ECHANTILLONNAGE

I. LE MODELE D'INFERENCE STATISTIQUE

I.1. Situation statistique.

[illegible] statistique, on appelle population un ensemble fini $\mathcal{P}$ dont le nombre d'éléments est noté N. Les éléments de $\mathcal{P}$ seront appelés individus et N est la taille de la population. Cette population peut être de nature très variée : population animale, végétale, agents économiques, pièces d'un échéancier d'un certain type, logements dans une ville donnée, etc.

Pour la population considérée on s'intéresse à un caractère statistique $\mathcal{X}$ concernant les individus ; à chaque individu de $\mathcal{P}$ on va associer une modalité du caractère $\mathcal{X}$; celles-ci sont appelées valeurs des modalités sont appelées valeurs.

Supposons que le caractère $\mathcal{X}$ est quantitatif discret et prend [illegible] distinctes $c_1, \dots, c_k$. On notera x_i la valeur du caractère pour le $i^{ème}$ individu de la population $\mathcal{P}$. Les x_i ne sont pas nécessairement distincts.

Si on note n_j le nombre d'individus de $\mathcal{P}$ pour lesquels la valeur de $\mathcal{X}$ est c_j, la **distribution (ou répartition)** du caractère $\mathcal{X}$ dans la population $\mathcal{P}$ consiste en la donnée du tableau 1 :

valeurs de X	c_1	c_2	...	c_k
effectifs	n_1	n_2	...	n_k

avec $\sum_{i=1}^{k} n_i = N$, ou du tableau 2 :

valeurs de X	c_1	c_2	...	c_k
fréquences	f_1	f_2	...	f_k

avec $f_j = \frac{n_j}{N}$, pour $j = 1 \dots k$ et $\sum_{j=1}^{k} f_j = 1$.

On peut remarquer que le tableau 2 précédent est analogue à celui définissant la loi de probabilité d'une variable aléatoire discrète finie.

Le problème du statisticien :
En général, la distribution du caractère $\mathcal{X}$ dans $\mathcal{P}$ est inconnue et le statisticien souhaite la connaître au moins partiellement ; il peut s'intéresser par exemple à la moyenne m du caractère quantitatif $\mathcal{X}$ dans la population encore appelée moyenne théorique définie par :

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{j=1}^{k} n_j c_j = \sum_{j=1}^{k} f_j c_j$$

il peut s'intéresser à la variance du caractère $\mathcal{X}$ dans la population, ou variance théorique,

$$v = \frac{1}{N}\sum_{i=1}^{N} (x_i - m)^2 = \frac{1}{N}\sum_{j=1}^{k} n_j (c_j - m)^2 = \sum_{j=1}^{k} f_j (c_j - m)^2$$

I.2 Echantillon aléatoire.

Pour connaître exactement la distribution de $\mathcal{X}$ il suffirait de noter la valeur du caractère pour chaque individu de la population $\mathcal{P}$. En pratique, cette solution est rarement réalisable car elle conduit à un coût d'observation et/ou à un délai de réalisation trop élevé voire même absurde (dans le cas où l'observation est destructrice, par exemple lorsqu'il s'agit du seuil de rupture d'une pièce mécanique).

C'est pourquoi, généralement on limite l'observation à un sous ensemble $\mathcal{E}$ de $\mathcal{P}$ que l'on appelle échantillon. Le choix de cet échantillon peut-être fait de façon raisonnée ou aléatoire. Seul le choix aléatoire, c'est-à-dire par tirage au hasard, permet d'utiliser les résultats du calcul des probabilités. C'est dans ce cadre que nous nous placerons dorénavant.

Supposons pour commencer que l'on tire au hasard un seul individu dans $\mathcal{P}$ et que le tirage est réalisé avec équiprobabilité. La probabilité affectée à chaque individu est donc égale à 1/N. Désignant par X la valeur du caractère $\mathcal{X}$ pour l'individu choisi, on voit que X est une variable aléatoire puisque sa valeur dépend directement du tirage. La loi de probabilité de X est donnée par :

$$P(X=c_j) = \frac{n_j}{N} = f_j \text{ pour tout } j = 1, \ldots k.$$

En effet, le tirage étant équiprobable, pour calculer $P(X=c_j)$, on peut appliquer la règle : "nombre de cas favorables divisé par le nombre de cas possibles" (ici n_j et N). On constate alors que la loi de la v.a. X est donnée par le même tableau que celui de la distribution du caractère statistique $\mathcal{X}$ dans la population ; on a alors : $E(X) = m$ et $Var(X) = v$.

notion d'échantillon :
Plus généralement, si l'on tire n individus dans la population $\mathcal{P}$, alors l'échantillon est de taille n. Au $i^{ème}$ individu de l'échantillon on associe la v.a. X_i égale à la valeur du caractère $\mathcal{X}$ observée pour le $i$ème individu.

