CHAPITRE I

ECHANTILLONNAGE

A - Rappels de cours

1. Lois de probabilités

1.1 Définitions et caractérisations

Les principales lois de probabilités, leurs conditions de validité, et leurs paramètres représentatifs sont rappelées dans le tableau ci-dessous:

Loi	Nature	Définition	Caractérisation	E(X)	Var(X)
BERNOULLI	Discrète	Variable indicatrice d'un caractère au cours de n épreuves de BERNOULLI (*)	Valeurs : $\{0,1\}$ $\Pr ob(X=0)=q$ $\Pr ob(X=1)=p$	p	q
Binomiale B(n,p)	Discrète	Occurrence d'un caractère au cours de n épreuves de BERNOULLI indépendantes	Valeurs : $\{0,1,2....,n\}$ $\Pr ob(X=x)=C_n^x p^x q^{n-x}$	$n.p$	$n.p.q$
Hypergéométrique	Discrète	Occurrence d'un caractère au cours de n épreuves de BERNOULLI dépendantes (à savoir le tirage sans remise d'un échantillon de taille n dans une population de taille N)	Valeurs : $\{0,1,2....,n\}$ $\Pr ob(X=x)=\frac{C_{N.p}^x.C_{N.q}^{n-x}}{C_N^n}$	$n.p$	$\frac{N-n}{N-1}.npq$
POISSON P(a)	Discrète	Occurrence des événements relativement rares	Valeurs : N $\Pr ob(X=x)=e^{-a}.\frac{a^x}{x!}$	a	a

(*) Pour rappel, l'épreuve de BERNOULLI est une épreuve dans laquelle, seuls sont possibles, les résultats C (avec la probabilité p) et $\overline{C}$ (avec la probabilité complémentaire $q=1-p$).

Loi	Nature	Définition	Caractérisation	E(X)	Var(X)
Géométrique	Discrète	Nombre de tentatives nécessaires jusqu'à l'obtention du caractère C à travers des épreuves de BERNOULLI indépendantes	Valeurs : N^* $\Pr ob(X = x) = q^{x-1}.p$	$\frac{1}{p}$	$\frac{q}{p^2}$
Binomiale négative	Discrète	Nombre de tentatives jusqu'à l'obtention r fois d'un caractère C à travers des épreuves de BERNOULLI indépendantes	Valeurs : $[r, +\infty[$ $\Pr ob(X = x) = C_{x-1}^{r-1} p^r .q^{x-r}$	$\frac{r}{p}$	$\frac{r.q}{p^2}$
Uniforme $U_{[a,b]}$	Continue	Probabilité uniforme sur $[a,b]$	Valeurs : $[a,b]$ $f(x) = \frac{1}{b-a}.1_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponentielle	Continue	Caractéristique des durées de vie des équipements qui ne vieillissent pas (loi « sans mémoire »)	Valeurs : R^+ $f(x) = \lambda.e^{-\lambda.x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma n	Continue	Loi de la somme de n variables aléatoires exponentielles indépendantes	Valeurs : R^+ $f(x) = \frac{\lambda^n .e^{-\lambda.x}.x^{n-1}}{(n-1)!}$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$
Normale $N(m,\sigma)$	Continue	Loi « universelle » vers laquelle convergent une large part des autres lois	Valeurs : R $f(x) = \frac{1}{\sigma.\sqrt{2\pi}}.e^{\frac{(x-m)^2}{2.\sigma^2}}$ (tables de valeurs en annexes)	m	σ^2

