.

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

MIHI Samah

Titre :

Théorie de l'échantillonnage et application

Membres du Comité d'Examen :

Dr.	**BENATIA Fateh**	UMKB	Président
Dr.	**SAYAH Abdallah**	UMKB	Encadreur
Dr.	**BERKANE Hassiba**	UMKB	Examinateur

Juin 2019

Dédicace

Je dédie ce travail :

A ma chère mére

A mon cher père

A mes fréres et soeurs

Qui m'ont toujours soutenu et encouragé

A mes amis et collégues et tous ceux qui m'a aidé

Et à tous ceux qui m'ont soutenu

Samah

REMERCIEMENTS

Tout d'abord, je remercie Dieu le tout-puissant qui m'a donné la force et le savoir afin d'accomplir ce travail.

Un grand merci pour mon encadreur **Sayah Abdallah** pour son encouragement et son suivi attentif pour la réalisation de ce travail.

Je tiens aussi à remercier également tous les membres de jury **Benatia Fateh** et **Berkane Hassiba** pour avoir accepter d'évaluer mon travail.

Je remercie tous les enseignants qui ont contribués à ma formation, ainsi que tous les employés du département de Mathématiques.

A mes familles, surtout mes parents qui nous ont épaulés, soutenus et suivis tout au long de ce projet.

A mes chéres amis qui ont toujours été présents et fidéles.

Enfin, pour toute personne qui a contribué, de prés ou de loin, à l'élaboration de ce mémoire. Veuillez bien trouver ici l'expression de mes sincéres remerciements.

Table des matières

Table des figures

Introduction

Pour effectuer une étude statistique, on se sert généralement d'un échantillon. Celui-ci doit refléter le plus exactement possible l'image de la population.

L'échantillonnage c'est choisier une partie d'un population pour représenter l'ensemble de la population, est fondamental et résulte de l'impossibilité de collecter des données sur tous les éléments d'une population, souvent pour des raisons pratiques, techniques ou économiques. L'échantillonnage consiste essentiellement à tirer des informations d'une fraction d'un grand groupe ou d'une population, de façon à en tirer des conclusions au sujet de l'ensemble de la population. Son objet est donc de fournir un échantillon qui représentera la population et reproduira aussi fidèlement que possible les principales caractéristiques de la population étudiée.

les principaux avantages de la technique d'échantillonnage sont le moindre coût et gain de temps, la rapidité, la portée et la précision accrues, en effet avec un échantillon on peut obtenir des résultats plus exacts car il est plus facile de contrôler les sources d'erreurs liées à la fiabilité, à la clarté des instructions aux mesures et à l'enregistrement et au traitement et à l'analyse des données, définir les modalités de l'échantillonnage consiste à définir la localisation, préciser les objectifs de recherche, identification de la population d'origine à partir de laquelle on sélectionne l'échantillon, détermination des caractéristiques de la population, sélectionner la taille de l'échantillon.

Ce mémoire est organisé en trois chapitres qui sont structurés selon la maniére suivante : Le premier chapitre intitulé notions de base en statistique et échantillonnage, compor-

tant les définition de bases nécessaires à la bonne compréhension dans la statistique, puis nous avons parlé sur les défférentes méthodes d'échantillonnage et le modéle d'échantillonnage,nous avons présenté les deux théorémes fondamentaux de la statistique asymptotique qui sont les plus importantes dans notre étude.

Le deuxiéme chapitre nommé distributions des caractéristiques d'un échantillon porte une étude focalisée sur distributions des certaines caractéristiques et propriétés d'un échantillon aléatoire, et les plus importantes (la moyenne empirique, la variance empirique, la fonction de répartition empirique, le quantile empirique), ainsi que leurs comportements asymptotique, enfin nous donnons un aperçu de l'échantillon issus d'une variable normale.

Dans le chapitre trois dénommé application sous R, nous prouvons les comportements asymptotique des (la moyenne empirique, la fonction de répartition empirique, le quantile empirique) à l'aide du logiciel d'analyse statistique R.

Chapitre 1

Notions de base en statistique et échantillonnage

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, la partie de la population que l'on va examiner s'appelle l'échantillon.

1.1 Définitions de base

Population de taille finie

Définition 1.1.1 *Soit E un ensemble, que nous appellerons population mère, contenant un nombre fini N d'éléments, on note e_i chaque élément de E,*

$$E = \{e_1, ..., e_N\}.$$

L'échantillon aléatoire

Définition 1.1.2 *Un échantillon ξ_n est un sous ensemble de n individus extraits d'un population E composée de N élément nous écrivons :*

$$\xi_n = \{e_{i_k} \ ; i_k \in U\}.$$

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

Statistique

Définition 1.1.3 *La statistique est une v.a, définie comme une fonction de l'échantillon aléatoire, qui donne une information sur un paramètre de la population .*

$$S = f(X_1, X_2, ..., X_n).$$

Exemple 1.1.1 *Les statistique les plus utilisées :*

1. La moyenne empirique : $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. La variance empirique : $\widetilde{S_n}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2$.
3. La fonction de répartition empirique : $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}$.

1.2 Modes de convergence

On considére une suite de $v.a$ $(X_1, X_2, ..., X_n)$ $i.i.d$, on peut définir plusieurs modes de convergence pour une telle suite on notera F_{X_n} la fonction de répartition de X_n.

Convergence en probabilité

On dit que (X_n) converge en probabilité (ou converge faiblement) vers la $v.a$ X si, $\forall \varepsilon > 0$:

$$\lim_{n \to +\infty} P\left(\mid X_n - X \mid < \varepsilon\right) = 1,$$

et l'on note : $X_n \underset{n \to +\infty}{\xrightarrow{p}} X$.

Proposition 1.2.1 *Si (X_n) est une suite de v.a telle que :*

$$\begin{cases} \mathbb{E}\left(X_n\right) \underset{n \to +\infty}{\to} a \\ \mathbb{V}ar\left(X_n\right) \underset{n \to +\infty}{\to} 0 \end{cases},$$

alors : $X_n \underset{n \to +\infty}{\xrightarrow{p}} a$.

Convergence en loi

On dit que (X_n) converge en loi vers la $v.a$ X si l'on a, en tout x où sa fonction de répartition F_X est continue.

$$\lim_{n \to +\infty} F_{X_n} = F_X,$$

et l'on note : $X_n \underset{n \to +\infty}{\xrightarrow{\mathcal{L}}} X$.
Si la $v.a$ est discrétes, alors X_n converge en loi vers X ssi : $\forall x \in \mathbb{R}$, $\lim\limits_{n \to +\infty} P\left(X_n = x\right) = P\left(X = x\right)$.

Convergence presque sûre

On dit que (X_n) converge presque sûrement (ou converge fortement) vers la $v.a$ X si, $\forall \varepsilon > 0$:

$$\lim_{n \to +\infty} P\left(\sup_{m \geq n} \left\{\mid X_m - X \mid\right\} < \varepsilon\right) = 1,$$

et l'on note :$X_n \underset{n \to +\infty}{\xrightarrow{p.s}} X$.

Proposition 1.2.2 *Soit (X_n) une suite de v.a telle que $X_n \overset{p.s}{\to} X$ et g une fonction continue alors :*

$$g(X_n) \underset{n\to+\infty}{\overset{p.s}{\to}} g(X).$$

Proposition 1.2.3 *Soit (X_n) telle que $X_n \overset{p.s}{\to} X$ et (Y_n) telle que $Y_n \overset{p.s}{\to} Y$ si f est continue dans $\mathbb{R}^2$ alors :*

$$f(X_n, Y_n) \overset{p.s}{\to} f(X, Y).$$

Ces deux propositions sont également vraies pour la convergence en probabilité, elles s'étendent également à des fonction de k $v.a$ où $k > 2$.

Convergence en moyenne d'ordre p

On dit que la suite de $v.a$ (X_n) converge en moyenne d'ordre p, avec $0 < p < +\infty$, vers la $v.a$ X si :

$$\mathbb{E}\,|X_n - X|^p \to 0 \quad \text{quand } n \to +\infty,$$

et l'on note : $X_n \underset{n\to+\infty}{\overset{m.p}{\to}} X$.
Dans le cas particulier $p = 2$, la convergence en moyenne d'ordre 2 s'appelle convergence en moyenne quadratique $(m.q)$.

Remarque 1.2.1 $\left.\begin{array}{c} \textit{La convergence presque sûrement} \\ \textit{La convergence en moyenne d'ordre } p \end{array}\right\} \Longrightarrow$ *la convergence en probabilité* $\Longrightarrow$ *la convergence en loi.*

1.3 Lois fondamentales d'échantillonnage

1.3.1 Lois des grands nombres

Elles sont deux types : loi faible mettant en jeu la convergence en probabilité et loi forte relative à la convergence presque sûre. Nous considérons ici une suite de $v.a$ $X_1, X_2, ..., X_n$ non nécessairement de même loi.

Loi faible des grands nombres

Soit $X_1, X_2, ..., X_n$ indépendantes d'espérance $\mu_1, \mu_2, ..., \mu_n$ finies et de variance $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$ finies. si $\frac{1}{n}\sum_{i=1}^{n}\mu_i \to \mu$ et si $\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 \to 0$ alors $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n}X_i$ est tel que :

$$\overline{X_n} \xrightarrow{p} \mu.$$

Loi forte des grands nombres

Soit $X_1, X_2, ..., X_n$ indépendantes telles que $\frac{1}{n}\sum_{i=1}^{n}\mu_i \to \mu$ et $\sum_{i=1}^{n}\frac{\sigma_i^2}{i^2}$ est convergente alors :

$$\overline{X_n} \xrightarrow{p.s} \mu.$$

1.3.2 Théorème central limite

Le théorème central limite ou théorème de la limite centrale établit la convergence en loi de la somme de $v.a$ indépendantes de même loi vers la loi normale.

