Petit cours (attractif) de probabilités et statistiques

I. Introdution.

Il s'agit dans ce petit cours, de faire le point sur quelques notions élémentaires de probabilités et de statistiques, d'expliquer le lien entre ces deux branches des mathématiques et d'avoir un petit socle théorique permettant un enseignement cohérent des probabilités et des statistiques dans les classes du collège et du lycée.

Aujourd'hui, du point de vue mathématique (au sens moderne) la théorie des probabilités est une branche de la théorie de la mesure. Néanmoins, l'étude des probabilités est née de l'étude des jeux de hasard et tout son vocabulaire en découle.

La statistique, quant à elle déborde largement la théorie mathématique et couvre aujourd'hui à peu près tous les champs d'étude : médecine, économie, sciences sociales...

L'étude d'un problème statistique peut se décomposer en quatre étapes :

- Le recueil des données.
- Le classement et la réduction de ces données : c'est la statistique descriptive.
- L'analyse des données visant à les rattacher à un modèle probabiliste.
- La déduction de prévisions.

C'est le troisième point qui nous intéresse ici et que nous allons plus particulièrement développer, introduisant le vocabulaire (abstrait) des probabilités pour l'appliquer à l'étude (concrète) de séries statistiques.

Dans un cours, la question se pose d'introduire la théorie des probabilités à partir d'études de séries statistiques de la vie courante, ou de jeux de hasard, ou bien de commencer par donner un vocabulaire théorique pour ensuite l'appliquer à la modélisation de séries statistiques.

Bien entendu, dans le secondaire, c'est le premier choix qui paraît naturel. Néanmoins, dans ce cours, c'est le second choix qui sera fait.

II. Expérience aléatoire. Espace probabilisé.

Il s'agit d'étudier la réalisation d'évenements dont l'issue n'est pas connue à l'avance, ce sont des *expériences aléatoires*.

L'ensemble des résultats possibles d'une telle expérience va nous fournir ce que l'on appelle *l'univers des possibles*, noté Ω. Sur cet univers, on définira une *probabilité* qui sera une application, définie pour tout *évenement*, ou partie, de Ω, à valeur dans l'ensemble $[0, 1]$.

Par exemple :

- Le jeu de pile ou face avec une pièce parfaitement équilibrée : Ω={pile,face}, avec une probabilité de $1/2$ pour chacun des résultats.

- Le lancer d'un dé non-pipé : $\Omega = \{1, 2, 3, 4, 5, 6\}$, avec une probabilité de $1/6$ pour chacun des résultats.

Ces deux exemples sont théoriques, et leur vérification pratique nécessite un "grand" nombre d'expériences pour que la fréquence observée des résultats obtenus se rapproche de la probabilité théorique.

- La naissance d'un garçon : il s'agit là, par contre, d'une probabilité statistique, obtenue à partir de l'observation d'un très grand nombre de naissances.

Nous allons, avec l'introduction du vocabulaire des probabilités, fixer un sens rigoureux à des mots souvent utilisés dans le langage courant avec un sens plus flou.

Définitions. Considérons l'univers des possibles, Ω, et $\mathcal{P}(\Omega)$ l'ensemble de ses parties. On appelle *événement* un élément de $\mathcal{P}(\Omega)$.

Par exemple si $\Omega = \{1, 2, 3, 4, 5, 6\}$, l'ensemble $\{2, 4, 6\}$ correspond à l'évenement "le résultat obtenu est pair".

Lorsque l'ensemble Ω est fini, ou dénombrable, on peut considérer comme événement tout élément de $\mathcal{P}(\Omega)$. Sinon on notera $\mathcal{A}$ l'ensemble des événements, dits observables sur Ω, et l'on supposera que cet ensemble $\mathcal{A} \subset \mathcal{P}(\Omega)$ vérifie les propriétés suivantes :

Propriétés. *Pour tout $A, B \in \mathcal{A}$,*
- le complémentaire $\overline{A}$ de A dans Ω est dans $\mathcal{A}$.
- la réunion $A \cup B$ est dans $\mathcal{A}$.

Ce qui implique que $\emptyset$ et Ω sont des événements et que si A et B sont dans $\mathcal{A}$, alors $A \cap B \in \mathcal{A}$. Un tel ensemble $\mathcal{A}$ est appelé une algèbre de Boole. Bien que dans la pratique, au collège ou au lycée, on ait essentiellement affaire à des ensembles finis, il est bon d'avoir en tête ces propriétés ensemblistes.
On demandera aussi, raisonnablement, que les singletons soient dans $\mathcal{A}$ et que la réunion et l'intersection d'une suite finie ou infinie d'événements soit encore un événement.

Nous allons maintenant définir ce que l'on appelle une *probabilité*, c'est-à-dire ce qui permet de *mesurer* les événements de l'univers des possibles.

