Jakob I Bernoulli (1654-1708) et la loi des grands nombres

Monte Verità, 20-23 septembre 2016

JP Gabriel
Département de mathématiques de l'Université
Chemin du Musée 23
1700 Fribourg

jean-pierre.gabriel@unifr.ch

Les aspects historiques concernant le travail de Jakob I Bernoulli ont été présentés par Martin Mattmüller dans sa splendide conférence. Pour cette raison, nous aborderons ici uniquement les aspects mathématiques de son travail.

Rappels de la théorie des probabilités.

Dans le but de mieux apprécier le travail de Jakob I Bernoulli, nous commençons par rappeler quelques notions de la théorie des probabilités, dont le modèle actuel d'une épreuve aléatoire.

L'axiomatique de la géométrie euclidienne décrit des relations entre points, droites, plans, etc...sans définir les points. Une perception intuitive de ceux-ci suffira pour proposer les axiomes dont seront déduits les théorèmes. Une situation analogue se présente avec la théorie des probabilités, car bien que tout un chacun en ait une perception intuitive, nous ne disposons pas d'une définition satisfaisante du hasard. La notion centrale de cette théorie est celle d'épreuve aléatoire: une épreuve est un ensemble d'issues (ou résultats) et nous dirons qu'elle est aléatoire, si la réalisation de ses issues est gouvernée par le hasard. Le modèle actuel de cet objet est le triple de Kolmogorov (1933):

$$(\Omega, \mathfrak{F}, P)$$

où Ω est un ensemble en bijection avec l'ensemble des issues (par abus de langage nous le nommerons ensemble des issues). A priori chaque sous-ensemble de Ω est interprétable comme événement, à savoir celui qui est réalisé si et seulement si l'une des issues qu'il contient est réalisée. En ce sens Ω devient l'événement certain car il est toujours réalisé puisqu'il contient

toutes les issues. L'élément noté $\mathfrak{F}$ est la famille des événements associée à l'épreuve et est donc constituée de sous-ensembles de Ω:

$$\mathfrak{F} \subset \mathfrak{P}(\Omega) = \text{ la famille de tous les sous-ensembles de } \Omega.$$

Les opérations ensemblistes $A^c, A \cup B$ et $A \cap B$ deviennent respectivement les événements **contraire**, A **ou** B est réalisé et A **et** B sont réalisés. Certaines fois, les opérations $\cup$ et $\cap$ devront être répétées une infinité dénombrable de fois, par exemple dans des discussions de convergence. On demandera alors que $\mathfrak{F}$ soit stable pour le complémentaire et pour les réunions et les intersections dénombrables. Plus précisément, $\mathfrak{F}$ sera une σ-algèbre, c'est-à-dire:

1. $\Omega \in \mathfrak{F}$,
2. $\mathfrak{F}$ est stable pour l'opération ^{c} (i.e. le complémentaire),
3. $\mathfrak{F}$ est stable pour les réunions dénombrables.

La stabilité pour les intersections dénombrables découle de la loi de Morgan.

Une mesure μ sur une σ-algèbre $\mathfrak{F} \subset \mathfrak{P}(\Omega)$ est une fonction vérifiant :

1. $\mu : \mathfrak{F} \to [0, +\infty]$,
2. $\mu(\emptyset) = 0$,
3. μ est σ-additive c'est-à-dire, pour toute suite $(A_n)_{n \geq 1}$ d'éléments de $\mathfrak{F}$ disjoints deux à deux i.e. $A_m \cap A_n = \emptyset$ pour tout $m, n \geq 1, m \neq n$:

$$P(\bigcup_{n=1}^{+\infty} A_n) = \sum_{n=1}^{+\infty} P(A_n).$$

Une mesure μ est dite finie si $\mu(\Omega) < +\infty$.

Dans $(\Omega, \mathfrak{F}, P)$, P sera la probabilité, à savoir une mesure sur $\mathfrak{F}$ avec $P(\Omega) = 1$, donc une mesure finie.

Il est clair que $\mathfrak{P}(\Omega)$ est une σ-algèbre et nous souhaiterions tous, pour des raisons de simplicté, pouvoir choisir $\mathfrak{F} = \mathfrak{P}(\Omega)$ comme famille des événements

dans $(\Omega, \mathfrak{F}, P)$. Ce choix est toujours possible dans les cas où Ω est fini ou infini dénombrable. Mais la chose se complique si Ω est non-dénombrable, car la probabilité peut devenir un obstacle. Rappelons que deux ensembles A et B sont dits équipotents s'il existe une bijection de A sur B.

On trouvera sur notre chemin le théorème suivant:

Théorème d'Ulam (1930)

Soit Ω un ensemble infini équipotent à $\mathbb{R}$. Il n'existe pas de probabilité P définie sur $\mathfrak{P}(\Omega)$ qui s'annule sur les singletons i.e. $\forall\ \omega \in \Omega,\ P(\{\omega\}) = 0$.

Pourrions-nous alors nous restreindre aux cas Ω fini et Ω dénombrable ? Malheureusement (ou heureusement ?) la réponse est non. Considérons l'épreuve aléatoire comportant uniquement deux issues, par exemple le jet d'une pièce de monnaie (symétrique pour simplifier i.e. p=probabilité de pile $=\frac{1}{2}$). Il s'agit de l'épreuve aléatoire la plus simple possible car une épreuve avec une seule issue n'aurait pas de sens; celle-ci serait en effet toujours réalisée ! Il est nécessaire, en théorie des probabilités, de pouvoir répéter indéfiniment une épreuve aléatoire. Faisons-le avec le jet de la pièce en supposant l'indépendance des jets. Représentons pile avec 1 et face avec 0. Une issue est donc une suite infinie formée de zéros et de uns. L'ensemble des issues:

$$\Omega = \text{ensemble des suites infinies formées de 0 et de 1} = \{0,1\}^{\mathbb{N}^*}$$

est équipotent à $\mathbb{R}$ (voir la représentation dyadique des réels de l'intervalle $[0,1]$). Cette épreuve aléatoire est centrale pour la théorie des probabilités et ne peut en aucun cas être ignorée. Soit $(a_n)_{n\geq 1}$ une suite d'éléments de $\{0,1\}$. Notons:

$$A = (a_1, a_2, ...a_n, a_{n+1}, a_{n+2}, ...)$$

l'issue ayant a_n comme résultat au n-ième jet, $n \geq 1$. Considérons l'événement A_n pour lequel seuls les résultats des n premiers jets sont imposés et coïncident avec $a_1, a_2, ..., a_n$, la queue étant complètement libre. Par l'indépendance des jets et la symétrie de la pièce, $P(A_n) = \frac{1}{2^n}$. Il est clair que, comme ensembles, pour tout $n \geq 1, A \subset A_n$. Une probabilité étant monotone (i.e. $A \subset B \Rightarrow P(A) \leq P(B)$),

$$\forall n \geq 1,\ \ 0 \leq P(A) \leq P(A_n) = \frac{1}{2^n}.$$

On en déduit que $P(A) = 0$ et donc P s'annule sur les singletons. Le théorème d'Ulam empêche l'existence d'une probabilité définie sur tous les

sous-ensembles de $\Omega = \{0,1\}^{\mathbb{N}^*}$. Par conséquent, une réduction du domaine de définition de P s'impose et donc $\mathfrak{F} \neq \mathfrak{P}(\Omega)$. Nous sommes ainsi condamnés à accepter que certains sous-ensembles de Ω ne puissent pas représenter des événements associés à l'épreuve aléatoire.

Cette formalisation de la notion d'épreuve aléatoire a permis de donner un sens mathématique précis aux objets qui intéressent les probabilistes. Par exemple, une variable aléatoire (v.a.) X était auparavant *une grandeur qui prend ses valeurs au hasard*, notion assez floue. Notons $\mathbb{B}$ la σ-algèbre de Borel (i.e. la plus petite σ-algèbre qui contient les intervalles de $\mathbb{R}$ ou ausi la plus petite qui contient les ouverts). Quelques réflexions conduisent à définir une v.a. comme une fonction:

$$X : \omega \in \Omega \mapsto X(\omega) \in \mathbb{R}$$

$(\mathfrak{F}, \mathbb{B})$-mesurable, c'est-à-dire:

$$\forall B \in \mathbb{B},\ X^{-1}(B) := \{\omega \in \Omega\ ; X(\omega) \in B\} \in \mathfrak{F},$$

$X^{-1}(B)$ désigne donc l'image inverse de B par X. Une v.a. devient donc une fonction au sens ordinaire du terme, c'est-à-dire un objet mathématique tout-à-fait classique.

Par analogie avec la topologie, si $\mathfrak{F}$ et $\mathbb{B}$ étaient les familles des ouverts de deux topologies, alors cette définition nous fournirait la continuité de X.

Moyenne empirique et estimation

Vous avez mesuré (observé) n fois une grandeur inconnue et obtenu les résultats suivants:

$$x_1, x_2, ..., x_n \in \mathbb{R}.$$

Les résultats diffèrent en raison des erreurs de mesure. Que faites vous de ces nombres pour estimer la grandeur inconnue ? *Tout le monde* vous répondra: on utilise la moyenne arithmétique (ou empirique):

$$\overline{x} := \frac{1}{n} \sum_{k=1}^{n} x_k.$$

Question: pourquoi ?

Quelques repères historiques

La moyenne arithmétique est une très vieille notion qui était utilisée, par exemple, dans le partage des héritages. Toutefois, la moyenne de valeurs

observées (ou moyenne empirique) pour estimer une grandeur inconnue est une idée plus récente. Toutefois, l'utilisation de la moyenne empirique pour obtenir une meilleure estimation que celle donnée par une seule observation est plus récente [3, 8].

En 1638, Galilée (1564-1642) mesure les temps de descente de sphères de bronze. Il s'est intéressé aux valeurs extrêmes (min et max) mais n'a jamais "combiné" ses valeurs observées pour tenter d'améliorer son estimation. Comme Galilée était l'un des plus grands scientifiques de son époque, on peut supposer qu'il était au courant des méthodes pratiquées alors. On peut donc admettre que la moyenne empirique n'était pas encore utilisée systématiquement en 1638. Par contre Simpson, en 1755, se fait l'avocat de la moyenne empirique pour estimer une grandeur inconnue mesurée plusieurs fois. Cette méthode est donc apparue de façon explicite entre 1638 et 1755.

Des manifestations implicites sont toutefois survenues avant 1638. On trouve chez Cardan (1501-1576) une problématique très proche: si p est la probabilité d'un succès et que l'épreuve est répétée n fois, Cardan utilisait déjà la formule $\mu = np$ pour approcher le nombre de succès.

De même avec *l'incroyable* découverte de Kepler (1571-1630): le poids total d'un grand nombre de pièces de monnaie de même type ne dépend pas des différences de poids des pièces entre elles. Il y a compensation des différences dans la sommation.

Première réponse à la question

Une première réponse interprète les valeurs obtenues:

$$x_1, x_2, ..., x_n$$

comme un nuage de points sur la droite réelle. Nous décidons de remplacer ce nuage par un seul point noté a. Nous introduisons une fonction de coût (mesure d'erreur ou pénalité) lorsque l'on remplace $x \in \mathbb{R}$ par a, notée $C(x, a)$, le coût total pour le nuage étant donné par:

$$C_{tot} := \sum_{k=1}^{n} C(x_k, a).$$

Si, en suivant Gauss, on choisit le coût quadratique:

$$C_2(x, a) := (x - a)^2,$$

alors le coût quadratique total devient:

$$C_2(a) = \sum_{k=1}^{n} C_2(x_k, a) = \sum_{k=1}^{n} (x_k - a)^2.$$

Il est alors logique de chercher la valeur de a qui minimise $C_2(a)$. Nous annulons la dérivée:

$$\frac{d}{da}C_2(a) = (-2)\sum_{k=1}^{n}(x_k - a) = (-2)(\sum_{k=1}^{n} x_k - na) = 0.$$

L'unique solution est

$$a = \frac{1}{n}\sum_{k=1}^{n} x_k. = \overline{x}$$

c'est-à dire la moyenne empirique. Puisque:

$$\frac{d^2}{da^2}C_2(a) = (-2)\sum_{k=1}^{n}(-1) = 2n > 0,$$

$\overline{x}$ réalise bien le minimum. Cette propriété nous donne une raison de choisir $\overline{x}$ pour remplacer notre nuage de points par un seul.