Ainsi au tirage d'un échantillon $\mathcal{E}$ de taille n dans $\mathcal{P}$ est associé le n-uplet de variables aléatoires $(X_1, \ldots, X_n)$ où

(1) Les v.a. $X_1, \ldots, X_n$ ont même loi que la v.a. X associée au caractère $\mathcal{X}$,

(2) Les v.a. $X_1, \ldots, X_n$ sont indépendantes en probabilité :
- de façon exacte, si les tirages ont lieu avec remise,
- de façon approchée, si les tirages ont lieu sans remise, mais que le taux de sondage $\frac{n}{N}$ reste inférieur à 10% $(\frac{n}{N} < \frac{1}{10})$.

Ceci nous amène à dépasser la notion d'échantillon comme sous-ensemble de $\mathcal{P}$ et à donner une deuxième définition de l'échantillon.

Définition d'un échantillon :
Etant donnée une v.a. X, on appellera **n-échantillon** de X le n-uplet $(X_1, \ldots X_n)$ de v.a. **indépendantes et de même loi** que X.

CamScanner

En pratique : les échantillons sont toujours tirés sans remise, on sera alors amené dans les exercices à vérifier ou à supposer que la condition $\frac{n}{N} < \frac{1}{10}$ est bien satisfaite.

I.3 Statistiques d'échantillonnage.

La loi de probabilité de l'échantillon $(X_1, ..., X_n)$ est la loi, non pas d'une variable aléatoire, mais d'un n-uplet de variables aléatoires.; on dit aussi que $(X_1, ..., X_n)$ est un vecteur aléatoire. La loi d'un vecteur aléatoire n'est pas facile à manier en pratique on considérera l'échantillon $(X_1, ..., X_n)$ dans sa globalité est souvent inutile. Ainsi dans la pratique on s'intéressera à une fonction des variables $X_1, ..., X_n$ ou statistique d'échantillonnage

Définition d'une statistique :
On appelle statistique d'échantillonnage ou plus simplement statistique toute variable aléatoire fonction des v.a. $X_1, ..., X_n$ de l'échantillon ; autrement dit, une statistique T est de la forme : $T = g(X_1, ..., X_n)$ où g est une fonction de n variables réelles

Exemples de statistiques usuelles :

La moyenne empirique :
Elle est notée $\bar{X}_n$ ou simplement $\bar{X}$ et elle est définie par :

$$\bar{X} = \frac{X_1 + .. + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

La variance empirique :
Il s'agit de la variance du caractère quantitatif $\mathcal{X}$ dans l'échantillon soit

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

Les valeurs extrêmes de l'échantillon :
La plus petite des valeurs prises par $X_1, ..., X_n$ est notée $Y_1 = \min (X_1, ..., X_n)$ et la plus grande $Y_n = \max (X_1, ..., X_n)$.

I.4 Inférence statistique et échantillonnage

Les notions introduites précédemment vont nous permettre d'expliquer ce qu'est l'**inférence statistique** et comment l'**échantillonnage** intervient dans cette démarche.

L'Inférence :
L'objectif du statisticien est d'obtenir des informations sur la distribution du caractère $\mathcal{X}$ dans la population $\mathcal{P}$ sans pour celà observer la modalité prise par $\mathcal{X}$ sur chaque individu de la population.

A partir de l'observation du caractère $\mathcal{X}$ dans un échantillon $\mathcal{E}$, prélevé au hasard dans $\mathcal{P}$, on s'attache à construire une image approximative de la distribution du caractère dans la population ; par exemple, on peut connaître une valeur approchée de certains paramètres de cette distribution, tels que la moyenne m ou la variance v.
On verra par exemple au chapitre traitant de l'estimation statistique que l'on peut déterminer une valeur approchée de la moyenne m du caractère $\mathcal{X}$ dans la population, à partir d'une réalisation de la moyenne empirique $\bar{X}$ observée sur un échantillon.