Définition. - *Si Ω est fini, on appelle* probabilité *sur Ω une application P de $\mathcal{P}(\Omega)$ dans $[0,1]$ telle que*
(i) $P(\Omega) = 1$.
(ii) Si $A, B \in \mathcal{P}(\Omega)$ vérifient $A \cap B = \emptyset$, alors $P(A \cup B) = P(A) + P(B)$.
- *Une* probabilité *sur un espace probabilisable $(\Omega, \mathcal{A})$ est une application P de $\mathcal{A}$ dans $[0,1]$ telle que*
(i) $P(\Omega) = 1$.
(ii) Pour toute suite d'événements $(A_i)_{i>0} \in \mathcal{A}$ deux à deux disjoints, on a

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

On appelle espace probabilisé *la donnée d'un triplet $(\Omega, \mathcal{A}, P)$.*

Le problème consiste maintenant à définir cette application P, c'est-à-dire, à attribuer une probabilité à chaque événement de l'univers des possibles.

Le cas le plus simple est celui d'un univers Ω fini, dans lequel tous les singletons, c'est-à-dire tous les évènements élémentaires, sont équiprobables. Ainsi, si l'expérience aléatoire admet n issues possibles, la probabilité de chacune des ces issues, ou de chacun de ces événements élémentaires est égale à $1/n$. Par additivité de la probabilité, la probabilité d'un événement $E \in \mathcal{P}(\Omega)$ est donc

$$P(E) = \frac{\text{card}(E)}{\text{card}(\Omega)} = \frac{\text{nombre d'issues favorables à E}}{\text{nombre d'issues possibles}}.$$

Le calcul des probabilités est alors ramené à un problème de dénombrement.

Nous ne parlerons pas dans ce cours, forcément incomplet, de probabilités conditionnelles, notre choix est plutôt de faire rapidement le lien avec les statistiques.

III. Variables aléatoires réelles. Séries statistiques.

L'univers des possibles est un ensemble qui peut contenir les résultats d'épreuves les plus variées, résultats non nécessairement numériques. Aux événement de Ω, on peut vouloir associer un nombre réel. Par exemple, le gain obtenu dans un jeu de hasard et l'on aimerait éventuellement pouvoir estimer l'espérance d'un tel gain. C'est ce qui nous amène à introduire la notion de variable aléatoire réelle. Nous nous concentrerons plus particulièrement sur les variables aléatoires réelles discrètes et leurs lois de probabilité.

Définition. *Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé. On appelle variable aléatoire discrète sur $(\Omega, \mathcal{A}, P)$ toute application*

$$\begin{aligned} X : \Omega &\longrightarrow \mathbb{R} \\ \omega &\longmapsto X(\omega) \end{aligned}$$

vérifiant
(ii) L'ensemble $X(\Omega) = \{X(\omega),\ \omega \in \Omega\}$ est une partie dénombrable (ou finie) de $\mathbb{R}$,

$$X(\Omega) = \{x_1, x_2, \ldots, x_n, \ldots\}.$$

(ii) Pour tout $x_k \in X(\Omega)$ l'ensemble $A_k = \{\omega \in \Omega,\ X(\omega) = x_k\}$ est dans $\mathcal{A}$, c'est à dire est un événement (mesurable par P).

La variable aléatoire X va permettre de considérer la probabilité P sur des parties de $\mathbb{R}$, c'est ce que l'on appellera la loi de probabilité de la variable X.

Définition. *Soit X une variable aléatoire discrète sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. On note $X(\Omega) = \{x_1, x_2, \ldots, x_n, \ldots\}$. La fonction*

$$\begin{array}{rccc} P_X : & \mathcal{P}(\mathbb{R}) & \longrightarrow & [0,1] \\ & \{x_k\} & \longmapsto & P(A_k) = p_k \\ & B & \longmapsto & \displaystyle\sum_{x_k \in B} p_k \end{array}$$

est une probabilité sur $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$, on l'appelle loi de probabilité de la variable aléatoire X.

On note

$$p_k = P(A_k) = P(\{\omega \in \Omega,\ X(\omega) = x_k\}) = P(X = x_k) = P_X(\{x_k\}).$$

Une variable aléatoire peut aussi prendre n'importe quelle valeur réelle, elle peut-être continue, sa loi admet alors une densité de probabilité et s'exprime par une expression du type

$$P(a \leq X \leq b) = \int_a^b f(t)dt$$

où f est une fonction qui vérifie $\int_{-\infty}^{+\infty} f(t)dt = 1$, mais nous sortons là de notre cadre et nous nous limiterons aux lois discrètes.

L'étude des lois de probabilités des variables aléatoires va permettre d'appréhender un certain nombre de problèmes aléatoires ou statistiques et de faire ainsi par des approximations, des estimations et des prédictions.

