الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

N° Réf :……………

Centre Universitaire de Mila

Institut des Sciences et de la Technologie **Département de Mathématiques et Informatique**

Mémoire préparé En vue de l'obtention du diplôme de licence

en : Filière mathématiques fondamentales

Thème

STATISTIQUE INFERENTIELLE ET ECHANTION

Préparé par :

1. *Djamaa AHLAM*
2. *Bazoula SOUHILA*
3. *Laib SARRA*

***En cadré par :*Zerari AMEL**

Année universitaire : 2013/2014

Remerciements

Nous tenons à remercier dieu le plus puissant pour nous avoir donné le courage d'accomplir ce modeste travail.

Nous remercions tous ceux qui ont participé à la réalisation de ce travail qui est le fruit de notre dur labeur. Nous citons avec respect notre encadrant maitre assistanat **Zerari AMEL** pour sa disponibilité et ses judicieuses orientations ainsi que l'intérêt qu'il a manifesté à notre projet.

De même , nous tenons à remercier tous ceux qui ont contribué de prés ou de loin à la réalisation de ce travail.

Table des matières

Introduction Générale

La statistique connut, au début des XX^e siécle, une véritable révolution, celle de l'inférece.Il s'agit, à partir de résultats statistique limités, observés sur un échantillon,d'inférer à toute une population pour laquelle on induira des estimation, à partir des observations.Cette démarche trouve sa motivation dans des domaines appliqués spécifiques essentiellement en Grand Bretagne et aux Etats-Unis mais les outils ainsi crées trouveront, de part leur efficacité, des applications quasi universelles. La France restera à l'écart de ce mouvement. A l'époque du bourbakisme montant, si de grands noms comme Emile Borel, Maurice Fréchet, Paul Levy, s'illustrent dans la théorie (formelle) des probabilités, les méthodes statistiques ont mauvaise presse dans les milieux universitaires. Il faudra attendre les années cinquante (et encore plus tard dans l'enseignement) pour que, face à la pression économique des applications et à l'efficacité des nouvelles méthodes statistiques, la France commence à refaire son retard dans ce domaine.

Depuis le $XVIII^e$ siècle, deux approches des probabilités coexistent : le point de vue subjectif (représenté par Bayes et Laplace), où l'on considère la probabilité en termes de "raisons de croire", d'estimation du degré de confiance dans la réalisation d'un événement aléatoire et le point de vue fréquentiste, s'appuyant sur la loi des grands nombres de Bernoulli et les théorèmes limites (Laplace), où l'on conçoit la probabilité comme stabilisation limite de la fréquence lors de la répétition, un grand nombre de fois, de l'évènement. Au XIX^e siècle, Quételet et ses successeurs ne retiendront de la synthèse de Laplace que l'aspect fréquentiste, laissant en sommeil les techniques d'estimations amorcées au $XVIII^e$.

Dans un article publié en 1935 dans la "Revue de l'Institut International de la Statistique", Willcox dénombrait 115 définitions de la statistique. En étant plus raisonnable, voici quelques exemples :

1. Dans le style dénigrement (par ignorance ?), celle attribuée à Bismarck : "Il y a trois formes de mensonge, le mensonge ordinaire, le damné mensonge et la statistique."

2. Achenwall (1749) : "La statistique est la connaissance approfondie de la situation

respective et comparative des Etats."

Définition correspondant à l'origine de la statistique, liée à la notion d'Etat.

3. Sinclair (1785) : "La statistique a pour but de constater la somme de bonheur dont jouit une population et les moyens de l'augmenter".

Ce mémoire est répartit sur l'introduction générale et deux chapitres .Dans le premier chapitre, nous avons présenté statistique inférentielle et échantillon celle qui contient statistique descriptive, l'échantillon et statistique inférentielle.

Dans le deuxième chapitre on s'est intéressé de échantillon aléatoire, caractéristique d'un échantillon aléatoire, distribution du chi deux, distribution de fisher snédecor ,distribution de student et cas particulier d'échantillons gaussiens.

Chapitre 1

statistique inférentielle et échantillon

La synthése entre les observations et la population étudiée s'obtient à l'aide des procédés d'inférence statistique, c'est à dire d'estimation de tests statistiques. L'inférence utilise les modéles mathématiques et le calcul des probabilités.

1.1 statistique déscriptive

Définition 1.1.1 *1.* ***Statistique déscriptive*** *est la branche des statistiques qui regroupe les nombreuses techniques utilisées pour décrire un ensemble relativement important de données.*

2. ***La population statistique*** *est le groupe formé par toutes les personnes ou tous les objets à propos desquels on souhaite obtenir de l'information .*

3. ***La taille de la population statistique*** *est l'effectif de cette population c'est- a-dire le nombre d'individus statistiques dont elle est constituee. Une population finie est une population dont la taille est finie. Une population infinie est une population dont la taille est infinie. Dans la pratique, on peut considérer une population finie comme etant infinie si elle est d'effectif trés grand.*

4. ***Les individus statistiques(*** *ou unites statistiques) sont les élements de la population.*

5. ***L'ensemble fondamental*** Ω *appelé l'univers définissant l'ensemble des résultats possibles de* ε *appelés événements élémentaires.*

6. ***Effectif total*** *c'est le nombre total d'individus.*

7. ***Caractère*** *c'est la propriété étudiée. On distingue les caractères discréts qui ne peuvent prendre qu'un nombre fini de valeurs et les caractères continus dont on regroupe les valeurs par intervalles .*

8. ***Les modalités*** *ce sont les différentes valeurs qui peut prendre un caractére et on noté* x_i.

1.1.1 Séries statistiques associées à un caractère discrèt

Définition 1.1.2 *On appelle série statistique la donnée simultanée (dans un tableau) des valeurs du caractère étudié (noté* x_i*), rangées dans l'ordre croissant, et des effectifs (notés* n_i*) de ces valeurs.*

Remarque 1.1.3 *A la place des effectifs* (n_i)*, on peut aussi utiliser les fréquences* $f_i = \frac{n_i}{N}$*(N représente l'effectif total) et les fréquences en pourcentages* $f_i = \frac{n_i}{N} \times 100$.

Exemple 1.1.4 *Les notes sur 20 obtenues lors d'un devoir de mathématiques dans une classe de seconde sont les suivantes :*

10, 8, 11, 9, 12, 10, 8, 10, 7, 9, 10, 11, 12, 10, 8, 9, 10, 9, 10, 11.

La population étudiée est la classe et les individus sont les élèves. L'effectif total est égal à 20 *et la note obtenue au devoir est le caractère discret que l'on étudie.*

La série statistique définie par les effectifs est la suivante :

Valeurs du caractère (notes) x_i	7	8	9	10	11	12
Effectifs (nb d'élèves ayant la note) n_i	1	3	4	7	3	2

La série statistique définie par les fréquences en pourcentage est la suivante :

Définition 1.1.5

Valeurs du caractère (notes) x_i	7	8	9	10	11	12
Fréquences en % $f_i = \frac{n_i}{N} \times 100$	5%	15%	20%	35%	15	10%

Effectifs cumulés

Définition 1.1.6 1. *L'effectifes cumulés croissante* $(n_i \nearrow)$ *est le nombre d'individus ayant la valeur de la modalité* $\preceq x_i$ *et définie par :*

$$n_i \nearrow = n_1 + n_2 + ... + n_i = \sum_{s=1}^{i} n_s$$

2. *L'effectifes cumulés décroissante ($n_i \searrow$) est le nombre d'individus ayant la valeur de la modalité $\succeq x_2$ et définie par :*

$$n_i \searrow = N - (n_1 + n_2 + ... + n_{i-1}) = N - \sum_{s=1}^{i} n_s$$

Fréquences cumulés

Définition 1.1.7 *1. Fréquences cumulé croissant est la fréquence d'individus ayant la valeur de la modalité ($f_i \nearrow$) et définie par :*

$$f_i \nearrow = f_1 + f_2 + ... + f_i = \sum_{j=1}^{i} f_j$$

2. *Fréquences cumulé décroissant est la fréquence d'individus ayant la valeur de la modalité ($f_i \searrow$) et définie par :*

$$f_i \searrow = 1 - (f_1 + f_2 + ... + f_{i-1}) = 1 - \sum_{j=1}^{i-1} f_j$$

Exemple 1.1.8 *Avec l'exemple* (1.1.3) *on a :*

Valeurs x_i	7	8	9	10	11	12
Effectif cumulé croissant	1	4	8	15	18	20
Effectif cumulé décroissant	19	16	12	5	2	0

Paramétres de position

Les paramétres de position, aussi appelés valeure centrales, servent à caractériser l'ordre de grandeur des données.

Définition 1.1.9 (Moyenne arithmitique) *On appelle moyenne d'une série statistique d'éffectif total N, le réel :*

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i$$

tel que k représente le nombre de valeurs prises par le caractère.

Exemple 1.1.10 *Avec l'exemple* (1.1.3), *on a* $\overline{X} = 9,7$

1. Si on ajouté à toutes les valeurs d'une série statistique le même nombre b, on augmente la moyenne de cette série par b.

2. Si les valeurs d'une série statistique sont multipliées ou divisées par un même nombre a, la moyenne de cette série est aussi multipliée ou divisée par a.