Théorème 1.3.1 *Soit (X_n) une suite de v.a indépendantes de même loi, et de carré intégrable d'espérance μ et d'écart-type σ alors :*

$$\frac{G_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad \textit{quand } n \to +\infty \qquad \textit{où } G_n = \sum_{i=1}^{n} X_i.$$

Preuve. On pose : $K_n = \frac{G_n - n\mu}{\sigma\sqrt{n}}$, soit K est un $v.a$ suit la loi normale centrée réduite, on va montre $K_n \xrightarrow{\mathcal{L}} K$ quand $n \to +\infty$

$\iff \forall t \in \mathbb{R} : \varphi_{K_n}(t) \underset{n\to+\infty}{\to} \varphi_K(t) = e^{-\frac{1}{2}t^2}$.

$$\varphi_{K_n}(t) = \mathbb{E}\left(e^{itK_n}\right) = \mathbb{E}\left(e^{it\frac{G_n - n\mu}{\sigma\sqrt{n}}}\right) = \mathbb{E}\left(e^{it\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma\sqrt{n}}\right)}\right),$$

on pose :$W_i = X_i - \mu$, avec $\mathbb{E}(W_i) = 0$.

$$\varphi_{K_n(t)} = \mathbb{E}\left(e^{it\sum_{i=1}^{n}\frac{W_i}{\sigma\sqrt{n}}}\right) = \prod_{i=1}^{n}\varphi_{W_i}\left(\frac{t}{\sigma\sqrt{n}}\right) = \left[\varphi_{W_1}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n,$$

en utilisons le développement de Taylor au voisinage de zéro :

$\varphi_{W_1}(u) = \varphi_{W_1}(0) + u\acute{\varphi}_{W_1}(0) + \frac{u^2}{2}\varphi''_{W_1}(0) + O(u^2)$.

On a :

$$\varphi_{W_1}(0) = 1 \text{ et } \acute{\varphi}_{W_1}(0) = 0 \text{ et } \varphi''_{W_1}(0) = -\sigma^2,$$

alors :

$$\varphi_{W_1}(u) = 1 - \frac{u^2\sigma^2}{2} + O\left(u^2\right).$$

Pour n assez grand $u = \frac{t}{\sigma\sqrt{n}} \to 0$.

$$\begin{aligned}\varphi_{W_1}\left(\frac{t}{\sigma\sqrt{n}}\right) &= 1 - \frac{t^2}{2n\sigma^2}\sigma^2 + O\left(u^2\right)\\ &= 1 - \frac{t^2}{2n} + O\left(u^2\right).\end{aligned}$$

Alors :

$$\varphi_{K_n(t)} = \mathbb{E}\left(e^{itK_n}\right) = \left(1 - \frac{t^2}{2n} + O\left(u^2\right)\right)^n \backsim \left(1 - \frac{t^2}{2n}\right)^n.$$

$$\lim_{n\to+\infty}\varphi_{K_n(t)} = e^{-\frac{1}{2}t^2} = \varphi_K(t).$$

■

1.4 L'estimation

On appelle estimateur du paramétre θ, toute fonction aléatoire des valeurs observées $X_1, X_2, ..., X_n$, susceptibles de servir à estimer θ.

$$T_n = f(X_1, X_2, ..., X_n).$$

1.4.1 Qualité d'un estimateur

Estimateur sans biais ou non biaisé

On appelle biais de l'estimateur T_n du paramétre θ la valeur :

$$b_\theta(T_n) = \mathbb{E}(T_n) - \theta.$$

On dit que cet estimateur est sans biais si : $b_\theta(T_n) = 0$, c-à-d : $\mathbb{E}(T_n) = \theta$.

Exemple 1.4.1 *Soit* $(X_1, X_2, ..., X_n)$ *un échantillon et* $\mathbb{E}(X_i) = \mu \quad \forall i = 1, .., n$, $\overline{X_n}$ *est un estimateur sans biais, en effet :* $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

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

Remarque 1.4.1 *Si* $b_\theta(T_n) \neq 0$ *on dit que cet estimateur est avec biais.*

Estimateur asymptotiquement sans biais

On dit que l'estimateur T_n de θ est un estimateur asymptotiquement sans biais si :

$$\lim_{n\to+\infty} b_\theta(T_n) = 0 \quad \text{c-à-d} : \lim_{n\to+\infty}\mathbb{E}(T_n) = \theta.$$

Exemple 1.4.2 *Soit* $\mathbb{V}ar(X) = \sigma^2 < +\infty$, *la variance empirique est asymptotiquement*

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

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

Remarque 1.4.2 *Entre deux estimateurs sans biais, le meilleur sera celui dont la variance est minimale.*

1.5 Étapes pour la sélection de l'échantillon

Détermination des objectifs de la recherche : Avant de commencer toute recherche, on détermine d'abord l'objectif principal de la recherche, qui est une étape essentielle du succès de toutes les étapes. Par exemple, si l'on veut étudier le problème de l'abandon scolaire, l'échantillon doit représenter l'ensemble du secteur.

Identifier la population d'origine dans laquelle nous avons choisi l'échantillon : Identifier la population est l'étape la plus importante car les résultats de l'étude lui seront présentés, par exemple : les étudiants de l'université Mohammed Khiedr Biskra, et excluent donc toute personne qui n'applique pas ces caractéristiques .(étudiant, université de Mohammed Khiedr Biskra).

Détermination des caractéristiques de la population : Les caractéristiques de la population sont déterminées par une liste des variables couvertes par l'étude, telles que : (âge, sexe, état matrimonial, lieu de résidence ...).

Spécification de la taille de l'échantillon : La taille de l'échantillon a deux types : soit petit pour être facile à manipuler, soit grand, et il faut faire attention ici à la difficulté d'ajuster les variables pour leur multiplication. Méthodes de sélection de l'échantillon.

1.6 Méthodes d'échantillonnage

L'échantillonnage peut se faire avec ou sans remise et une population peut être considérée comme finie ou infinie, une population finie dans laquelle on procède à un échantillonnage avec remise peut être théoriquement considérée comme infinie.

1.6.1 Méthodes probabilistes (aléatoires)

Echantillonnage aléatoire simple

Un échantillon aléatoire simple est un échantillon sélectionnée de manière à ce que chaque échantillon possible de taille n ait la même probabilité d'être sélectionné, on prélève dans la population des individus au hasard, tous les individus ont la même probabilité d'être prélevés, ce choix peut se faire avec ou sans remise.

Tirage avec remise Un individu peut être choisi plusieurs fois et la population reste la même après chaque tirage, et 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.

Tirage sans remise Un individu peut être choisi au plus une fois, pour chaque tirage la taille de population diminue par une unité, le processus de tirage des individus de la population devient non indépendant l'un de l'autre dans ce cas, il y a $C_N^n = \frac{N!}{n!(N-n)!}$ échantillons possible.

Exemple 1.6.1 *Une population comprend les nombres* $\{2, 3, 6, 8, 11\}$*, alors le nombre des échantillons possibles :*

Si le tirage avec remise :

Il y a $N^n = 5^2 = 25$, échantillons qui peuvent être tirés. Ces échantillons sont :

$$\begin{aligned}\xi_{1n} = \{&(2,2);(2,3);(2,6);(2,8);(2,11);(3,2);(3,3);(3,6);(3,8);(3,11)\\ &(6,2);(6,3);(6,6);(6,8);(6,11);(8,2);(8,3);(8,6);(8,8);(8,11)\\ &(11,2);(11,3);(11,6);(11,8);(11,11)\}\end{aligned}$$

Si le tirage est sans remise :

Il y a $C_N^n = C_5^2 = 10$, échantillons de taille 2 qui peuvent être tirés. Ces échantillons sont :

$$\xi_{2n} = \{(2,3);(2,6);(2,8);(2,11);(3,6);(3,8);(3,11);(6,8);(6,11);(8,11)\}$$

Echantillonnage stratifié

L'échantillonnage stratifié est une technique qui consiste à subdiviser une population non homogéne, d'effectif N, en H sous populations ou « strates » plus homogènes d'effectif Ni, (ex : stratification par tranche d'âge), de telle sorte que $N = N1 + N2 + \ldots\ldots + N_H$, dans chaque strate, on fait un échantillonnage aléatoire simple sans remise, de taille proportionnelle à la taille de strate dans la population, les individus de la population n'ont pas tous la même probabilité d'étre tirés, telle que l'échantillon final $n = n_1 + n_2 + \ldots. + n_H$, le nombre d'échantillons possibles est : $\prod_{h=1}^{H} C_{N_h}^{n_h}$.

Exemple 1.6.2 *Supposons que* 60% *des étudiants de l'école sont des filles et* 40% *des garçons ,pour former un échantillon de* 120 *étudiants en respectant ces strates, on devrait chosir au hasard* $60\% * 120 = 72$ *filles et* $40\% * 120 = 48$ *garçons .*

Echantillonnage par grappe

La population est divisée en G grappes, pas forcément de même taille, on tire au hasard sans remise g grappes ou familles d'individus et on examine tous les individus de la grappe (ex : on tire des immeubles puis on interroge tous les habitants), le nombre d'échantillon

possibles est C_G^g .

Exemple 1.6.3 *Les étudiants de première année Master sont répartis en* 5 *groupes, les groupes sont numérotés de* 1 *à* 5*. Supposons que on tire au hasard les groupes* 2, 3 *tous les étudiants de ces* 2 *groupes feront partie de l'échantillon.*

Echantillonnage systématique

L'échantillonnage systématique est une technique qui consiste à prélever des unités d'échantillonnage situées à intervalles égaux. Le choix du premier individu détermine la composition de tout l'échantillon. si on connaît l'effectif total de la population N et qu'on souhaite prélever un échantillon d'effectif n, l'intervalle entre deux unités successives à sélectionner est donné par :

$$K = \frac{N}{n}.$$

Connaissant K, on choisit le plus souvent, pour débuter, un nombre aléatoire i compris entre 1 et K, le rang des unités sélectionnées est alors i, $i + K$, $i + 2K$, $i + 3K$, ...
l'échantillonnage systématique est facile à préparer et en général facile à exécuter, il réduit le temps consacré à la localisation des unités sélectionnées .

Exemple 1.6.4 *On veut sélectionner un échantillon de* 30 *entreprises au sein d'une population de* 1800 *entreprises .* $K = \frac{1800}{30} = 60$*, ainsi on va tirer une entreprise toutes les* 60 *en partant d'un nombre tiré aléatoirement entre* 1 *et* 60*, supposons ce nombre est le* 15 *on va donc sélectionner la* $15^{éme}$ *entreprise puis la* $75^{éme}$ *la* $135^{éme}$ *jusqu' à la* $1755^{éme}$ *ce qui nous donnera l'échantillon de* 30 *entreprises donc :* $\xi_n = \{15, 75, 135, ..., 1755\}$.