IV. Espérance. Variance. Ecart-type.

Nous allons maintenant définir les moments (espérance et variance) d'une variable aléatoire et faire le lien avec les séries statistiques. Une étude statistique commence par le recueil de données concernant un caractère à étudier. Nous nous intéresserons uniquement aux caractères quantitatifs, nous parlerons alors de variable statistique. Nous resterons dans le cas discret, et même fini, dans la mesure où, en statistiques, on ne recueillera jamais qu'un nombre fini de données ; qu'il s'agisse de lancers de dés, ou d'étude de populations humaines.

Avant d'expliquer ces notions de variance et d'écart-type, nous allons présenter les définitions dans le cas des lois de probabilités et dans le cas des séries statistiques sous forme de tableau.

Variable aléatoire discrète	**Série statistique**
Univers des possibles	Valeurs observées
$X(\Omega) = \{x_1, \ldots, x_k\}$	$x_1, x_2, \ldots, x_k$
	Effectifs correspondants
	$n_1, \ldots, n_k, \quad n = \sum_{i=1}^{k} n_i =$ effectif total
Probabilités	Fréquences
$p_i = P(X = x_i)$, pour $1 \leq i \leq k$	$f_i = \dfrac{n_i}{n}$, pour $1 \leq i \leq k$
Espérance	Moyenne
$E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$	$\overline{x} = \dfrac{n_1 x_1 + \cdots + n_k x_k}{n}$
$= x_1 p_1 + \cdots + x_k p_k$	$= x_1 f_1 + \cdots + x_k f_k$
Variance	Variance
$V(X) = \sum_{i=1}^{k} p_i (x_i - E(X))^2$	$V = \dfrac{1}{n} \sum_{i=1}^{k} n_i (\overline{x} - x_i)^2$
$= \sum_{i=1}^{k} p_i x_i^2 - E(X)^2$	$= \sum_{i=1}^{k} f_i (\overline{x} - x_i)^2 = \sum_{i=1}^{k} f_i x_i^2 - \overline{x}^2$
Ecart-type	Ecart-type
$\sigma(X) = \sqrt{V(X)}$	$\sigma = \sqrt{V}$
Ici, en théorie on connait les valeurs exactes des probabilités	Ici, les fréquences fournissent une approximation des probabilités.

Les variables aléatoires et leurs lois de probabilité pourront servir de modèle théorique pour l'étude de phénomènes quantitatifs observés en statistiques.

Le mot *espérance* provient de l'espérance de gain dans un jeu d'argent de hasard.

Par exemple, supposons que, dans le jeu de pile ou face, le joueur gagne 10 euros lorque sa pièce indique face et perde 2 euros lorsque sa pièce tombe sur pile. La probabilité de chaque face étant supposée égale à 1/2. On considère la variable aléatoire qui à chaque tirage fait correspondre la somme gagnée ou perdue, on a alors

$$E(X) = \frac{1}{2}(10 - 2) = 4.$$

Quatre euros est alors le gain moyen d'un joueur qui jouerait un très grand nombre de parties.

Par ailleurs, on sait bien que la connaissance de la moyenne d'une série statistique ne nous renseigne pas de manière satisfaisante. C'est pourquoi, il faut pouvoir mesurer la *dispersion* autour de cette moyenne, c'est le rôle de la variance et de l'écart-type dont nous verrons plus loin les intéressantes propriétés. L'écart-type a aussi l'avantage de s'exprimer dans les mêmes unités que la variable aléatoire et son espérance.

V. Lois de probabilités.

Nous allons donner ici trois exemples de loi aléatoire discrète, qui sont celles rencontrées dans les programmes du secondaire. Nous ne parlerons de lois continues que dans un prochain paragraphe, pour approcher certaines lois discrètes.

1) Loi de Bernouilli. C'est la loi qui intervient dans une épreuve à deux résultats possibles, *succès/échec*. Si p est un réel de l'intervalle $[0,1]$, la loi de Bernouill de paramètre p, notée $\mathcal{B}(p)$ est la loi de la variable aléatoire X telle que

$$X(\Omega) = \{0,1\}, \quad P(X=1) = p, \quad P(X=0) = 1-p = q.$$

On a alors $E(X) = p$, $V(X) = p(1-p)$ et $\sigma(X) = \sqrt{p(1-p)}$.