On peut objecter qu'une autre fonction de coût pourrait nous fournir un autre résultat. Suivons Legendre en choisissant $C_1(x, a) = |x - a|$. La valeur absolue n'étant pas dérivable partout, la méthode précédente n'est pas applicable pour déterminer un point qui réalise le minimum. On peut toutefois démontrer que le minimum est maintenant réalisé par n'importe quelle médiane empirique $a = m_e$ du nuage de points [10, p.81]. Cette première réponse, quoique digne d'intérêt, n'est donc pas vraiment satisfaisante.

Fournir une bonne réponse à la question initiale est plus difficile qu'il n'y paraît à première vue. Il faut invoquer l'un des théorèmes les plus profonds de la théorie des probabilités, à savoir la *loi des grands nombres*. Le théorème de Bernoulli qui va nous intéresser ici en est la toute première version. Ainsi, dès le début du XVIIIe siècle (peut-être même dès la fin du XVIIe siècle), son travail donne non seulement une justification à la méthode d'estimation ci-dessus, mais apporte un support théorique à la régularité statistique (asymptotique) observée expérimentalement. L'idée d'estimer une grandeur à l'aide de la moyenne empirique des observations sera complètement acceptée au XIXe siècle seulement.

Modèle usuel des observations d'une v.a. en statistique

Notons par exemple X le poids d'une personne choisie au hasard dans une population donnée. On tire au hasard (avec remise) et sans influences mutuelles, n personnes dont on mesure les poids. On obtient n nombres réels positifs $x_1, x_2, ..., x_n$ qui sont les résultats des mesures. Ce ne sont pas des v.a. mais

des observations (ou réalisations) de v.a.. Si l'on effectue un second tirage aléatoire de n personnes, il faut s'attendre à obtenir des nombres différents des premiers car d'autres personnes auront vraisemblablement été choisies. En statistique, le *modèle des observations*, consiste en un $n-uple$ de v.a.:

$$(X_1, X_2, ..., X_n)$$

où X_i est le résultat de la i-ème observation, $1 \leq i \leq n$.. Les hypothèses suivantes semblent alors raisonnables:

1. Comme les v.a. X_i sont toutes des observations de X elles auront la même loi que X (i.e. même fonction de répartition). On dit alors qu'elles sont identiquement distribuées comme X (ou ont la même loi que X).

2. Si les observations ne s'influencent pas mutuellement, on supposera l'indépendance des n v.a..

Nous supposerons donc que les v.a. $(X_1, X_2, ..., X_n)$ sont *i.i.d. comme* X, c'est-à-dire indépendantes et identiquement distribuées comme la v.a. X. Un tel n-uple de v.a. est aussi appelé *n-échantillon issu de* X. Cette notion apparaît déjà implicitement dans le théorème de Bernoulli qui compare la fréquence empirique des succès obtenus en répétant l'épreuve, à la probabilité d'un succès. Il s'agit vraisemblablement du premier modèle mathématique incluant à la fois un phénomène et ses observations.

Rappel: l'espérance d'une v.a.

L'espérance d'une v.a. X, notée $E(X)$, est sa moyenne théorique. L'espérance d'une v.a. $X \geq 0$ est toujours définie car la valeur $+\infty$ ne fait pas problème et est donc admise. Si X prend des valeurs positives et négatives, on utilise alors la représentation:

$$X = X^+ - X^- \quad \text{où} \quad X^+ = \max(X, 0) \geq 0 \text{ et } X^- = -\min(X, 0) \geq 0.$$

Une v.a. X sera dite intégrable si $E(X^+) < +\infty$ **et** $E(X^-) < +\infty$ et dans ce cas l'espérance est définie par $E(X) = E(X^+) - E(X^-)$. On peut montrer que l'intégrabilité de X équivaut à $E(|X|) < +\infty$.

Seconde réponse à la question

Notons θ la valeur inconnue (vraie valeur) que nous désirons estimer et $\mathfrak{E}$ l'erreur aléatoire commise lors d'une observation. Avec Gauss, nous supposons que la valeur obtenue lors d'une observation, notée X, est de la forme:

$$X = \theta + \mathfrak{E}.$$

Nous admettons que $E(\mathfrak{E}) = 0$, la moyenne théorique de l'erreur est nulle, car nous imaginons que les erreurs de signes différents ont tendance à se compenser. Ainsi, l'espérance d'une somme étant égale à la somme des espérances et $E(\theta) = \theta$ (espérance d'une constante):

$$E(X) = E(\theta + \mathfrak{E}) = E(\theta) + E(\mathfrak{E}) = \theta.$$

Supposons que l'on observe indéfiniment X. La suite qui représente ces observations est maintenant infinie et sera notée $(X_n)_{n\geq 1}$. Par hypothèse, les v.a. de cette suite seront i.i.d. comme X. La loi dite **forte** des grands nombres (voir plus loin) nous assure que, si X est intégrable, alors la suite des moyennes empiriques:

$$\overline{X}(n) := \frac{1}{n}\sum_{k=1}^{n} X_k$$

convergera presque sûrement (i.e. avec probabilité 1) vers $E(X)$, lorsque $n \to +\infty$. Or $E(X) = \theta$, c'est-à-dire la vraie valeur θ cherchée. La loi des grands nombres justifie ainsi l'utilisation de la moyenne empirique pour estimer la vraie valeur θ. Notons également que, plus n sera grand, meilleure sera en principe l'estimation.

Variables aléatoires de Bernoulli et binômiale

Nous rappelons quelquess résultats classiques qui nous seront utiles. Une variable aléatoire (v.a.) X est dite de Bernoulli de paramètre $p, (0 < p < 1)$ si:

$$X = \begin{cases} 1 & : \text{avec probabilité p,} \\ 0 & : \text{avec probabilité q=1-p.} \end{cases}$$

Notation: $X \sim Ber(p)$. La valeur 1 est souvent interprétée comme succès et 0 comme échec.
Une v.a. X est dite binômiale de paramètres $(n, p), n \in \mathbb{N}^*$ et $0 < p < 1$ si:

$$\text{Pour } k \in \{0, 1, 2, ..., n\}, P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}; \text{Notation : } X \sim Bin(n, p).$$

Fonctions génératrices

Rappelons qu'une probabilité sur $\mathbb{N}$ peut être caractérisée par sa densité $\underline{p} = (p_k)_{k\in\mathbb{N}}$ vérifiant (1) $p_k \in [0,1]$ et (2) $\sum_{k=0}^{+\infty} p_k = 1$. Laplace a eu l'idée géniale de leur associer la série suivante:

$$t \in [0,1] \mapsto h_{\underline{p}}(t) := \sum_{k=0}^{+\infty} p_k t^k$$

dénommée fonction génératrice de la densité de probabilité $(p_k)_{k\in\mathbb{N}}$. L'objet mathématique est une série dite "entière". La série converge en $t = 1$ car $\sum_{k=0}^{+\infty} p_k = 1$ et par conséquent sur tout l'intervalle $[0,1]$ puisque $\sum_{k=0}^{n} p_k t^k$ est non décroissante en n et bornée:

$$\forall n \geq 1, \quad 0 \leq \sum_{k=0}^{n} p_k t^k \leq \sum_{k=0}^{n} p_k \leq 1.$$

On peut bien sûr choisir t dans les nombres complexes $\mathbb{C}$, auquel cas la convergence pour $t = 1$ nous garantit que son rayon de convergence R satisfait $R \geq 1$.

Si X est une v.a. à valeurs dans $\mathbb{N}$, alors $p_k = P(X = k)$ est une densité de probabilité sur $\mathbb{N}$ et on définit:

$$t \in [0,1] \mapsto h_X(t) := \sum_{k=0}^{+\infty} P(X = k) t^k$$

appelée fonction génératrice de X. Si X est une v.a. à valeurs dans $\mathbb{N}$, alors son espérance est définie par:

$$E(X) := \sum_{k=0}^{+\infty} kP(X = k) \leq +\infty.$$

Pour une fonction $f : \mathbb{N} \to \mathbb{R}_+$, de la définition précédente on peut déduire [15, p. 209]:

$$E(f(X)) = \sum_{k=0}^{+\infty} f(k)P(X = k).$$

Nous obtenons ainsi une autre représentation de $h_X(t)$ qui s'avère très souvent plus pratique:

$$E(t^X) = \sum_{k=0}^{+\infty} t^k P(X = k) = h_X(t).$$

Notons que, pour $i = \sqrt{-1}$:

$$t \in \mathbb{R} \mapsto \phi_X(t) := E(e^{itX}) \in \mathbb{C}$$

est appelée fonction caractéristique de X et s'applique à toutes les v.a réelles. Il s'agit en fait de la transformée de Fourier de la loi de X.

Exemples

1. $X \sim Ber(p)$: $h_X(t) = q + pt$.
2. $X \sim Bin(n, p)$: $h_X(t) = (q + pt)^n$.
   En effet:
   $$h_X(t) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} t^k = \sum_{k=0}^{n} \binom{n}{k} (pt)^k (1-p)^{n-k} = (q + pt)^n.$$
3. $X \sim Poisson(\lambda)$, *i.e.* $\exists \lambda > 0$ tel que $P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, k \in \mathbb{N}$. Alors:
   $$h_X(t) = e^{-\lambda} \sum_{k=0}^{+\infty} \frac{(\lambda t)^k}{k!} = e^{\lambda(t-1)}.$$

Nous aurons besoin des résultats suivants:

(a) Soient X, Y deux v.a. indépendantes et deux fonctions (mesurables) f, g de $\mathbb{R}$ dans $\mathbb{R}$. Alors $f(X)$ et $g(Y)$ sont encore des v.a. indépendantes.

(b) Si X et Y sont des v.a. indépendantes intégrables, alors XY est aussi intégrable et $E(XY) = E(X)E(Y)$.

(c) Soit X, Y deux v.a. indépendantes à valeurs dans $\mathbb{N}$. Des points (a) et (b) découle:

$$h_{X+Y}(t) = E(t^{X+Y}) = E(t^X \cdot t^Y) = E(t^X)E(t^Y) = h_X(t) \cdot h_Y(t).$$

La convolution est transformée en produit ordinaire de fonctions.

(d) Les densités de probabilité sur $\mathbb{N}$ sont en bijection avec leurs fonctions génératrices.

A l'aide des fonctions génératrices, on démontre facilement le théorème suivant:

Théorème 1

Si les v.a. $X_1, X_2, ...X_n$ sont indépendantes et identiquement distribuées (i.i.d.) comme $X \sim Ber(p)$, alors leur somme est une v.a. binômiale:

$$S_n = \sum_{i=1}^{n} X_i \sim Bin(n, p).$$

Démonstration

Pour $1 \leq i \leq n$, $h_{X_i}(t) = (q + pt)$ et en vertu de leur indépendance:

$$h_{S_n}(t) = h_{X_1}(t) \cdots h_{X_n}(t) = (q + pt)^n.$$

En vertu de (d), la loi de S_n est une binômiale de paramètres n et p, donc $S_n \sim Bin(n, p)$.