CamScanner

Le passage de l'échantillon à la population constitue la démarche d'inférence ou inductive statistique.

L'Echantillonnage :
La démarche inverse qui, connaissant la distribution du caractère $\mathcal{X}$ dans la population, examine comment varient les échantillons aléatoires qui y sont prélevés, fait l'objet de la théorie de **l'échantillonnage**.

Échantillonnage et inférence utilisent les mêmes outils probabilistes. C'est pourquoi nous appellerons la théorie de l'inférence la situation stochastique modélisée sur laquelle ces deux démarches opèrent.

I.3 Détermination du modèle d'inférence

Préalablement à toute étude relevant de l'échantillonnage ou de l'inférence statistique, il convient d'identifier avec précision le modèle d'inférence dans lequel on s'astreint à travailler. Il est spécifié :

(1) la population concernée, éventuellement sa taille N,

(2) le caractère statistique étudié $\mathcal{X}$, les caractéristiques de sa distribution dans la population

(3) l'échantillon et sa taille n, les hypothèses faites sur le prélèvement d'échantillon

(4) X la variable du modèle, les hypothèses sur la loi de X et les paramètres du modèle.

II. QUELQUES STATISTIQUES USUELLES ET LEURS PROPRIETES

Dans ce paragraphe, on se propose d'étudier quelques statistiques usuelles ; on étudiera notamment la moyenne empirique $\bar{X}$, la fréquence empirique F_n, la variance empirique S^2 ou la variance empirique corrigée S'^2.

On rappelle que la moyenne du caractère quantitatif $\mathcal{X}$ dans la population est notée m, la variance v et l'écart-type σ ($\sigma^2 = v$).

Le fait de tirer l'échantillon au hasard dans la population permet d'associer à $\mathcal{X}$ une variable aléatoire notée X, vérifiant : $E(X) = m$ et $Var(X) = v = \sigma^2$, X est la variable du modèle.

II.1 Moyenne empirique.

Définition de la moyenne empirique :

Elle est notée $\bar{X}_n$ ou simplement $\bar{X}$ et elle est définie par :

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

ses réalisations sont de la forme : $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ où $(x_1, x_2, \ldots x_n)$, est une réalisation du n-uplet de v.a. $(X_1, X_2, \ldots X_n)$.

La moyenne empirique $\bar{X}$ est la moyenne arithmétique des variables de l'échantillon aléatoire $(X_1, X_2, \ldots X_n)$.

CamScanner

Il faut bien insister sur le fait que la moyenne empirique $\bar{X}$ est une variable aléatoire tandis que $\bar{x}$ est une réalisation de $\bar{X}$, $\bar{x}$ est une valeur numérique.

La v.a. $\bar{X}$, sa loi, ses caractéristiques numériques dépendent de n taille de l'échantillon, ce qui est précisé dans la proposition suivante :

Propriétés de la moyenne empirique $\bar{X}$:

Soit $(X_1, ..., X_n)$ un n-échantillon de la v.a. X. Si l'on suppose que $E(X^2)$ existe alors, pour tout entier $n \geq 1$, la moyenne empirique $\bar{X}$ vérifie :

$$E(\bar{X}) = m \quad \text{et} \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$

preuve :

En utilisant la linéarité de l'espérance mathématique on obtient d'abord :

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

Comme les v.a. $X_1, ..., X_n$ sont de même loi, on a pour tout $i = 1, ..., n$, $E(X_i) = m$ et donc

$$E(\bar{X}) = \frac{n\,m}{n} = m.$$

Pour la variance de $\bar{X}$ on a tout d'abord :

$$Var(\bar{X}) = \frac{1}{n^2} Var(X_1 + ... + X_n)$$

Les v.a. $X_1, ..., X_n$ étant indépendantes, on a ensuite :

$$Var(\bar{X}) = \frac{1}{n^2} (Var(X_1) + ... + Var(X_n))$$

de plus elles ont même loi que X, il vient alors :

$$Var(\bar{X}) = \frac{1}{n^2} n\,Var(X) = \frac{Var(X)}{n} = \frac{\sigma^2}{n}$$

On peut interpréter ces résultats en disant que, "en moyenne", la v.a. $\bar{X}$ est égale à m avec une dispersion d'autant plus faible que n est grand.