2) Loi uniforme. Dans le cas d'une variable suivant une loi uniforme, on a

$$X(\Omega) = \{x_1, \ldots, x_n\}, \quad P(X = x_i) = \frac{1}{n}, \quad E(X) = \frac{x_1 + \cdots + x_n}{n}.$$

3) Loi binomiale. C'est la loi suivie par une variable aléatoire X égale au nombre de succès dans une suite répétées de n épreuves de Bernouilli indépendantes, où p est la probabilité d'un succès. On a alors

$$X(\Omega) = \{0,1,\ldots,n\}, \quad P(X=k) = C_n^k p^k (1-p)^{n-k}, \text{ où } C_n^k = \frac{n!}{k!(n-k)!}$$

On dit que X suit la loi binomiale de paramètres n et p notée $X \sim \mathcal{B}(n,p)$ et on a

$$E(X) = np, \quad V(X) = np(1-p), \quad \sigma(X) = \sqrt{np(1-p)}.$$

Ce qui va maintenant nous intéresser, c'est, par exemple, le comportement de la loi binomiale, lorsque le nombre d'épreuves est "*grand*".

VI. La loi des grands nombres.

On se place dans la situation d'épreuves répétées caractérisées par la donnée d'une suite $X_1, \ldots, X_n$ de n variables aléatoires qui ont même loi et donc même espérance, notée μ, même variance, notée σ^2 et même écart-type, noté σ. On définit alors deux nouvelles variables aléatoires qui sont, la *somme*

$$S_n = X_1 + X_2 + \cdots + X_n,$$

et la *moyenne*

$$M_n = \frac{S_n}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

On a alors

$$E(S_n) = n\mu, \quad V(S_n) = n\sigma^2, \quad \sigma(S_n) = \sigma\sqrt{n},$$

et

$$E(M_n) = \mu, \quad V(M_n) = \frac{\sigma^2}{n}, \quad \sigma(M_n) = \frac{\sigma}{\sqrt{n}}.$$

Ces formules sont à la base des principaux estimateurs en statistiques.

La *loi faible des grands nombres* nous dit que "pour une expérience donnée, dans le modèle défini par une probabilité P, les distributions des fréquences calculées sur des séries de taille n se rapprochent de P quand n devient grand"(Maths 1$^{\text{ère}}$ S, repères, Hachette). Ce qui signifie que si l'on considère une expérience ayant k issues possibles

$$e_1, \ldots, e_k,$$

si l'on répète cette expérience n fois et que l'on note $f_i(n)$ la fréquence d'apparition de l'issue e_i au cours des n expériences et $p_i = P(e_i)$ la probabilité de l'issue e_i, alors

$$\lim_{n \longrightarrow +\infty} f_i(n) = p_i.$$

Cette loi des grands nombres que nous énoncerons plus rigoureusement, découle de *l'inégalité de Bienaymé-Tchebychev*, cette dernière traduit quantitativement le fait que, plus l'écart-type d'une variable aléatoire est faible, plus sa distribution (loi) de probabilité est concentrée autour de son espérance mathématique.

Inégalité de Bienaymé-Tchebychev. *Soit X une variable aléatoire, d'espérance $E(X) = \mu$ et d'écart-type σ, alors*

$$\forall t > 0,\ P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Ce qui se traduit, en posant $t = k\sigma$, par

$$\forall k > 0,\ P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Autrement dit, la probabilité d'observer une déviation par rapport à l'espérance d'au moins k unités d'écart-type est majorée par $1/k^2$.

Théorème (loi faible des grands nombres). *Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires deux à deux indépendantes, de même loi, de même espérance $E(X_i) = \mu$, pour tout i et de même écart-type σ. On définit leurs moyennes :*

$$M_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Alors, pour tout $\varepsilon > 0$, on a

$$\lim_{n \longrightarrow +\infty} P(|M_n - \mu| \geq \varepsilon) = 0.$$

Il s'agit d'une convergence "en probabilité", c'est-à-dire qu'il est toujours possible qu'un écart ε soit dépassé pour n grand, mais cela devient de plus en plus improbable. La conclusion du théorème peut encore s'écrire

$$P(\mu - \varepsilon < M_n < \mu + \varepsilon) \longrightarrow 1 \ \text{ quand } n \longrightarrow +\infty.$$

Considérons une suite $(X_n)_{n\geq 1}$ de variables de Bernouilli indépendantes, de même paramètre p, alors pour tout $\varepsilon > 0$,

$$\lim_{n\longrightarrow+\infty} P\left(\left|\frac{X_1+\cdots+X_n}{n} - p\right| \geq \varepsilon\right) = 0.$$

En effet, d'après l'inégalité de Bienaymé-Tchebychev, on a

$$P(|M_n - p| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} = \frac{p(1-p)}{n\epsilon^2}$$

qui tend vers 0 quand n tend vers l'infini, pour ε fixé.

Cette inégalité pourra s'écrire selon ce l'on connaît ou que l'on cherche la probabilité p,

$$P(p - \varepsilon < M_n < p + \varepsilon) \geq 1 - \frac{p(1-p)}{n\epsilon^2}$$

ou, en remarquant que pour $p \in [0,1]$, on a $p(1-p) \leq 1/4$

$$P(M_n - \varepsilon < p < M_n + \varepsilon) \geq 1 - \frac{1}{4n\varepsilon^2}.$$

Ainsi, la loi faible des grands nombres justifie l'approche fréquentiste qui attribue comme probabilité d'un événement une valeur autour de laquelle la fréquence se stabilise lorsque le nombre d'expériences indépendantes devient grand. Cependant, il n'est pas toujours possible de réaliser de telles expériences et on pourra être conduit à fixer à priori la valeur de la probabilité d'un événement et de valider ce choix à postériori.