Le grand résultat de Bernoulli.

Nous nous permettons de moderniser la forme des énoncés et des arguments de Bernoulli, tout en respectant au mieux sa démarche.Voici une première formulation simplifiée pour prendre connaissance du résultat de Bernoulli:

A la seule condition de répéter un nombre de fois suffisamment grand une même expérience aléatoire, il y a une probabilité aussi voisine de 1 que l'on veut pour que la fréquence empirique d'un événement donné soit aussi proche que l'on veut de la probabilité de cet événement.

Bernoulli considère une urne contenant $r \geq 1$ boules blanches et $s \geq 1$ noires, $t = r + s$. En fait, la démonstration de son théorème demandera des divisions par $r - 1$ et $s - 1$ donc $r \geq 2, s \geq 2$ et par conséquent $t = r + s \geq 4$. Le succès sera l'extraction d'une boule blanche (échec pour une boule noire) et Bernoulli qualifie de "féconde" l'épreuve produisant un succès. Il introduit les notations:

$$p = \frac{r}{t}, \quad q = \frac{s}{t}$$

où p et q sont les probabilités respectives d'un succès et d'un échec. Pour $n \in \mathbb{N}^*$, Bernoulli s'intéresse à nt répétitions (indépendantes) de l'épreuve pour n grand. Aujourd'hui une épreuve aléatoire consistant en succès-échec est appelée *épreuve de Bernoulli*. Les résultats successifs des nt répétitions seront modélisés par des v.a. $X_1, X_2, ..., X_{nt}$. Chaque $X_i \sim Ber(p)$ et $X_i = 1$ signifie qu'un succès est réalisé lors de la i-ème répétition tandis que $X_i = 0$

correspond à un échec. L'indépendance supposée des répétitions est traduite par l'indépendance des $X_i, 1 \leq i \leq nt$. Le modèle des observations devient donc $X_1, X_2, ..., X_{nt}$ v.a i.i.d. comme $X \sim Ber(p)$. Le nombre de succès dans les nt répétitions est le nombre de 1 dans $X_1, X_2, ..., X_{nt}$ et donc donné par::

$$S_{nt} = \sum_{i=1}^{nt} X_i.$$

En vertu du théorème 1, $S_{nt} \sim Bin(nt, p)$. Le quotient $\frac{S_{nt}}{nt}$ sera dénommé *fréquence ou moyenne empirique* des succès dans les nt répétitions.

Selon [6], la dénommination "loi des grands nombres" n'apparaît qu'en 1837 chez S. D. Poisson et n'a jamais été utilisée par Jakob Bernoulli, contrairement à ce qu'affirme E. Borel dans son livre Le Hasard (PUF, p. 27, 1948). Cette loi mettra deux siècles à devenir "faible" par opposition à la version "forte" de Borel (1909) qui introduisit la notion de convergence presque sûre. Nous utiliserons néanmoins cette terminologie pour désigner le théorème de Bernoulli.

Théorème: Loi des grands nombres de Bernoulli

Soient r, s des entiers plus grands ou égaux à 2, $t = r + s \geq 4, n \in \mathbb{N}^*$ et c un nombre réel positif. Soit une épreuve de Bernoulli avec probabilité de succès $p = \frac{r}{t}$ qui est répétée nt fois, les répétitions étant supposées indépendantes. Pour t fixé que l'on peut choisir arbitrairement grand, si S_{nt} est le nombre de succès après nt répétitions de l'épreuve, alors, quel que soit $c > 0$:

$$P(|\frac{S_{nt}}{nt} - p| \leq \frac{1}{t}) > \frac{c}{c+1}$$

pour n suffisamment grand.

Remarque

Nous constatons que le théorème concerne seulement les épreuves de Bernoulli avec en plus un paramètre p rationnel.

Démonstration

Nous nous inspirons de l'approche de M. Henry [2].

S_{nt} est une v.a. binômiale de paramètres nt et $p = \frac{r}{t}$ $(q = 1 - p = \frac{s}{t})$ et donc:

$$P(S_{nt} = k) = \binom{nt}{k} p^k q^{nt-k},\ 0 \leq k \leq nt.$$

Rappelons que pour $0 \leq k \leq nt$, $\binom{nt}{nt-k} = \binom{nt}{k}$.

Notation de Bernoulli

$$M_k := \binom{nt}{k} r^{nt-k} s^k,\quad 0 \leq k \leq nt.$$

Il est clair que:

$$M_k = \binom{nt}{nt-k} r^{nt-k} s^k \text{ et donc } \sum_{k=0}^{nt} M_k = \sum_{k=0}^{nt} \binom{nt}{nt-k} r^{nt-k} s^k = (r+s)^{nt} = t^{nt},$$

De plus:

$$P(S_{nt} = nt-k) = \binom{nt}{nt-k} p^{nt-k} q^k = \binom{nt}{k} p^{nt-k} q^k = \frac{1}{t^{nt}} \binom{nt}{k} r^{nt-k} s^k = \frac{M_k}{t^{nt}}.$$

Monotonie des $\mathbf{M_k}$

Pour $0 \leq k < nt$, considérons les quotients:

$$\frac{M_{k+1}}{M_k} = \frac{\binom{nt}{k+1} r^{nt-k-1} s^{k+1}}{\binom{nt}{k} r^{nt-k} s^k} = \frac{(nt)!}{(k+1)!(nt-k-1)!} \frac{k!(nt-k)!}{(nt)!} \cdot \frac{s}{r}$$

$$= \frac{nt-k}{k+1} \cdot \frac{s}{r}.$$

Traitons d'abord le cas $0 \leq k < ns$. Alors $nt-k > nt-ns = nr$ et $ns \geq k+1$:

$$\frac{M_{k+1}}{M_k} = \frac{nt-k}{k+1} \cdot \frac{s}{r} > \frac{nr}{k+1} \cdot \frac{s}{r} = \frac{ns}{k+1} \geq 1.$$

Ainsi, pour $0 \leq k < ns, M_k < M_{k+1}$ et donc M_k est strictement croissante.

Si $ns \leq k < nt$, nous avons $-k \leq -ns, nt-k \leq nt-ns = nr$ et $ns < k+1$, d'où:

$$\frac{M_{k+1}}{M_k} = \frac{nt-k}{k+1} \cdot \frac{s}{r} \leq \frac{nr}{k+1} \cdot \frac{s}{r} = \frac{ns}{k+1} < 1.$$

Ainsi, pour $ns \leq k < nt, M_k > M_{k+1}$ et donc M_k est strictement décroissante.

Bernoulli a ainsi démontré que, pour $0 \leq k < nt, M_k$ passe par un maximum pour $k = ns$ et que les autres termes décroissent de part et d'autre de M_{ns}. Nous avons donc:

$$M_0 < M_1 < ... < M_{ns-1} < M_{ns}$$

et

$$M_{ns} > M_{ns+1} > ... > M_{ns+nr-1} > M_{ns+nr} = M_{nt}.$$

Monotonie des quotients

Dans la relation $\frac{M_{k+1}}{M_k} = \frac{nt-k}{k+1} \cdot \frac{s}{r}$, nous étendons la fonction

$$k \in \{0, 1, \cdots nt\} \mapsto \frac{nt-k}{k+1} \cdot \frac{s}{r}$$

à l'intervalle $x \in [0, nt]$:

$$x \in [0, nt] \mapsto \frac{nt-x}{x+1} \cdot \frac{s}{r}.$$

Puisque:

$$\frac{\partial}{\partial x}(\frac{nt-x}{x+1} \cdot \frac{s}{r}) = -\frac{(1+nt)}{(x+1)^2} \cdot \frac{s}{r} < 0$$

la fonction est strictement décroissante sur $\{0, 1, \cdots nt\}$. Nous en concluons que $k \mapsto \frac{M_{k+1}}{M_k}$ est strictement décroissante sur $\{0, 1, \cdots nt - 1\}$, ce qui implique:

$$\text{pour } \ 0 < k \leq ns \ \text{ et } \ j < k - 1, \ \ \frac{M_k}{M_{k-1}} < \frac{M_{k-j}}{M_{k-j-1}}.$$

Par contre, il est clair que $\frac{M_k}{M_{k+1}}$ est strictement croissante sur $\{0, 1, \cdots nt-1\}$, ce qui implique:

$$\text{pour } \ ns \leq k < nt \ \text{ et } \ j < nt - k, \ \ \frac{M_k}{M_{k+1}} < \frac{M_{k+j}}{M_{k+j+1}}.$$

Bernoulli s'intéresse ensuite aux trois éléments suivants:

$$L = M_{ns-n}, \ \ M = M_{ns}, \ \ \Lambda = M_{ns+n},$$

et va montrer que, pour r et s fixés:

$$\lim_{n\to+\infty}\frac{M}{L}=+\infty \quad \text{et} \quad \lim_{n\to+\infty}\frac{M}{\Lambda}=+\infty.$$

L'auteur donne d'abord un très mauvais argument heuristique qui ne résiste pas à une analyse sérieuse. Pressentant la difficulté, Bernoulli propose alors une argumentation très soignée, montrant sa parfaite maîtrise des limites un bon siècle avant les définitions rigoureuses de Cauchy.

Le rapport $\frac{M}{L}$ peut s'écrire de la façon suivante:

$$\frac{M}{L}=\frac{M_{ns}}{M_{ns-n}}=\frac{M_{ns}}{M_{ns-1}}\cdot\frac{M_{ns-1}}{M_{ns-2}}\cdots\frac{M_{ns-n+1}}{M_{ns-n}} \quad (\mathfrak{R}).$$

Pour $0\leq k\leq nt-1$, nous avons:

$$\frac{M_{k+1}}{M_k}=\frac{nt-k}{k+1}\cdot\frac{s}{r}$$

et dans le produit $(\mathfrak{R})$, l'indice k doit satisfaire $ns\geq k+1\geq ns-n+1$ et donc $ns-1\geq k\geq ns-n$. Par conséquent:

$$\frac{M}{L}=\frac{nt-(ns-1)}{(ns-1)+1}\cdot\frac{s}{r}\cdot\frac{nt-(ns-2)}{(ns-2)+1}\cdot\frac{s}{r}\cdots\frac{nt-(ns-n)}{(ns-n)+1}\cdot\frac{s}{r}.$$

En utilisant $nt-ns=nr$:

$$\frac{M}{L}=\frac{nr+1}{ns}\cdot\frac{nr+2}{ns-1}\cdots\frac{nr+n}{ns-n+1}\cdot(\frac{s}{r})^n=\prod_{j=1}^{n}\frac{(nr+j)s}{(ns-(j-1))r}$$

et finalement:

$$\frac{M}{L}=\prod_{j=1}^{n}\frac{nrs+js}{nrs-(j-1)r}.$$

Vérifions que, pour $1\leq j\leq n$, chaque facteur de ce produit est bien défini et strictement supérieur à 1.

$$nrs-(j-1)r=r(ns-j+1)\geq r(ns-n+1)=r(n(s-1)+1)>0$$

et:

$$\frac{nrs+js}{nrs-(j\text{–}1)r}>\frac{nrs}{nrs}=1.$$

De plus, $\frac{nrs+js}{nrs-(j-1)r}$ est croissante en j car pour $x\in[0,n]$:

$$\frac{\partial}{\partial x}(\frac{nrs+xs}{nrs-(x-1)r}) = \frac{nrs^2+rs+nr^2s}{(nrs-(x-1)r)^2} > 0.$$