Proposition 1.1.11 *Si une population d'effectif N est composée d'une partie d'effectif N_1 et de moyenne $(\overline{x_1})$ et d'une autre partie d'effectif N_2 et de moyenne $\overline{x_2}$ alors la moyenne $\overline{X}$ de la population totale est telle que : $\overline{X} = \frac{N_1\overline{x_1}+N_2\overline{x_2}}{N}$.*

Exemple 1.1.12 *Si dans une classe, les 15 garçons d'une classe mesurent en moyenne 182cm et si les 20 filles mesurent en moyenne 168cm, alors la taille moyenne d'un élève de cette classe est égale à $\frac{15\times182+20\times168}{15+20} = 174cm$.*

Définition 1.1.13 (Médiane) *La médiane M_e d'une série statistiques est la valeur qui permet de partager le groupe etudie en deux sous groupe de même effectif tel que, tous les éléments du premier groupe (resp. second groupe) ont des valeur inférieures (resp. supérieures) ou égales à M_e.*

1. Si n_i est paire $(N = 2l)$ alors :

$$M_e = \frac{x_l + x_{l+1}}{2}$$

2. Si n_i est impaire $(N = 2l + 1)$ alors :

$$M_e = x_{l+1}$$

Exemple 1.1.14 *On considère la série statistique suivante :*

Valeurs du caractère x_i	7	8	9	10	11	14	16
éffectifs n_i	2	1	1	1	2	1	2

L'effectif total est pair : la médiane M_e est la demi-somme des 2 valeurs situées au milieu.

D'où : $M_e = \frac{10+11}{2} = 10,5$.

Exemple 1.1.15 *On considère la série statistique suivante :*

Valeurs du caractère x_i	6	8	9	12	13	17
éffectifs n_i	3	1	2	1	3	3

L'effectif total est impair : la médiane M_e est la valeur située au milieu.

D'où : $M_e = 12$.

Définition 1.1.16 (Mode) *On appelle le mode d'une série statistique la modalité la plus fréquente c'est-à-dire celle dont l'éffectif est le plus grand.*

population étudie : hommes francais.
caractére étudie : âge en 2005.

âge	37	38	39	40	41	42	43
éffectif	436554	437247	449879	449881	456985	450554	435859

mode (M_o) : $41ans$.

Paramètres de dispersion

Les paramètres de dispersion (étendue, variance,...) sont calculés pour les variables statistiques quantitatives.

Définition 1.1.17 (L'entendue) *L'entendue d'une serie statistique est la diférence entre la plus grand valeur et la plus petite valeur des caractére etudie.*

Exemple 1.1.18 *d'aprés l'éxemple (1.1.8), l'entendue est :*6

Définition 1.1.19 (Variance) *Elle définie par la formule suivante :*

$$var(x) = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2 = \sum_{i=1}^{k} f_i(x_i - \overline{x})^2$$

Définition 1.1.20 (L'ecart type) *Est la racine carrée de la variance ,définie par la formule suivante :*

$$\sigma_x = \sqrt{var(x)}$$

La fonction de repartition

Définition 1.1.21 *La fonction de répartition d'un série statistique discrét est l'application F définie par :*

$$\begin{array}{rcl} F & : & \mathbb{R} \to [0;1] \\ x & \longmapsto & F(x) = \sum\limits_{x_{i \leq x}} f_i \end{array}$$

Tel que :

$$F(X) = \begin{cases} 0 & x \prec x_1 \\ \sum_{j=1}^{i} f_j & x_i \leq x \prec x_{i+1} \\ 1 & x \leq x_k \end{cases}$$

1.1.2 Séries statistiques associées à un caractère continue

Définition 1.1.22 *Soit la classe* $[e_{i,}e_{i+1}[$.

1. *Les nombre* $e_{i,}e_{i+1}$ *s'appellent les bornes de la classe.*
2. *La demi somme* $c_i = \frac{e_i+e_{i+1}}{2}$ *s'appelle le centre de la classe.*
3. *La différence* $e_{i+1} - e_i$ *s'appelle l'entendue.*
4. *L'effectifs* n_i *de la classe est la somme des effectifs partiels des valeurs de la série appartenant a la classe.*
5. *La fréquence* f_i *de la classe est la rapport :* $f_i = \frac{n_i}{N}$.

Paramètres d'une série statistique continue

Définition 1.1.23 (La classe modale) *C'est la classe dont les effectif le plus grand.*

Définition 1.1.24 (Le mode) *On déterminer la classe modale* $[e_{i,}e_{i+1}[$, *on pose* $h_i = n_i - n_{i-1}$ *et* $h_{i+1} = n_{i+1} - n_i$.

$$M_o = \frac{h_i e_{i+1} + h_{i+1} e_i}{h_i + h_{i+1}}$$

Définition 1.1.25 (Médiane) *Graphéquement on définie par la formule suivante :*

$$\tan\alpha = \frac{f_i \nearrow - f_{i-1} \nearrow}{e_{i+1} - e_i} = \frac{\frac{1}{2} - f_{i-1} \nearrow}{M_e - e_i}$$

Définition 1.1.26 (La moyenne) $\overline{X}$ *se calculer par la formule suivant :*

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{k} n_i c_i$$

Remarque 1.1.27 $Var(x)$ *et* σ_x *sont les même dans le cas discréte, on remplacer les modalité* x_i *par* c_i.

La fonction de repartition

Définition 1.1.28 *La fonction de répartition d'un série statistique continue est définie par :*

$$\begin{array}{ccc} F & : & \mathbb{R} \longrightarrow [0;1] \\ x & \longmapsto & F(x) \end{array}$$

Telle que :

$$F\left(x\right)=\left\{\begin{array}{lll} 0 & \text{si} & x\prec e_{1} \\ \sum_{j=1}^{i-1}f_{j}+\frac{x-e_{i}}{e_{i+1}-e_{i}}f_{i} & \text{si} & x\in\left[e_{i,}e_{i+1}\right[. \\ 1 & \text{si} & x\succeq e_{i+1} \end{array}\right.$$

Exemple 1.1.29 *Soit la distribution :*

Classe	$[8,10[$	$[10,12[$	$[12,14[$	$[14,16[$	$[16,18[$	$[18,20[$
c_i	9	11	13	15	17	19
n_i	1	2	4	6	5	2
f_i	0,05	0,1	0,2	0,3	0,25	0,1
$f_{i\nearrow}$	0,05	0,15	0,35	0,65	0,9	1
n_ic_i	9	2	52	90	85	38

La classe modale est : $[14,16[$.

La classe médiane est : $[14,16[$.

$M_{_O}=15,33$ $\quad M_e=15$ $\quad \overline{X}=14,8$ $\quad var\left(x\right)=6,76$ $\quad \sigma_x=2,6$

1.1.3 Variable aléatoire

Définition 1.1.30 *Une variable statistique X est une application de Ω dans un ensemble E.*

Types des variables aléatoires

Il ya deux types des variables aléatoires :

Quantitative

Définition 1.1.31 *Une variable aléatoire X est dite quantitative si elle prend des valeurs dans $\mathbb{R}$, il s'agit d'une variable représentée par une quantité (une valeur).*

Exemple 1.1.32 *L'âge, le poids et la taille,... etc.*

– Il existe deux types de variables quantitatives : variables discrètes et variables continues.

Variables discrètes

Définition 1.1.33 *Une variable aléatoire est dite discrète si elle ne prend que des valeurs discontinues dans un intervalle fini ou dénombrable.*

Exemple 1.1.34 *Le résultat du jet d'un dé, le nombre d'enfants dans une famille...*

Variables continues

Définition 1.1.35 *Une variable aléatoire est dite continue si elle ne prend ses valeurs dans $\mathbb{R}$ ou une partie ou un ensemble de parties de $\mathbb{R}$.*

Exemple 1.1.36 *La moyenne des étudiants, la taille...*

Qualitative

Définition 1.1.37 *Une variable qualitative est quant à elle une variable representée par une qualité, (modalité plutôt que d'une valeur).*

Exemple 1.1.38 *Le sexe, encore l'état civil, le degré de satisfaction d'un service quelconque...*

– Il existe également deux type de variables : variables nominales et variables ordinales.

Variables nominales

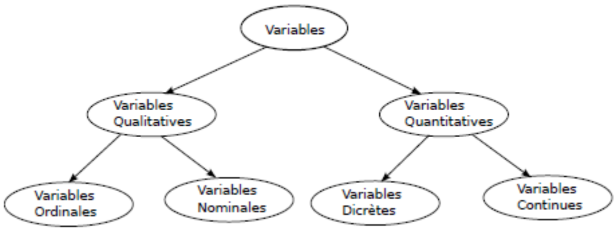
Définition 1.1.39 *Une variable aléatoire qualitative nominale est une variable qui correspondent à des noms, il n'y a aucun ordre précis sur les modalités.Ce sont seulement des mots dans le désordre.*

Exemple 1.1.40 *La variable sexe est une variable qualitative nominale qui a deux modalités possibles : féminin ou masculin et dont l'ordre n'importe pas.*

Variables ordinales

Définition 1.1.41 *Une variable aléatoire qualitative ordinale est une variable qui correspondent à un ordre. Il y a un ordre sur les modalités.*

Exemple 1.1.42 *Le degré de satisfaction par rapport à un service : très satisfait, satisfait, insatisfait...*



les types de variable aléatoire

Fonction de densité de probabilité

Définition 1.1.43 *On appelle fonction de densité de probabilité toute fonction continue par morceaux (intégrable) vérifier les deux conditions suivante :*

1. $\forall x \in \mathbb{R} \quad f(x) \succeq 0$.
2. $\int_{-\infty}^{+\infty} f(x)\, dx = 1$.