Nous verrons dans des exemples, comment s'utilisent les inégalités ci-dessus, comment jouer sur les paramètres ε et n selon ce que l'on veut estimer.

Avant de passer aux problèmes d'estimation et d'échantillons, nous allons faire un petit détour par la célèbre "*courbe en cloche*", la loi normale, ou loi de Gauss.

VII. Approximation par la loi normale. Théorème de la limite centrée.

La *loi normale* contrairement aux lois discrètes que nous avons vues, est une loi continue définie par une densité. C'est-à-dire par une fonction f telle que la loi de probabilité de la variable aléatoire suivant cette loi s'écrive

$$P(a \leq X \leq b) = \int_a^b f(t)dt = F(b) - F(a).$$

Définition. *Soit $m \in \mathbb{R}$ et $\sigma \in]0,+\infty[$. On appelle densité gaussienne ou normale la fonction*

$$\begin{aligned} f_{m,\sigma} : \mathbb{R} &\longrightarrow \mathbb{R}^+ \\ t &\longmapsto \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-m)^2}{2\sigma^2}\right) \end{aligned}$$

Les courbes représentatives $C_{m,\sigma}$ de ces fonctions se déduisent toutes de la courbe $C_{0,1}$, par translations et changement d'échelle. La courbe $C_{0,1}$ est appelé courbe en cloche de Gauss.
Lorsqu'une variable aléatoire suit une loi normale, centrée et réduite, c'est-à-dire quand $m = 0$ et $\sigma = 1$, on note $X \sim \mathcal{N}(0, 1)$ et on lit les valeurs des probabilités dans des tables.

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt = \Pi(b) - \Pi(a).$$

La loi normale intervient dans la modélisation de phénomènes aléatoires possédant de nombreuses causes indépendantes dont les effets s'additionnent, sans qu'aucun d'entre eux ne domine. De nombreuses distributions "naturelles" sont ainsi approchées par une loi normale. Compte tenu de la complexité des phénomènes économiques et sociaux, la loi normale intervient dans tous les domaines.

On remarque, par ailleurs, que pour n grand, les diagrammes en batons représentant une variable aléatoire suivant une loi binomiale $\mathcal{B}(n, p)$ peuvent être approchés par des courbes en cloche, ce qui fait penser qu'une loi binomiale peut être approchée par une loi normale.

Théorème (Moivre-Laplace). *Soit S_n une variable aléatoire de loi $\mathcal{B}(n, p)$. On note*

$$S_n^* = \frac{S_n - E(S_n)}{\sigma(S_n)} = \frac{S_n - np}{\sqrt{np(1-p)}}$$

la variable centrée et réduite associée. Alors, pour tout réels $a < b$, on a

$$\lim_{n \longrightarrow +\infty} P(a < S_n^* < b) = \Pi(b) - \Pi(a) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt.$$

(de même avec des inégalités larges)

Mais ce phénomène n'est pas lié à la loi binomiale, il est général pour des épreuves répétées. Si $X_1 \ldots, X_n$ sont des variables aléatoires indépendantes suivant la même loi, de même espérance μ et de même écart-type σ, alors, pour n "grand" la variable aléatoire moyenne

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

suit approximativement une loi normale $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$. Plus précisément, on a le théorème suivant.

Théorème central limite. *(ou de la limite centrée) Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires suivant toutes la même loi, de même espérance μ et de même écart-type σ. Notons*

$$S_n = X_1 + X_2 + \cdots + X_n, \text{ et } M_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

On a alors

$$E(S_n) = n\mu, V(S_n) = n\sigma^2, \sigma(S_n) = \sigma\sqrt{n}, \text{ et } E(M_n) = \mu, V(M_n) = \frac{\sigma^2}{n}, \sigma(M_n) = \frac{\sigma}{\sqrt{n}}.$$

Notons Z_n les variables centrées réduites correspondantes, $Z_n = \dfrac{S_n - n\mu}{\sigma\sqrt{n}} = \dfrac{M_n - \mu}{\sigma/\sqrt{n}}$. *Alors*

$$\lim_{n \longrightarrow +\infty} P(Z_n < a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) = \Pi(a).$$

En pratique, on admet que l'on a une bonne approximation dès que $n \geq 50$ (et même 30) avec $np(1-p) \geq 9$.

Comme nous le verrons dans des exemples à la fin de ce cours, la loi faible des grands nombres conduit en théorie à choisir des valeurs de n beaucoup trop grandes, c'est pourquoi, on lui préfèrera l'approximation par des lois normales.