1.7 Modéle d'échantillonnage

Définition 1.7.1 *Soit une expérience aléatoire définie par la v.a X telle que :*

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

On appelle modéle 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 a_n est la tribu produit des événements de $\mathcal{L}^n$ et P_Θ^n la loi jointe des expériences et $\mathcal{L}^n$ la produit des espaces des valeur de X, on notera X_i la $v.a$ de même loi qui X, associée à la $i^{éme}$ expérience et x_i sa réalisation. A l'espace $(\mathcal{L}, a, P_X)^n$ correspondre donc une séquence de $v.a$ $(X_1, X_2, ..., X_n)$ independantes et identiquement distribuées $(i.i.d)$ de méme loi P_X.

Remarque 1.7.1 *Le fait que les modéles d'échantillonnage soient composés d'expériences indépendantes les rend trés simples à manipuler.*

Si X est une variable discréte :

$$P_{(X_1,X_2,...,X_n)}(x_1, x_2, ..., x_n) = P\left(X_1 = x_1, X_2 = x_2, ..., X_n = x_n\right) = \prod_{i=1}^{n} P_{X_i}\left(x_i\right) = \prod_{i=1}^{n} P_X\left(x_i\right).$$

Si X est continue de densité f :

$$f\left(x_1, x_2, ..., x_n\right) = \prod_{i=1}^{n} f\left(x_i\right).$$

Chapitre 2

Distributions des caractéristiques d'un échantillon

La notion de distribution d'échantillonnage est à la base des méthodes d'inférence statistique, l'étude des échantillons nous permet d'obtenir des résultats avec les mêmes caractéristiques que la population d'origine et donc les résultats peuvent être généralisés à la population dans son ensemble. Le mode de tirage le plus simple et aussi le plus important est l'échantillonnage aléatoire simple correspondant à des tirages équiprobables et indépendants les uns des autres.

2.1 Moyenne empirique

Définition 2.1.1 *On appelle moyenne de l'échantillon ou moyenne empirique la statistique, notée* $\overline{X_n}$*, définie par :*

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

Proposition 2.1.1 *Soit* $(X_1, X_2, ..., X_n)$ *un échantillon aléatoire d'un v.a* X *de moyenne* $\mu = \mathbb{E}(X)$ *et variance* $\sigma^2 = \mathbb{V}ar(X)$ *pour calculer* $\mathbb{E}\left(\overline{X_n}\right)$ *et* $\mathbb{V}ar\left(\overline{X_n}\right)$*, il convient de distinguer le mode de tirage.*

1. Tirage avec remise :

   $\mathbb{E}\left(\overline{X_n}\right) = \mu$ et $\mathbb{V}ar\left(\overline{X_n}\right) = \frac{\sigma^2}{n}$.

2. Tirage sans remise :

   $\mathbb{E}\left(\overline{X_n}\right) = \mu$ et $\mathbb{V}ar\left(\overline{X_n}\right) = \left(\frac{N-n}{N-1}\right).\frac{\sigma^2}{n}$.

Preuve. 1. La moyenne et la variance de $\overline{X_n}$ dans le cas d'un échantillon de $v.a$ $i.i.d$ (tirage avec remise) sont :

L'espérance de $\overline{X_n}$:

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

La variance de $\overline{X_n}$:

$$\mathbb{V}ar\left(\overline{Xn}\right) = \mathbb{V}ar\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}ar\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n.\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

2. La moyenne et la variance de $\overline{X_n}$ dans le cas d'un échantillon de $v.a$ ne sont pas indépendantes (tirage sans remise) sont :

L'espérance de $\overline{X_n}$ est μ.

La variance de $\overline{X_n}$:

$$\mathbb{V}ar\left(\overline{X_n}\right) = \mathbb{V}ar\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathbb{V}ar\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathbb{V}ar\left(X_i\right) + \sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\mathbb{C}ov\left(X_i, X_j\right)\right],$$

avec $\sum_{i=1}^{n} \mathbb{V}ar\left(X_i\right) = n\sigma^2$. Et

$$\begin{aligned}
\mathbb{C}ov\left(X_i, X_j\right) &= \mathbb{E}\left[\left(X_i - \mu\right)\left(X_j - \mu\right)\right] \\
&= \sum_{k=1}^{N}\sum_{z=1}^{N}\left(x_k - \mu\right)\left(x_z - \mu\right) P\left(X_i = x_k, X_j = x_z\right) \\
&= \sum_{k=1}^{N}\sum_{z=1}^{N}\left(x_k - \mu\right)\left(x_z - \mu\right) P\left(X_i = x_k\right) P\left(X_j = x_z \mid X_i = x_k\right) \\
&= \sum_{k=1}^{N}\sum_{z=1}^{N}\left(x_k - \mu\right)\left(x_z - \mu\right) \frac{1}{N} P\left(X_j = x_z \mid X_i = x_k\right),
\end{aligned}$$

alors

$$\mathbb{C}ov\left(X_i, X_j\right) = \begin{cases} \sum_{k=1}^{N}\sum_{z=1}^{N}\left(x_k - \mu\right)\left(x_z - \mu\right)\frac{1}{N}\frac{1}{N-1} & \text{si } k \neq z \\ 0 & \text{si } k = z \end{cases},$$

donc

$$\mathbb{C}ov\left(X_i, X_j\right) = \frac{1}{N}\frac{1}{N-1}\sum_{\substack{k,z=1 \\ k\neq z}}^{N}\left(x_k - \mu\right)\left(x_z - \mu\right),$$

comme

$$\left[\sum_{k=1}^{N}\left(x_k - \mu\right)\right]^2 = \sum_{k=1}^{N}\left(x_k - \mu\right)^2 + \sum_{\substack{k,z=1 \\ k\neq z}}^{N}\left(x_k - \mu\right)\left(x_z - \mu\right),$$

alors

$$\begin{aligned}
\sum_{\substack{k,z=1 \\ k\neq z}}^{N}\left(x_k - \mu\right)\left(x_z - \mu\right) &= \left[\sum_{k=1}^{N}\left(x_k - \mu\right)\right]^2 - \sum_{k=1}^{N}\left(x_k - \mu\right)^2 \\
&= \left[N\mu - N\mu\right]^2 - N\sigma^2 = -N\sigma^2,
\end{aligned}$$

donc

$$\mathbb{C}ov\left(X_i, X_j\right) = \frac{1}{N}\frac{1}{N-1}\left(-N\sigma^2\right) = \frac{-\sigma^2}{N-1}.$$

On obtient finalement

$$\begin{aligned}\mathbb{V}ar\left(\overline{X_n}\right) &= \frac{1}{n^2}\left[n\sigma^2 + n\left(n-1\right)\left(\frac{-\sigma^2}{N-1}\right)\right] \\ &= \frac{\sigma^2}{n}\left[1 - \frac{(n-1)}{N-1}\right] \\ &= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).\end{aligned}$$

■

Remarque 2.1.1 *1. Dans le cas tirage sans remise on :* $\mathbb{V}ar\left(\overline{X_n}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \simeq \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}$*, si* $N \to +\infty$ *donc* $\mathbb{V}ar\left(\overline{X_n}\right) \to \frac{\sigma^2}{n}$*, donc il n'y a pas de différence entre les deux modes de tirage.*

2. Si la taille n *des échantillons est assez grande (en pratique* $n \geq 30$*), la distribution d'échantillonnage de la moyenne s'approche de la distribution normale quelle que soit la distribution de la population.*

Exemple 2.1.1 *Le même exemple précédent (Exemple 1.11.1). La moyenne de la population est :*

$$\mu = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6.$$

L'écart-type de la population est :

$$\sigma^2 = \frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2}{5} = \frac{16+9+4+25}{5} = 10.8,$$

et $\sigma = 3.29$.

Si le tirage avec remise : les moyennes correspondant à ces échantillons ξ_{1n} sont :

$$\{(2);(2.5);(4);(5);(6.5);(2.5);(3);(4.5);(5.5);(7);(4);(4.5);(6);(7);(8.5);(5);(5.5) \tag{2.1}$$

$$(7);(8);(9.5);(6.5);(7);(8.5);(9.5);(11)\},$$

et la moyenne de l'échantillonnage des moyennes est :

$$\mu_{\overline{X_n}} = \frac{\text{somme des moyennes de tous les échantillons dans (2.1) ci-dessus}}{25} = \frac{150}{25} = 6.$$

Illustrant le fait que $\mu_{\overline{X_n}} = \mu$.

La variance $\sigma^2_{\overline{X_n}}$ de la distribution d'échantillonnage des moyennes est obtenue en soustrayant la moyenne 6 de chaque valeur dans (2.1), en élevant le résultat au carré, en additionnant les 25 nombres ainsi obtenus et en divisant par 25. le résultat final est :

$$\sigma^2_{\overline{X_n}} = \frac{135}{25} = 5.40 \quad \text{et} \quad \sigma_{\overline{X_n}} = \sqrt{5.40} = 2.32.$$

Ceci illustre le fait que $\sigma^2_{\overline{X_n}} = \frac{\sigma^2}{n}$ puisque le membre de droite est $\frac{10.8}{2} = 5.40$, en accord avec la valeur ci-dessus.

Si le tirage est sans remise : les moyennes correspondant à ces échantillons ξ_{2n} sont :

$$\{(2.5)\,;(4)\,;(5)\,;(6.5)\,;(4.5)\,;(5.5)\,;(7)\,;(7)\,;(8.5)\,;(9.5)\}\,.$$

Et la moyenne de la distribution d'échantillonnage des moyennes est :

$$\mu_{\overline{X_n}} = \frac{2.5+4+5+6.5+4.5+5.5+7+7+8.5+9.5}{10} = 6.$$

Illustrant le fait que $\mu_{\overline{X_n}} = \mu$.

La variance de la distribution d'échantillonnage des moyennes est :

$$\sigma^2_{\overline{X_n}} = \frac{(2.5-6)^2+(4-6)^2+(5-6)^2+...+(9.5-6)^2}{10} = 4.05,$$

et $\sigma_{\overline{X_n}} = 2.01$.

Ce qui illustre $\sigma^2_{\overline{X_n}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$, puisque le membre de droite vaut $\frac{10.8}{2}\left(\frac{5-2}{5-1}\right) = 4.05$, comme on l'a obtenu ci-dessus.