II.2 La variance empirique

Définition de la variance empirique :

Il s'agit de la variance dans l'échantillon, elle est notée S_n^2 ou plus simplement S^2 :

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

Cette deuxième expression résulte de la formule de Huyghens vue en statistique descriptive : $s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$.

Une autre statistique sera utile en estimation :

la **variance empirique corrigée** S'^2_n ou plus simplement notée S'^2 définie par :

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n}{n-1} S^2$$

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

CamScanner

Propriétés de la variance empirique et de la variance

Soit X une v.a. telle que $m_4 = E(X^4)$ existe, alors, pour tout $n > 1$, on a :

(a) $E(S^2) = \frac{(n-1)v}{n}$ et $E(S'^2) = v$

(b) $Var(S^2)$ et $Var(S'^2)$ sont majorées par une expression de la forme $\frac{c}{n}$, où c est une constante faisant intervenir les moments d'ordre 4 de la v.a. X_i (c = ...)

Preuve : on se contentera de montrer (a).

Pour commencer on remarque que si l'on remplace chaque v.a. X_i par $X_i - m$, on ne modifie ni S^2, ni S'^2 ce qui permet de supposer que les X_i sont centrées.

Pour montrer (a), on utilise la relation :

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

et l'on prend l'espérance membre à membre, ce qui donne

$$E(S^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\bar{X}^2)$$

Or, on a supposé que les X_i sont centrées, donc pour tout $i = 1, ..., n$

$$E(X_i) = 0 \quad \text{et} \quad E(X_i^2) = Var(X_i) = v$$

Alors :

$$E(\bar{X}) = 0 \quad \text{et} \quad E(\bar{X}^2) = Var(\bar{X}) = \frac{v}{n}$$

On en déduit que :

$$E(S^2) = v - \frac{v}{n} = \frac{(n-1)v}{n} \quad \text{et} \quad E(S'^2) = v.$$

II.3 La fréquence empirique

Définition :

On considère un caractère qualitatif $\mathcal{X}$ à deux modalités A et $\bar{A}$ que l'on peut coder par 1 et 0 ; on note p la proportion des individus qui possèdent la modalité A dans la population ; on dit aussi que p est la fréquence théorique de A. Dans ces conditions la variable aléatoire Y égale au nombre d'individus de l'échantillon qui présentent la modalité A, s'écrit : $Y = X_1 + ... + X_n$ où les v.a. $X_1, ..., X_n$ sont des v.a. de Bernoulli de paramètre p et indépendantes. Pour tout $i = 1, ..., n$ la v.a. X_i prend la valeur 1 si le ième individu de l'échantillon possède la modalité A et prend la valeur 0 sinon.

On pose alors : $F = \frac{Y}{n}$

La v.a. F est appelée **fréquence empirique** ; elle représente la **proportion dans l'échantillon**, des individus pour lesquels la modalité de $\mathcal{X}$ est A.

Remarques: F est un cas particulier de moyenne empirique; la v.a. F, sa loi, ses caractéristiques numériques dépendent de n, taille de l'échantillon.

Propriétés :

La fréquence empirique vérifie : $E(F) = p$ et $V(F) = \frac{1}{n} p(1-p)$.

Par ailleurs, compte-tenu de l'indépendance des v.a. $X_1, ..., X_n$, la v.a. $Y = nF$ suit une loi binomiale de paramètres n et p.

CamScanner

II.4 La distribution empirique

On appelle **distribution empirique** la distribution du caractère $\mathcal{X}$ dans l'échantillon. L'échantillon étant prélevé au hasard, les effectifs ou les fréquences qui interviennent dans cette distribution sont aléatoires.

On notera $c_1, ..., c_k$ les classes obtenues après regroupement des valeurs de $\mathcal{X}$ et pour tout $j = 1, ..., k$ on notera N_j l'effectif de la classe c_j et $F_j = \frac{N_j}{n}$ la fréquence de la classe c_j.

N_j et F_j sont des **variables aléatoires**; on peut remarquer que F_j est la fréquence empirique associée au caractère qualitatif dont les deux modalités possibles sont :

A_j = "appartenir à la classe c_j" et $\bar{A}_j$ = "ne pas appartenir à la classe c_j".