Loi	Nature	Définition	Caractérisation	E(X)	Var(X)
Chi-deux $\chi^2(n)$	Continue	Loi de la somme $\sum_{i=1}^{i=n} X_i^2$ où les X_i sont des variables normales, centrées, réduites, et indépendantes	Valeurs : R^+ $f(x)=\frac{x^{\frac{n}{2}-1}.e^{-\frac{x}{2}}}{2^{\frac{n}{2}}.\Gamma(\frac{n}{2})}$ avec $\Gamma(n)=\int_0^{+\infty} t^{n-1}.e^{-t}.dt$ (tables de valeurs en annexes)	n	$2n$
STUDENT T(n)	Continue	Loi de $T=\frac{X}{\sqrt{\frac{Y}{n}}}$ où X est normale centrée réduite et où Y suit la loi du chi -deux $\chi^2(n)$	Valeurs : R^+ $f(x)=\frac{\Gamma(\frac{n+1}{2}).(1+\frac{x^2}{n})^{-\frac{(n+1)}{2}}}{\sqrt{n.\pi}.\Gamma(\frac{n}{2})}$ (tables de valeurs en annexes)	0 n ≻ 1 (indéterminée pour n=1)	$\frac{n}{n-2}$ n ≻ 2 (infinie pour n ≤ 2)
FISHER SNEDECOR F(n,p)	Continue	Loi de $F=\frac{X/n}{Y/p}$ où X et Y suivent respectivement les lois $\chi^2(n)$ et $\chi^2(p)$	Valeurs : R^+ $f(x)=\frac{n^{\frac{n}{2}}.p^{\frac{p}{2}}.\Gamma(\frac{n+p}{2}).x^{\frac{n}{2}-1}}{\Gamma(\frac{n}{2}).\Gamma(\frac{p}{2}).(n.x+p)^{\frac{n+p}{2}}}$ (tables de valeurs en annexes)	$\frac{p}{p-2}$ p ≻ 2	Voir renvoi (*) ci-dessous.

(*) La variance de la loi de FISHER SNEDECOR est égale à $(\frac{p}{p-2})^2.\frac{2.(n+p-2)}{n.(p-4)}$ pour $p \succ 4$.

1.2 Propriétés de convergence

• Le **théorème central limite** tient une place fondamentale dans la justification desdites convergences. Pour rappel, son énoncé est le suivant :

Soit $X_n, n \in N,$ une suite de variables aléatoires indépendantes de même loi d'espérance m et de variance σ^2 finies. Alors, la somme $Z=\sum_{i=1}^{i=n} X_i$ converge pour n assez grand (en pratique à partir de n=30) vers la loi normale de moyenne $n.m$ et d'écart -type $\sigma.\sqrt{n}$.

• *Sur un plan plus général*, les lois de probabilités mentionnées dans le paragraphe précédent satisfont à un **ensemble de convergences**, essentielles pour les applications en statistique, et qui s'énoncent comme suit :

- La **loi hypergéométrique** *converge*, pour N grand, vers la **loi binomiale** $B(n,p)$ (condition le plus souvent satisfaite dès lors qu'on est amené à pratiquer un sondage).

Pratiquement, cette convergence est satisfaite pour $\frac{N}{n} \geq 10$.

- La **loi binomiale** $B(n,p)$ *converge*, pour n assez grand et p ni trop voisin de 1 ni de 0 vers la **loi normale** $N(m = n.p, \sigma^2 = n.p.q)$.

C'est le **théorème de MOIVRE- LAPLACE** qui résulte de l'application du théorème central limite au cas particulier de la somme de n variables aléatoires de BERNOULLI indépendantes.

Au plan pratique, plusieurs conditions de validité de cette convergence sont applicables. On peut retenir entre autres, $n \geq 30$ **et** $n.p \succ 5$ **et** $n.q \succ 5$, ou, $n \geq 30$ **et** $n.p \geq 15$ **et** $n.p.q \succ 5$.

- La **loi binomiale** $B(n,p)$ *converge*, pour n assez grand, et p faible (ou voisin de 1) vers la **loi de POISSON** de paramètre $a = n.p$.

Au plan pratique, on peut citer, entre autres, la condition $n \geq 30$ **et** $p \leq 0{,}1$ **et** $n.p \prec 15$.