2.1.1 Comportement asymtotique de $\overline{X_n}$

Théorème 2.1.1 *Soit $(X_1, X_2, ..., X_n)$ un échantillon d'un v.a X de moyenne empirique $\overline{X_n}$, en appliquant le T.C.L :*

$$\sqrt{n}\frac{\overline{X_n}-\mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \qquad \text{quand } n \to +\infty.$$

Loi faible des grands nombres :

$$\overline{X_n} \xrightarrow{p} \mathbb{E}(X) \qquad \text{quand } n \to +\infty.$$

Loi forte des grands nombres :

$$\overline{X_n} \xrightarrow{p.s} \mathbb{E}(X) \qquad \text{quand } n \to +\infty.$$

Proposition 2.1.2 *Le troisiéme et le quatriéme moment centré de $\overline{X_n}$ sont données par :*

$$\mu_3\left(\overline{X_n}\right) = \frac{\mu_3(X)}{n^2} \quad , \quad \mu_4\left(\overline{X_n}\right) = \frac{\mu_4(X) + 3(n-1)\sigma^4}{n^3}.$$

Preuve. Le troisiéme moment centrée de $\overline{X_n}$:

$$\begin{aligned}\mu_3\left(\overline{X_n}\right) &= \mathbb{E}\left(\overline{X_n}-\mu\right)^3 \\ &= \frac{1}{n^3}\mathbb{E}\left(\sum_{i=1}^{n}(X_i-\mu)\right)^3 = \frac{1}{n^3}\sum_{i=1}^{n}\mathbb{E}(X_i-\mu)^3 = \frac{\mu_3(X)}{n^2}.\end{aligned}$$

Le quatriéme moment centrée de $\overline{X_n}$:

$$\begin{aligned}\mu_4\left(\overline{X_n}\right) &= \mathbb{E}\left(\overline{X_n}-\mu\right)^4 \\ &= \frac{1}{n^4}\mathbb{E}\left(\sum_{i=1}^{n}(X_i-\mu)\right)^4 \\ &= \frac{1}{n^4}\sum_{i=1}^{n}\mathbb{E}(X_i-\mu)^4 + C_4^2\frac{1}{n^4}\sum_{i<j}\mathbb{E}\left[(X_i-\mu)^2(X_j-\mu)^2\right] \\ &= \frac{\mu_4(X)+3(n-1)\sigma^4}{n^3}.\end{aligned} \tag{2.2}$$

■

Coefficients d'asymetrie et d'aplatissement

Définition 2.1.2 *Soit $(X_1, X_2, ..., X_n)$ est un échantillon d'une v.a X de moyenne μ et d'écart type σ, alors :*

1. Le coefficient d'asymétrie de $\overline{X_n}$ noté :

$$cd\left(\overline{X_n}\right) = \frac{cd(X)}{\sqrt{n}}.$$

2. Le coefficient d'aplatissement de $\overline{X_n}$ noté :

$$ca\left(\overline{X_n}\right) = 3 + \frac{ca(X)-3}{n}.$$

Preuve. On commence par la cofficient d'asymétrie de $\overline{X_n}$:

$$cd\left(\overline{X_n}\right) = \frac{\mu_3\left(\overline{X_n}\right)}{\left(\sigma\left(\overline{X_n}\right)\right)^3} = \frac{\frac{\mu_3(X)}{n^2}}{\left(\frac{\sigma}{\sqrt{n}}\right)^3} = \frac{\mu_3(X)}{\sigma^3}\frac{n^{\frac{"3}{2}}}{n^2} = \frac{\mu_3(X)}{\sigma^3}n^{\frac{-1}{2}} = \frac{cd(X)}{\sqrt{n}}.$$

Le coefficient d'aplatissement de $\overline{X_n}$:

$$\begin{aligned} ca\left(\overline{X_n}\right) &= \frac{\mu_4\left(\overline{X_n}\right)}{\left(\sigma\left(\overline{X_n}\right)\right)^4} = \frac{\frac{\mu_4(X)+3(n-1)\sigma^4}{n^3}}{(\frac{\sigma}{\sqrt{n}})^4} \\ &= \frac{\mu_4\left(X\right)+3\left(n-1\right)\sigma^4}{\sigma^4}\frac{n^2}{n^3} \\ &= \frac{\mu_4\left(X\right)+3\left(n-1\right)\sigma^4}{\sigma^4 n} = \frac{\mu_4\left(X\right)}{\sigma^4 n} + \frac{3\left(n-1\right)}{n} \\ &= \frac{\mu_4\left(X\right)}{\sigma^4 n} + \frac{3n-3}{n} = 3 + \frac{\frac{\mu_4(X)}{\sigma^4}-3}{n} \\ &= 3 + \frac{ca\left(X\right)-3}{n}. \end{aligned}$$

■

Remarque 2.1.2 *On voit que* $cd\left(\overline{X_n}\right) \underset{n\to+\infty}{\to} 0$ *et* $ca\left(\overline{X_n}\right) \underset{n\to+\infty}{\to} 3$, *ce qui traduit la normalité asymptotique de* $\overline{X_n}$.

2.2 Variance empirique

Définition 2.2.1 *On appelle variance empirique la statistique, notée* $\widetilde{S_n}^2$, *définie par :*

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

Proposition 2.2.1 *Soit* X *une v.a de moyenne* μ *et variance* σ^2.

L'espérance de $\widetilde{S_n}^2$:

1. Tirage avec remise :
$$\mathbb{E}\left(\widetilde{S_n}^2\right) = \frac{n-1}{n}\sigma^2.$$
2. Tirage sans remise :
$$\mathbb{E}\left(\widetilde{S_n}^2\right) = (\frac{n-1}{n})(\frac{N}{N-1})\sigma^2.$$

La variance de $\widetilde{S_n}^2$:

$$\mathbb{V}ar\left(\widetilde{S_n}^2\right) = \frac{n-1}{n^3}\left[(n-1)\,\mu_4 - (n-3)\,\mu_2^2\right].$$

Preuve. L'espérance de la variance empirique $\widetilde{S_n}^2$:

La décomposition de $\widetilde{S_n}^2$:

$$\begin{aligned}
\widetilde{S_n}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2 \qquad (2.3)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\mu) - \left(\overline{X}_n - \mu\right)\right]^2\\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\overline{X_n}-\mu\right)^2 - \frac{2}{n}\sum_{i=1}^{n}(X_i-\mu)\left(\overline{X_n}-\mu\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 + \left(\overline{X_n}-\mu\right)^2 - \frac{2}{n}\left(\overline{X_n}-\mu\right)\sum_{i=1}^{n}(X_i-\mu)\\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \left(\overline{X_n}-\mu\right)^2.
\end{aligned}$$

L'espérance est linéaire donc :

$$\begin{aligned}
\mathbb{E}\left(\widetilde{S_n}^2\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left((X_i-\mu)^2\right) - \mathbb{E}\left(\left(\overline{X_n}-\mu\right)^2\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{V}ar\left(X_i\right) - \mathbb{V}ar\left(\overline{X_n}\right),
\end{aligned}$$

c'est-à-dire :

$$\mathbb{E}\left(\widetilde{S_n}^2\right) = \sigma^2 - \mathbb{V}ar\left(\overline{X_n}\right).$$

Nous en déduisons :

1.Tirage avec remise :

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

2.Tirage sans remise :

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

La variance de la variance empirique $\widetilde{S_n}^2$:

D'aprés la décomposition (2.3) on a $\widetilde{S_n}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \left(\overline{X_n}-\mu\right)^2$.

Posons $Y_i = (X_i-\mu)^2$, on donc :

$$\begin{aligned}
\mathbb{V}ar\left(\widetilde{S_n}^2\right) &= \frac{1}{n}\mathbb{V}ar\left(Y_1\right) + \mathbb{V}ar\left(\left(\overline{X}_n-\mu\right)^2\right) - \frac{2}{n}\sum_{i=1}^{n}\mathbb{C}ov\left(Y_i, \left(\overline{X_n}-\mu\right)^2\right) \\
&= \frac{1}{n}\mathbb{V}ar\left(Y_1\right) - 2\mathbb{C}ov\left(Y_1, \left(\overline{X_n}-\mu\right)^2\right) + \mathbb{V}ar\left(\left(\overline{X_n}-\mu\right)^2\right) \\
&= U_n + V_n - 2W_n.
\end{aligned}$$

On a d'abord

$$U_n = \frac{1}{n}\left[\mathbb{E}\left((X_1-\mu)^4\right) - \mathbb{E}^2\left((X_1-\mu)^2\right)\right] = \frac{\mu_4-\sigma^4}{n}. \tag{2.4}$$

D'autre part

$$\begin{aligned}
W_n &= \mathbb{E}\left[(X_1-\mu)^2\left(\overline{X_n}-\mu\right)^2\right] - \mathbb{E}\left((X_1-\mu)^2\right)\mathbb{E}\left(\left(\overline{X_n}-\mu\right)^2\right) \\
&= \mathbb{E}\left[(X_1-\mu)^2\left(\overline{X_n}-\mu\right)^2\right] - \frac{\sigma^4}{n},
\end{aligned}$$

avec

$$\begin{aligned}
\mathbb{E}\left[(X_1-\mu)^2\left(\overline{X_n}-\mu\right)^2\right] &= \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathbb{E}\left((X_1-\mu)^2(X_i-\mu)^2\right) + \sum_{j\neq k}\mathbb{E}\left[(X_1-\mu)^2(X_j-\mu)(X_k-\mu)\right]\right] \\
&= \frac{1}{n^2}\left[\mathbb{E}\left((X_1-\mu)^4\right) + \sum_{i=2}^{n}\mathbb{E}\left[(X_1-\mu)^2(X_j-\mu)^2\right] + 0\right] \\
&= \frac{\mu_4 + (n-1)\sigma^4}{n^2}.
\end{aligned}$$

D'ou

$$W_n = \frac{\mu_4+(n-1)\sigma^4}{n^2} - \frac{\sigma^2}{n} = \frac{\mu_4-\sigma^4}{n^2}. \tag{2.5}$$

Enfin

$$V_n = \mathbb{V}ar\left(\left(\overline{X_n}-\mu\right)^2\right) = \mathbb{E}\left(\left(\overline{X_n}-\mu\right)^4\right) - \mathbb{E}^2\left(\left(\overline{X_n}-\mu\right)^2\right) = \mathbb{E}\left(\left(\overline{X_n}-\mu\right)^4\right) - \frac{\sigma^4}{n^2},$$

d'aprie (2.2) on

$$\mathbb{E}\left(\left(\overline{X_n}-\mu\right)^4\right) = \frac{\mu_4 + 3\left(n-1\right)\sigma^4}{n^3},$$

alors

$$V_n = \frac{\mu_4 - 3\sigma^4}{n^3} + \frac{2\sigma^4}{n^2}. \tag{2.6}$$

On obtient d'aprés (2.4),(2.5),(2.6) :

$$\begin{aligned}\mathbb{V}ar\left(\widetilde{S_n}^2\right) &= \frac{\mu_4 - \sigma^4}{n} + \frac{\mu_4 - 3\sigma^4}{n^3} - \frac{2\left(\mu_4 - 2\sigma^4\right)}{n^2} \\ &= \frac{n-1}{n^3}\left[\left(n-1\right)\mu_4 - \left(n-3\right)\sigma^4\right].\end{aligned}$$