Examinons la faiblesse de la loi des grands nombres, supposons que X soit une variable aléatoire suivant la loi $\mathcal{N}(\mu, \sigma)$ et considérons, pour $t > 0$, la probabilité

$$P_t = P(\mu - t\sigma < X < \mu + t\sigma).$$

Si $t = 2$ ou 3, la lecture des tables de la loi normale nous donne $P_2 = 0,95$ et $P_3 = 0,99$ alors que l'inégalité de Bienaymé-Tchebychev nous donne les minorations suivantes $P_2 \geq 3/4$ et $P_3 \geq 8/9$.

Nous allons maintenant utiliser ces théorèmes pour traiter les problèmes d'estimations et d'intervalles de confiance.

VIII. Echantillonnage. Estimations

Le problème de l'échantillonnage consiste, connaissant les propriété d'une population, à évaluer les propriétés d'échantillons aléatoires. En réalité ce sera plutôt le problème inverse qui nous intéressera, c'est-à-dire estimer les propriétés d'une population à partir d'observations d'échantillons.

Dans tous ces problèmes, c'est le théorème de la limite centrée qui permettra l'étude d'estimations de moyennes et de fréquences.

Si l'on considère une population de *moyenne* μ et d'écart-type σ, et si $\overline{X}$ est la variable aléatoire qui à tout échantillon d'effectif n associe sa moyenne, alors lorsque n est "grand", la variable $\overline{X}$ suit approximativement la loi normale $\mathcal{N}(\mu, \sigma/\sqrt{n})$.

Si l'on considère une population dans laquelle une proportion p possède une certaine propriété. Si F est la variable qui à tout échantillon d'effectif n associe le pourcentage d'éléments ayant cette propriété, ou la *fréquence* d'apparition cette propriété, alors lorsque n est "grand", la variable F suit approximativement la loi normale $\mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

Avant d'étudier quelques exemples, tentons de cerner le problème et d'obtenir quelques résultas généraux.

Etant donnée une série statistique, une fois fait le choix d'une loi de probabilité, il s'agit d'estimer ses paramètres à partir des observations d'échantillons d'effectif n. Nous nous contenterons ici de l'estimation d'une moyenne et d'une probabilité, l'estimation de la variance nécessitant un facteur $n/(n-1)$ correctif, nous ne la traiterons pas ici.

Considérons donc une population d'effectif très grand sur laquelle on étudie un caractère quatitatif de moyenne μ et d'écart-type σ. On considère des échantillons $E_1, E_2, \ldots, E_k$ d'effectif n sur lesquelles la moyenne observée est $\overline{x_k}$. L'ensemble

$$\overline{X} = \{\overline{x}_1, \ldots, \overline{x}_k\}$$

est une série statistique d'effectif k appelée distribution des moyennes. On a

$$E(\overline{X}) = \mu \quad \text{et} \quad \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

En effet, la variable aléatoire suit la loi normale $\mathcal{N}(\mu, \sigma/\sqrt{n})$.

Connaissant la moyenne $\overline{x}$ d'un échantillon, il s'agit maintenant d'estimer la moyenne inconnue μ de la population. L'estimation peut se faire de manière ponctuelle ou par intervalle de confiance.

Ponctuellement on considère $\overline{x}$ comme estimation de la moyenne μ, de même, s'il s'agit de fréquence, on considère le pourcentage f observé dans un échantillon comme estimation ponctuel de la proportion p inconnue d'individus possédant la propriété observée dans la population.

Intervalles de confiance.

Dans le cas de la moyenne, on a $\overline{X} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$, ainsi, si T est la variable centrée, réduite associée $T = \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu)$, alors $T \sim \mathcal{N}(0,1)$, ainsi

$$\forall t \geq 0, \quad P(-t \leq T \leq t) = 2\Pi(t) - 1.$$

Par exemple, si l'on veut $2\Pi(t) - 1 = 0,95$ alors la table de la loi normale nous donne comme valeur de t, $t = 1,96$, on a donc

$$P\left(\mu - 1,96\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95.$$

C'est-à-dire qu'avant de prélever un échantillon de taille n dans la population, il y a 95 chances sur 100 pour que la variable aléatoire $\overline{X}$ se trouve dans l'intervalle

$$\left[\mu - 1,96\frac{\sigma}{\sqrt{n}}, \mu + 1,96\frac{\sigma}{\sqrt{n}}\right].$$

Cependant, comme μ est inconnu, on va plutôt utiliser l'inégalité sous la forme

$$P\left(\overline{X} - 1,96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95.$$

Ainsi, avant le prélèvement de l'échatillon, il y a 95 chances sur 100 que la variable aléatoire $\overline{X} - 1,96\frac{\sigma}{\sqrt{n}}$ prenne une valeur inférieure à μ et que la variable aléatoire $\overline{X} + 1,96\frac{\sigma}{\sqrt{n}}$ prenne une valeur supérieure à μ.