Ces variables vérifient les relations : $N_1 + ... + N_k = n$ et $F_1 + ... + F_k = 1$ et ne sont donc pas indépendantes.

Enfin si p_j est la probabilité de la classe c_j, la variable N_j suit une loi $B(n,p_j)$.

III. PROPRIETES ASYMPTOTIQUES DES STATISTIQUES USUELLES

Dans ce paragraphe, on va donner sans démonstration deux théorèmes fondamentaux de la théorie de l'échantillonnage. Ces théorèmes permettent d'étudier le comportement asymptotique des principales statistiques usuelles, autrement dit ils fournissent des propriétés des statistiques d'échantillonnage applicables seulement si la taille de l'échantillon est suffisamment grande. Ils permettent également de justifier l'approximation normale des lois binomiale et de Poisson faite au chapitre I.

III.1 Convergence en proba des moments empiriques d'ordre 1 et 2.

Dans ce paragraphe, les statistiques d'échantillonnage seront indicées par la taille de l'échantillon, puisque désormais on s'intéresse aux échantillons de grande taille.

III.1.1 La convergence en probabilité d'une suite de v.a.

On considère une suite de v.a. $(X_n)_{n\in\mathbb{N}}$ et un nombre réel a.

Définition :
On dit que la suite $(X_n)_{n\in\mathbb{N}}$ converge en probabilité vers la constante a, ssi pour tout nombre ε positif arbitrairement fixé, on a :

$$\lim_{n\to\infty} P(|X_n - a| > \varepsilon) = 0 \text{ ou ce qui est équivalent :}$$
$$\lim_{n\to\infty} P(|X_n - a| \leq \varepsilon) = 1$$

On note alors : $\lim_{n\to\infty} X_n = a$ (Pr) ou $X_n \xrightarrow{Pr} a$

Critère de convergence en probabilité :
Lorsque $E(X_n)$ tend vers a et $V(X_n)$ tend vers 0, quand $n \to \infty$ alors la suite $(X_n)_{n\in\mathbb{N}}$ converge en probabilité vers a.

exemple :
Une expérience aléatoire consiste à lancer n fois une pièce de monnaie non truquée. On note X_n le nombre de piles obtenus au cours des n lancers. Alors $F_n = \frac{X_n}{n}$ est la fréquence d'apparition de pile au cours de n lancers. On montrera que F_n converge en probabilité

CamScanner

vers $\frac{1}{2}$. Autrement dit, la fréquence d'apparition de pile au cours de n lancers d'une pièce non truquée, converge en probabilité vers $\frac{1}{2}$ quand n tend vers l'infini.

Définition :
On dit que la suite $(X_n)_{n\in\mathbb{N}}$ converge en probabilité vers une variable aléatoire X ssi la suite $(X_n - X)_{n\in\mathbb{N}}$ converge en probabilité vers 0

II.1.2 Propriétés de la convergence en probabilité

Le théorème suivant montre que une fonction continue au sens usuel, reste continue pour la convergence en probabilité des suites de variables aléatoires.

Théorème de Slutsky :
On considère une fonction continue g, de $\mathbb{R}^2$ dans $\mathbb{R}$ et deux suites de v.a. $(Y_n)_{n\in\mathbb{N}}$ et $(Z_n)_{n\in\mathbb{N}}$ convergeant en probabilité respectivement vers les v.a. Y et Z. Alors la suite des v.a. $(g(Y_n, Z_n))_{n\in\mathbb{N}}$ tend en probabilité vers la v.a. g(Y, Z).