■

Remarque 2.2.1 *Si la taille n de l'échantillon est grande, on a :*

$$\mathbb{V}ar\left(\widetilde{S_n}^2\right) \simeq \frac{\mu_4 - \sigma^4}{n}.$$

2.2.1 Variance empirique corrigée S_n^2

Définition 2.2.2 *On appelle variance empirique corrigée, la statistique, notée S_n^2, définie par :*

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

Proposition 2.2.2 *S_n^2 est un estimateur sans biais de σ^2 car :*

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

Exemple 2.2.1 *Le même exemple précédent (Exemple 1.11.1), on calcule la moyenne et l'écart-type de la distribution d'échantillonnage des variances.*

Si le tirage avec remise : les variances correspondant à ces échantillons ξ_{1n} sont :

$$\begin{gathered}\{(0)\,;(0.25)\,;(4)\,;(9)\,;(20.25)\,;(0.25)\,;(0)\,;(2.25)\,;(6.25)\,;(16)\,;(4)\,;(2.25)\,;(0)\,;(1)\\ (6.25)\,;(9)\,;(6.25)\,;(1)\,;(0)\,;(2.25)\,;(20.25)\,;(16)\,;(6.25)\,;(2.25)\,;(0)\}\end{gathered} \tag{2.7}$$

La moyenne de la distribution d'échantillonnage des variances est :

$$\mu_{\widetilde{S_n}^2} = \frac{\text{somme de toutes les variances ci-dessus}}{25} = \frac{135}{25} = 5.4,$$

qui illustre le fait que $\mu_{\widetilde{S_n}^2} = \frac{(n-1)}{n}\sigma^2$, puisque pour $n = 2$ et $\sigma^2 = 10.8$, le membre de droite est $\left(\frac{1}{2}\right)(10.8) = 5.4$.

Ce résultat indique également pourquoi une expression de la variance corrigée des échantillons est souvent définie comme $S_n^2 = \frac{n}{n-1}\widetilde{S_n}^2$; il en résulte, en effet, que $\mu_{S_n^2} = \sigma^2$.

La variance de la distribution d'échantillonnage des variances $\sigma^2_{\widetilde{S_n}^2}$ s'obtient en soustrayant la moyenne 5.4 de chaque valeur dans (2.7), en élevant le résultat au carré, puis en faisant la somme que l'on divise par 25 et, $\sigma^2_{\widetilde{S_n}^2} = \frac{575.75}{25} = 23.03$ ou $\sigma_{\widetilde{S_n}^2} = 4.8$.

Si le tirage est sans remise : les variances correspondant à ces échantillons ξ_{2n} sont :

$$\{(0.25)\,;(4)\,;(9)\,;(20.25)\,;(2.25)\,;(6.25)\,;(16)\,;(1)\,;(6.25)\,;(2.25)\} \tag{2.8}$$

La moyenne de la distribution d'échantillonnage des variances est :

$$\mu_{\widetilde{S_n}^2} = \frac{0.25 + 4 + 9 + 20.25 + 2.25 + 6.25 + 16 + 1 + 6.25 + 2.25}{10} = 6.75,$$

cas particulier du résultat général $\mu_{\widetilde{S_n}^2} = \left(\frac{N}{N-1}\right)\left(\frac{n-1}{n}\right)\sigma^2$, comme on peut le vérifier en posant $N = 5, n = 2$ et $\sigma^2 = 10.8$, dans le membre de droite et on obtient $\mu_{\widetilde{S_n}^2} = \left(\frac{5}{4}\right)\left(\frac{1}{2}\right)(10.8) = 6.75$.

En soustrayant 6.75 de chaque valeur dans (2.8), en élevant les résultats au carré, en faisant leur somme et en divisant par 10, nous trouvons $\sigma^2_{\widetilde{S_n}^2} = 39.675$ ou $\sigma_{\widetilde{S_n}^2} = 6.3$.

2.2.2 Comportement asymtotique de $\widetilde{S_n}^2$ et S_n^2

Théorème 2.2.1 *Soit $(X_1, X_2, ..., X_n)$ un échantillon i.i.d d'une v.a X, telle que $\mathbb{E}(X^2) < +\infty$, on a :*

$$\widetilde{S_n}^2 \underset{n\to+\infty}{\overset{p.s}{\to}} \mathbb{V}ar(X) \qquad \text{et} \qquad S_n^2 \underset{n\to+\infty}{\overset{p.s}{\to}} \mathbb{V}ar(X).$$

Preuve. On a

$$\widetilde{S_n}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X_n}^2.$$

Donc

$$\overline{X_n} \overset{p.s}{\to} \mathbb{E}(X) \quad \Longrightarrow \quad \overline{X_n}^2 \overset{p.s}{\to} \mathbb{E}(X)^2,$$

et

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \overset{p.s}{\to} \mathbb{E}\left(X^2\right).$$

D'ou

$$\widetilde{S_n}^2 \overset{p.s}{\to} \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2 = \mathbb{V}ar(X).$$

De façon évidente $S_n^2 = \frac{n}{n-1}\widetilde{S_n}^2$ converge aussi $p.s$ vers $\mathbb{V}ar(X)$. ■

Théorème 2.2.2 *Soit $(X_1, X_2, ..., X_n)$ un échantillon i.i.d d'une v.a X, telle que $E(X^4) < +\infty$, on a :*

$$\sqrt{n}\frac{\left(\widetilde{S_n}^2 - \sigma^2\right)}{\sqrt{\mu_4 - \sigma^4}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \ \textit{quand } n \to +\infty.$$

Preuve. On pourra supposer que $\mathbb{E}(X) = 0$.

On a :

$$\mathbb{V}ar\left(\widetilde{S_n}^2\right) \simeq \frac{\mu_4 - \sigma^4}{n}.$$

Soit donc la quantité :

$$\sqrt{n}\frac{\left(\widetilde{S_n}^2 - \sigma^2\right)}{\sqrt{\mu_4 - \sigma^4}} = Y_n,$$

avec

$$Z_n = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \sigma^2\right)}{\sqrt{\mu_4 - \sigma^4}} \quad \text{et} \quad C_n = \frac{\sqrt{n}\overline{X_n}^2}{\sqrt{\mu_4 - \sigma^4}}.$$

Donc : $Y_n = Z_n - C_n$.

On commence par Z_n on a :

$$E\left(\overline{X_n^2}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) = E\left(X_1^2\right) = \sigma^2.$$

$$Var\left(\overline{X_n^2}\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i^2\right) = \frac{1}{n^2}\sum_{i=1}^{n}\left[\mathbb{E}\left(X_i^4\right) - \mathbb{E}^2\left(X_i^2\right)\right] = \frac{1}{n}\left(\mu_4 - \sigma^4\right).$$

Alors d'aprés le T.C.L on obtient :

$$Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

D'autre part , d'aprés le L.G.N, T.C.L et le théoréme de slutsky :

$$\left.\begin{array}{rcl} \overline{X_n} & \xrightarrow{p} & 0 \\ \sqrt{n}\overline{X_n} & \xrightarrow{\mathcal{L}} & \mathcal{N}(0,1) \end{array}\right\} \qquad \Longrightarrow \qquad \sqrt{n}\overline{X_n}^2 \xrightarrow{p} 0.$$

Et donc

$$C_n \quad \xrightarrow{p} \quad 0.$$

Alors

$$Y_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \qquad \text{quand } n \to +\infty.$$

■

Remarque 2.2.2 *La même convergence est vraie pour* S_n^2 *:*

$$\sqrt{n}\frac{(S_n^2-\sigma^2)}{\sqrt{\mu_4-\sigma^4}} \quad \xrightarrow{\mathcal{L}} \quad \mathcal{N}(0,1) \qquad\qquad \text{quand } n \to +\infty.$$

Proposition 2.2.3 *L'espérance de la moyenne empirique et de la variance empirique modifiée sont égales à l'espérance et la variance théorique de la variable* X*.*

2.3 Corrélation entre moyenne empirique et variance empirique modifiée

Théorème 2.3.1 *Soit* $(X_1, X_2, ..., X_n)$ *un échantillon i.i.d d'une v.a* X*, telle que* $\mathbb{E}(X^3) < +\infty$*, alors on a :*

$$\mathbb{C}ov\left(\overline{X_n}, S_n^2\right) = \frac{\mu_3}{n}.$$

Preuve. Calculons $\mathbb{C}ov\left(\overline{X_n}, \widetilde{S_n}^2\right)$:

$$\begin{aligned}
\mathbb{C}ov\left(\overline{X_n}, \widetilde{S_n}^2\right) &= \mathbb{E}\left(\overline{X_n}\widetilde{S_n}^2\right) - \mathbb{E}\left(\overline{X_n}\right)\mathbb{E}\left(\widetilde{S_n}^2\right) \\
&= \mathbb{E}\left(\overline{X_n}\widetilde{S_n}^2\right) \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} X_j^2 - \overline{X_n}^2\right)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left(X_i X_j^2\right) - \mathbb{E}\left(\overline{X_n}^3\right).
\end{aligned}$$

On suppose que $\mathbb{E}(X) = 0$, alors

$$\mathbb{E}\left(\overline{X_n}^3\right) = \frac{1}{n^2}\mathbb{E}\left(X^3\right) = \frac{\mu_3}{n^2}. \qquad \text{(car } X \text{ est centrée)} .$$
$$\mathbb{E}\left(X_i X_j^2\right) = 0 \text{ pour } i \neq j, \text{ à cause de l'indépendance.}$$

Donc :

$$\begin{aligned}\mathbb{C}ov\left(\overline{X_n}, \widetilde{S_n}^2\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(X_i^3\right) - \frac{\mu_3}{n^2} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mu_3 - \frac{\mu_3}{n^2} = \frac{\mu_3}{n} - \frac{\mu_3}{n^2} = \frac{n-1}{n^2}\mu_3,\end{aligned}$$

d'où

$$\mathbb{C}ov\left(\overline{X_n}, S_n^2\right) = (\frac{n}{n-1})(\frac{n-1}{n^2})\mu_3 = \frac{\mu_3}{n}.$$

■

Remarque 2.3.1 *1. Si la loi de X est symétrique, alors $\mu_3 = 0$ donc $\overline{X_n}$ et S_n^2 sont toujours non corrélés c-à-dire : $\mathbb{C}ov\left(\overline{X_n}, S_n^2\right) = 0$.*

2. On a : $\lim_{n\to+\infty}\mathbb{C}ov\left(\overline{X_n}, S_n^2\right) = 0$, alors $\overline{X_n}$ et S_n^2 sont asymptotiquement non corrélées.