L'intervalle $\left[\overline{X} - 1,96\frac{\sigma}{\sqrt{n}}; \overline{X} + 1,96\frac{\sigma}{\sqrt{n}}\right]$ est appelé *intervalle de confiance* de la moyenne de la population avec le coefficient de confiance 95%. Si l'on veut un coefficient de 99%, il faut choisir $t = 2,58$.

On remarque que cet intervalle fait intervenir l'écart-type, lorsque l'effectif est suffisamment grand, on pourra prendre pour valeur son estimation ponctuelle.

Dans l'étude d'une fréquence d'observation d'une certaine propriété de la population, la variable aléatoire F suit la loi normale $\mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ (approximation d'une loi binomiale de paramètre p par une loi normale), où p est le pourcentage inconnu. On a alors

$$P\left(F - 1,96\sqrt{\frac{p(1-p)}{n}} \leq p \leq F + 1,96\sqrt{\frac{p(1-p)}{n}}\right) = 0,95.$$

On remarque que l'on retrouve là l'intervalle de confiance à 95% qui apparaît, sans justifications, dans les cours de la classe de seconde, c'est-à-dire

$$\left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}}\right]$$

où f désigne la fréquence dans un échantillon d'effectif n.

En effet si l'on prend $t = 1,96 \simeq 2$ et si l'on majore $p(1-p)$ par $\sup_{x\in[0,1]} (x - x^2) = \frac{1}{4}$, l'approximation par la loi binomiale nous fournit cet intervalle.

Nous allons maintenant tester ces notions d'approximations et d'intervalles de confiances sur quelques exemples.

IX. Exemples.

Les exemples présentés ici sont empruntés au cours polycopié de Charles Suquet.

Exemple 1. On lance une pièce de monnaie, non truquée 800 fois. On note N le nombre d'apparitions de la face. Déterminer

$$P(390 \leq N \leq 420).$$

La variable aléatoire N suit une loi binomiale $\mathcal{B}(800, 1/2)$, on a donc

$$P(390 \leq N \leq 420) = \sum_{k=390}^{420} C_{800}^{k} \left(\frac{1}{2}\right)^{k} \left(\frac{1}{2}\right)^{n-k}.$$

Ce qui rend le calcul assez impraticable. On va alors utiliser l'approximation par la loi normale. Commenons par centrer la variable, on a

$$E(N) = 800.\frac{1}{2} = 400 \text{ et } \sigma(N) = \sqrt{800.\frac{1}{4}} = \sqrt{200},$$

en notant

$$Z = \frac{N - 400}{\sqrt{200}},$$

on obtient alors

$$P(390 \leq N \leq 420) = P(-0,707 \leq Z \leq 1,414) = \Pi(1,414) - \Pi(-0,707) = 0,6815.$$

Exemple 2. Une urne contient des boules rouges en proportion *inconnue* p et des boules vertes en proportion $q = 1 - p$. On veut estimer cette proportion. On effectue n tirages avec remise. On note X_i la variable aléatoire qui vaut 1 si la boule obtenue au $i-$ème tirage est rouge et 0 sinon. Soit, la moyenne

$$M_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Il est naturel d'estimer p par M_n. Afin d'obtenir une fourchette pour une telle approximation de p, on utilise l'inégalité de Bienaymé-Tchebycheff qui s'écrit

$$P(|M_n - p| \geq t) \leq \frac{\sigma(X_1)}{nt^2} = \frac{p(1-p)}{nt^2} \leq \frac{1}{4nt^2}.$$

En majorant la valeur inconnue $p(1-p)$ par $\sup_{x \in [0,1]} (x - x^2) = \frac{1}{4}$.

On a donc

$$P(M_n - t < p < M_n + t) \geq 1 - \frac{1}{4nt^2}.$$

On dit que l'intervalle $I =]M_n - t, M_n + t[$ est un intervalle de confiance pour t au niveau $\alpha \geq \frac{1}{4nt^2}$.

Exemple 3. *Sondage simplifié.* Une élection oppose deux candidats A et B. On note p la proportion d'électeurs, dans la population totale, décidés à voter pour le candidat A. On souhaite estimer cette proportion inconnue. Un sondage (assimilé à un tirage avec

remise) auprès de 1000 personnes donne une fréquence observée de $0,54$. L'inégalité de Bienaymé-Tchebycheff nout fournit un intervalle de confiance

$$I =]0,54 - t, 0,54 + t[\text{ avec un niveau } \alpha \geq 1 - \frac{1}{4nt^2}.$$

Ici, $n = 1000$ et l'on souhaite que le niveau de confiance soit au moins égal à 95%, ainsi il faudra choisir t tel que

$$1 - \frac{1}{4000t^2} \geq 0,95 \iff t \geq \frac{1}{10\sqrt{2}} \simeq 0,0707.$$

En prenant $t = 0,071$, on obtient l'intervalle $I =]0,469; 0,611[$ qui contient des $p < 1/2$, ce qui, bien que le sondage donne 54% d'intentions de votes, ne permet par de pronostiquer la victoire du candidat A avec une erreur inférieure à 5%.