- La **loi de POISSON** de paramètre a *converge*, pour n assez grand, vers la **loi normale** $N(m = a, \sigma^2 = a)$.

Au plan pratique, la convergence en question devient satisfaisante dès que $a \succ 15$.

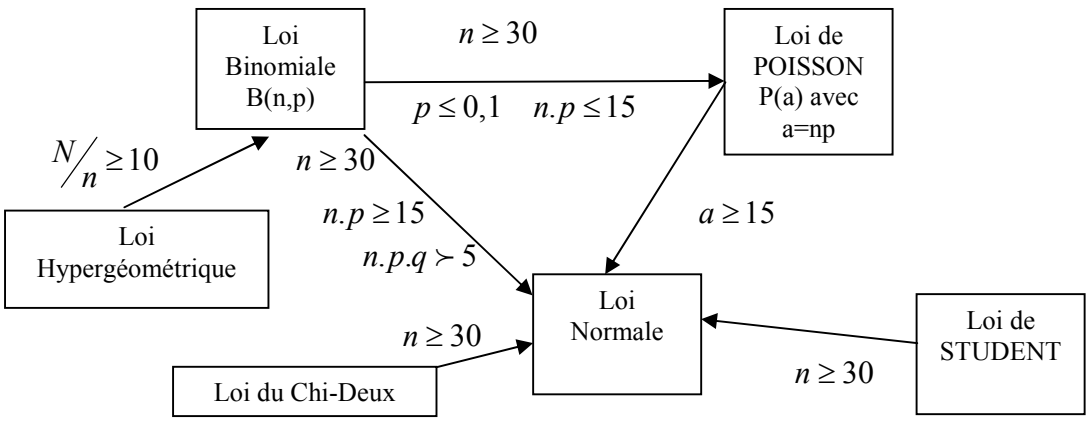
- La **loi de STUDENT**, $T(n)$, *converge*, pour n assez grand, vers la **loi normale** centrée réduite $N(0,1)$.

Au plan pratique, cette approximation devient satisfaisante dès que $n \geq 30$.

- La **loi du chi-deux**, $\chi^2(n)$, *converge*, pour n assez grand, vers la **loi normale** $N(m = n, \sigma^2 = 2n)$.

Ici encore, cette approximation est vérifiée à partir de $n = 30$.

Le schéma ci-dessous résume les propriétés de convergence susmentionnées :



2. Statistiques et distributions d'échantillonnage

2.1 Le principe de l'inférence statistique

L'objectif est d'évaluer la valeur inconnue d'un paramètre caractéristique déterminé au sein d'une *population*, à travers le *prélèvement d'un échantillon* et une expression du paramètre en question en fonction des observations faites (**principe de l'inférence statistique**). Il faut distinguer dans ce processus :

a) les *techniques de prélèvement de l'échantillon* $(X_1, X_2, ..., X_n)$ dont la forme la plus simple est celle d'un tirage aléatoire avec remise (échantillons dits de « BERNOULLI » non exhaustifs) ;

b) les *données* $(x_1, x_2, ..., x_n)$ fournies par un échantillon particulier et *l'estimation* qui en résulte pour le paramètre θ inconnu, soit $\hat{\theta} = T_n(x_1, x_2, ..., x_n)$;

c) l'étude des variations aléatoires de l'estimation $T_n(x_1, x_2, ..., x_n)$ en fonction des divers échantillons $(x_1, x_2, ..., x_n)$ que l'on peut extraire de la population, c'est-à-dire la caractérisation de la loi de la *statistique* $T_n(X_1, X_2, ..., X_n)$ (dite encore « *estimateur* »), loi formant la *distribution d'échantillonnage*.

Il est précisé qu'on appelle « statistique » toute fonction des observations faites.