3. Si X suit la loi normale, alors $\overline{X_n}$ et S_n^2 sont indépendantes.

2.4 Moments empiriques

Définition 2.4.1 *On appelle moment empirique d'ordre k la statistique, noté m_n^k, définie par :*

$$\forall k \in \mathbb{N}^* \qquad m_n^k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Proposition 2.4.1 *Si X une v.a admet un moment m_k d'ordre k, alors :*

1. *L'espérance de m_n^k est :*

$$\mathbb{E}\left(m_n^k\right) = m_k.$$

2. *La variance de* m_n^k *est* :

$$\mathbb{V}ar\left(m_n^k\right) = \frac{1}{n}\left(m_{2k} - m_k^2\right).$$

Preuve. On a :

$$\mathbb{E}\left(m_n^k\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i^k\right) = \mathbb{E}\left(X^k\right) = m_k.$$

et

$$\mathbb{V}ar\left(m_n^k\right) = \mathbb{V}ar\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}ar\left(X_i^k\right) = \frac{1}{n}\left[\mathbb{E}\left(X^{2k}\right) - \mathbb{E}^2\left(X^k\right)\right] = \frac{1}{n}\left(m_{2k} - m_k^2\right).$$

■

Définition 2.4.2 *On appelle moment empirique centré d'ordre k la statistique, noté* μ_n^k*, définie par :*

$$\forall k \in \mathbb{N}^* \quad \mu_n^k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^k$$

2.5 Fonction de répartition empirique

Définition 2.5.1 *Soit* $(X_1, X_2, ..., X_n)$ *un échantillon i.i.d d'une v.a* X *définie sur un espace de probabilité* $(\Omega, \mathcal{A}, P)$ *à valeurs danc* $\mathbb{R}$*. La fonction de répartition empirique* F_n *est définie par :* $\forall x \in \mathbb{R}$*,*

$$F_n(x) = \frac{\textit{nombre d'éléments dans l'échantillon } \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}.$$

Avec : $1_{\{X_i \leq x\}} = \begin{cases} 1 & \textit{si } X_i \leq x \\ 0 & \textit{si } X_i > x \end{cases}$.

2.5.1 Comportement asymptotique de F_n

Ces trois théorémes sont fondamentaux et justifient la convergence de F_n vers F.

Théorème 2.5.1 *Soit X une v.a de fonction de répartition $F(x)$, alors on a :*

$$F_n(x) \xrightarrow{p.s} F(x) \qquad quand\ n \to \infty.$$

Preuve. $F_n(x)$ étant une moyenne empirique de $v.a$ indépendantes, on d'aprés la loi forte des grands nombres :

$$F_n(x) \xrightarrow{p.s} \mathbb{E}\left(1_{\{X_i \leq x\}}\right) = \mathbb{E}\left(1_{\{X \leq x\}}\right),$$

avec

$$\mathbb{E}\left(1_{\{X \leq x\}}\right) = F(x).$$

■

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

$$\frac{\sqrt{n}\left(F_n(x) - F(x)\right)}{\sqrt{F(x)\left(1 - F(x)\right)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \qquad quand\ n \to +\infty.$$

Preuve. Pour chaque X, la $v.a$ $1_{\{X_i \leq x\}}$ suit la loi de bernoulli, car prendre la valant soit 0 soit 1 de paramétre $p = \mathbb{E}\left(1_{\{X \leq x\}}\right) = P(X \leq x) = F(x)$. Par conséquent, la variable $nF_n(x)$ est distribuée selon une loi binomaile de moyenne $nF(x)$ et de variance $nF(x)(1 - F(x))$. Alors $\mathbb{E}(F_n(x)) = F(x)$ et $\mathbb{V}ar(F_n(x)) = \frac{F(x)(1-F(x))}{n}$. Alors on dit que $F_n(x)$ est un estimateur sans biaisé de $F(x)$. On appliquant le T.C.L et on trouve le résultat. ■

Théorème 2.5.3 (Glivenko-cantelli) *La convergence de F_n vers F est presque surement uniforme c'est-à-dire que :*

$$\sup_{x \in \mathbb{R}} \mid F_n(x) - F(x) \mid \xrightarrow{p.s} 0 \quad quand \ \ n \to +\infty.$$

2.6 Les quantiles empiriques

On appelle quantile ou fractile d'ordre p de la population, le nombre x_p est l'inverse généralisée de la distribution F, il est définie pour tout $p \in [0,1]$, par :

$$x_p = F^{-1}(p) = \inf\{x \in \mathbb{R};\ F(x) \geq p\}.$$

Si F est strictement croissante et continue, alors x_p est l'unique nombre réel tel que :

$$F(x_p) = p.$$

La fonction des quantiles empirique de l'échantillon $X_1, X_2, ..., X_n$, est donnée par :

$$Q_n(p) = F_n^{-1}(p) = \inf\{x \in \mathbb{R};\ F_n(x) \geq p\},\ 0 < p < 1.$$

Définition 2.6.1 *Soit $(X_1, X_2, ..., X_n)$ un n-échantillon issu d'une loi F et $(X_{(1,n)}, X_{(2,n)}..., X_{(n,n)})$ l'échantilon ordonné. Soit $p \in]0,1[$. La statistique d'ordre $X_{([np]+1,n)}$ (où $[np]$ désigne la partie entière de np) s'appelle le quantile empirique d'ordre p de l'échantillon $Q_n(p)$.*

2.6.1 Comportement asymptotique d'un quantile empirique

Soient n variables aléatoires $X_1, X_2, ..., X_n$ *i.i.d* de fonction de répartition commune F et de densité f.

Théorème 2.6.1 *Pour tout $p \in]0,1[$, si F posséde un unique quantile d'ordre p, qui est alors égal à x_p (c'est-à-dire que F^{-1} est continue en p), alors :*

$$Q_n(p) \underset{n \longrightarrow +\infty}{\overset{p.s}{\longrightarrow}} x_p$$

Théorème 2.6.2 *Soit $0 < p < 1$ et supposons que F posséde une dérivée au voisinage de*

x_p avec $0 < f(x_p) < 1$ et f continue au point x_p, alors :

$$\sqrt{n}(Q_n(p) - x_p) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \frac{p(1-p)}{[f(F^{-1}(p))]^2}).$$

2.7 Echantillon d'une loi normale

Nous allons étudier le cas particulier ou la $v.a$ X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ telle que $\mu \in \mathbb{R}$, $\sigma > 0$, soit $(X_1, X_2, ..., X_n)$ un échantillons $i.i.d$ de la $v.a$ X .

Loi de $\overline{X_n}$: La moyenne empirique étant une combinaison linéaire de $v.a$ normales indépendantes suit aussi une loi normale, on a :

$$\overline{X_n} \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Ou :

$$\sqrt{n}\frac{\left(\overline{X_n} - \mu\right)}{\sigma} \rightsquigarrow \mathcal{N}(0,1). \tag{2.9}$$

2.7.1 Corrélation entre $\overline{X_n}$ et $X_i - \overline{X_n}$

On a $\overline{X_n} \rightsquigarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, et $X_i - \overline{X_n}$ suit une loi normale,
on calcule l'espérance et la variance de $X_i - \overline{X_n}$:

$$\mathbb{E}\left(X_i - \overline{X_n}\right) = 0$$

Et

$$\begin{aligned}
\mathbb{V}ar\left(X_i-\overline{X_n}\right) &= \mathbb{V}ar\left(X_i\right)+\mathbb{V}ar\left(\overline{X_n}\right)-2\mathbb{C}ov\left(X_i,\overline{X_n}\right)\\
&= \sigma^2+\frac{\sigma^2}{n}-2\mathbb{C}ov\left(X_i,\frac{1}{n}\sum_{j=1}^{n}X_j\right)\\
&= \sigma^2+\frac{\sigma^2}{n}-2\frac{1}{n}\sum_{j=1}^{n}\mathbb{C}ov\left(X_i,X_j\right)\\
&= \sigma^2+\frac{\sigma^2}{n}-2\frac{1}{n}\left[\mathbb{V}ar\left(X_i\right)+(n-1)\,\mathbb{C}ov\left(X_i,X_j\right)\right]\\
&= \sigma^2-\frac{\sigma^2}{n}\\
&= \frac{(n-1)}{n}\sigma^2.
\end{aligned}$$

Alors :

$$X_i-\overline{X_n}\rightsquigarrow\mathcal{N}\left(0,\frac{(n-1)}{n}\sigma^2\right).$$

Covariance entre $\overline{X_n}$ et $X_i-\overline{X_n}$:

$$\begin{aligned}
\mathbb{C}ov\left(\overline{X_n},X_i-\overline{X_n}\right) &= \mathbb{C}ov\left(\overline{X_n},X_i\right)-\mathbb{C}ov\left(\overline{X_n},\overline{X_n}\right)\\
&= \mathbb{C}ov\left(\overline{X_n},X_i\right)-\mathbb{V}ar\left(\overline{X_n}\right)\\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbb{C}ov\left(X_j,X_i\right)-\frac{\sigma^2}{n}\\
&= \frac{1}{n}\left[\mathbb{V}ar\left(X_i\right)+(n-1)\,\mathbb{C}ov\left(X_i,X_j\right)\right]-\frac{\sigma^2}{n}\\
&= \frac{\sigma^2}{n}-\frac{\sigma^2}{n}=0.
\end{aligned}$$

Donc $\overline{X_n}$ et $X_i-\overline{X_n}$ ne sont pas corrélé.

Théorème 2.7.1 *$\overline{X_n}$ et $X_i-\overline{X_n}$ $\forall i=1,..,n$, sont indépendantes. On déduit que $\overline{X_n}$ est indépendant de tous les $\left(X_i-\overline{X_n}\right)^2$, et donc :*
Dans le cas d'un échantillon gaussien, $\overline{X_n}$ et $\widetilde{S_n}^2$ sont des v.a indépendantes.

