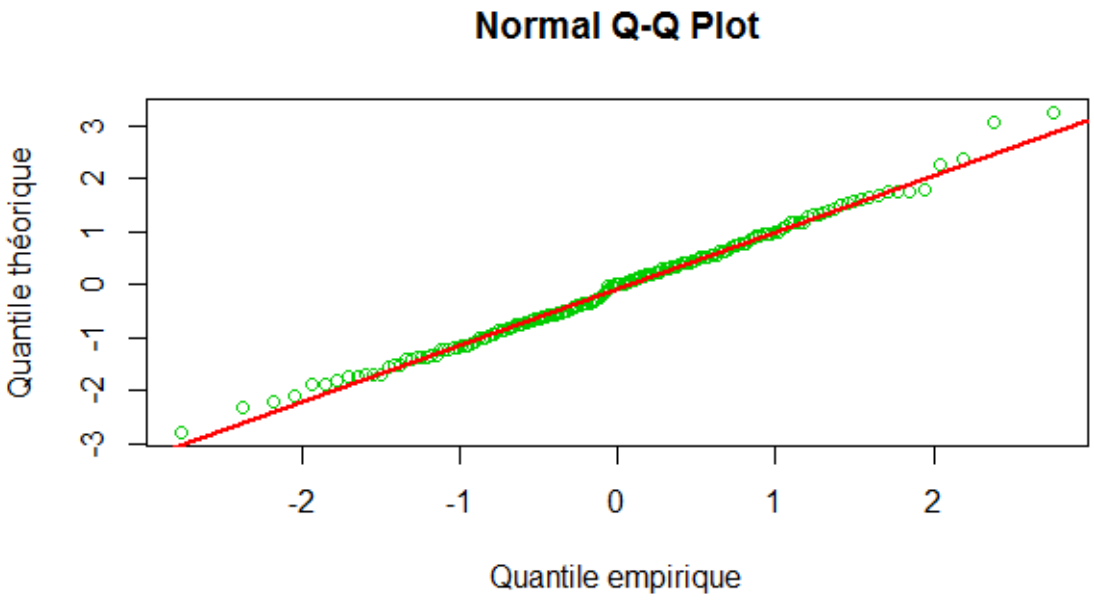

FIG. 3.6 – La convergence de la distribution empirique vers la distribution normale

Nous remarquons dans la graphe 3.6 la distribution empirique $\frac{\sqrt{n}(F_n(x)-F(x))}{\sqrt{F(x)(1-F(x))}}$ s'approche d'une distribution normale $\mathcal{N}(0,1)$ quand n est grand. Chaque fois que nous changeons t nous obtenons le même résultat.

Conclusion

L'objectif de l'échantillonnage est une bonne connaissance de la population à travers une étude d'échantillon avec les mêmes caractéristiques que la population et généralisation des résultats à la population. Dans ce mémoire nous avons étudié la théorie d'échantillonnage et ses applications.

Dans ce travail, nous avons présenté quelques méthodes d'échantillonnage en mettant l'accent sur l'étude de la distribution statistique des caractéristiques de l'échantillon (moyenne empirique, variation empirique, fonction de distribution empirique) et les comportements asymptotiques de chaque distribution. Enfin, nous soutenons notre étude par le processus de simulation numérique pour vérifier les comportements asymptotiques étudiés dans ce mémoire.

Nous avons conclu de cette étude simple que l'échantillonnage est une étape fondamentale dans toutes les expériences scientifiques et peut être utilisé pour résoudre de nombreux problèmes scientifiques. Dans un proche avenir, nous espérons que les chercheurs approfondiront leurs études de cette stratégie.

Bibliographie

[1] A A Borovkov-Mathematical statistics-CRC Press (1999).

[2] Barra J -R , Link Yu V-Notions fondamentales de statistique mathematique-Dunod (1971).

[3] Borovkov, A -Statistique Mathématique-Edition Mir Moscou (1987) .

[4] (Cours 1ère année) Thérèse Phan-Probabilités et statistique. 2-Ecole Centrale Paris (2005).

[5] Dagnelie, Pierre. Statistique théorique et appliquée. Belgium : Presses agronomiques de Gembloux, 1992.

[6] Hosmer, D. W., and S. Lemeshow. "Wiley series in probability and statistics. Texts and references section." (2000).

[7] Jean-Pierre Lecoutre-Statistique et probabilités _ Travaux dirigés-Dunod (2008).

[8] Lejeune, Michel. "Tests d'hypothèses paramétriques." Statistique-La thèorie et ses applications (2010) : 201 − 250.

[9] http ://math.univ-lyon1.fr/~gelineau/files/loi_forte_grands_nombres.pdf.

[10] O. Wintenberger. Statistique Mathématique http ://wintenberger fr/cours/L3MAS/2011/StatMathPoly2011 pdf.

[11] Probabilités et statistique avec R https ://ljk imag fr/membres/Bernard Ycart/mel/dr/node11 html.