Rappelons qu'un produit infini de facteurs tous plus grands que 1 peut très bien ne pas tendre vers $+\infty$. Par exemple $\prod_{k=1}^{+\infty}(1+\frac{1}{k^\alpha})$ converge pour $\alpha > 1$. En effet, l'inégalité $1+x \leq e^x$ implique

$$\prod_{k=1}^{n}(1+\frac{1}{k^\alpha}) \leq e^{\sum_{k=1}^{n}\frac{1}{k^\alpha}} < e^{\sum_{k=1}^{+\infty}\frac{1}{k^\alpha}} < +\infty$$

car $\sum_{k=1}^{+\infty}\frac{1}{k^\alpha} < +\infty$ si $\alpha > 1$. La suite des produits finis est donc croissante et bornée, d'où sa convergence dans $\mathbb{R}$. Par contre, si $\alpha = 1$:

$$\prod_{k=1}^{n}(1+\frac{1}{k}) = \prod_{k=1}^{n}(\frac{k+1}{k}) = \frac{2}{1}\cdot\frac{3}{2}\cdots\frac{n+1}{n} = n+1 \to +\infty \text{ si } n \to +\infty.$$

Par ailleurs, si $a > 1$, alors a^n tend vers $+\infty$ avec n car $\ln(a^n) = n\ln(a)$ et $\ln(a) > 0$. L'idée de Bernoulli consiste, pour n donné, à minorer le plus grand nombre possible, noté $m^*(n)$, de facteurs de ce produit par $\frac{r+1}{r} > 1$ (nombre indépendant de l'indice) sans modifier les autres qui, comme nous le savons sont tous > 1 et donc leur produit aussi. Si $m^*(n)$ tend vers $+\infty$ avec n, alors pour $\gamma > 0$ arbitrairement grand fixé et n suffisamment grand, nous pourrons réaliser les inégalités suivantes:

$$\frac{M}{L} > (\frac{r+1}{r})^{m^*(n)} \geq \gamma.$$

En suivant Bernoulli qui choisit d'utiliser les logarithmes décimaux, la seconde inégalité ci-dessus équivaut à:

$$m^*(n) \geq \frac{Log(\gamma)}{Log(r+1)-Log(r)} =: m_L.$$

Déterminons maintenant $m^*(n)$, le nombre de facteurs plus grands que $\frac{r+1}{r}$ dans le produit:

$$\frac{M}{L} = \prod_{j=1}^{n}\frac{(nr+j)s}{(ns-(j-1))r}.$$

Nous avons vu précédemment que $\frac{(nr+j)s}{(ns-(j-1))r}$ est croissant en j. Il suffit donc de déterminer le plus petit j ayant cette propriété. La minoration ci-dessus

n'est possible que pour les j vérifiant:

$$\frac{nrs+js}{nrs-(j-1)r} \geq \frac{r+1}{r}.$$

Or, cette inégalité équivaut à chacune des suivantes:

$$nr^2s+jrs \geq nr^2s+nrs-(j-1)r^2-(j-1), \quad jrs+(j-1)r^2+(j-1)r \geq nrs,$$

$$jrs+(j-1)r^2+(j-1)r \geq nrs, \;\; js+jr-r+j-1 \geq ns, j(r+s+1) \geq ns+r+1$$

et finalement à:

$$\frac{ns+r+1}{s+r+1} \leq j \leq n.$$

Par conséquent, pour $n \geq 1$ donné, le premier élément étant inclu, le nombre de valeurs de j permettant cette minoration est:

$$m^*(n) = n+1-\frac{ns+r+1}{s+r+1} = \frac{n(r+1)+s}{s+r+1} \geq 1 \;\; \text{car} \;\; n \geq 1.$$

Remarquons que pour r et s fixés, $m^*(n)$ tend vers $+\infty$ avec n. Cette fonction de $\mathbb{R}$ sur $\mathbb{R}$ est strictement croissante et donc inversible, sa fonction inverse (après calcul) étant:

$$n^*(m) = \frac{m(r+s+1)-s}{r+1}.$$

Elle possède les mêmes propriétés que $m^*(n)$ et la condition $m^*(n) \geq m_L$ équivaut à $n = n^*(m^*(n)) \geq n^*(m_L)$ d'où:

$$n \geq n^*(m_L) = \frac{m_L(r+s+1)-s}{r+1} = \frac{\frac{Log(\gamma)}{Log(r+1)-Log(r)}(r+s+1)-s}{r+1} := n_L.$$

Ainsi, pour r, s et $\gamma > 0$ fixés, si $n \geq n_L$, alors:

$$\frac{M}{L} > (\frac{r+1}{r})^{m^*(n)} \geq \gamma.$$

Par conséquent:

$$\forall \gamma > 0 \;,\; \exists \;\; n_L \;\; \text{tel} \;\; \text{que} \;\; \forall \;\; n \geq n_L, \;\; \frac{M}{L} \geq \gamma.$$

et donc:

$$\frac{M}{L} \to +\infty \;\; \text{pour} \;\; n \to +\infty.$$

Le même procédé s'applique aux termes situés à droite de M en échangeant toutefois s et r dans les formules puisque:

$$\frac{M}{\Lambda} = \frac{ns+1}{nr} \cdot \frac{ns+2}{nr-1} \cdots \frac{ns+n}{nr-n+1} \cdot (\frac{r}{s})^n = \prod_{j=1}^{n} \frac{nrs+jr}{nrs-(j-1)s}.$$

Notons $m_*(n)$ le nombre de facteurs qui sont minorés par $\frac{s+1}{s}$ dans le produit ci-dessus:

$$m_*(n) = n+1-\frac{nr+s+1}{s+r+1} = \frac{n(s+1)+r}{s+r+1}.$$

On se donne $\delta > 0$ au lieu de $\gamma > 0$ et nous cherchons maintenant à réaliser les inégalités:

$$\frac{M}{\Lambda} > (\frac{s+1}{s})^{m_*(n)} \geq \delta.$$

La seconde est réalisée si et seulement si:

$$m_*(n) \geq \frac{Log(\delta)}{Log(s+1)-Log(s)} =: m_\Lambda,$$

et:

$$m_*(n) = \frac{n(s+1)+r}{r+s+1}.$$

Sa fonction inverse est:

$$n_*(m) = \frac{m(r+s+1)-r}{s+1}.$$

La condition $m_*(n) \geq m_\Lambda$ équivaut alors à $n = n_*(m_*(n)) \geq n_*(m_\Lambda)$ d'où:

$$n \geq n^*(m_\Lambda) = \frac{m_\Lambda(r+s+1)-r}{s+1} = \frac{\frac{Log(\delta)}{Log(s+1)-Log(s)}(r+s+1)-r}{s+1} := n_\Lambda.$$

Ainsi, pour r, s et $\delta > 0$ fixés, si $n \geq n_\Lambda$, alors:

$$\frac{M}{\Lambda} > (\frac{s+1}{s})^{m_*(n)} \geq \delta.$$

La conclusion est, pour r et s fixés:

$$\forall \delta > 0 \ , \quad \exists \ \ n_\Lambda \ \ \text{tel} \quad \text{que} \ \ \forall \ \ n \geq n_\Lambda, \quad \frac{M}{\Lambda} \geq \delta$$

ce qui implique:

$$\frac{M}{\Lambda} \to +\infty \quad \text{pour} \quad n \to +\infty.$$

Pour satisfaire toutes ces conditions, une borne inférieure pour le nombre de répétitions de l'épreuve sera:

$$nt = \max(n_L t, n_\Lambda t).$$

Et maintenant quelques petits "coups de génie"!

Revenons à:

$$M_0, M_1, ..., M_{ns-2n}, ..., M_{ns-n-1}, M_{ns-n} = \mathbf{L}, ..., M_{ns-1}, M_{ns} = \mathbf{M},$$

$$M_{ns+1}, ..., M_{ns+n} = \mathbf{\Lambda}, ., ., M_{nt-1}, M_{nt}.$$

Dans une première étape, Bernoulli compare les quotients du type $\frac{M_k}{M_{k-1}}$ pour k descendant de ns à $ns-n+1$ à ceux indexés par k de $ns-n$ à $ns-2n+1$. En vertu de la "monotonie des quotients" établie précédemment, nous avons:

$$\frac{M}{M_{ns-1}} = \frac{M_{ns}}{M_{ns-1}} < \frac{M_{ns-n}}{M_{ns-n-1}} = \frac{L}{M_{ns-n-1}}; \ \frac{M_{ns-1}}{M_{ns-2}} < \frac{M_{ns-n-1}}{M_{ns-n-2}};$$

$$\frac{M_{ns-2}}{M_{ns-3}} < \frac{M_{ns-n-2}}{M_{ns-n-3}}; ...; \frac{M_{ns-n+1}}{M_{ns-n}} < \frac{M_{ns-2n+1}}{M_{ns-2n}}.$$

Ceci implique:

$$\frac{M}{L} < \frac{M_{ns-1}}{M_{ns-n-1}}; \frac{M_{ns-1}}{M_{ns-n-1}} < \frac{M_{ns-2}}{M_{ns-n-2}}; ...; \frac{M_{ns-n+1}}{M_{ns-2n+1}} < \frac{M_{ns-n}}{M_{ns-2n}} = \frac{L}{M_{ns-2n}},$$

d'où:

$$\frac{M}{L} < \frac{M_{ns-1}}{M_{ns-n-1}} < \frac{M_{ns-2}}{M_{ns-n-2}} < \frac{M_{ns-3}}{M_{ns-n-3}} < ... < \frac{M_{ns-n}}{M_{ns-2n}} = \frac{L}{M_{ns-2n}}.$$

Lemme

Soient $a_1, a_2, ..., a_n, b_1, b_2, ..., b_n$ des nombres réels positifs tels que:

$$\frac{a_1}{b_1} < \frac{a_2}{b_2} < \frac{a_3}{b_3} < ... < \frac{a_n}{b_n}.$$

Alors:

$$\frac{a_2 + a_3 + \cdots + a_n}{b_2 + b_3 + \cdots + b_n} > \frac{a_1}{b_1}.$$

Démonstration

Les hypothèses impliquent:

$$a_1b_2 < b_1a_2,\ a_1b_3 < b_1a_3, ..., a_1b_n < b_1a_n$$

d'où l'on déduit:

$$b_1(a_2 + a_3 + \cdots + a_n) > a_1(b_2 + b_3 + \cdots + b_n)$$

et finalement le résultat cherché.

Pour appliquer le lemme, l'indice du haut va de $ns - n$ à $ns - 1$ et celui du bas de $ns - 2n$ à $ns - n - 1$. L'inégalité (voir avant) $\frac{M}{L} \geq \gamma$ pour $n \geq n_L$ et le dernier lemme fournissent:

$$\frac{\sum_{k=ns-n}^{ns-1} M_k}{\sum_{k=ns-2n}^{ns-n-1} M_k} > \frac{M}{L} \geq \gamma.$$

Bernoulli va remplacer $\sum_{k=ns-2n}^{ns-n-1} M_k$ par la somme depuis $k = 0$. Il constate que $\sum_{k=0}^{ns-n-1} M_k$ contient $ns - n = (s-1)n$ termes. Il l'exprime comme somme de $s - 1$ paquets formés chacun de la somme de n termes adjacents. Comme M_k est croissante en k pour $0 \leq k \leq ns$, chacune de ces $s - 1$ sommes est majorée par $\sum_{k=ns-2n}^{ns-n-1} M_k$ et donc:

$$\sum_{k=0}^{ns-n-1} M_k \leq (s-1) \sum_{k=ns-2n}^{ns-n-1} M_k.$$

Par conséquent, pour c aussi grand que l'on veut, c, s et r étant fixés dans le raisonnement, $\gamma = (s-1)c$ et $n \geq n_L$:

$$\frac{\sum_{k=ns-n}^{ns-1} M_k}{\sum_{k=0}^{ns-n-1} M_k} \geq \frac{\sum_{k=ns-n}^{ns-1} M_k}{(s-1)\sum_{k=ns-2n}^{ns-n-1} M_k} \geq \frac{\gamma}{s-1} = c.$$

Une démarche "symétrique" avec $\frac{M}{\Lambda}$ et les quotients du type $\frac{M_k}{M_{k+1}}$ à droite de ns, fournit le résultat suivant avec $c > 0$ arbitrairement grand, $\delta = (r-1)c$ et $n \geq n_\Lambda$:

$$\frac{\sum_{k=ns+1}^{ns+n} M_k}{\sum_{k=ns+n+1}^{nt} M_k} \geq \frac{\sum_{k=ns+1}^{ns+n} M_k}{(r-1)\sum_{k=ns+n+1}^{ns+2n} M_k} \geq \frac{\delta}{r-1} = c.$$

Nous constatons que les raisonnements précédents exigent $r \geq 2$ et $s \geq 2$.