Loi de $\widetilde{S_n}^2$: D'aprés la décomposition (2.3) de $\widetilde{S_n}^2$:

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

On multiplie les deux cotés de l'égalité par $\frac{n}{\sigma^2}$ on trouve :

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

La somme de n carrés de $v.a$ qui suivent la loi normale centrées réduites indépendantes est une variable suivant la loi de Khi-deux à n degrés de liberté $\mathcal{X}^2_n$, alors on déduit que :

$$\frac{n\widetilde{S_n}^2}{\sigma^2} \rightsquigarrow \mathcal{X}^2_{n-1}. \tag{2.10}$$

2.7.2 Théoréme de Fisher

Soit $(X_1, X_2, ..., X_n)$ un échantillon $i.i.d$ d'une $v.a$ X, qui suivent la loi normale $\mathcal{N}(0,1)$ les variables :

$$\sqrt{n}\,\overline{X_n} \quad \text{et} \quad \sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)^2 = (n-1)\,S_n^2 = n\widetilde{S_n}^2,$$

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

Preuve. Voir [10] ■

2.7.3 Conséquences du Théoréme de Fisher

1. **Calcul de $\mathbb{V}ar\left(S_n^2\right)$ et $\mathbb{V}ar\left(\widetilde{S_n}^2\right)$:**
   On a :$(n-1)\frac{S_n^2}{\sigma^2} \rightsquigarrow \mathcal{X}^2_{n-1}$ et $\frac{n\widetilde{S_n}^2}{\sigma^2} \rightsquigarrow \mathcal{X}^2_{n-1}$.

Alors :

$$\mathbb{V}ar\left((n-1)\frac{S_n^2}{\sigma^2}\right) = 2(n-1) = \frac{(n-1)^2}{\sigma^4}\mathbb{V}ar\left(S_n^2\right)$$
$$\mathbb{V}ar\left(S_n^2\right) = \frac{2\sigma^4}{n-1}\ .$$
$$\mathbb{V}ar\left(\frac{n\widetilde{S_n}^2}{\sigma^2}\right) = 2(n-1) = \frac{n^2}{\sigma^4}\mathbb{V}ar\left(\widetilde{S_n}^2\right)$$
$$\mathbb{V}ar\left(\widetilde{S_n}^2\right) = \frac{2(n-1)\sigma^4}{n^2}.$$

2. **La loi de student :**

Définition 2.7.1 *Soit Z une v.a de loi normale centrée réduite et soit U une variable indépendante de Z et distribuée suivant la loi de $\mathcal{X}_n^2$, par définition, la variable*

$$T = \frac{Z}{\sqrt{\frac{U}{n}}}\ ,$$

suit une loi de student à n degrés de liberté.

D'aprés précédemment (2.9),(2.10) :

$$\sqrt{n}\frac{\left(\overline{X_n}-\mu\right)}{\sigma} \rightsquigarrow \mathcal{N}(0,1) \quad et \quad \frac{n\widetilde{S_n}^2}{\sigma^2} \rightsquigarrow \mathcal{X}_{n-1}^2.$$

On a :

$$T = \frac{\sqrt{n}\frac{\left(\overline{X_n}-\mu\right)}{\sigma}}{\sqrt{\frac{n\widetilde{S_n}^2}{\sigma^2(n-1)}}} = \sqrt{(n-1)}\frac{\left(\overline{X_n}-\mu\right)}{\widetilde{S_n}},$$

où T est une variable de student à $(n-1)$ degrés de liberté et ne dépend pas de σ, on écrit :

$$\sqrt{(n-1)}\frac{\left(\overline{X_n}-\mu\right)}{\widetilde{S_n}} \rightsquigarrow \mathcal{T}_{n-1}.$$

Chapitre 3

Application sous R

R est un système, communément appelé langage et logiciel, qui permet de réaliser des analyses statistiques. Plus particulièrement, il comporte des moyens qui rendent possible la manipulation des données, les calculs et les représentations graphiques. 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.

3.1 Loi des grands nombres

Pour un échantillon de la loi uniforme sur $[0, 1]$ de taille $n = 1000$, calculons les moyennes empiriques successives et traçons la moyenne empirique en fonction de la taille de l'échantillon et la droite horizontale qui réprésente la moyenne théorique.

Programme sous R :

```
n=1000
X=runif(n)
Y=cumsum(X)
x=seq(1,n,by=1)
plot(x,Y/N,xlab="Taille d'échantillon",ylab="Moyenne empirique",main="Loi des grands
```

nombres")

abline(μ=0.5,col="red")

text(200,0.6,expression(mu==1/2),col="red")

legend(300,0.6,legend=c("Moyenne empirique","Moyenne théorique"),lwd=c(1,2),lty=c(3,1),

col=c('1','red'))

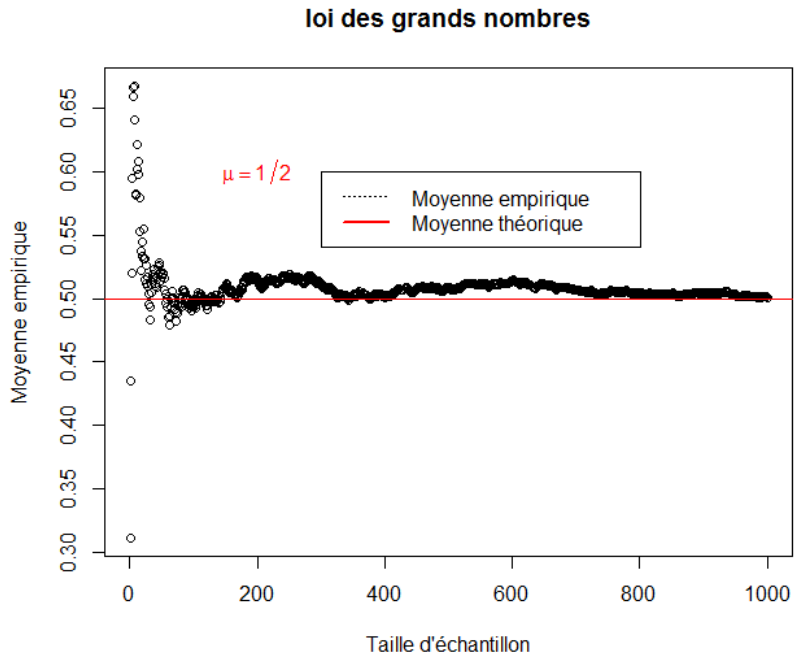

FIG. 3.1 – La convergence de la moyenne empirique vers la moyenne théorique

On observe que la moyenne empirique converge vers la moyenne théorique, actrice par une droite horizontale rouge sur le graphique 3.1 lorsque n est trés grand.

3.2 Théorème central limite

On prend un échantillon de la loi uniforme sur $[0, 1]$ de tailles $(n = 50, n = 500)$, Dans la figure 3.2 nous avons comparé l'histogramme d'une somme de variables aléatoires indépendantes de la loi uniforme par la densité de la loi normale de moyenne $\frac{n}{2}$ et de variance $\frac{n}{12}$.

Programme sous R :

```
par(mfrow=c(1,2))
R=100
n=50
X=numeric(R)
for (i in 1 : R)
X[i]=sum(runif(n))
hist(X, col="blue",main="Taille d'échantillon 50", probability=T)
lines(density(X), col="red", lwd=2)
x=X
sigma2=n/12
curve(dnorm(x,mean=n/2,sd=sqrt(sigma2)), add=T, col="green", lwd=2)
```

La même chose pour $n = 500$.

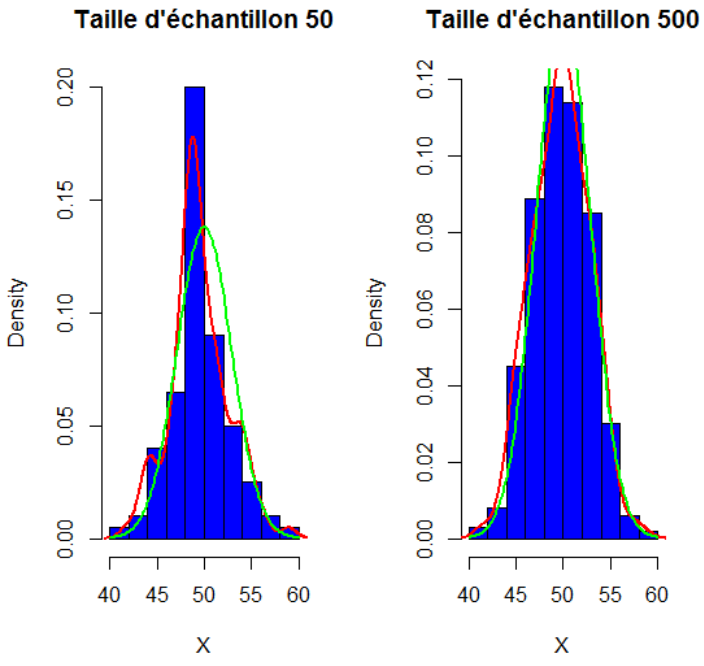

FIG. 3.2 – T.C.L. la convergence en loi de la moyenne empirique.

On remarque que la convergence est très rapide si n est grand. Ceci justifie la pratique qui revient à considérer que la loi d'un estimateur est gaussienne lorsque n est suffisamment grand.

3.3 Fonction de répartition empirique

On prend un échantillon suit une loi de Khi-deux à 1 degrés de liberté de taille $n = 1000$,nous allons tracer la fonction de répartition empirique et théorique par taille d'échantillon,

puis on étudie le comportement asymptotique de la fonction de répartition empirique.