2.2 Le cas d'une moyenne

• Considérant une variable aléatoire X (de moyenne m inconnue et de variance σ^2 connue ou non) et un échantillon $(X_1, X_2, ..., X_n)$ de n valeurs indépendantes prises par X (échantillons de type « BERNOULLI »), la transposition de l'expression probabiliste de $E(X)$ conduit, pour ce qui est de la moyenne, à la **statistique** $\overline{X} = \dfrac{\sum_{i=1}^{i=n} X_i}{n}$ dont il découle immédiatement $E(\overline{X}) = m,\ Var(\overline{X}) = \dfrac{\sigma^2}{n}$.

La linéarité de l'espérance mathématique entraîne immédiatement $E(\overline{X}) = \dfrac{\sum_{i=1}^{i=n} E(X_i)}{n}$, soit $E(\overline{X}) = \dfrac{n.m}{n} = m$. Par ailleurs, l'indépendance des X_i entraîne $Var(\overline{X}) = \dfrac{\sum_{i=1}^{i=n} Var(X_i)}{n^2}$, étant entendu, par ailleurs que $Var(a.X) = a^2.Var(X)$. Finalement, on obtient bien $Var(\overline{X}) = \dfrac{n.\sigma^2}{n^2} = \dfrac{\sigma^2}{n}$.

• Pour ce qui est de la **distribution d'échantillonnage**, le *théorème central limite* entraîne, pour $n \geq 30$, la convergence de $\overline{X}$ vers la **loi normale** $N(m, \dfrac{\sigma}{\sqrt{n}})$. En d'autres termes, la variable $\dfrac{\overline{X} - m}{\dfrac{\sigma}{\sqrt{n}}}$ suit la loi normale centrée réduite $N(0,1)$.

Plus encore, désignant par $\widehat{S^2}$ l'estimateur ponctuel de la variance σ^2 lorsque cette dernière est inconnue ($\widehat{S^2} = \frac{1}{n-1}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$ suivant résultats présentés dans le chapitre II), la variable $\frac{\overline{X} - m}{\frac{\hat{S}}{\sqrt{n}}}$ suit la **loi de STUDENT**, $T(n-1)$, à $\nu = n-1$ degrés de libertés.

La démonstration de ce résultat est présentée dans l'application 1.1 du présent chapitre.

2.3 Le cas d'une proportion

- Considérant la fréquence inconnue p d'un caractère C dans une population et la variable aléatoire X qui décrit l'occurrence de C dans des échantillons de taille n aléatoires, indépendants (prélèvements avec remise), la transposition de l'expression probabiliste de $E(X)$ conduit, pour ce qui est de la fréquence inconnue p, à la **statistique** $F_n = \frac{X}{n}$ dont il est évident que $E(F_n) = p$ et $Var(F_n) = \frac{p.q}{n}$.

  En effet, X suit la loi binomiale $B(n, p)$ de moyenne $n.p$ et de variance $n.p.q$. La linéarité de l'espérance entraîne $E(F_n) = \frac{E(X)}{n} = \frac{n.p}{n} = p$. Par ailleurs, $Var(F_n) = \frac{1}{n^2}.Var(X)$, soit $Var(F_n) = \frac{p.q}{n}$.

  On remarquera que p représente aussi l'espérance de la loi de BERNOULLI associée à chaque élément prélevé de l'échantillon. Dès lors et par application des résultats du paragraphe 4 susmentionné pour ce qui concerne les moyennes, la statistique représentative de p est fournie par $\overline{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$ où les X_i forment une suite de n variables aléatoires de BERNOULLI indépendantes.

  La somme $\sum_{i=1}^{i=n} X_i$ constituant la variable X de loi binomiale $B(n, p)$, on on retrouve ainsi l'expression $\frac{X}{n}$ qui caractérise F_n.

  Cette analogie d'une proportion avec une moyenne sera couramment utilisée par la suite.