En remplaçant successivement la fonction g par l'addition, la multiplication et la division, considérées comme des fonctions de deux variables, on en déduit les résultats suivants :

Corollaire 1 :
Si (Y_n) et (Z_n) sont deux suites de v.a. convergeant en probabilité respectivement vers les v.a. Y et Z, alors on a :

(i) $Y + Z = \lim_{n\to\infty} (Y_n + Z_n)$ (Pr)

(ii) $YZ = \lim_{n\to\infty} Y_n Z_n$ (Pr)

(iii) $\frac{Y}{Z} = \lim_{n\to\infty} \frac{Y_n}{Z_n}$ (Pr) à condition que $Pr(Z=0) = 0$

Corollaire 2 :
Dans le cas où g est une fonction continue d'une seule variable, le théorème de Slutsky reste valable et permet d'affirmer que si (X_n) converge en probabilité vers X alors $g(X_n)$ converge en probabilité vers g(X) autrement dit, on a :

$$g(X) = \lim_{n\to\infty} g(X_n) \quad (Pr)$$

II.1.3 La loi des grands nombres :

La loi des grands nombres affirme une convergence de la moyenne empirique $\bar{X}_n$ vers la moyenne théorique $m = E(X)$, lorsque la taille n de l'échantillon tend vers l'infini.

- **Théorème 1 : Loi des Grands Nombres (LGN)**
  Si X est une v.a. telle que $m = E(X)$ existe, alors pour toute suite infinie $(X_n)_{n\in\mathbb{N}}$ de v.a. **deux à deux indépendantes** et de même loi que X, la suite des v.a. :

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (1)$$

  converge en probabilité vers $m = E(X)$ lorsque n tend vers l'infini.

Interprétation :
Ce théorème montre que la probabilité pour que $\bar{X}_n$ prenne une valeur voisine de m est d'autant plus grande que la taille n de l'échantillon est élevée.

CamScanner

Dans le *cas particulier* d'une population où l'on étudie un caractère qualitatif à deux modalités A et $\bar{A}$, on voit donc que la fréquence empirique d'observation de la modalité A, soit F_n, converge en probabilité vers la fréquence théorique p de cette modalité, autrement dit, on a : $\lim_{n\to\infty} F_n = p$ (Pr.)

III.1.4 Convergence en probabilité de la variance empirique et de l'écart type empirique :

Les propriétés de la convergence en probabilité permettent de montrer que la variance empirique S_n^2 et la variance empirique corrigée S'^2_n tendent en probabilité vers la variance théorique $v = \sigma^2$, lorsque n tend vers l'infini, à condition que $E(X^2)$ (et donc aussi v) soit finie.

En effet, d'après la loi des grands nombres, on a :

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i^2 = E(X^2) \text{ (Pr).}$$

De plus, en utilisant le corollaire 1(ii) du thm de Slutsky avec $X_n - Z_n = X_n$ on voit que :

$$\lim_{n\to\infty} (\bar{X}_n)^2 = m^2 = [E(X)]^2 \text{ (Pr).}$$

Il en résulte que la variance empirique :

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

tend en probabilité vers $v = E(X^2) - [E(X)]^2$ lorsque n tend vers l'infini.
La convergence de S'^2_n vers v s'en déduit immédiatement.

En utilisant le corollaire 2 du théorème de Slutsky avec la fonction $g : u \to g(u) = \sqrt{u}$, on peut aussi montrer que S_n et S'_n convergent en probabilité vers σ, l'écart-type de X.

III.1.5 La convergence en moyenne quadratique

Définition :
Soit (X_n) une suite de variables aléatoires admettant des moments d'ordre 1 et 2. La suite (X_n) converge en moyenne quadratique vers la variable X et on note $X_n \xrightarrow{mq} X$ ssi la suite des nombres réels ≥ 0, $(E(X_n - X)^2)$ converge vers 0, quand n tend vers l'infini.

Critère de convergence en m.q. :
Si (X_n) est une suite de v.a. vérifiant : $\lim_{n\to\infty} E(X_n) = a$ et $\lim_{n\to\infty} V(X_n) = 0$ alors (X_n) converge vers a en moyenne quadratique.

exemple : dans l'exemple des n lancers de la pièce de monnaie, on a $E(F_n) = \frac{1}{2}$

et $V(F_n) = \frac{1}{4n}$ donc $F_n \xrightarrow{mq} \frac{1}{2}$.

Définition : $(X_n)_{n\in\mathbb{N}}$ converge vers X en moyenne quadratique ssi la suite $(X_n - X)_{n\in\mathbb{N}}$ converge en moyenne quadratique vers 0.

Remarque :
La convergence de (X_n) vers X en moyenne quadratique implique la convergence de (X_n) vers X en probabilité.

CamScanner

III.2 Loi asymptotique des moments empiriques d'ordre 1 et 2.