Programme sous R :

```
n=1000
k=r=numeric(n)
X=rchisq(n,1)
x=seq(min(X),max(X),0.1)
Fn=function(x){
for(i in 1 :n){
if (X[i]<=x) k[i]=1 else k[i]=0 }
mean(k)}
for(i in 1 :n){
r[i]=Fn(X[i])}
plot(X,r,xlab="x",ylab="Fn(x)",main="Loi des grands nombres",col=1,lwd=2)
lines(x,pchisq(x,1),col=3,lwd=4)
legend(3,0.2,legend=c("Distribution empirique Fn"," Distribution théorique F"),lwd=c(1,2),
lty=c(3,1), col=c(1,3))
```

On remarque dans la graphe 3.3 que le deux fonctions de répartition sont trés proches, donc nous concluons que la fonction de répartition empirique converge bien vers la fonction

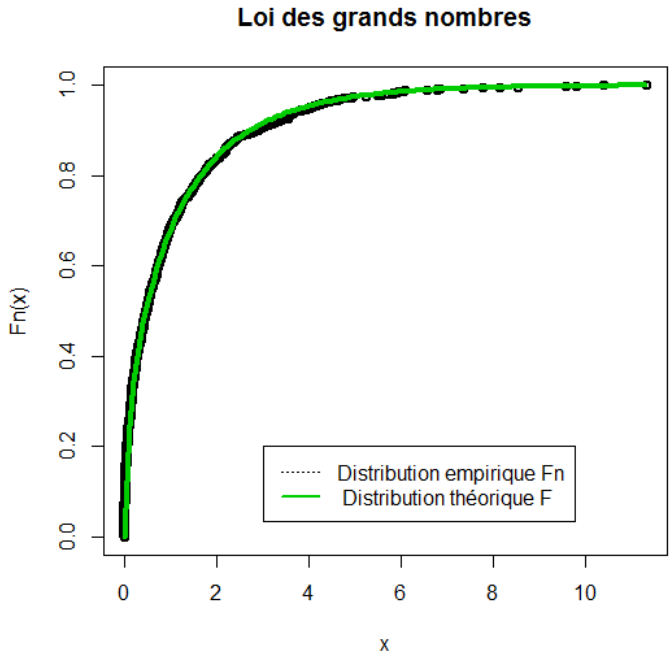

FIG. 3.3 – La convergence de la distribution empirique vers la distribution théorique

de répartition théorique.

3.4 Quantile empirique

On prend un échantillon qui suit la loi exponentielle de paramètre 2 de tailles ($n = 50, n = 500$), nous allons tracer la quantile empirique et la théorique par taille d'échantillon, puis on étudie le comportement asymptotique de la quantile empirique.

```
par(mfrow=c(1,2))
n=50
X=rexp(n,2)
p=seq(0,1,length.out=n)
Y=sort(X)
Z=numeric(n)
q=function(p){m=floor(p*n)
```

```
s=m+1
Y[s]}
for(i in 1 :n){
Z[i]=q(p[i])
}
qt=-(1/2)*log(1-p)
plot(qt,p,main="Taille d'échantillon 50",type="l")
lines(Z,p,xlab="p",col=3,lwd= 2)
legend(0.25,0.25,legend=c("quantil empirique","quantil théorique"),lwd=c(1,2),lty=c(3,1),
col=c(3,1))
```

La même chose pour $n = 500$.

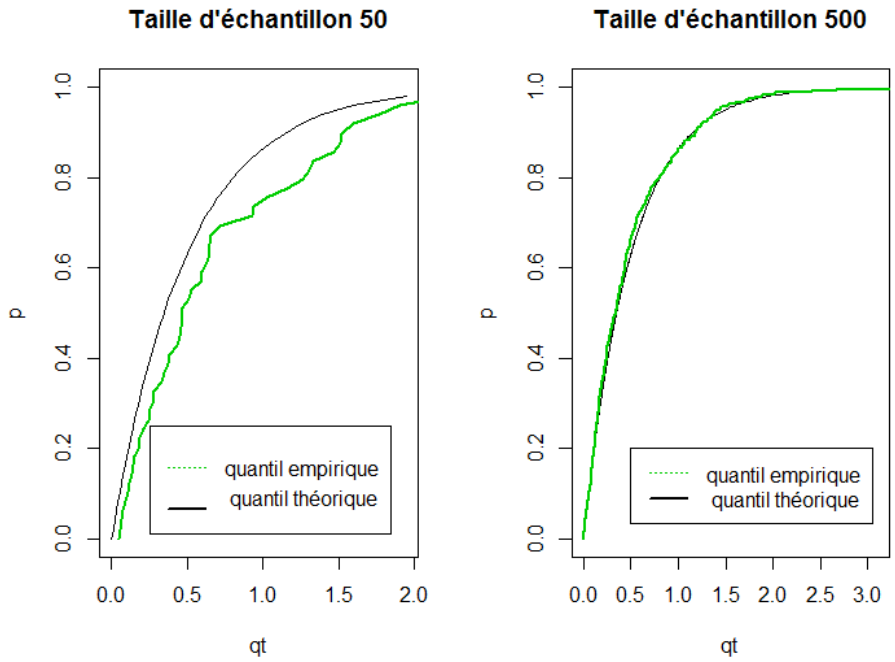

FIG. 3.4 – Quantile empirique vs quantile théorique d'une loi exponentielle

Nous remarquons dans la graphe 3.4 la quantile empirique s'approche a la quantile théorique lorsque n est grand.

Conclusion

Dans ce mémoire nous avons étudié la théorie d'échantillonnage et ses applications. L'objectif de l'échantillonnage est de tirer des conclusions sur les caractéristiques de la population inconnue sur la base d'informations sur certaines caractéristiques d'un échantillon sélectionné, la stratégie de cette méthode constitue une étape essentielle de la conception des expériences scientifiques, 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 (exemple : dans les laboratoires de chimie pour déduire les phénomènes ou les transformations de la matière, prendre des échantillons de sang pour détecter le groupe sanguin et certaines maladies, ...)

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.

Bibliographie

[1] Adil Elmarhoum.(2013).Echantillonnage et estimations.Univarsite Mohamed V-Agdal.

[2] Jean-Pierre (2008).Statistique et probabilités_travaux dirigés-Dunod.

[3] Khaldi Khaled.(2008).Methodes statistiques.Office des publicatons universitaires. Alger.

[4] Kherri Abdenacer(2013/2014).Statistique de gestion.Ecole des hautes etudes commerciales.

[5] Lejeune Michel.(2010).Statistique La théorie et ses Applications.Deuxiéme édition.Springer,Paris.

[6] O.Wintenberger.Statistique mathématique.

[7] Pierre Dusart.(2017).cours de statistiques inférentielles.Licence 2-S4 SI-Mass.

[8] Saporta,Gilbert.(2006) .Probabilités,analyse des données et statistique.Editions technip.Paris.

[9] Spiegel Murray R(1987).Probabilités et statistique.serie Schaum.

[10] Tassi Philippe.(1989).Méthodes Statistiques.Deuxiéme édition.Economica,Paris.

[11] Vaillant Jean .(2005).Initiation à la théorie de l'échantillonnage.

[12] Veysseyre Renée.(2006).Aide-mémoire statistique et probabilités pour l'ingénieur.deuxiéme édition.Dunod,Paris.

Annexe A : Logiciel *R*

R est un système, communément appelé langage et logiciel, qui permet de réaliser des analyses statistiques. Plus particulièrement, il comporte des moyens qui rendent possible la manipulation des données, les calculs et les représentations graphiques. *R* a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes et comporte un grand nombre de procédures statistiques appelées paquets. Ces derniers permettent de traiter assez rapidement des sujets aussi variés que les modèles linéaires (simples et généralisés), la régression (linéaire et non linéaire), les séries chronologiques, les tests paramétriques et non paramétriques classiques, les différentes méthodes d'analyse des données,... Plusieurs paquets, tels *ade4, FactoMineR, MASS, multivariate, scatterplot3d et rgl* entre autres sont destinés à l'analyse des données statistiques multidimensionnelles.

Il a été initialement créé, en 1996, par *Robert Gentleman* et *Ross Ihaka* du département de statistique de l'Université d'Auckland en Nouvelle Zélande. Depuis 1997, il s'est formé une équipe "*R Core Team*" qui développe *R*. Il est conçu pour pouvoir être utilisé avec les systèmes d'exploitation *Unix*, *Linux*, *Windows* et *MacOS*.

Un élément clé dans la mission de développement de *R* est le *Comprehensive R Archive Network* (CRAN) qui est un ensemble de sites qui fournit tout ce qui est nécessaire à la distribution de R, ses extensions, sa documentation, ses fichiers sources et ses fichiers binaires. Le site maître du CRAN est situé en Autriche à Vienne, on peut y accéder par l'URL : "*http ://cran.r-project.org/*". Les autres sites du CRAN, appelés sites miroirs, sont répandus partout dans le monde.

R est un logiciel libre distribué sous les termes de la "GNU Public Licence". Il fait partie intégrante du projet GNU et possède un site officiel à l'adresse "*http ://www.R-project.org*". Il est souvent présenté comme un clone de *S* qui est un langage de haut niveau développé par les *AT&T Bell Laboratories* et plus particulièrement par *Rick Becker, John Chambers* et *Allan Wilks*. *S* est utilisable à travers le logiciel *S-Plus* qui est commercialisé par la société Insightful (*http : //www.splus.com/*).

Annexe B : Abréviations et Notations

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

$v.a$	Variable aléatoire.
$E(.)$	Espérence mathématique.
$Var\,(.)$	Variance.
σ	l'écart type
m_k,μ_k	Moment et moment centré d'ordre k.
$\varphi_X\,(.)$	Fonction caractéristique de X.
E	Population.
ξ_n	Echantillon de E.
$\xrightarrow{p}$	Convergence en probabilité.
$\xrightarrow{\mathcal{L}}$	Convergence en loi.
$\xrightarrow{p.s}$	Convergence presque sûre.
$\xrightarrow{m.p}$	Convergence en moyenne d'order p
LGN	Lois des grands nombres.
TCL	Théoréme centrale limite.
G_n	La somme de v.a indépendantes.
$b_\theta\,(T_n)$	Biais de l'estimateur.

$b_\theta(T_n)$	Biais de l'estimateur.
θ	Paramètre à estimer.
$\overline{X_n}$	Moyenne empirique.
$cd(.)$	La coefficient d'asymétrie.
$ca(.)$	La coefficient d'aplatisement.
$\widetilde{S_n}^2$	Variance empirique.
S_n^2	Variance empirique corrigée
$\mathbb{C}ov\left(\overline{X_n}, S_n^2\right)$	Covariance de $\overline{X_n}$ et S_n^2.
m_n^k,μ_n^k	Moment empirique et moment centré empirique d'ordre k.
$F_n(.)$	Fonction de répartition empirique
$1_{\{X_i \leq x\}}$	Fonction indicatrice de l'ensemble $\{X_i \leq x\}$.
$i.i.d$	Indépendantes et identiquement distribuées.
x_p	Quantile d'order p.
$Q_n(.)$	Quantile empirique d'order p.
$\mathcal{N}(0,1)$	Loi normale centré réduite.
$\mathcal{N}(\mu, \sigma^2)$	Loi normale de moyenne μ et de variance σ^2
$\mathcal{X}_n^2$	Loi khi-deux à n degrés de liberté.
$\mathcal{T}_n$	Loi Student à n degrés de liberté.
$\mathbb{R}$	Ensemble des nombres réels.
$\mathbb{N}^*$	Ensemble des entiers naturels non nul.
$\mathbb{C}$	Ensemble des nombres complexes.