Si, maintenant, l'institut de sondage veut une fourchette de 1% et un niveau de confiance de 95%, on prend alors $t = 0,01$ ce qui va nous imposer un effectif n de l'échantillon sondé vérifiant

$$\frac{1}{4n(0,01)^2} \leq 0,05,$$

soit $n = 50000$, ce qui est évidemment un peu énorme.

Cet exemple montre bien comment les inégalités fournissant des intervalles de confiances qui dépendent de la fourchette de confiance et du niveau de confiance souhaités ainsi que de la taille des échantillons, et que l'on ne peut pas gagner sur tous les tableaux.

Exemple 4. On lance 3600 fois un dé non pipé. On veut minorer la probabilité que le nombre d'apparitions du 1 soit compris entre 540 et 660. On note S la variable aléatoire correspondant à ce nombre. La variable S suit une loi binomiale $\mathcal{B}(3600, 1/6)$, la valeur exacte de cette probabilité est

$$P(540 < S < 660) = \sum_{k=541}^{659} C_{3600}^{k} \left(\frac{1}{6}\right)^{k} \left(\frac{5}{6}\right)^{n-k}.$$

Nous allons donc, dans un premier temps utiliser l'inégalité de Bienaymé-Tchebycheff, puis dans un second temps l'approximation par la loi normale.

Bienaymé-Tchebycheff : Sachant que $E(S) = 600$ et $\sigma(S)^2 = 500$ et que $540 - 600 = -60$ et $660 - 600$, on obtient

$$540 < S < 660 \iff -60 < S - 600 < 60 \iff |S - 600| < 60.$$

Or, pour tout $t > 0$, on a

$$P(|S - 600| \geq t) \leq \frac{500}{t^2},$$

ainsi, pour $t = 60$, on a

$$P(|S - 600| \geq t) \leq \frac{500}{3600},$$

c'est-à-dire

$$P(540 < S < 660) = P(|S - 600| < 60) = 1 - \frac{5}{36} \geq 0,8611.$$

Approximation par la loi normale : On centre la variable, en notant

$$Z = \frac{S - E(S)}{\sigma(S)} = \frac{S - 600}{\sqrt{500}}\,.$$

Ainsi, on a

$$P(540 < S < 660) = P\left(\frac{540 - 600}{\sqrt{500}} < Z < \frac{660 - 600}{\sqrt{500}}\right) = P\left(\frac{-60}{10\sqrt{5}} < Z < \frac{60}{10\sqrt{5}}\right).$$

En approchant par la loi binomiale, on obtient

$$P(540 < S < 660) = P\left(\frac{-60}{10\sqrt{5}} < Z < \frac{60}{10\sqrt{5}}\right) \simeq 2\Pi\left(\frac{6}{\sqrt{5}}\right) - 1 \simeq 2\Pi(2,68) - 1 \simeq 0,9926.$$

Comparaison : Pour pouvoir affirmer que l'approximation par la loi normale donne un meilleur résultat que l'inégalité de Bienaymé-Tchebycheff, il faut pouvoir vérifier que l'erreur d'approximation est inférieure à $0,99 - 0,86 = 0,13$.

Nous utiliserons ici un résultat (Théorème d'Uspensky) qui nous dit que l'erreur commise en approchant une loi binomiale $\mathcal{B}(n,p)$ par la loi normale est majorée par

$$\frac{0,588}{\sqrt{npq}}.$$

Dans notre cas l'erreur Δ est donc majorée par

$$\Delta \leq \frac{0,588}{\sqrt{500}} < 0,0263.$$

On peut alors affirmer que

$$P(540 < S < 660) \geq 0,9926 - 0,0263 > 0,9662.$$

L'approximation gaussienne donne donc, dans ce cas, une bien meilleure approximation que l'inégalité de Bienaymé-Tchebycheff.

Bibliographie :

- Charles Suquet, *Introduction au Calcul des Probabilités, (à bac +2)*, polycopié Deug Mias et Mass, USTL, 2002-2003.

- Bernard Bigot, Bernard Verlant, *Statistiques et Probabilités*, Enseignement supérieur, Editions Foucher, 1990.

- Jean Trignan, *Probabilités, Statistiques et leurs applications*, BTS, IUT, Bréal 1990.

- Bernard Lannuzel, *Probabilités et statistique*, CAPES, Dunod 1999.