III.2.1 La convergence en loi

Cette convergence est plus faible que les deux précédentes. Cependant, en pratique elle permet d'approcher une loi par une autre.



Définition

Soit (X_n) une suite de variables aléatoires, on note G_n la fonction de répartition de X_n. On dit que la suite (X_n) converge en loi vers la variable X de fonction de répartition G si pour tout $x \in \mathbb{R}$ où G est continue la suite $(G_n(x))$ converge vers $G(x)$ dans $\mathbb{R}$, et on note :

$X_n \xrightarrow{L} X$

Remarques :

(1) Pour des variables discrètes la convergence en loi vers une variable discrète s'exprime par la convergence des probabilités ponctuelles autrement dit par la convergence de $P(X_n=x)$ vers $P(X=x)$ quand n tend vers l'infini.
C'est ainsi qu'on a établi la convergence de la loi binomiale vers la loi de Poisson.
Cependant une suite de v.a. discrètes peut converger en loi vers une variable continue (voir plus loin).

(2) Pour des variables continues, on montre que si X_n admet pour densité f_n et X pour densité f alors $(X_n)_{n\in\mathbb{N}}$ converge en loi vers X, si pour tout réel x, $f_n(x)$ converge vers $f(x)$.

(3) La convergence en probabilité implique la convergence en loi mais la réciproque est fausse en général.
Néanmoins lorsque la limite est une valeur certaine, les deux convergences sont équivalentes.

(4) En résumé, on peut schématiser la hiérarchie des convergences comme suit :

convergence en m.q. $\Rightarrow$ convergence en probabilité $\Rightarrow$ convergence en loi



III.2.2 Convergence en loi de la moyenne empirique :

On vient de voir que la moyenne empirique tend en probabilité vers la moyenne théorique lorsque n tend vers l'infini. Ce résultat va être précisé par le théorème central limite qui donne des informations sur la loi de $\bar{X}_n$ quand n est grand.



• Théorème central limite :

Si X est une v.a. de loi quelconque telle que $E(X^2)$ existe et si $(X_n)_{n\geq 1}$ est une suite infinie de v.a. indépendantes et de même loi que X, alors la suite des v.a. :

$$U_n = \frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \quad \text{où} \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \ , \ m = E(X) \ , \ \sigma^2 = V(X)$$

converge en loi vers la loi normale $\mathcal{N}(0,1)$ lorsque n tend vers l'infini.
Autrement dit, pour tout u on a : $F_0(u) = \lim_{n\to\infty} F_n(u)$ où F_n désigne la fonction de répartition de la v.a. U_n et F_0 la fonction de répartition de la loi $\mathcal{N}(0,1)$.

En pratique : on pourra considérer que, si la taille n de l'échantillon est suffisamment grande ($n \geq 50$), la loi de U_n peut être assimilée à la loi $\mathcal{N}(0,1)$. Donc le TCL nous permet

CamScanner

d'affirmer que si n est grand ($n \geq 50$), la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ suit façon approchée la loi normale $\mathcal{N}\left(m, \frac{\sigma}{\sqrt{n}}\right)$ où $m = E(X)$, $\sigma^2 = V(X)$.

Autrement dit la moyenne empirique $\bar{X}_n$ est asymptotiquement normalement distribuée.

III.3 Convergence de la loi binomiale vers la loi normale :

• Théorème de De Moivre-Laplace :
Soit (X_n) une suite de variables de lois binomiales $\mathcal{B}(n,p)$, alors quand n tend vers l'infini, la suite des v.a. $\frac{X_n - np}{\sqrt{np(1-p)}}$ converge en loi vers la loi normale $\mathcal{N}(0,1)$.

Dans la pratique :
Pour n assez grand, on pourra approximer la loi binomiale $\mathcal{B}(n,p)$ par une loi normale $\mathcal{N}(np, \sqrt{np(1-p)})$.

Remarque :
Dans le cours de Stat I on a retenu comme conditions d'approximations :
$n \geq 50$ et $0,4 < p < 0,6$ ou avec $np(1-p) \geq 18$ (dans le cours de Stat II, on pourra également retenir comme conditions np et $n(1-p) > 5$ cf. SAPORTA).
(On n'effectuera pas de correction de continuité).
Le théorème de De Moivre-Laplace qui justifie l'approximation normale des lois binomiales résulte immédiatement d'une version plus générale du TCL énoncée au paragraphe suivant.