Fonction de répartition

Définition 1.1.44 *La fonction de répartition d'une variable aléatoire réelle X, notée $F_X(x)$, caractérise la loi de probabilité de cette variable. Elle représente la probabilité que la variable aléatoire réelle X prenne une valeur inférieure ou égale à x :*

$$F_X(x) = P(X \leq x)$$

Telle que :

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$$

La fonction de répartition $F_X(x)$ est donc l'une des primitives de la fonction de densité de probabilités.

Paramétre de variable aléatoire

cas discrète

Définition 1.1.45 (L'espérance) *Soit X une variable aléatoire discrète sur Ω telle que : $X(\Omega) = \{x_1, ..., x_n\}$,avec $x_1 \prec .... \prec x_n$. On appelle espérance mathématique de X le réel :*

$$E(X) = \sum_{i=1}^{n} x_i p_i = \sum_{i=1}^{n} x_i p(X = x_i)$$

Définition 1.1.46 (La variance) *Soit X une variable aléatoire discrète sur Ω telle que : $X(\Omega) = \{x_1, ..., x_n\}$, avec $x_1 \prec .... \prec x_n$. On appelle variance de X le réel :*

$$var(X) = E\left((X - E(X))^2\right) = E\left(X^2\right) - E^2(X)$$

Définition 1.1.47 (L'ecart type) *On appelle écart type le réel :*

$$\sigma_X = \sqrt{var(X)}$$

Définition 1.1.48 (Covariance) *Soient X et Y deux variables aléatoires. La covariance de X et Y est le réel :*

$$Cov(X, Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

cas continue Soit X une variable aléatoire continue admettant pour densité la fonction f. Sous réserve de convergence des intégrales, on définit :

- **L'espérance mathématique** de X par :

$$E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx$$

- **La variance** de X par :

$$var(X) = \int_{-\infty}^{+\infty} x^2 f(x)\, dx - E^2(X)$$

- **l'écart type** de X par :

$$\sigma(X) = \sqrt{var(X)}$$

Lois de probabilité

lois discrète

Loi uniforme

Définition 1.1.49 *Une distribution de probabilité suit une loi uniforme lorsque toutes les valeurs prises par la variable aléatoire sont équiprobables. Si n est le nombre de valeurs différentes prises par la variable aléatoire.*

$$\forall i,\ P(X = x_i) = \frac{1}{n}$$

Exemple 1.1.50 *La distribution des chiffres obtenus au lancer de dé suit une loi uniforme dont la loi de probabilité est la suivante :*

X	1	2	3	4	5	6
$P(X = x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$E(x) = 3,5 \qquad Var(x) = 2,92$

Loi de Bernoulli

Définition 1.1.51 *On effectue un tirage avec remise dans une urne contenant des boules blanches et noires, les premiéres en proportion p, les secondes en proportion $q = 1 - p$. Soit X le nombre des boules blanches tirées.*

Notation 1.1.52 $X \sim B(1, p)$.

Proposition 1.1.53 1. *L'espérance de la variable de Bernoulli est :*

$$E(X) = p$$

avec $p = P(X = 1)$.

2. *La variance de la variable de Bernoulli est :*

$$V(X) = pq$$

avec $q = P(X = 0) = 1 - p$.

Loi binomiale

Définition 1.1.54 *On effectue n tirage avec remise dans une urne contenant des boules blanches et noires, les premiéres en proportion p, les secondes en proportion $q = 1 - p$ .Soit X le nombre des boules blanches tirées .Elle définie par :*

$$P(X = k) = C_n^k P^k (1 - P)^{n-k}$$

pour $k \in \{0, 1, ..., n\}$.

Notation 1.1.55 $X \sim B(n; p)$.

Proposition 1.1.56
1. $E(X) = np$.
2. $Var(X) = npq$.

Loi de Poisson

Définition 1.1.57 *Soit X le nombre d'apparitions d'un événement rare sur un intervalle de temps donné .*

$$P(X = k) = \frac{\lambda^K}{K!} \exp(-\lambda)$$

Notation 1.1.58 $X \sim P(\lambda)$.

Proposition 1.1.59
1. $E(X) = \lambda$.
2. $Var(X) = \lambda$.

Loi géométrique

Définition 1.1.60 *On effectue des tirages avec remise dans une urne contenant des boules blanches et noire , les premiéres en proportion p, les secondes en proportion $q = 1-p$ .Soit X le nombre des tirages nécessaires pour obtenir une boule blanche .*

$$p(X = k) = pq^{k-1}$$

avec : $k \in \mathbb{N}^*$

Notation 1.1.61 $X \sim G(p)$.

Proposition 1.1.62
1. $E(X) = \frac{1}{P}$.
2. $Var = \frac{q}{p^2}$.

lois continues

loi exponentielle

Définition 1.1.63 *Une variable aléatoire continue X suit la loi exponentiel de paramètre $\lambda\,(\lambda \succ 0)$ si elle admet pour densité la fonction f définie sur $\mathbb{R}$ par :*

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{si} \quad x \succeq 0 \\ 0 & \text{sinon} \end{cases}$$

Notation 1.1.64 $X \sim \exp(\lambda)$.

Proposition 1.1.65
1. $E(X) = \frac{1}{\lambda}$.
2. $Var(X) = \frac{1}{\lambda^2}$.

Loi uniforme

Définition 1.1.66 *Une variable aléatoire continue X suit la loi uniforme sur l'intervalle $[a;b]$ si elle admet pour densité la fonction f définie sur $\mathbb{R}$ par :*

$$f(x) = \begin{cases} \frac{1}{b-a} & si\ x \in [a;b] \\ 0 & sinon \end{cases}$$

Notation 1.1.67 $X \sim U(a,b)$.

Proposition 1.1.68
1. $E(X) = \frac{a+b}{2}$.
2. $Var(X) = \frac{(b-a)^2}{12}$.

Loi normale

Définition 1.1.69 *Une variable aléatoire continue X suit la loi normale de paramètre μ et σ $(\sigma \succ 0)$ si elle admet pour densité la fonction f définie sur $\mathbb{R}$ par :*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Notation 1.1.70 $X \sim N(\mu, \sigma)$.

Proposition 1.1.71
1. $E(X) = \mu$.
2. $Var(X) = \sigma^2$.

Loi du khi-deux

Définition 1.1.72 *Si* $(X_1, ..., X_n)$ *sont un v.a indépendantes suivant toutes la loi normale centrée réduite, alors la quantité* $X = X_1^2 + ... + X_n^2$ *est une variable aléatoire qui suit une loi de khi-deux à* n *degrés de liberté.*

Notation 1.1.73 $X \sim \chi_n^2$.

Proposition 1.1.74

1. $\chi_n^2 \succeq 0$, *cette loi n'est donc pas symétrique.*
2. χ_n^2 *admet une densité (dificile a retenir).*
3. $$\begin{aligned} E[\chi_n^2] &= n \\ Var(\chi_n^2) &= 2n. \end{aligned}$$
4. *Pour* $n \succeq 30$ *:* $$\sqrt{2\chi_n^2} - \sqrt{2n-1} \sim N(0;1)$$

Loi de Student

Définition 1.1.75 *Si* X *et* Y *sont deux variables aléatoires indépendantes,* $X \sim N(0,1)$ *et* $Y \sim \chi_n^2$, *alors la quantité* $T = \frac{X\sqrt{n}}{Y}$ *est une variable aléatoire qui suit la loi de student à* n *degrés de liberté.*

Notation 1.1.76 $T \sim T_n$.

Proposition 1.1.77

1. $T(n)$ *admet une densité paire, cette loi est donc symétrique*
2. $$E(T) = 0 \ si \ n \succ 1$$
3. $$Var(T) = \frac{n}{n-2} \ si \ n \succ 2$$

Loi de fisher-Snédécor

Définition 1.1.78 *Si* $X_1 \sim \chi_{n_1}^2$, $X_2 \sim \chi_{n_2}^2$, *et si* X_1 *et* X_2 *sont indépendantes, alors la quantité* $F = \frac{X_1/n_1}{X_2/n_2}$ *est une variable aléatoire qui suit la loi de fisher -Snédécor à* n_1*et* n_2*degrés de libertés.*

Notation 1.1.79 $X \sim F_{n_1,n_2}$.

Loi normale réduite

Définition 1.1.80 *Il s'agit de la loi normale obtenue pour $\mu = 0$ et $\sigma = 1$. Sa densité f est alors définie par :*

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

Notation 1.1.81 $X \sim N(0;1)$.