Pour démontrer le prochain théorème nous aurons besoin du résultat suivant:

Lemme

Soient A, B, C, D et α cinq réels positifs tels que $\frac{A}{B} \geq \alpha$ et $\frac{C}{D} \geq \alpha$. Alors:

$$\frac{A+C}{B+D} \geq \alpha.$$

Démonstration

Des hypothèses découlent $A \geq \alpha B, C \geq \alpha D$ et donc $A + C \geq \alpha(B + D)$, d'où le résultat attendu.

Théorème

$$\frac{\sum_{k=ns-n}^{ns+n} M_k}{t^{nt}} \to 1 \quad \text{lorsque} \quad n \to +\infty.$$

Démonstration

Rappelons que $\sum_{k=0}^{nt} M_k = (r+s)^{nt}$ et $M_{ns} = M > 0$. Le dernier lemme fournit la dernière inégalité si $n \geq \max(n_L, n_\Lambda)$:

$$\frac{\sum_{k=ns-n}^{ns+n} M_k}{(r+s)^{nt} - \sum_{k=ns-n}^{ns+n} M_k} = \frac{\sum_{k=ns-n}^{ns-1} M_k + M_{ns} + \sum_{k=ns+1}^{ns+n} M_k}{\sum_{k=0}^{ns-n-1} M_k + \sum_{k=ns+n+1}^{nt} M_k} >$$

$$> \frac{\sum_{k=ns-n}^{ns-1} M_k + \sum_{k=ns+1}^{ns+n} M_k}{\sum_{k=0}^{ns-n-1} M_k + \sum_{k=ns+n+1}^{nt} M_k} \geq c.$$

Ainsi, pour $n \geq \max(n_L, n_\Lambda)$:

$$\frac{\sum_{k=ns-n}^{ns+n} M_k}{(r+s)^{nt} - \sum_{k=ns-n}^{ns+n} M_k} > c.$$

Notons $D_n = \sum_{k=ns-n}^{ns+n} M_k$. Puisque $0 < D_n < t^{nt}$, nous avons $0 < \frac{D_n}{t^{nt}} < 1$ et pour n suffisamment grand:

$$\frac{\sum_{k=ns-n}^{ns+n} M_k}{t^{nt} - \sum_{k=ns-n}^{ns+n} M_k} = \frac{D_n}{t^{nt} - D_n} = \frac{\frac{D_n}{t^{nt}}}{1 - \frac{D_n}{t^{nt}}} > c.$$

Ceci équivaut à:

$$\frac{D_n}{t^{nt}} > c(1 - \frac{D_n}{t^{nt}}) \quad \text{et} \quad \text{donc} \quad \text{à} \quad (c+1)\frac{D_n}{t^{nt}} > c$$

et finalement:

$$\frac{D_n}{t^{nt}} > \frac{c}{c+1}.$$

Ainsi, quel que soit $c > 0$ pour n suffisamment grand.:

$$\frac{c}{c+1} < \frac{D_n}{t^{nt}} < 1.$$

Puisque $c > 0$ peut être choisi arbitrairement grand, il s'ensuit que:

$$\frac{D_n}{t^{nt}} \to 1 \text{ lorsque } n \to +\infty.$$

Rappelons que $t = r + s, p = \frac{r}{t}$, $P(S_{nt} = nt - k) = \frac{M_k}{t^{nt}}$ et que t est fixé.Ainsi:

$$\frac{D_n}{t^{nt}} = \sum_{k=ns-n}^{ns+n} \frac{M_k}{t^{nt}} = \sum_{k=ns-n}^{ns+n} P(S_{nt} = nt - k)$$

$$= P(nt - ns - n \leq S_{nt} \leq nt - ns + n) = P(n(t-s) - n \leq S_{nt} \leq n(t-s) + n)$$

$$= P(nr - n \leq S_{nt} \leq nr + n) = P(n(r-1) \leq S_{nt} \leq n(r+1))$$

$$= P(\frac{r-1}{t} \leq \frac{S_{nt}}{nt} \leq \frac{r+1}{t}) = P(p - \frac{1}{t} \leq \frac{S_{nt}}{nt} \leq p + \frac{1}{t}) = P(-\frac{1}{t} \leq \frac{S_{nt}}{nt} - p \leq \frac{1}{t})$$

$$= P(|\frac{S_{nt}}{nt} - p| \leq \frac{1}{t}) \to 1 \text{ pour } n \to +\infty.$$

Le théorème est démontré.

La conclusion est aussi équivalente à:

$$\text{Pour } n \to +\infty, \quad P(|\frac{S_{nt}}{nt} - p| > \frac{1}{t}) \to 0.$$

On peut aussi formuler le résultat ainsi: pour $\varepsilon > 0$ arbitrairement petit, on peut choisir t tel que $\frac{1}{t} < \varepsilon$ et ensuite fixer r, s avec $p = \frac{r}{t}$ et $s = t - r$. Alors, pour t fixé, quel que soit $c > 0$, pour n suffisamment grand, nous avons:

$$P(|\frac{S_{nt}}{nt} - p| \leq \varepsilon) > \frac{c}{c+1}.$$

Estimations numériques

Bernoulli a appliqué son théorème pour évaluer le nombre de répétitions nt nécessaires pour estimer p avec une précision imposée à priori. Plus précisément, p étant fixé, pour une précision ε et un nombre c donnés, déterminer nt pour que la fréquence relative des succès (observée) $\frac{S_{nt}}{nt}$ soit éloignée de p de moins de ε avec une probabilité supérieure à $\frac{c}{c+1}$.

Bernoulli propose les valeurs suivantes: $p = \frac{3}{5}, \varepsilon = 0.02, t = 50, r = 30$ (et donc $s = 20$) et $c = 1000$ (i.e. $\frac{c}{c+1} = 0.999$), $\gamma = 19'000$ et $\delta = 29'000$. Le théorème affirme que, pour n assez grand:

$$P(p - \varepsilon \leq \frac{S_{nt}}{nt} \leq p + \varepsilon) = P(|\frac{S_{nt}}{nt} - p| \leq \varepsilon) > \frac{c}{c+1} = 0.999.$$

Rappelons que $n \geq \max(n_L, n_\Lambda)$ où:

$$m_L = \frac{Log(\gamma)}{Log(r+1) - Log(r)}, \quad n_L = \frac{m_L(r+s+1)}{r+1} \quad \text{et} \quad \gamma = c(s-1)$$

$$m_\Lambda = \frac{Log(\delta)}{Log(s+1) - Log(s)}, \quad n_\Lambda = \frac{m_L(r+s+1)}{s+1} \quad \text{et} \quad \delta = c(r-1).$$

Numériquement nous obtenons

$$m_L = 301, n_L = 495, m_\Lambda = 211, n_\Lambda = 511$$

et donc $n \geq \max(495, 511) = 511$,. La condition devient $nt \geq 25'550$, nombre difficilement réalisable en pratique. Pour 0.99 il faudra $19'800$ répétitions et pour 0.9, $14'000$ répétitions. Si l'on choisit $\varepsilon = 0.01, t = 100$ et $c = 1'000$, alors $nt = 109'500$. Ces grandeurs sont évidemment inutilisables en pratique. Nous partageons le sentiment de M. Henry [2]: ce problème pourrait être l'une des raisons ayant poussé Bernoulli a laissé Ars Conjectandi inachevé, cherchant peut-être à améliorer son résultat. J'aimerais toutefois rappeler que Martin Mattmüller n'est pas de cet avis; selon lui, aucun indice dans les écrits de Bernoulli ne laisse penser que ce dernier ait jamais envisagé une réalisation pratique de son théorème. N'oublions pas que Martin est un grand connaisseur de Jakob Bernoulli, tandis que l'autre avis participe plus d'une vision sentimentale que d'une véritable connaissance historique.

Une autre raison aurait été la recherche d'applications des probabilités à tous les domaines des activités humaines [4]. Il est surprenant que Leibniz ait eu une opinion opposée, convaincu que le calcul des probabilités devait se limiter aux jeux de hasard et qu'il ne voyait pas l'intérêt d'un tel calcul pour l'étude concernant d'autres activités humaines. Et puis, disait-il:

Ce n'est pas un nombre fini d'expériences qui va permettre d'estimer ce qui dépend d'une infinité de circonstances. Et d'ailleurs, dans les affaires humaines, les conditions générales peuvent changer, de nouvelles maladies apparaître, il est donc impossible de faire des prévisions valables pour l'avenir.

Déroutant n'est-ce pas ?

La détermination du nombre de facteurs plus grands que $\frac{r+1}{r}$ (ou $\frac{s+1}{s}$) repose sur une approximation grossière des coefficients binômiaux. Celle-ci suffit pleinement à la démonstration du théorème de Bernoulli, mais pas à une évaluation fiable des valeurs de nt garantissant une précision donnée pour estimer p avec $\frac{S_{nt}}{nt}$. Malheureusement pour Bernoulli, les approximations numériques convenables de ces termes pour des grandes valeurs de n sont apparues plus tard (de Moivre et Stirling). Les estimations fournies par l'approximation normale du théorème de de Moivre (1718), pour $\varepsilon = 0.02, p = 0.6$ et $P(|\frac{S_n}{n} - p| < 0.02) = 0.999$, fournissent $n = 6'534$ répétitions, un nombre beaucoup plus accessible pratiquement. Bien sûr, les estimations décevantes de Bernoulli n'enlèvent absolument rien à l'énorme intérêt théorique de son travail.

La notion de convergence en probabilité

Dans la littérature, en conclusion du travail de Bernoulli, on peut trouver l'affirmation:

$\frac{S_{nt}}{nt}$ *converge en probabilité vers* p *pour* $n \to +\infty$.

Rappelons d'abord la définition de cette notion.

Définition

Une suite $(X_n)_{n\geq 1}$ de v.a.converge en probabilité vers la v.a. X lorsque $n \to +\infty$ si:

$$\forall \varepsilon > 0 \ , \lim_{n\to+\infty} P(|X_n - X| > \varepsilon) = 0$$

ou, de façon équivalente:

$$\forall \varepsilon > 0 \ , \lim_{n\to+\infty} P(|X_n - X| \leq \varepsilon) = 1.$$

Notation

$$X_n \xrightarrow{P} X \ \text{ lorsque } \ n \to +\infty.$$

Conséquence

Si une suite converge en probabilité, alors toutes ses sous-suites convergent en probabilité vers la même limite.

Jakob a-t-il vraiment démontré la convergence en probabilité dans son approche? Ceci n'est pas évident. En effet, changer ε implique changer t et donc la sous-suite $(\frac{S_{nt}}{nt})_{n\geq 1}$, d'où le problème. De plus, même si pour tout $t \geq 4, (\frac{S_{nt}}{nt})_{n\geq 1}$ convergeait en probabilité vers la même limite, nous ne pourrions pas en déduire, sans argument supplémentaire, la convergence en probabilité de $(\frac{S_n}{n})_{n\geq 1}$. L'exemple qui suit fournit une suite $(X_n)_{n\geq 1}$ qui ne converge pas en probabilité, bien que toutes ses sous-suites de la forme $(X_{nt})_{n\geq 1}$, avec $t \geq 4$, convergent en probabilité vers la même limite. Bien sûr, il contient moins de structure que la suite étudiée par Bernoulli.