• **Seconde version du Théorème central limite** : Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes, toutes de même loi, d'espérance m et de variance σ^2, si on note X_i^* la v.a. centrée réduite associée à X_i, i.e. $X^*_i = \frac{X_i - m}{\sigma}$ alors la suite $T_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i^*$ converge en loi vers la loi normale $\mathcal{N}(0,1)$.

en pratique : si la suite $(X_n)_{n\in\mathbb{N}}$ vérifie les conditions du théorème ci-dessus alors la v.a. $S_n = \sum_{i=1}^{n} X_i$ suit approximativement une loi normale $\mathcal{N}(nm, \sigma\sqrt{n})$ pourvu que n soit assez grand ($n>50$). Autrement dit toute <u>somme</u> **de v.a. indépendantes** de même espérance et de même variance est asymptotiquement normalement distribuée.

remarque :
Dans le cas particulier où les variables X_i sont des variables de Bernoulli de paramètre p, on retrouve la convergence de la loi $\mathcal{B}(n,p)$ vers la loi normale $\mathcal{N}\left(np, \sqrt{np(1-p)}\right)$.

IV. PETITS ECHANTILLONS ISSUS D'UNE LOI NORMALE

Lorsque la taille de l'échantillon est petite ($n < 50$) les résultats asymptotiques ne sont plus applicables. Par contre si on fait l'hypothèse que la v.a. du modèle suit une loi normale, on peut alors préciser les lois des principales statistiques d'échantillonnage.

CamScanner

$\mathcal{N}(E(x), \sqrt{V(x)}^2)$.

Dans tout le paragraphe X désigne une v.a. normale d'espérance m et d'écart-type σ.
On rappelle que X admet pour densité la fonction f, définie par

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

et la v.a. $T = \frac{X-m}{\sigma}$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$, de densité

$f_0(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})$ $(t \in \mathbb{R})$. Si on note F_0 la fonction [illegible] :

$F_0(t) = \int_{-\infty}^{t} f_0(u)\,du$ $(t \in \mathbb{R})$ et $F_0(-t) = 1 - F_0(t)$, $(t \in \mathbb{R})$.

[illegible] la loi normale empirique d'un échantillon normal
[illegible] le résultat suivant concernant la stabilité de la loi normale par addition indépendante [illegible] que la somme de deux variables indépendantes suivant une loi normale

[illegible] v.a. indépendantes de lois respectives $\mathcal{N}(m_i, \sigma_i)$, i variant de 1 à n
alors la v.a. $X_1 + X_2 + \ldots + X_n$ suit la loi normale d'espérance $m = m_1 + \ldots + m_n$
et d'écart-type $\sigma = \sqrt{\sigma_1^2 + \ldots + \sigma_n^2}$

Corollaire :

Si $(X_1, \ldots, X_n)$ est un n-échantillon de la v.a. X de loi $\mathcal{N}(m, \sigma)$, alors la moyenne empirique $\bar{X}_n$ suit la loi normale d'espérance m et d'écart-type $\sigma/\sqrt{n}$, ce que l'on note : $\bar{X}_n \sim \mathcal{N}(m ; \sigma/\sqrt{n})$.

Cette proposition montre que, lorsque l'échantillon $(X_1, \ldots, X_n)$ est normal, la moyenne empirique suit une loi normale, **quelle que soit la taille** de l'échantillon.

IV.2 Variance empirique d'un échantillon normal.

Le résultat suivant est utile pour estimer v sans connaître la moyenne théorique m.

Proposition :
Soit $(X_1, \ldots, X_n)$ un échantillon normal. On rappelle que :

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \quad \text{et que} \quad S'^2 = \frac{n}{n-1} S^2$$

Alors, pour tout entier $n > 1$, la statistique $\frac{n S^2}{\sigma^2} = \frac{(n-1) S'^2}{\sigma^2}$ suit la loi χ^2_{n-1}.

En effet, si on pose :

$$W_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - m)^2$$

On voit que la v.a. $\frac{nW_n}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - m}{\sigma}\right)^2$ suit la loi χ^2_n

et la v.a. S^2 se déduit de W_n en remplaçant m par $\bar{X}$, d'où le résultat.

CamScanner