- Pour ce qui est de la **distribution d'échantillonnage**, le *théorème de* MOIVRE LAPLACE justifie, pour $n \geq 30$ et p ni trop faible, ni trop voisin de 1 (critères pratiques rappelés précédemment), la possibilité d'approcher la loi de F_n par la **loi normal**e de moyenne p et de variance $\frac{p.q}{n}$, soit la loi $N(p, \sqrt{\frac{p.q}{n}})$.

  Il est précisé que dans l'hypothèse contraire où p est faible, voire n petit, on pourra mener des calculs directs à partir des lois binomiales et de POISSON et déterminer ainsi la distribution d'échantillonnage de $F_n = \frac{X}{n}$.

2.4 Le cas d'une variance

• Soient X une variable aléatoire (de moyenne m connue ou non et de variance σ^2 inconnue) et $(X_1, X_2, ..., X_n)$ un échantillon de n valeurs indépendantes prises par X (échantillon de type « Bernoullien »). La transposition de l'expression probabiliste de Var(X) conduit, pour ce qui est de la variance, à la **statistique** $S^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(X_i - m)^2$ (resp. la statistique $S'^2 = \frac{1}{n}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$ lorsque m est inconnue).

Le lecteur se méfiera néanmoins que, dans l'hypothèse où m est inconnue, c'est l'estimateur « non biaisé », $\widehat{S^2} = \frac{1}{n-1}.\sum_{i=1}^{i=n}(X_i - \overline{X})^2$ qu'il faudra retenir (et non S'^2) -> se reporter pour cela au chapitre II.

On montre dans l'application 1.2 proposée ci-après, que $E(S^2) = \sigma^2$ (resp. $E(\widehat{S^2}) = \frac{n-1}{n}.\sigma^2$ lorsque m est inconnue). Par ailleurs, il est montré également dans la même application que $Var(S^2) = \frac{\mu_4}{n} - \frac{\sigma^4}{n}$ et $Var(\widehat{S^2}) = \frac{\mu_4}{n} - \frac{n-3}{n.(n-1)}.\sigma^4$ (μ_4 désignant le moment d'ordre h de la variable centrée $X - E(X)$, soit $\mu_4 = E\left[(X - E(X))^4\right]$.

• Pour ce qui est de la **distribution d'échantillonnage,** et sous l'hypothèse de la *normalité* de la loi de X (échantillons dits « gaussiens »), la variable $\frac{n.S^2}{\sigma^2} = \frac{\sum_{i=1}^{i=n}(X_i - m)^2}{\sigma^2}$ suit la **loi du chi- deux** à n degrés de libertés, soit $\chi^2(n)$.

De même, la variable $\frac{(n-1).\widehat{S^2}}{\sigma^2} = \frac{n.S'^2}{\sigma^2} = \frac{\sum_{i=1}^{i=n}(X_i - \overline{X})^2}{\sigma^2}$ suit la **loi du chi- deux** à $n-1$ degrés de libertés, soit $\chi^2(n-1)$.

Le premier de ces résultats est immédiat puisque la loi du chi- deux, $\chi^2(n)$, caractérise la somme des carrés de n variables aléatoires, normales, centrées, réduites, indépendantes, ce qui est le cas pour les variables $\frac{X_i - m}{\sigma}$. Quant au second résultat, sa démonstration est proposée dans l'application 1.1 ci-après.

• Pour $n \geq 30$, on pourra *approcher* la **loi du chi- deu**x, soit $\chi^2(n)$, par la **loi normale** de moyenne n et de variance $2n$, soit $N(n, \sqrt{2n})$, et ceci conformément au *théorème central limite*.

Cette convergence est assez simple à établir. On rappelle tout d'abord que si $(U_1, U_2, ..., U_n)$ forment une suite de n variables aléatoires indépendantes, il en est de même de la suite $(U_1^2, U_2^2, ..., U_n^2)$. En effet, partant d'un n- uplet $(U_1, U_2, ..., U_n)$ de densité de probabilité $f(u_1, u_2, ..., u_n)$, il est évident que l'indépendance des U_i entraîne, pour cette densité sur R^n une expression égale au produit $\prod_{i=1}^{i=n}\varphi(u_i)$ des densités $\varphi(u_i)$ de chacune des variables U_i.