Exemple.

Nous considérons la suite (croissante) des nombres premiers notée $(n_k)_{k\geq 1}$ et une suite de v.a. $(X_{n_k})_{k\geq 1}$ qui converge en probabilité vers 0 lorsque $k \to +\infty$. Notons $(m_j)_{j\geq 1}$ la suite (croissante) des entiers qui ne sont pas premiers et $(X_{m_j})_{j\geq 1}$ une suite de v.a. qui converge en probabilité vers 1 lorsque $j \to +\infty$. La réunion de ces deux suites nous fournit une suite indexée par $\mathbb{N}^*$ qui sera notée $(X_n)_{n\geq 1}$. Puisque toutes les sous-suites d'une suite qui converge en probabilité convergent en probabilité vers la même limite, ceci exclut la convergence en probabilité de $(X_n)_{n\geq 1}$. Or, pour $t \geq 4$ la suite (croissante) $(nt)_{n\geq 1}$ est une sous-suite de $(m_j)_{j\geq 1}$ car aucun des nombres qui la constitue n'est premier. Par conséquent $(X_{nt})_{n\geq 1}$ converge en probabilité vers 1 lorsque $n \to +\infty$.

Nous savons aujourd'hui (voir plus loin) que les hypothèses de Bernoulli entraînent la convergence en probabilité de $(\frac{S_n}{n})_{n\geq 1}$. Même si, au sens strict, Bernoulli n'a pas démontré cette convergence, il a bien évidemment ouvert largement la porte à celle-ci.

Remarques sur le travail de Bernoulli

Bernoulli démontre son théorème **seulement** pour des épreuves de Bernoulli avec une argumentation qui paraît aujourd'hui inutilement compliquée. Elle occupe en effet 16 pages dans *Ars Conjectandi* alors que les techniques

actuelles permettent non seulement d'englober un cadre beaucoup plus large de cas, mais aussi de fournir des démonstrations qui se réduisent à quelques lignes (voir plus loin). La démonstration de Bernoulli conserve cependant tout son intérêt. Sa preuve, qui repose sur une analyse des coefficients binômiaux, est un chef-d'oeuvre d'intelligence et de finesse et son contenu restera toujours le travail fondateur de la théorie moderne des probabilités et de la statistique mathématique.

Le théorème de Bernoulli est d'abord le fruit d'une intuition extraordinaire. Un grand théorème ne peut pas être un produit uniquement technique, même si cette dernière est très complexe.

Einstein disait:

La logique n'a jamais rien créé.

Et Euler:

La science, c'est ce qu'on fait après avoir deviné juste.

Comme l'indique la liste ci-dessous, l'auteur a donné naissance à un véritable carrefour innovateur:

1. Il est vraisemblablement le premier à avoir modélisé, en utilisant le même cadre mathématique, un phénomène d'une part et les observations de celui-ci d'autre part.

2. Il est le premier à justifier mathématiquement l'utilisation de la moyenne empirique pour estimer une probabilité, en fait une espérance car pour une v.a. $X \sim Ber(p), p = E(X)$.

3. Il est le premier à avoir donné un cadre mathématique à la régularité statistique asymptotique (grand nombre de répétitions) observée expérimentalement.

4. Il invente la notion d'intervalle de confiance et fonde ainsi la théorie de l'estimation et donc la statistique mathématique. Il évalue même le nombre de répétitions qui garantit une précision donnée à priori. Il utilise malheureusement des approximations très grossières débouchant sur des nombres de répétitions beaucoup trop élevés pour être réalisables.

5. Il introduit une nouvelle notion de convergence qui conduira à celle de convergence en probabilité.

6. Ce travail est la charnière qui a permis le passage du traditionnel *calcul des probabilités développé essentiellement pour la discussion des jeux de hasard* à la *théorie des probabilités*.

7. Comme P. Picard le fait remarquer [4], ce théorème est le précurseur d'une longue lignée de théorèmes asymptotiques, qui conjugueront élégance mathématique et utilité pratique. En introduisant la notion de limite, Bernoulli aura offert au calcul des probabilités un champ immense pour de futurs développements.

8. Bien que son champ d'application soit limité au cas des épreuves de Bernoulli avec paramètres rationnels, ce théorème reste fondamental par ses gigantesques conséquences.

Etat actuel de la loi des grands nombres.

Loi faible et forte des grands nombres

L'adjectif "faible" désigne la convergence en probabilité et "forte" la convergence (p.s.) qui entraîne la précédente.

Définition

Une suite $(X_n)_{n\geq 1}$ de v.a. définies sur l'ensemble Ω converge presque sûrement vers la v.a. X pour $n \to +\infty$, s'il existe $N \in \mathfrak{F}$ tel que $P(N) = 0$ et

$$\forall \ \ \omega \in N^c, \lim_{n\to+\infty} X_n(\omega) = X(\omega).$$

Notation

$$X_n \overset{\text{(p.s.)}}{\longrightarrow} X \ \ \text{lorsque} \ \ n \to +\infty.$$

Voici deux théorèmes qui lient convergence en probabilité et convergence (p.s.).

Théorème 2

La convergence (p.s.) d'une suite de v.a. entraîne sa convergence en probabilité. L'inverse est en général faux.

Le théorème 2 est une conséquence immédiate du théorème suivant:

Théorème 3 (sans démonstration)

Une suite $(X_n)_{n\geq 1}$ de v.a. converge en probabilité vers la v.a. X lorsque $n \to +\infty$ si et seulement si toute sous-suite contient une sous-suite (donc une sous-sous-suite de la suite initiale) qui converge (p.s.) vers X.

Corollaire du théorème 3

Soient deux suites de v.a. telles que $X_n \xrightarrow{P} X$ et $Y_n \xrightarrow{P} Y$ lorsque $n \to +\infty$. Alors $X_n + Y_n \xrightarrow{P} X + Y$. lorsque $n \to +\infty$.

Démonstration

Il suffit d'appliquer le théorème 3. Soit $(n_k)_{k\geq 1}$ une sous-suite quelconque. Il existe alors une sous-suite de cette dernière telle que $X_{n_{k_j}} \xrightarrow{(p.s.)} X$ et une sous-sous-suite $Y_{n_{k_{j_l}}} \xrightarrow{(p.s.)} Y$. Par conséquent:

$$X_{n_{k_{j_l}}} + Y_{n_{k_{j_l}}} \xrightarrow{\text{(p.s.)}} X + Y \quad \text{lorsque} \quad l \to +\infty.$$

Inégalités de Markov et de Chebyshev

Soit une v.a. $X \geq 0$. Nous avons l'inégalité de Markov:

$$\forall \;\; \lambda > 0, \quad P(X \geq \lambda) \leq \frac{E(X)}{\lambda}.$$

En effet:

$$E(X) = \int_\Omega X dP \geq \int_\Omega I_{\{X\geq\lambda\}} X dP \geq \lambda \int_\Omega I_{\{X\geq\lambda\}} dP = \lambda P(X \geq \lambda).$$

L'inégalité de Chebyshev:

$$\forall \;\; \lambda > 0, \quad P(|X - E(X)| \geq \lambda) \geq \frac{Var(X)}{\lambda^2}$$

en est un cas particulier. En effet:

$$P(|X-E(X)| \geq \lambda) = P((X-E(X))^2 \geq \lambda^2) \leq \frac{E((X - E(X))^2)}{\lambda^2} = \frac{Var(X)}{\lambda^2}.$$

Rappelons qu'une v.a. de carré intégrable est automatiquement intégrable.

En effet, cadeau d'une probabilité i.e. d'une mesure finie: pour $p \geq 1$, les espaces $\mathbf{L}_p$= ensemble des fonctions mesurables de Ω dans $\mathbb{R}$ de p-ième puissance intégrables, sont emboîtés en décroissant avec p. Ainsi une v.a. de carré intégrable est automatiquement intégrable.

Théorème 4: Loi faible des grands nombres

Soit $(X_n)_{n\geq 1}$ une suite de v.a. i.i.d. comme X de carré intégrable, $E(X) = \mu$, et $S_n = \sum_{k=1}^{n} X_k$. Alors $(\frac{S_n}{n})_{n\geq 1}$ converge en probabilité vers μ lorsque $n \to +\infty$ i.e.

$$\forall\ \varepsilon > 0, \lim_{n\to+\infty} P(|\frac{S_n}{n} - \mu| > \varepsilon) = 0.$$

Démonstration

$Var(X) = E(X^2) - (E(X))^2 = \sigma^2$ est donc finie. De plus, nos hypothèses impliquent $E(\frac{S_n}{n}) = \mu$ et $Var(\frac{S_n}{n}) = \frac{1}{n^2}Var(S_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$, la deuxième égalité découlant de l'indépendance des v.a.. L'inégalité de Chebyshev fournit:

$$\forall\ \ \varepsilon > 0, \quad P(|\frac{S_n}{n} - \mu| > \varepsilon) \leq \frac{Var(\frac{S_n}{n})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \ \text{ si } \ n \to +\infty.$$

Le théorème 4 contient bien sûr le cas de Bernoulli avec $X \sim Ber(p), \mu = p = E(X), \sigma^2 = p(1-p) = Var(X)$ et $S_n \sim Bin(n,p)$. Que de progrès réalisés depuis le travail original de Bernoulli. Rappelons que la démonstration de Jakob Bernoulli, parue dans la publication posthume de Ars Conjectandi (1713), s'étend sur 16 pages et traite uniquement le cas des épreuves de Bernoulli. Or l'inégalité de Chebyshev nous donne la loi des grands nombres en une ligne et ceci pour toutes les v.a. de carré intégrable. Heureusement que Bernoulli n'ait pas vu cette démonstration ! Mais celle-ci n'enlève rien à l'importance de son résultat ni à notre admiration pour son intuition et sa preuve.

Affaiblissement des hypothèses

Il est encore possible de réduire les exigences. L'hypothèse d'indépendance dans le théorème 4 peut être affaiblie car la démonstration s'appuie sur la propriété $Var(\sum_{k=1}^{n} X_k) = \sum_{k=1}^{n} Var(X_k)$. Cette dernière est déjà réalisée par la condition plus faible $X_1, X_2, ..., X_n, ...$ sont non-corrélées i.e.:

$$\forall\ i, j \geq 1, i \neq j, E(X_i X_j) = E(X_i)E(X_j),$$

en supposant bien sûr l'existence de toutes ces espérances. De plus, on peut admettre des v.a. qui ne sont pas identiquement distribuées. Le théorème 4 peut être reformulé ainsi:

Théorème 5

Soit $(X_n)_{n\geq 1}$ une suite de v.a. de carré intégrables, non-corrélées, d'espérances et de variances respectives $(\mu_n)_{n\geq 1}$ et $(\sigma_n^2)_{n\geq 1}$ (variances finies). Si:

$$\lim_{n\to+\infty} \frac{\sum_{k=1}^{n} \sigma_k^2}{n^2} = 0,$$

alors,

$$\forall\ \ \varepsilon > 0, \quad \lim_{n\to+\infty} P(|\frac{X_1 + \cdots + X_n}{n} - \frac{\mu_1 + \cdots + \mu_n}{n}| > \varepsilon) = 0$$

i.e.

$$\frac{X_1 + \cdots + X_n}{n} - \frac{\mu_1 + \cdots + \mu_n}{n} \xrightarrow{P} 0 \ \text{ lorsque } \ n \to +\infty.$$

Démonstration

L'inégalité de Chebyshev et nos hypothèses impliquent:

$$P(|\frac{X_1 + \cdots + X_n}{n} - \frac{\mu_1 + \cdots + \mu_n}{n}| > \varepsilon) \leq \frac{\sum_{k=1}^{n} \sigma_k^2}{n^2\varepsilon^2}$$

et la conclusion suit immédiatement.