1.2 L'échantillon

Définition 1.2.1 *On considère une population Ω de taille N. On appelle échantillon un sous ensemble de cette population. . L'objectif est d'obtenir une meilleure connaissance de la population par l'étude du seul échantillon. Le recours à un échantillon répond en général a une contrainte pratique (manque de temps, de place, évaluation destructive d'une production, coût financier...) interdisant l'étude exhaustive de la population.*

– L'echantillon satisfait les deux conditions suivantes :

1. Tous les individus sont séléctionnés dans la même population.
2. Les individus sont séléctionnés de facon indépendante.

Définition 1.2.2 (La taille de l'échantillon) *La taille de l'échantillon est l'effectif de cet échantillon c'est- à-dire le nombre d'individus statistiques observes dans la population statistique.*

Définition 1.2.3 (Sondage) *On appelle sondage toute observation partielle d'une population statistique c'est- à-dire l'observation d'une partie de cette population.*

Définition 1.2.4 (Fluctuation d'échantillonnage) *On appelle fluctuation d'échantillonnage la variation de la distribution des fréquences d'un échantillon à l'autre.*

1.2.1 Méthodes de sondage

Définition 1.2.5 *Une méthode de sondages (ou d'échantillonnage) d'écrit la façon dont la population statistique sera observée partiellement a travers un de ses sous-ensembles appelé échantillon. Et il ya deux méthodes :*

Méthode probabiliste :

Définition 1.2.6 *Sélection de l'échantillon par tirage aléatoire dans la population-mère. Chaque individu statistique doit avoir exactement la même chance que les autres de participer à l'enquête.*

On distingue quatre méthodes :

Échantillonnage aléatoire simple

Définition 1.2.7 *Dans un échantillon aléatoire simple, les éléments constituant l'échantillon sont extraits au hasard*

(à l'aide d'une table de nombres au hasard, par exemple) d'une liste de la population.

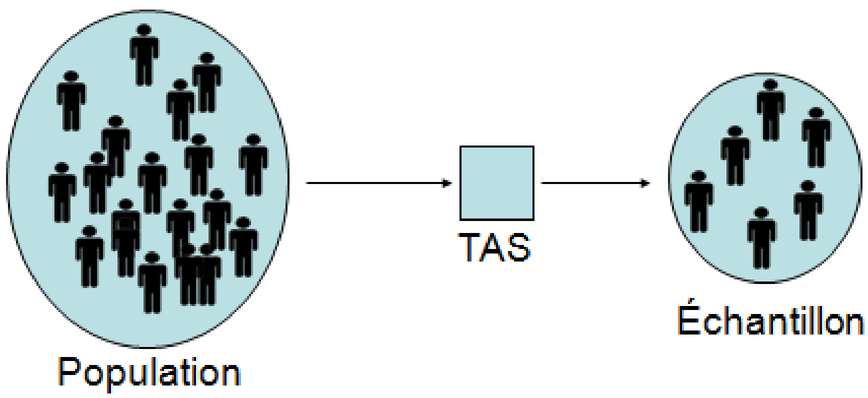
On extrait ainsi n individus d'une population de taille N.

Le tirage peut s'effectuer avec ou sans remise, renvoyant ainsi généralement à un modèle de loi binomiale (avec remise), ou hypergéométrique (sans remise).

Remarque 1.2.8 *Si le tirage s'effectue avec remise, l'échantillon aléatoire simple est dit indépendant*

(EASI = Echantillon Aléatoire Simple et Indépendant).

Exemple 1.2.9 *Dans une urne, on place autant de boules que d'individus dans la population. Chaque boule correspond à un individu de la population, on mélange les boules et on en tire une au hasard. On sélectionne ainsi un individu. On ne remet pas la boule dans l'urne (tirage sans remise) ; Retour à l'étape 2 jusqu'à ce que l'échantillon soit complet.*



Échantillonnage aléatoire simple

Echantillonnage aléatoire systématique

Définition 1.2.10 *Les individus de la population* Ω *sont numérotés de* 1 *à* N*. Pour sélectionner* n *individus, nous partageons la population en* $k = \frac{N}{n}$ *groupes :* $\{1, ..., k\}$, $\{1 + k, ..., 2k\}$, ... , $\{1 + (n-1)k, ..., N\}$.

Nous choisissons au hasard l'individu i par les individus numérotés de 1 *à* k.

Nous constituons notre échantillon des individus $\{i, i + k, i + 2k, ..., i + (n-1)k\}$.

Le choix de l'individu i détermine entièrement la constitution de l'échantillon.

Exemple 1.2.11 $\Omega = \{1, ..., 20\}$, $k = 4$.

Les échantillons possibles sont : $\{1, 5, 9, 13, 17\}$, $\{2, 6, 10, 14, 18\}$, $\{3, 7, 11, 15, 19\}$,$\{4, 8, 12, 16, 20\}$

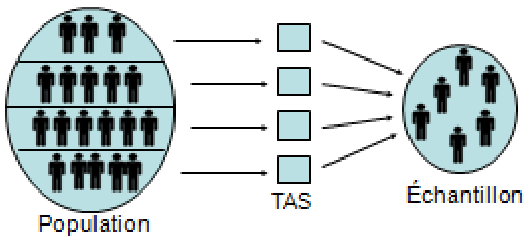
Echantillonnage stratifi

Définition 1.2.12 *La population étudiée* Ω *est partitionnée en q sous-populations* $\Omega_1, \Omega_2, ..., \Omega_q$, *appelées "strates".L'échantillon est constitué de la réunion de q échantillons choisis au hasard, un par strate : nous effectuons dans chaque strate un échantillonnage simple.*

Exemple 1.2.13 $\Omega = \{1, 2, 3, 4, 5\}$, $\Omega_1 = \{1, 2\}$, $\Omega_2 = \{3, 4, 5\}$.

Nous sélectionnons deux individus, dans Ω_1 *et trois dans* Ω_2.

Nous obtenons l'un des six échantillons possibles.



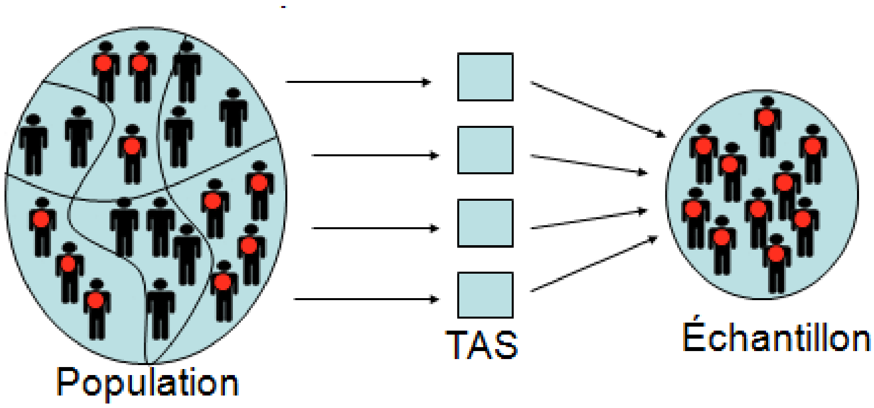
Echantillonnage stratifi

Echantillonnage en grappes et à plusieurs degrés

Définition 1.2.14 *La population* Ω *est divisée en sous-populations appelées unités primaires. Chaque unité primaire est divisée en unités secondaires, etc.*

Nous effectuons des tirages au hasard en cascade : nous tirons des unités primaires, dans chaque unité primaire, nous tirons une unité secondaire, etc.

Exemple 1.2.15 *L'INSEE effectue des échantillonnages à quatre niveaux : départements, cantons, communes, ménages*



Echantillonnage en grappes et à plusieurs degrés

Méthode non-probabiliste :

Définition 1.2.16 *Identifier dans la population-mère, quelques critères de répartition significatifs puis d'essayer de respecter cette répartition, dans l'échantillon d'individus interrogés. On utilise cette méthode lorsqu'il n'est pas possible de constituer une liste exhaustive de toutes les unités du sondage.*

Echantillonnage par quotas

Définition 1.2.17 *C'est une méthode d'échantillonnage qui consiste à s'assurer de la représentativité d'un échantillon, en lui affectant une structure similaire à celle de la population mère, au titre de plusieurs critères que sont, dans le cas d'une étude grand public, le sexe, l'âge, la profession, la région et la catégorie d'agglomération (critères détaillés dans la rubrique échantillon) puis à calculer le pourcentage de personnes appartenant à chaque catégorie selon les données du recensement de l'INSEE.*

Exemple 1.2.18 *Si les ouvriers hommes, âgés de* 30 *à* 40 *ans et urbains, représentent* 2% *de la population des électeurs, un échantillon de* 1000 *individus établis selon la méthode des quotas doit comprendre* 20 *personnes appartenant à cette catégorie. Lors d'une enquête téléphonique par quotas, un logiciel est utilisé pour décompter les individus déjà interrogés pour chaque catégorie de répondants. Des questions d'identification sont préalablement posées par les sondeurs pour savoir si l'individu peut être interrogé en fonction de l'état d'avancement de l'obtention des quotas. Lorsqu'une catégorie de répondants est* *difficile à obtenir, les instituts de sondages utilisent parfois des techniques de redressement d'échantillon.*

La méthode des itinéraires

Définition 1.2.19 *On impose à l'enquêteur :*

1. *Un point de départ dans une commune.*
2. *Un itinéraire à suivre avec tirage systématique des logements dans lesquels effectuer les interviews.*

– Le but est reproduire un certain tirage aléatoire des enquêtés, sans donner explicitement des noms et adresses à l'enquêteur.