Dès lors, le changement de variables ($Y_1 = U_1^2, Y_2 = U_2^2, ..., Y_n = U_n^2$) conduit, pour le n- uplet $(U_1^2, U_2^2, ..., U_n^2)$ à la densité de probabilité élémentaire :

$$f(\sqrt{y_1}, \sqrt{y_2}, ..., \sqrt{y_n}).|J|.dy_1.dy_2....dy_n$$

où, le jacobien J est égal au déterminant :

$$J = \begin{vmatrix} \frac{1}{2.\sqrt{y_1}} & 0 & ... & 0 \\ 0 & \frac{1}{2.\sqrt{y_2}} & ... & 0 \\ . & . & ... & . \\ 0 & 0 & ... & \frac{1}{2.\sqrt{y_n}} \end{vmatrix} = \frac{1}{2^n.\sqrt{y_1}.\sqrt{y_2}....\sqrt{y_n}}$$

Or, la décomposition de $f(y_1, y_2, ..., y_n)$ en fonction des produits des densités $\varphi(y_i)$ conduit, pour la densité du n- uplet $(Y_1, Y_2, ..., Y_n)$ au produit ci-dessous :

$$\frac{\varphi(\sqrt{y_1}).dy_1}{2.\sqrt{y_1}}.\frac{\varphi(\sqrt{y_2}).dy_2}{2.\sqrt{y_2}}....\frac{\varphi(\sqrt{y_n}).dy_n}{2.\sqrt{y_n}}$$

qui est le produit des densités de probabilités de chacune des variables $U_1^2, U_2^2, ..., U_n^2$. Ainsi l'indépendance des U_i^2 est-elle établie.

Si on considère désormais la suite des variables normales, centrées, réduites, et indépendantes, soient $U_i = \frac{X_i - m}{\sigma}$ (loi $N(0,1)$ de densité de probabilité $\varphi(u) = \frac{1}{\sqrt{2\pi}}.\exp(-\frac{u^2}{2})$), on remarque que $E(U_i^2) = Var(U_i) + \left[E(U_i)\right]^2 = 1$ (puisque $E(U_i) = 0$).

D'autre part, $Var(U_i^2) = E(U_i^4) - \left[E(U_i^2)\right]^2$ avec $E(U_i^4) = \frac{1}{\sqrt{2\pi}}.\int_{-\infty}^{+\infty} t^4.\exp(-\frac{t^2}{2}).dt$,

soit $E(U_i^4) = \left[-\frac{1}{\sqrt{2\pi}}.t^3.\exp(-\frac{t^2}{2})\right]_{-\infty}^{+\infty} + 3.\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} t^2.\exp(-\frac{t^2}{2}).dt$ (suivant intégration par parties). Suite à la nullité du premier des deux termes ci-dessus, il reste $E(U_i^4) = 3.\frac{1}{\sqrt{2\pi}}.\int_{-\infty}^{+\infty} t^2.\exp(-\frac{t^2}{2}).dt = 3.E(U_i^2) = 3.$

Ainsi obtient-on, le résultat, $Var(U_i^2) = 3 - 1^2 = 2$.

En résumé, le théorème central limite appliqué aux n variables aléatoires indépendantes U_i^2 de moyenne égale à 1 et de variance égale à 2, entraîne la convergence de la somme $\sum_{i=1}^{i=n} U_i^2$ vers la loi normale de moyenne n et de variance $2n$, ce qui forme le résultat annoncé.

2.5 Récapitulatif concernant espérance, proportion, et variance

- Tous les résultats précédents qui, rappelons le, correspondent au cas d'un échantillonnage aléatoire élémentaire avec replacement (tirages non exhaustifs), sont résumés dans le tableau présenté ci-après.