[12] (Statistique et probabilités appliquées) Michel Lejeune-Statistique. La théorie et ses applications, Deuxième dition-Springer (2010).

[13] Saporta, Gilbert. Probabilités, analyse des données et statistique. Editions Technip, (2006).

[14] Schmuller_Statistical Analysis with R For Dummies_FD-2017.

[15] Tassi, Philippe. "Méthodes statistiques." (2004).

[16] Théorème_de_Glivenko-Cantelli.pdf
"https ://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_de_Glivenko-Cantelli"

[17] Veysseyre, Renée. "Aide-mémoire Statistique et probabilités pour l'ingénieur, Dunod,2006." Détection CFCAR en Milieux Non-Gaussiens Corrélés.

[18] Ihaka, R., Gentleman, R. (1996) *R : A Language for Data Analysis and Graphics.* Journal of Computational and Graphical Statistics **5** : 299 − 314.

Annexe A : Logiciel *R*

R est un logiciel permettant de faire des analyses statistiques et de produire des graphiques. Il dispose d'une bibliothèque très large de fonctions et de procédures statistiques, d'autant plus large qu'il est possible d'en intégrer de nouvelles par le système des packages. Il existe plusieurs versions de R telle que R.3.3.2 que l'on a utilisé dans l'application de ce mémoire.

Dans notre travail on utilise des fonctions et des procédures pour des simulations numériques des résultats de la théorie de l'échantillonnage et pour présente des graphiques.

Fonctions et procédures utilisées :

plot(x,y) : permet de dessiner la relation entre les variable X et Y.

abline : Tracer un droite.

text : Ajouter du texte à une figure.

legend : cette fonction peut être utilisée pour ajouter des légendes au plot.

qqnorm : comparer la quantile empirique avec celui de la loi normale graphiquement.

qqline : Tracer un droite entre le premier et le troisième quartiles.

lines(X,Y) : pour dessiner une ligne relié les point de l'echantillon (X,Y).

ylim=c(,) : Cette fonction définit les coordonnées du monde pour une fenêtre graphique au niveau de l'axe vertical.

curve : Dessine une courbe correspondant à une fonction sur l'intervalle [a, b].

hist : Dessiner une fonction dans un forme histogramme.

Annexe B : Abréviations et Notations

Les différentes abrévariations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous.

E	population de taille finie N.
ξ_n	échantillon de E.
e_i	unité de la population E.
e_{i_k}	unité de la l'échantillon ξ_n.
$(X_1, ..., X_n)$	échantillon de taille n de v.a's.
$X_1, ..., X_n \sim X$	un échantillon aléatoire d'une v.a. X.
$\bar{X}_n$,$E(X)$ ou μ	moyenne empirique et espérance mathématique.
S_n^2 ,$Var(X)$ ou σ^2	variance empirique et variance mathématique.
S_n^{*2}	variance empirique corrigée.
$b_\theta(T)$	biais de l'estimateur T.
θ	paramètre à estimer.
X $V.a.$	variables aléatoires.
(Ω, B, P)	espace de probabilité.
$(\mathcal{L}^n, \alpha_n, P_\theta^n)$	espace de probabilité produit de n expériences.
$\xrightarrow{P}$	convergence en probabilité.
$\xrightarrow{Loi}$	convergence en loi, convergence en distribution.

$\xrightarrow{p.s}$	convergence presque sûre.
$\xrightarrow{L^p}$, $\xrightarrow{L^2}$	convergence en moyenne quadratique d'ordre p. et d'ordre 2.
$i.i.d.$	indépendantes et identiquement distribués.
$T.L.C$	théorème limite central.
$L.G.N.$	lois des grands nombres.
μ_k, $\mu_k^{(c)}$	moment simple et moment centré d'ordre k respectivement.
M_n^k, $M_n^{k\star}$	moment empirique et moment centré empirique d'ordre k respectivement.
$B(p)$	loi bernoulli de paramètre p.
$F(X)$	la distribution théorique.
$F_n(X)$	la distribution empirique.
$1\{X_i \leq x\}$	fonction indicatrice de l'ensemble $\{X_i \leq x\}$.
$\varphi_X(.)$	fonction caractéristique de X .
$\mathcal{N}(\mu, \frac{\sigma^2}{n})$	loi normale d'espérence μ et d'écart type $\frac{\sigma^2}{n}$.
$\simeq$	à peu près égale
$\chi^2(\nu)$	loi khi-deux à ν degrés de liberté.
$\mathcal{T}(\nu)$	loi Studentà ν degrés de liberté.
Z_i	variables aléatoires de loi normale centré réduite.
$\mathbb{R}$	ensemble des nombres réels.
$\mathbb{N}^*$	ensemble des entiers naturelle non nule.
$\rightsquigarrow$	La distribution se rapproche d'une autre distribution.
$:=$	éngallité par définition.