Échantillonnage de convenance ou au jugé

Définition 1.2.20 *On prélève un échantillon en se fondant sur certains jugements au sujet de l'ensemble de la population.*

Échantillonnage sur place

Définition 1.2.21 *L'échantillon étudié est définie par un lieu. Cette méthode est utilisé dans l'échantillonnage de populations mobiles, rares ou spécifiques. Avec cette méthode il faut faire attention à :*

1. *Ne pas sur-représenter les individus passant + de temps sur place.*
2. *Les périodes d'enquête.*
3. *Les pondérations a posteriori pour tenir compte de la probabilité de présence.*

1.3 Statistique inférentielle

Définition 1.3.1 *Statistique inférentielle est l'ensemble des méthodes permettant de tirer des conclusions sur un groupe détermine a partir des données provenant d'un échantillon choisi dans cette population.*

1.3.1 Objectif

L'objectif de la statistique inférentielle est :

1. Fournir des résultats relatifs à une population à partir de mesures statistiques réalisées sur d'échantillons.

2. Déduire des renseignements sur un échantillon à partir de la connaissance de la population (échantillonnage).
3. Déduire des renseignements sur une population à partir de la connaissance d'un échantillon (estimation).

– Le but de statistique inférentielle est de faire des previsions et prendre des décision au vu des données.

Nous avons vu deux grandes categories de méthodes :
L'estimation ponctuelle et par intervalles de confiance, avec la méthode
des moments , la méthode du maximum de vraisemblance et les tests d'hypothèses.

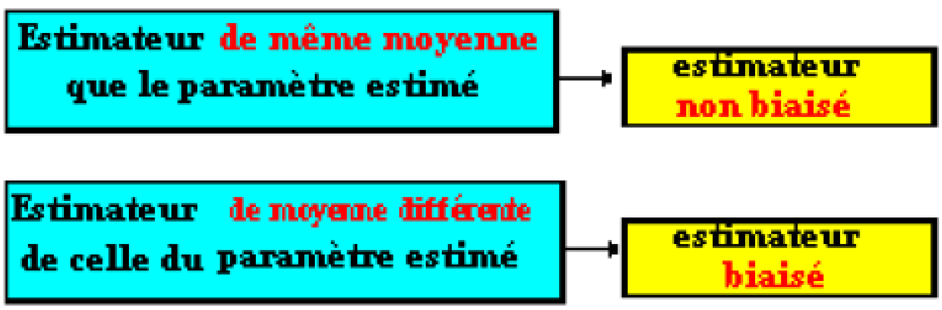
1.3.2 Estimation

Définition 1.3.2 *L'estimation statistique consiste à donner une valeur approchée à une caractéristique d'une population, à partir d'un échantillon d'observations issus de cette population.*

1.3.3 Estimateur

Définition 1.3.3 *Un estimateur est une variable aléatoire (c'est une fonction de variable aléatoire). Ou bien est un paramètre d'échantillon utilisé pour "estimer" la valeur d'un paramètre statistique de la population.*

Proposition 1.3.4
1. *Si l'estimateur a même moyenne que le paramètre à estimer, on dit que cet estimateur est non biaisé. Dans le cas contraire, on dit qu'il est dit biaisé.*
2. *Si l'on prend la moyenne d'échantillon X_i comme estimateur de μ, on dira que c'est un estimateur non biaisé puisque la moyenne des X_i est égale à μ (théorème des moyennes). par contre, si on utilise l'écart type d'échantillon.*



estimation ponctuelle d'un paramétre

Moyenne La valeur moyenne m_e d'un paramètre observé sur un échantillon de population, dont la taille est fixée, fournit une estimation $\overline{x}$ de la moyenne réelle de ce paramètre sur la population considérée.

Exemple 1.3.5 *Une usine produit des vis cruciformes. On souhaite estimer la moyenne des longueurs des vis dans la production de la journée qui s'élève à* 10000 *pièces.*

*On choisit un échantillon de*150 *vis et on obtient une moyenne de* $m_e = 4.57cm$. *On en déduit donc que la longueur moyenne des vis de la production journalière est* $\overline{x} = 4.57cm$.

Écart-type Le problème est toujours le même, mais cette fois-ici, l'estimation de l'écart-type est moins intuitive

Proposition 1.3.6 *L'écart-type σ_e d'un paramètre observé sur un échantillon de population, dont la taille est fixée, fournit une estimation faussée de l'écart-type de ce paramètre dans toute la population considérée.*

Une meilleure estimation de l'écart-type réel est obtenue en considérant le nombre

$$\sigma = \sigma_e \sqrt{\frac{n}{n-1}}$$

où n est la taille de l'échantillon servant au calcul de σ_e.

Remarque 1.3.7 *La correction devient assez rapidement minime lorsque la taille de l'échantillon augmente car :*

$$\lim_{n \longrightarrow +\infty} \sqrt{\frac{n}{n-1}} = 1$$

La correction est ainsi de l'ordre de 0.5% *pour des échantillons de taille* 100, *et de l'ordre de* 0.05% *pour des échantillons de taille*1000.

Exemple 1.3.8 *La mesure de la longueur des vis produites dans l'échantillon précédent de* 150 *pièces conduit à relever un écart-type de* $3mm$.

La meilleure estimation possible de l'écart-type de la production journalière n'est pas de $3mm$ *comme dans le cas précédent pour la moyenne, mais de* $\sigma = 3\sqrt{\frac{150}{149}} \simeq 3.01$

Fréquence La fréquence d'apparition f_e d'un paramètre observé sur un échantillon de population, dont la taille est fixée, fournit une estimation f de la fréquence réelle d'apparition de ce paramètre sur la population considérée.

Exemple 1.3.9 *Dans l'exemple précédent, On prélève un échantillon de* 150 *vis et on relève* 3 *pièces défectueuses. On peut alors donner une estimation de la fréquence de vis défectueuses dans la production journalière : on a* $f_e = \frac{3}{150} = 0.02$ *donc* $f = 0.02$.

Estimation par intervalle de confiance d'un paramètre

Définition 1.3.10 *Les estimations ponctuelles dépendent directement de l'échantillon prélevé au hasard. Dans de très nombreux cas, l'importance attribuée au hasard est grande, cela conduit à s'interroger avant d'utiliser ces estimations pour prendre des décisions dont les conséquences peuvent être lourdes !Aussi, sans rejeter les informations fournies par l'étude d'un échantillon, est on amené à chercher un nouveau type d'estimation de la fréquence et de la moyenne d'une population, en utilisant le calcul de probabilités qui permet de « contrôler » l'influence d'un échantillon particulier.*

Moyenne On souhaite, à partir des observations faites sur un échantillon, déterminer un intervalle de confiance contenant la valeur moyenne avec un risque d'erreur décidé à l'avance.

On suppose que les conditions sont réunies pour faire l'approximation que la loi d'échantillonnage de la moyenne $\overline{x}$ est la loi normale $N(m, \frac{\sigma}{\sqrt{n}})$

On pose :

$$\begin{aligned} T &= \frac{\overline{x} - m}{\frac{\sigma}{\sqrt{n}}} \\ T &\sim N(0;1) \end{aligned}$$

Soit la probabilité fixée à l'avance, pour que T n'appartienne pas à l'intervalle $[-t;t]$.On peut écrire :

$$\begin{aligned} p\left(|T| \succ t\right) &= \alpha \Leftrightarrow 1 - p\left(|T| \leq t\right) = \alpha \\ &\Leftrightarrow p\left(|T| \leq t\right) = 1 - \alpha \\ &\Longleftrightarrow p\left(-t \leq T \leq t\right) = 1 - \alpha \\ &\Longleftrightarrow p\left(-t \leq \frac{\overline{x} - m}{\frac{\sigma}{\sqrt{n}}} \leq t\right) = 1 - \alpha \\ &\Longleftrightarrow p\left(\overline{x} - t\frac{\sigma}{\sqrt{n}} \leq m \leq \overline{x} + t\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \end{aligned}$$

Autrement dit, m appartient à l'intervalle $\left[\overline{x} - t\frac{\sigma}{\sqrt{n}}; \overline{x} + t\frac{\sigma}{\sqrt{n}}\right]$, pour $100(1-\alpha)\%$ d'échantillons.

- Cet intervalle est appelé intervalle de confiance.
- α est le risque d'erreur ou le seuil de risque.
- $1-\alpha$ est le coefficient de confiance.

Proposition 1.3.11 *L'intervalle* $\left[\overline{x}-t\frac{\sigma}{\sqrt{n}};\overline{x}+t\frac{\sigma}{\sqrt{n}}\right]$ *est l'intervalle de confiance de la moyenne* m *de la population avec le coefficient de confiance* $2\pi(t)-1=1-\alpha$.

Fréquence A l'aide d'un échantillon, nous allons définir, avec un coefficient de confiance choisi à l'avance, un intervalle de confiance de la fréquence p des éléments de la population possédant une certaine propriété.

On se place dans le cas où on peut approximer la loi par la loi normale $N(p;\sqrt{\frac{p(1-p)}{n}})$.