Inspiré par le théorème 5, nous élargissons quelque peu la notion de loi faible des grands nombres.

Définition

Soit $(X_n)_{n\geq 1}$ une suite de v.a. et $S_n = \sum_{k=1}^{n} X_k$- Nous dirons que la suite $(X_n)_{n\geq 1}$ satisfait une loi faible des grands nombres s'il existe une suite de nombres $(\mu_n)_{n\geq 1}$ telle que:

$$\frac{S_n}{n} - \mu_n \xrightarrow{P} 0 \text{ si } n \to +\infty,$$

Condition nécessaire et suffisante, dans le cas i.i.d., pour avoir une loi faible des grands nombres

Résultat remarquable, le théorème 6 fournit une condition nécessaire et suffisante garantissant la validité d'une loi faible des grands nombres pour le cas i.i.d. [5, 11].

Théorème 6 (sans démonstration)

Soit $(X_n)_{n\geq 1}$ une suite de v.a. i.i.d. comme X et $S_n = \sum_{k=1}^{n} X_k$. Les deux assertions suivantes sont équivalentes:

(a) $(X_n)_{n\geq 1}$ satisfait une loi faible des grands nombres,

(b) $tP(|X| > t) \to 0$ si $t \to +\infty$,

Si tel est le cas, on peut toujours choisir $\mu_n = E(XI_{\{|X|\leq n\}})$.

Théorème de la convergence dominée de Lebesgue (sans démonstration)

Si la suite de v.a. $(X_n)_{n\geq 1}$ converge (p.s.) vers X lorsque $n \to +\infty$ et s'il existe une v.a. Y non-negative et intégrable telle que, pour tout $n \geq 1, |X_n| \leq Y$, alors:

$$\lim_{n\to+\infty} E(X_n) = E(\lim_{n\to+\infty} X_n) = E(X).$$

Le théorème reste valable pour une suite indexée par le continu.

Lemme

Soit X une v.a. intégrable avec $E(X) = \mu$. Alors:

$$(\alpha) \quad tP(|X| > t) \to 0 \ \text{ si } \ t \to +\infty.$$

et

$$(\beta) \quad \mu_n = E(XI_{\{|X| \leq n\}}) \to \mu \ \text{ si } \ n \to +\infty.$$

Démonstration

(α) X intégrable entraîne que X est finie (p.s.) et donc $|X|I_{\{|X|>t\}}$ converge vers 0 (p.s.). Il suffit alors d'utiliser le théorème de la convergence dominée car $|X|I_{\{|X|>t\}} \leq |X|$ et $|X|$ est intégrable. Nous avons donc:

$$tP(|X| > t) \leq E(|X|I_{\{|X|>t\}}) \to 0 \ \text{ si } \ t \to +\infty.$$

(β) X intégrable entraîne que X est finie (p.s.) et donc $XI_{\{|X|\leq n\}}$ converge (p.s.) vers X si $n \to +\infty$. Il suffit d'appliquer à nouveau le théorème de la convergence dominée.

Remarque

Comme le montre le théorème suivant, la loi faible des grands nombres peut être obtenue sans contrainte sur les variances qui peuvent même être infinies. Nous verrons plus loin que l'hypothèse "X est intégrable" (i.e. $E(|X|) < +\infty$) entraîne déjà la "loi forte des grands nombres", c'est-à-dire la convergence (p.s.) qui est plus forte que la convergence en probabilité et donc plus difficile à démontrer.

Théorème 7

Soit $(X_n)_{n\geq 1}$ une suite de v.a. i.i.d. comme X et $S_n = \sum_{k=1}^{n} X_k$. Si X est intégrable et $E(X) = \mu$, alors:

$$\frac{S_n}{n} \xrightarrow{P} \mu \ \text{ lorsque } \ n \to +\infty.$$

Démonstration

Par le lemme ci-dessus, la condition $E(|X|) < +\infty$ avec la notation $E(X) = \mu$, garantit la validité du point (b) du théorème 6 avec $\mu_n = E(XI_{\{X\leq n\}}) \to \mu$. Il suffit ensuite d'écrire:

$$\frac{S_n}{n} - \mu = (\frac{S_n}{n} - \mu_n) + (\mu_n - \mu).$$

Chaque terme converge vers 0 en probabilité, le premier à cause du théorème 6 et le second car la convergence a lieu partout. Par le corollaire du théorème 3, nous savons que si deux suites de v.a. convergent en probabilité, alors leur somme converge en probabilité vers la somme des limites. Par conséquent: $\frac{S_n}{n} - \mu \xrightarrow{P} 0$ lorsque $n \to +\infty$ et donc:

$$\frac{S_n}{n} \xrightarrow{P} \mu \quad \text{lorsque} \quad n \to +\infty$$

Remarque

Il exsite toutefois des v.a.X avec $E(|X|) = +\infty$ telles que, pour $(X_n)_{n\geq 1}$ une suite de v.a. i.i.d. comme X et $S_n = \sum_{k=1}^{n} X_k$, il existe $\mu \in \mathbb{R}$ et:

$$\frac{S_n}{n} \xrightarrow{P} \mu \quad \text{lorsque} \quad n \to +\infty.$$

Lemme

Soit $Y \geq 0$ une v.a. définie sur l'ensemble Ω et $p > 0$. alors:

$$(*) \quad E(Y^p) = \int_0^{+\infty} pt^{p-1} P(Y > t) dt.$$

Démonstration

Il suffit d'utiliser le théorème de Tonelli qui autorise l'inversion de l'ordre des intégrales si les intégrants sont non-négatifs:

$$\int_0^{+\infty} pt^{p-1} P(Y > t) dt = \int_0^{+\infty} \int_\Omega pt^{p-1} I_{\{Y>t\}} dP dt$$

$$= \int_\Omega \int_0^{+\infty} pt^{p-1} I_{\{Y>t\}} dt dP = \int_\Omega \int_0^{Y} pt^{p-1} I_{\{Y>t\}} dt dP$$

$$= \int_\Omega Y^p dP = E(Y^p).$$

Corollaire

(a) Pour $p = 1$ et $Y \geq 0$, nous obtenons la formule très utile:

$$E(Y) = \int_0^{+\infty} P(Y > t) dt.$$

(b) Soit X une v.a.. Si $tP(|X| > t) \to 0$ pour $t \to +\infty$, alors:

$$\forall\ 0 < \varepsilon < 1, \quad E(|X|^{1-\varepsilon}) < +\infty.$$

Démonstration

(a) Choisir $p = 1$ dans $(*)$.

(b) Choisissons $p = 1 - \varepsilon$ dans $(*)$:

$$E(|X|^{1-\varepsilon}|) = \int_0^{+\infty} (1-\varepsilon)t^{-\varepsilon} P(|X| > t)dt = \int_0^{+\infty} \frac{1-\varepsilon}{t^{1+\varepsilon}} tP(|X| > t)dt.$$

L'hypothèse entraîne que l'intégrant est majoré par $\frac{1-\varepsilon.}{t^{1+\varepsilon}}$ dès que $tP(|X| > t) < 1$ et donc $\forall\ 0 < \varepsilon < 1, E(|X|^{1-\varepsilon}) < +\infty$.

L'exemple suivant confirme que le point (b) du corollaire est plus faible que l'intégrabilité de X?

Exemple

Soit $X \geq 0$ une v.a. telle $P(X > t) = \frac{2ln(2)}{(t+2)ln(t+2)}$ pour $t \geq 0$. Ainsi $P(X > 0) = 1$. Bien que, à l'évidence, $tP(X > t) = \frac{2ln(2)t}{(t+2)ln(t+2)} \to 0$ pour $t \to +\infty$, la v.a. X n'est pas intégrable.

Puisque X est non-négative:

$$E(X) = \int_0^{+\infty} P(X > t)dt = \int_0^{+\infty} \frac{2ln(2)}{(t+2)ln(t+2)}dt$$

$$= 2ln(2) \int_0^{+\infty} \frac{\frac{d(ln(t+2)}{dt}}{ln(t+2)}dt = 2ln(2) \int_0^{+\infty} \frac{d}{dt} ln(ln(t+2))dt$$

$$= 2ln(2) \lim_{T \to +\infty} \int_0^T \frac{d}{dt} ln(ln(t+2))dt$$

$$= 2ln(2)(\lim_{T \to +\infty} (ln(ln(T+2)) - ln(ln(2))) = +\infty.$$

Voici deux formulations de la loi forte des grands nombres:

Théorème 8: Loi forte des grands nombres (sans démonstration)

Soit $(X_n)_{n\geq 1}$ une suite de v.a. i.i.d. comme la v.a. X et $S_n = \sum_{k=1}^{n} X_k$. Alors:

1. Si $E(|X|) < +\infty$ alors $\frac{S_n}{n} \xrightarrow{p.s.} E(X)$ lorsque $n \to +\infty$.
2. Si $E(|X|) = +\infty$, alors $\overline{\lim}_{n\to+\infty} \frac{|S_n|}{n} \overset{p.s.}{=} +\infty$.

Le théorème suivant est une conséquence du précédent:

Théorème 9: Loi forte des grands nombres

Soit $(X_n)_{n\geq 1}$ une suite de v.a. i.i.d. comme la v.a. X et $S_n = \sum_{k=1}^{n} X_k$. Les deux assertions suivantes sont équivalentes:

1. Pour $n \to +\infty$, la suite $(\frac{S_n}{n})_{n\geq 1}$ converge (p.s.) dans $\mathbb{R}$.
2. $E(|X|) < +\infty$.

Dans ce cas $\frac{S_n}{n} \xrightarrow{p.s.} E(X)$ lorsque $n \to +\infty$.

Remarques

1. La propriété 1. dans le théorème 9 peut être remplacée par:

   "Pour $n \to +\infty$, la suite $(\frac{S_n}{n})_{n\geq 1}$ converge dans $\mathbb{R}$ sur un ensemble de probabilité positive".

2. A notre avis, la plus belle démonstration de la loi forte des grands nombres est celle qui utilise une martingale renversée [9].

3. Pour obtenir la loi forte des grands nombres, l'indépendance deux à deux des v.a. suffit (Etemadi dans [5]).

Conclusion

Selon le théorème 6(b), une suite de v.a. i.i.d. comme X satisfait une loi faible des grands nombres si et seulement si:

$$tP(|X| > t) \to 0 \text{ si } t \to +\infty.$$

Par la propriété (b) du corollaire, ceci entraîne:

$$\forall\ 0 < \varepsilon < 1, \quad E(|X|^{1-\varepsilon}) < +\infty.$$

Cette dernière condition laisse à penser que X est proche de l'intégrabilité. Et pourtant, ceci fait toute la différence entre loi faible et loi forte puisque, par le théorème 9, la validité de cette dernière équivaut à l'intégrabilité de X. Le dernier exemple ci-dessus montre que la loi faible peut être vraie sans que la loi forte le soit.

Lien avec la loi normale.

Rappelons tout d'abord qu'une v.a. X est dite normale de paramètres $\mu \in \mathbb{R}$ et $\sigma > 0$ si elle possède une densité de la forme:

$$x \in \mathbb{R} \mapsto \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

On peut démontrer que $E(X) = \mu$ et $Var(X) = \sigma^2$.