Proposition 1.3.12 *L'intervalle* $\left[f-t\sqrt{\frac{f(1-f)}{n-1}};f+t\sqrt{\frac{f(1-f)}{n-1}}\right]$ *est l'intervalle de confiance d'une fréquence* p *de la population avec le coefficient de confiance* $2\pi(t)-1=1-\alpha$*.ayant pour centre la fréquence* f *de l'échantillon considéré.*

Exemple 1.3.13 *Un sondage dans une commune révèle que sur les* 500 *personnes interrogées,* 42% *sont mécontentes de l'organisation des transport. On veut déterminer, au seuil de risque* 1%*, un intervalle de confiance du pourcentage* p *de personnes mécontentes dans la commun on a :*

$f=0.42$.

$n=500$.

$\alpha=1\%$ *donc :* $t=2.575$.

Un intervalle de confiance du pourcentage p *est :*

$$\left[0.42-2.58\sqrt{\frac{0.42*0.58}{499}};0.42-2.575\sqrt{\frac{0.42*0.58}{499}}\right]=[0.36;0.48]=[36\%;48\%]$$

- Le tableau ci-dessous regroupe toutes les situations dans lesquelles on doit savoir fournir une estimation ponctuelle ou par intervalle de confiance :

Paramètre de la population totale à estimer	Valeur du paramètre dans l'échantillon de taille n	Estimation ponctuelle pour la population totale	Estimation par intervalle de confiance au niveau de confiance $2\Pi(t)-1$ pour la population totale
Moyenne	m_e	$m = m_e$	$\left[m_e - t\frac{\sigma}{\sqrt{n}}; m_e + t\frac{\sigma}{\sqrt{n}}\right]$
Écart-type	σ_e	$\sigma = \sigma_e\sqrt{\frac{n}{n-1}}$	
Fréquence	f_e	$f = f_e$	$\left[f_e - a\sqrt{\frac{f_e(1-f_e)}{n-1}}; f_e + t\sqrt{\frac{f_e(1-f_e)}{n-1}}\right]$

Méthodes d'estimation

Méthode empirique On sélectionne un estimateur naturel et on étudie ses propriétés. On le modifie éventuellement pour que le nouvel estimateur ait de meilleures propriétés (par exemple, la variance corrigée).

Méthode des moments Si on a M paramètres à estimer (θ_m), tel que : $m = (1, ..., M)$ .On résout un système de M équations à M inconnues (les M paramètres) en égalant pour $m = (1, ..., M)$ le moment d'ordre m de la population m et le moment d'ordre m de l'échantillon :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i)^m$$

Exemple 1.3.14 *L'estimateur du moment de* σ^2 *est alors la variance non corrigée de l'échantillon.*

Méthode du maximum de vraisemblance Les meilleures valeurs du paramétre inconnu θ sont celles qui donnent a l'évènement observe $(x_1, ..., x_n)$ la plus grande probabilité.

On a vu que cette probabilité peut être représentée par la vraisemblance :

$$\theta L(x;\theta) = f(x_1;\theta); ...; f(x_n;\theta)$$

L'estimation "maximum de vraisemblance"de θ sera une fonction des observations qui rend $L(x;\theta)$ maximum.

Remarque 1.3.15 *L 'estimateur de maximum de vraisemblance n'existe pas toujours et n'est pas toujours unique.*

1.3.4 Test d'hypothèse

Définition 1.3.16 *Une hypothèse est un ensemble des valeurs des paramètres inconnus (paramètres de population).*

Proposition 1.3.17 *Une hypothèse est dite simple si elle ne contient qu'un seul élément, ce qui est généralement le cas pour*

$$H_0 : \theta = \theta_0$$

tel que : H_0 test d'hypothèse nulle et Θ_0 est l'ensemble de valeurs supposée du paramètre θ selon H_0. sinon elle est composite.

$$H_1 : \theta \in \Theta_1$$

avec : Θ_1 un sous ensemble de l'ensemble des paramètres disjoint de θ_0.

Proposition 1.3.18

H_1 se ramène souvent aux trois cas suivants :

1. $H_1 : \theta < \theta_0$.
2. $H_1 : \theta > \theta_0$.
3. $H_1 : \theta \neq \theta_0$.
   - *Dans les deux premiers cas, le test est dit unilatéral et dans le dernier le test est dit bilatéral .*

Types du test d'ypothéses

test paramétrique

Définition 1.3.19 *Un test paramétrique est un test pour lequel d'hypothèses sur la distribution des populations sont requises.*

test non paramétrique On qualifie les test non paramétriques les méthodes statistiques qui sont applicables dans les conditions genérales quant aux distributions des populations.

Région de rejet d'un test et erreurs

Définition 1.3.20 *Soit X une variable aléatoire et χ l'ensemble de ses valeurs. le test s'effectue en trouvant un sous ensemble $R \subseteq \chi$ appelé la région de rejet.*

Ainsi, si $X \in R$, l'hypothèse nulle (H_0) est rejetée, sinon, elle est retenue.

Cette région se définit sous la forme suivante :

$$R = \{x : T(x) \succ s\}$$

avec : T est une statistique de test et s un seuil.

- En effectuant un test d'hypothèses, on peut se tromper en rejetant l'hypothèse nulle ou en l'acceptant.

Erreurs de test hypothèse

Il existe deux types d'erreur : l'erreur de première espèce (dite aussi erreur de type I) et l'erreur de deuxième espèce (dite aussi erreur de type II).

Erreur de première espèce

Définition 1.3.21 *L'erreur de première espèce correspond au cas où l'on rejette H_0(décider H_1) alors que celle-ci est vraie.*

Erreur de deuxième espèce

Définition 1.3.22 *L'erreur de deuxième espèce correspond quant à elle au cas où l'on rejette H_1(décider H_0) alors que celle-ci est vraie.*

Décision \ Vérité	H_0	H_1
H_0	décision correcte	erreur de deuxième espèce
H_1	erreur de première espèce	décision correcte

Récapitulatif des décisions en test d'hypothèse

risque du test hypothèse

risque de première espèce

Définition 1.3.23 *Le risque de première espèce est notée α. Il représente le risque de rejeter H_0 à tort. des valeurs de ce risque sont* 1% , 5% , 10% *qui correspondent aux niveaux de confiance* 99% , 95% *et* 90%.

risque de deuxième espèce

Définition 1.3.24 *Le risque de deuxième espèce représente quant à lui le risque de retenir H_0 à tort. Il est noté β.*

niveau de confiance du test

Définition 1.3.25 *Le niveau de confiance du test est donc $1-\alpha$ qui correspond à retenir H_0 à raison.*

puissance d'un test

Définition 1.3.26 *La puissance d'un test est la probabilité de rejeter l'hypothèse nulle à raison la puissance du test est donc le complément de l'erreur de deuxième espèce et est donc égale à $1-\beta$*

Décision \ Vérité	H_0	H_1
H_0	niveau de confiance $1-\alpha$	risque β
H_1	risque α	puissance de test $1-\beta$

Récapitulatif sur les risques associés à un test d'hypothèses

Chapitre 2

Caractéristiques d'un échantillon

Etude statistique portant sur tous les éléments d'une population étant, soit impossible à réaliser (trop grand nombre d'individus à étudier), soit trop onéreuse, il faut obtenir des résultats fiables sur les caractéristiques d'une population en se limitant à l'étude des éléments ou unités d'un échantillon. Cet échantillon doit non seulement donner des estimations non biaisées des paramètres mais permettre, de plus, d'évaluer la marge d'erreurs dues aux fluctuations d'échantillonnage.

L'échantillon doit être représentatif de la population ; il en résulte, en particulier,que chaque unité doit avoir une probabilité non nulle d'être tirée, un tel échantillon est qualifié d'aléatoire.

Définition 2.0.27 *On étudie une caractéristique mesurable X d'une population de taille finie ou infinie. La composition de la population, vis à vis du caractère X, est entièrement définie par la connaissance des quantités F(x) tel que : $F(x) = Proportion$ des individus tels que $X < x$, pour toutes les valeurs de $x \in R$.Soit E l'expérience consistant à choisir au hasard un élément de la population. Avant le tirage, on se propose de prévoir la valeur du caractère X que l'on obtiendra. Ce caractère est une variable aléatoire X telle que $Pr(X < x) = F(x)$ pour toute valeur $x \in R$. À l'expérience E est associée une variable aléatoire X dont la fonction de répartition est F(x). On réalise n fois la même expérience E, dans des conditions indépendantes.*

Définition 2.0.28 *Un echantillon aléatoire de taille n de la variable aléatoire X est une suite de variables aléatoires indépendantes $X_1; X_2; ......; X_n$ ayant toutes la même distribution que X.*

Une suite $x_1; x_2; ....; x_n$ de valeurs prises par les v.a. X_i est une réalisation de l'echantillon.

2.0.5 Caractéristiques d'un échantillon aléatoire

Une statistique définie sur un échantillon aléatoire X de taille n est une fonction mesurable des variables X_i . Les principales caractéristiques d'un échantillon sont les statistiques $\overline{X}$ et S^2.

On suppose que les moments d'ordre 1 et 2 de la variable aléatoire parente X existent et on pose

$$\begin{aligned} E(X) &= m \\ Var(X) &= \sigma^2 \end{aligned}$$

Remarque 2.0.29 *1. moment centré empirique d'ordre r est définie par :*

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^r$$

2. moment non centré empirique d'ordre r est définie par :

$$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

Étude de la statistique $\overline{X}$

Définition 2.0.30 *La statistique $\overline{X}$ est la fonction mesurable des variables X_i, définie par :*

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

Proposition 2.0.31 *1. **Cas d'une population infinie :**Un calcul rapide donné les résultats suivants :*

$$E\left(\overline{X}\right) = m \text{ }, Var\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

*2. **Cas d'une population finie :**L'échantillon de taille n est extrait d'une population finie de taille N par un tirage sans remise , On obtient les résultats suivants :*

$$E\left(\overline{X}\right) = m \text{ }, Var\left(\overline{X}\right) = \frac{N-n}{N-1}\frac{\sigma^2}{n} \simeq \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}$$

Le coefficient $\frac{n}{N}$ représente la fraction de la population constituant l'échantillon.

*3. **Comportement asymptotique :***

- *Loi faible des grands nombres*
  $\overline{X}$ *converge en probabilité vers* m *quand* n *tend vers l'infini.*
- *Loi forte des grands nombres*
  $\overline{X}$ *converge presque sûrement vers* m *quand* n *tend vers l'infini, car la série :*

$$\sum_{i=1}^{n} \frac{\sigma_i^2}{i^2} = \sigma^2 \sum_{i=1}^{n} \frac{1}{i^2}$$

  est une série convergente.

Appliquons le théorème central limite à la variable aléatoire Y_n :

$$Y_n = \frac{\sum_{i=1}^{n} X_i - nm}{\sigma\sqrt{n}} = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$$

La variable Y_n converge en loi vers une variable suivant la loi normale $N(0;1)$ quand n tend vers l'infini.

Remarque 2.0.32 *Pour appliquer le théorème central limite, il faut que les moments d'ordre* 1 *et* 2 *existent.*

Étude de la statistique S^2 ou variance empirique

Définition 2.0.33 *La variance empirique* $\mathbf{S^2}$ *d'un échantillon de taille* n *est définie par :*

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

Proposition 2.0.34

1. ***Espérance mathématique de*** $\mathbf{S^2}$

$$E\left(\mathbf{S^2}\right) = \frac{n-1}{n}\sigma^2 = \left(1 - \frac{1}{n}\right)\sigma^2$$

   Si la taille n *de l'échantillon est grande, l'espérance de* S^2 *a pour valeur limite* σ^2*.*

2. ***Variance de*** $\mathbf{S^2}$ *(Calcul un peu plus long mais non difficile).*

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

   avec μ_4 *étant le moment centré d'ordre* 4 *de la variable* X*.*
   Si la taille n *de l'échantillon est grande, la variance de* S^2 *a pour valeur limite :*

$$\frac{\mu_4 - \sigma^4}{n}$$

Théorème central limite pour $\mathbf{S^2}$ La variable aléatoire :

$$\frac{S^2 - \left(1 - \frac{1}{n}\right)\sigma^2}{\sqrt{Var\ S^2}}$$

converge en loi vers une variable suivant la loi normale $N(0;1)$ quand n tend vers l'infini.

En prenant les limites de l'espérance et de la variance pour n grand, on obtient le résultat suivant. La variable aléatoire :

$$\frac{S^2 - \sigma^2}{\sqrt{\frac{\mu_4 - \sigma^4}{n}}}$$

converge en loi vers la loi $N(0;1)$.

Corrélation entre $\overline{\mathbf{X}}$ et S^2

Pour définir la corrélation entre $\overline{X}$ et S^2, on calcule la covariance entre ces deux variables aléatoires :

$$Cov\left(\overline{X}, S^2\right) = \frac{n-1}{n^2}\mu^3$$

μ_3 étant le moment centré d'ordre 3 de la variable X.

Proposition 2.0.35 *Si n tend vers l'infini, la covariance entre ces variables tend vers 0, les statistiques $\overline{X}$ et S^2 sont donc asymptotiquementnon corrélées.*

Si la distribution de la variable X est symétrique, le moment centré μ_3 est égal à 0, les statistiques $\overline{X}$ et S^2 sont donc non corrélées quelle que soit la valeur de n.

Si de plus, X suit une loi normale, les statistiques $\overline{X}$ et S^2 sont indépendantes quelle que soit la valeur de n.

2.0.6 Distribution du chi-deux

Définition 2.0.36 *La variable aléatoire, égale à la somme des carrés de ν variables aléatoires indépendantes, centrées, réduites, gaussiennes, suit la loi du chi-deux χ^2 à ν degrés de liberté :*

$$\chi^2(\nu) = \sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2$$

Cette distribution a été introduite par Karl Pearson en 1900.

Proposition 2.0.37 *1. La variable aléatoire $\chi^2(\nu)$ varie de 0 à $+\infty$.*

2. *Le paramètre ν est le nombre de degrés de liberté de la variable, il représente la dimension de l'espace dans lequel se trouve le point représentatif de l'échantillon X. Si les variables aléatoires X_i vérifient k relations linéaires, le nombre de degrés de liberté diminue de k.*
3. *La loi suivie par la somme de variables aléatoires indépendantes, suivant chacune des lois du chi-deux, est une loi du chi-deux dont le degré de liberté est la somme des degrés de liberté de chaque loi.*
4. *Moments :*

$$\begin{aligned} E\left(\chi^2(\nu)\right) &= v \\ Var\left(\chi^2(\nu)\right) &= 2v \end{aligned}$$

5. *Mode :*

$$M_0 = \nu - 2 \quad si \quad n > 2.$$

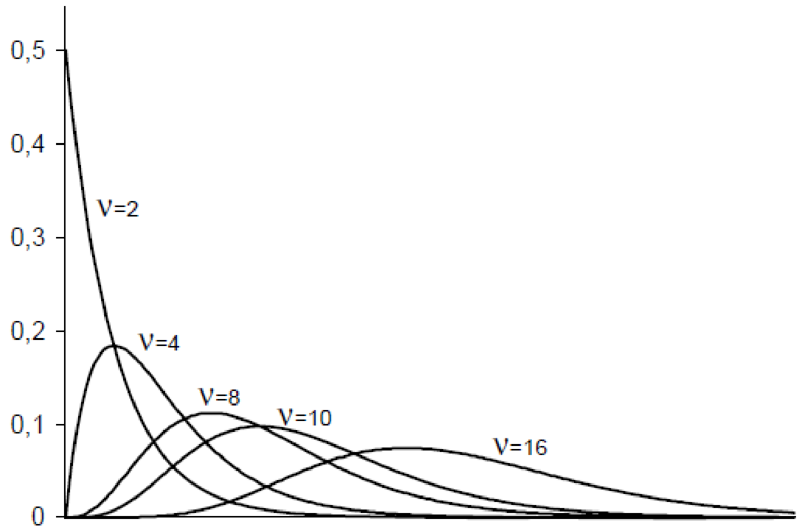
Fonction de densité

Définition 2.0.38 *La densité de probabilité de la variable aléatoire $\chi^2(\nu)$ est une expression mathématique compliquée et peu maniable :*

$$f\left(\chi^2(\nu)\right) = \frac{1}{2^{\nu/2}\Gamma\left(\nu/2\right)} \exp\left(\chi^2/2\right) \left(\chi^2\right)^{\nu/2-1}.$$

Γ *est la fonction eulérienne définie par :*

$$\Gamma(x) = \int_0^\infty \exp(-t)\, t^{x-1} dt \quad Pour\ x \succ 0$$



Densité de la loi du chi − deux pour différentes valeurs du paramètre ν.

Exemple 2.0.39 *on prend une valeur de degré de liberté par exemple*$(\nu = 4)$
donc $f\left(\chi^2(\nu)\right) = \frac{1}{2^2\Gamma(2)} \exp\left(\chi^2/2\right)\left(\chi^2\right)^{4/2-1}$ *avec* $\Gamma\left(2\right) = \int_0^\infty \exp\left(-t\right) t\ dt = 1$.
$E\left(\chi^2(4)\right) = 4$.
$Var\left(\chi^2(4)\right) = 8$.

2.0.7 Distribution de Fisher-Snédecor

Définition 2.0.40 *La distribution F de Fisher-Snédecor (ou plus simplement distribution F de Fisher) a été étudiée en 1924 par Fisher (statisticien anglais né en 1890) et calculée en 1934 par Snédecor ; elle joue un rôle important en analyse de la variance et en analyse de la régression.*

On considère deux variables aléatoires indépendantes suivant des lois du chi-deux à ν_1 *et* ν_2 *degrés de liberté respectivement.*

La variable aléatoire F *de Fisher est définie par :*

$$F(\nu_1; \nu_2) = \frac{\chi 2(\nu_1)/\nu_1}{\chi^2(\upsilon_2)/\upsilon_2}$$

Proposition 2.0.41 *Propriété évidente :* $F(\nu_1; \nu_2) = \frac{1}{F(\nu_2;\nu_1)}$

1. *La variable* $F(\upsilon_1; \upsilon_2)$ *varie de* 0 *à* $+\infty$.
2. *La loi de probabilité de la variable* F *dépend de deux paramètres, les degrés de liberté,*

ν_1 *et* ν_2.

3. *Moment :*
   $E(F) = \frac{\nu_2}{\nu_2 - 2} \qquad si\ \nu_2 \succ 2$
   $Var(F) = \left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{2(\nu_1 - \nu_2 - 2)}{\nu_1(\nu_2 - 4)} \quad si\ \ \nu_2 \succ 4$
4. *Mode :*
$$M_0 = \frac{\nu_2(\nu_1 - 2)}{\nu_1(\nu_2 - 2)} \qquad si\nu_1 > 2.$$

Fonction de densité

Définition 2.0.42 *La densité de probabilité de la variable aléatoire* $F(\nu_1;\nu_2)$ *est une expression mathématique compliquée et peu maniable :*

$$g(f) = \begin{cases} 0 & si\ f \preceq 0 \\ \frac{(\nu_1/\nu_2)^{\nu_1/2}}{\beta(\nu_1/2;\nu_2/2)} \times \frac{f^{(\nu_1/2-1)}}{\left(1+\frac{\nu_1}{\nu_2}f\right)^{(\nu_1+\nu_2)/2}} & sinon \end{cases}$$

avec β *est la fonction eulérienne définie par :*

$$\beta(p;q) = \int_0^1 t^{(p-1)} (1-t)^{(q-1)} dt.$$

Densité de la loi de Fisher − Snedecor pour différentes valeurs du paramètre ν_1, ν_2

Exemple 2.0.43 *on prend* $\nu_1 = 4$ *et* $\nu_2 = 6$.

$$donc\ g(f) \begin{cases} 0\ si\ f \preceq 0 \\ \frac{(4/6)^{4/2}}{\beta(4/2;6/2)} \times \frac{f^{4/2-1}}{\left(1+\frac{4}{6}f\right)^{(4+6)/2}}\ sinon \end{cases}$$

avec $\beta(2;3) = \int_0^1 t(1-t)^2 = 1/12$

$E(F) = \frac{6}{6-2} = 3/2 \qquad si\ \nu_2 \succ 2$

$Var(F) = \left(\frac{6}{6-2}\right)^2 \frac{2(4-6-2)}{4(6-4)} = -36/16\ \ si\ \ \nu_2 \succ 4$

2.0.8 Distribution de Student

Définition 2.0.44 *La distribution T de Student, pseudonyme du statisticien anglais W.S. Gosset (1876-1937), joue un rôle important dans l'étude de la statistique X pour une distribution normale dont on ne connaît pas la variance.*

La loi de Student est la loi de la variable aléatoire T définie par :

$$T^2(\nu) = \frac{U^2}{\chi^2(\nu)/\nu} = F(1;\nu) \ ou \ T(\nu) = \frac{U}{\sqrt{\chi^2(\nu)/\nu}}$$

où U est une variable aléatoire centrée réduite normale.

Proposition 2.0.45 *1. La variable $T(\nu)$ varie de $-\infty$ à $+\infty$.*

2. *La loi de probabilité de la variable $T(\nu)$ dépend d'un paramètre, le degré de liberté ν de la variable $\chi^2(\nu)$.*
3. *Moments :*

$$\begin{aligned} E[T(\nu)] &= 0 \\ Var[T(\nu)] &= \frac{\nu}{\nu-2} \quad si \nu > 2 \end{aligned}$$

4. *Mode :*

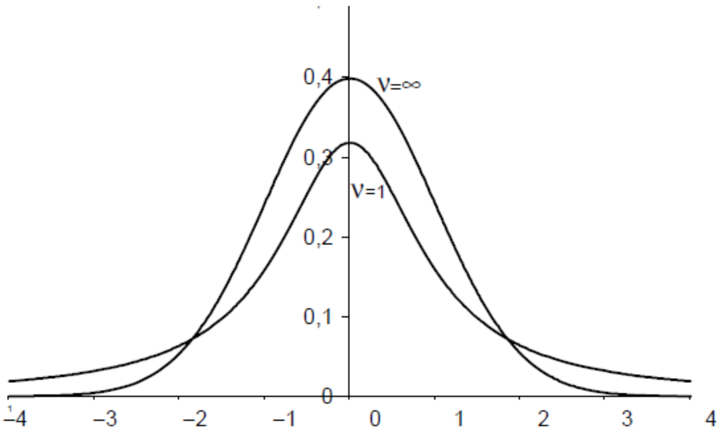
$$M_0 = 0$$

Fonction de densité

Définition 2.0.46 *La densité de probabilité de la variable aléatoireT est une expression mathématique compliquée et peu maniable :*

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \times \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} \\ &= \frac{1}{\sqrt{\nu}} \frac{1}{\beta\left(\frac{1}{2};\frac{\nu}{2}\right)} \left(1 + \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}\right) \end{aligned}$$

Γ *fonction eulérienne .*



Densité de la loi de Student.

Exemple 2.0.47 *on prend* $\nu = 2$

donc $f(t) = \frac{1}{\sqrt{2}} \frac{1}{\beta\left(\frac{1}{2};\frac{2}{2}\right)} \left(1 + \left(1 + \frac{t^2}{2}\right)^{-(2+1)/2}\right)$ *avec*$\beta\left(\frac{1}{2};\frac{1}{2}\right) = \int_0^1 t^{\frac{-1}{2}} (1-t)^0 = 2$

2.0.9 Cas particulier d'échantillons gaussiens

L'échantillons considérés dans ce paragraphe sont issus d'une population suivant la loi normale $N(m;\sigma)$ et les propriétés démontrées ne sont valables que sous cette hypothèse.

Étude de la statistique $\overline{X}$

La variable $\overline{X}$, combinaison linéaire de n variables aléatoires indépendantes gaussiennes, est une variable gaussienne.

Donc, $\forall n$: la loi de la variable X est la loi $N\left(m; \frac{\sigma}{\sqrt{n}}\right)$.

Étude de la statistique S^2

La décomposition de la statistique S^2

$$nS^2 = \left[\sum_{i=1}^{n} (X_i - m)^2\right] - n\left(m - \overline{X}\right)^2$$

et le théorème de Cochran sur la décomposition d'une forme quadratique conduisent au résultat suivant :

$$\sum_{i=1}^{n} \left(\frac{Xi - m}{\sigma}\right)^2 = \frac{n\ S^2}{\sigma^2} + \left(\frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}\right)^2$$

- Le premier membre, somme de n carrés de variables aléatoires centrées réduites, indépendantes, gaussiennes est une variable $\chi^2(n)$.
- Le deuxième membre est une somme de deux formes quadratiques :
  - la première est de rang $(n-1)$, car les variables vérifient la relation :

$$\sum_{i=1}^{n}\left(Xi-\overline{X}\right)=0$$

  - la deuxième est de rang 1.

On en déduit les deux résultats suivants :

1. $\frac{n\,S^2}{\sigma^2}$ est une variable $\chi^2(n-1)$.
2. $\overline{X}$ et S^2 sont deux variables indépendantes.

On démontre la réciproque suivante, qui est une propriété caractéristique des variables aléatoires gaussiennes

Si les statistiques $\overline{X}$ et S^2 sont indépendantes, la variable aléatoire X est une variable aléatoire gaussienne.

Conclusion Générale

En fin toute démarche statistique consiste :

A prélever un échantillon aléatoire de la population par des techniques appropriées. Les différentes méthodes utilisées pour obtenir un tel échantillon relèvent de la théorie de l'échantillonnage, quelques méthodes seront expliquées.

A étudier les principales caractéristiques d'un échantillon, issu d'une population dont on connaît la loi de probabilité, le cas particulier des échantillons issus d'une population normale . Les lois du chi-deux, de Fisher-Snedecor, de Student, lois dérivées de la loi normale, ayant de nombreuses applications dans la théorie de l'estimation et des tests.

Bibliographie

[1] **B. Fayolle** : (2010) Statistique inférentielle version complète , 7, 24, 26.

[2] **B. JOURDAIN** : (2013) Probabilites et statistique , 49.

[3] **D. Concordet** : Introduction a la statistique inferentielle , 7-35.

[4] **F. Chamroukhi** : (2013) Éléments de statistique inférentielle, 8-13, 144-150.

[5] **F. Gilbert et S. Digabel** : (2013) Distributions d'echantillonnage, 5.

[6] **I. Lyon** : (2010) Cours de probabilit´es et statistiques,15 , 23, 24 , 27-34.

[7] **J.-Christophe Breton** : (2008) Statistiques IUT Biotechnologie , 10, 12-13.

[8] **J.JACQUES** : Statistiques inférentielles département, 19-22, 25.

[9] **J. Vaillant** : (2005) Initiation a la théorie de l' echantillonnage , 1-3.

[10] **M. Gentes** : (2010) Cours de Probabilités et Statistiques , 13-28.

[11] **P.Brachet** : Statistique : Résumé de cours et méthodes , 1-6.

[12] **R. Veysseyre** : Statistique et probabilités pour l'ingénieur , 179-192.

[13] **S. Geffray** : (2009) Statistique descriptive et inferentielle.

[14] **S. Robin** : (2007) élements de la théorie de l'´echantillonnage , 7, 11, 17, 18.