La loi normale est plus ou moins liée au travail de J. Bernoullii. Ce dernier a calculé des probabilités contenant des facteurs binômiaux à l'aide de mauvaises approximations. Un peu plus tard, de Moivre (1667-1754) utilise la formule de Stirling pour rendre plus précis les calculs de Bernoulli. De Moivre sera le premier à dégager la loi normale comme limite d'une loi binômiale, travail qui préfigure le théorème de la limite centrée [14].

Par ailleurs, la loi normale est aussi apparue dans la théorie des erreurs de Gauss (1777-1855) où il propose une magnifique caractérisation basée sur la moyenne empirique des observations. Il nous paraît vraisemblable que sa démarche se soit inspirée du théorème de Bernoulli qui justifie l'estimation d'une probabilité (en fait une espérance) par la moyenne empirique des valeurs observées.

Rappel

Soient $X_1, X_2, ..., X_n$ des v.a. de densités respectives $g_1(x), g_2(x), ..., g_n(x)$. Si ces v.a. sont indépendantes, alors le vecteur aléatoire:

$$\underline{X} = (X_1, X_2, ..., X_n)$$

possède aussi une densité donnée par:

$$g_{\underline{X}}(x_1, x_2, ..., x_n) = g_1(x_1) \cdot g_2(x_2) \cdots g_n(x_n).$$

Caractérisation de la loi normale par Gauss

Dans le cadre de sa théorie des erreurs, Gauss a considéré un modèle dans lequel l'erreur de mesure est additive. Soit $\theta \in \mathbb{R}$ une grandeur (inconnue) à mesurer et X la valeur, avec erreur, obtenue lors d'une mesure. Gauss propose le modèle suivant (déjà rencontré dans la "seconde réponse à la question"):

$$X = \theta + \mathfrak{E}$$

où $\mathfrak{E}$ est l'erreur commise lors de la mesure. Par hypothèse, $\mathfrak{E}$ est une v.a. avec densité f définie sur $\mathbb{R}$, strictement positive et continûment dérivable. La fonction de répartition de X est:

$$x \in \mathbb{R} \mapsto F_X(x) := P(X \leq x) = P(\theta + \mathfrak{E} \leq x) = P(\mathfrak{E} \leq x - \theta)$$

$$= \int_{-\infty}^{x-\theta} f(t)dt = \int_{-\infty}^{x} f(u-\theta)du.$$

La dernière égalité provient du changement de variable $t = u - \theta$. On obtient alors la densité de X en dérivant l'intégrale par rapport à x, ce qui nous donne $f(x-\theta)$. Rappelons que, n observations indépendantes de X forment un n-échantillon issu de X, c'est-à-dire un n-uple:

$$\underline{X} = (X_1, X_2, ..., X_n) \text{ avec des v.a. i.i.d. comme X.}$$

Sous ces hypothèses, le rappel précédent nous assure que $\underline{X}$ aura pour densité sur $\mathbb{R}^n$:

$$f_{\underline{X}}(x_1, x_2, ..., x_n; \theta) = f(x_1 - \theta) \cdot f(x_2 - \theta) \cdots f(x_n - \theta).$$

Postulatum de Gauss

Gauss cherche les densités $f(x)$ de $\mathfrak{E}$ définies sur $\mathbb{R}$ pour lesquelles la densité du n-échantillon isssu de X, à savoir:

$$f(x_1 - \theta) \cdot f(x_2 - \theta) \cdots f(x_n - \theta)$$

admet son maximum en $\theta = \overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$, quel que soit $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$.

Il s'agit d'une démarche en quelque sorte inverse de celle du maximum de vraisemblance. En effet, dans cette dernière on se donne la densité $g(x;\theta)$ de X, fonction connue des variables x et θ. La densité $g_{\underline{X}}$ sur $\mathbb{R}^n$ du n-échantillon issu de X est donc:

$$g_{\underline{X}}(x_1, x_2, ..., x_n; \theta) = g(x_1; \theta) \cdot g(x_2; \theta) \cdots g(x_n; \theta)$$

et on cherche, en fonction de $x_1, x_2, ..., x_n$, la valeur $\theta(x_1, x_2, ..., x_n)$ qui maximise la densité $g_{\underline{X}}(x_1, x_2, ..., x_n; \theta)$. Chez Gauss on cherche la densité en imposant l'estimateur et dans le maximum de vraisemblance, on cherche l'estimateur en imposant la densité.

Le nombre $\theta = \overline{x}$ étant nécessairement à l'intérieur de $\mathbb{R}$, la dérivée de la densité:

$$f(x_1 - \theta) \cdot f(x_2 - \theta) \cdots f(x_n - \theta)$$

par rapport à θ doit s'annuler pour $\theta = \overline{x}$. Comme la dérivée d'une somme est plus conviviale que celle d'un produit, nous prenons d'abord le logarithme. Nous noterons $'$ la dérivée. Remarquons que la dérivée logarithmique de f, $\phi(x) = \frac{f'(x)}{f(x)}$, est définie partout et continue car, par hypothèse, $f(x) > 0$ pour tout x et f est continûment dérivable. De plus, les deux fonctions f' et ϕ s'annulent simultanément. Nous obtenons donc la condition:

$$\forall \underline{x} = (x_1, x_2, ..., x_n) \in \mathbb{R}^n,\ \frac{f'(x_1 - \overline{x})}{f(x_1 - \overline{x})} + \frac{f'(x_2 - \overline{x})}{f(x_2 - \overline{x})} + \cdots + \frac{f'(x_n - \overline{x})}{f(x_n - \overline{x})} = 0.$$

Nous cherchons donc toutes les fonctions $\phi(x)$ continues vérifiant:

$$(\mathbf{C})\ \ \forall \underline{x} = (x_1, x_2, ..., x_n) \in \mathbb{R}^n,\ \ \sum_{i=1}^{n} \phi(x_i - \overline{x}) = 0.$$

En notant $u_i = x_i - \overline{x}$, nous constatons que:

$$\sum_{i=1}^{n} u_i = \sum_{i=1}^{n} (x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0.$$

Introduisons les ensembles:

$$U := \{\underline{u} = (u_1, u_2, .., u_n) \in \mathbb{R}^n;\ \sum_{i=1}^{n} u_i = 0\},$$

$$V := \{(x_1 - \overline{x}, x_2 - \overline{x}, ..., x_n - \overline{x})\ ;\ (x_1, x_2, ..., x_n) \in \mathbb{R}^n\}$$

A l'évidence $U \subset V$ et $V \subset U$ et donc $U = V$. La condition $(\mathbf{C})$ entraîne donc:

$$\forall\ \ \underline{u} = (u_1, u_2, ..., u_n) \in U\ \ (\text{i.e.}\ \ \sum_{i=1}^{n} u_i = 0),\ \text{on}\ \ \text{a}\ \ \sum_{i=1}^{n} \phi(u_i) = 0.$$

Considérons d'abord le cas $n = 2$:

$$\forall\ (u_1, u_2) \in \mathbb{R}^2, u_1 + u_2 = 0,\ \phi(u_1) + \phi(u_2) = 0.$$

En posant $u = u_1$, alors $u_2 = -u$ et donc $\forall u \in \mathbb{R}, \phi(u) + \phi(-u) = 0$, ce qui équivaut à $\phi(-u) = -\phi(u)$. La fonction ϕ est donc impaire.

Considérons ensuite le cas $n = 3$:

$$\forall (u_1, u_2, u_3) \in \mathbb{R}^3, u_1 + u_2 + u_3 = 0,\ \phi(u_1) + \phi(u_2) + \phi(u_3) = 0.$$

Or, ϕ étant impaire et $-u_3 = u_1 + u_2$:

$$\phi(u_1) + \phi(u_2) = -\phi(u_3) = \phi(-u_3) = \phi(u_1 + u_2).$$

A condition que $u_3 = -(u_1 + u_2)$, nous pouvons bien sûr choisir librement u_1 et u_2. Par conséquent:

$$(E) \qquad \forall u_1, u_2 \in \mathbb{R},\ \phi(u_1 + u_2) = \phi(u_1) + \phi(u_2).$$

L'équation (E) est une équation dite fonctionnelle et ses solutions sont des fonctions dénommées "fonctions additives". Nous verrons que toute solution continue de (E) est de la forme $\phi(x) = \phi(1)x$. La question de l'existence de fonctions additives discontinues est restée longtemps ouverte. Les analystes ne parvenaient ni à prouver leur continuité, ni à donner un exemple de fonction additive discontinue. Hamel fut le premier à démontrer l'existence de solutions discontinues en ayant recours aux bases de Hamel. Dans [13], le théorème 2 de la page 277 nous apprend que le graphe d'une fonction additive discontinue de $\mathbb{R}$ dans $\mathbb{R}$ est **dense dans $\mathbb{R}^2$** !

Nous avons $\phi(0) = \phi(0 + 0) = \phi(0) + \phi(0) = 2\phi(0)$ et donc $\phi(0) = 0$.

Par induction finie, nous obtenons:

$$\forall u_1, u_2, ..., u_n \in \mathbb{R},\quad \phi(\sum_{i=1}^{n} u_i) = \sum_{i=1}^{n} \phi(u_i).$$

Par conséquent:

$$\forall\ n \in \mathbb{N}^*,\ \phi(1) = \phi(\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}) = n\phi(\frac{1}{n})$$

et donc:

$$\phi(\frac{1}{n}) = \frac{1}{n}\phi(1).$$

Si q est un rationnel positif, alors il existe $m, n \in \mathbb{N}^*$ tel que $q = \frac{m}{n}$. Ainsi:

$$\phi(q) = \phi(\frac{m}{n}) = m\phi(\frac{1}{n}) = \frac{m}{n}\phi(1) = \phi(1)q.$$

Si x est un réel positif, il existe une suite de rationnels positifs $(q_n)_{n\geq 1}$ telle que $\lim\limits_{n\to+\infty} q_n = x$. La continuité de ϕ nous fournit:

$$\phi(x) = \phi(\lim_{n\to+\infty} q_n) = \lim_{n\to+\infty} \phi(q_n) = \lim_{n\to+\infty} (\phi(1)q_n) = \phi(1) \lim_{n\to+\infty} q_n = \phi(1)x.$$

$\phi(x) = \phi(1)x$ étant impaire, la même forme reste valable sur tout $\mathbb{R}$:

$$x > 0, \quad \phi(-x) = -\phi(x) = -\phi(1)x = \phi(1)(-x)$$

et ainsi, puisque $\phi(0) = 0$:

$$\forall x \in \mathbb{R}, \quad \phi(x) = \phi(1)x.$$

Par intégration de l'équation $\frac{f'(x)}{f(x)} = \phi(1)x$, nous obtenons $\ln(f(x) = \phi(1)\frac{x^2}{2} + C$ (C étant une constante) d'où $f(x) = e^C e^{\frac{\phi(1)x^2}{2}}$ avec $e^C > 0$. Si $\phi(1) \geq 0$, alors $\int_{-\infty}^{+\infty} f(x)dx = +\infty$ au lieu de 1. Par conséquent $\phi(1)$ doit être négatif et donc de la forme $\phi(1) = -a^2 < 0$. Finalement, en notant $K = e^C > 0$:

$$f(x) = Ke^{-\frac{a^2x^2}{2}} \quad \text{et} \quad \text{comme} \quad \int_{-\infty}^{+\infty} Ke^{-\frac{a^2x^2}{2}} dx = K\frac{\sqrt{2\pi}}{a} = 1.$$

Ainsi $a > 0$ et:

$$f(x) = \frac{a}{\sqrt{2\pi}} e^{-\frac{a^2x^2}{2}}.$$

En posant $\sigma = \frac{1}{a}$:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}.$$

La v.a. $\mathfrak{E}$ est donc normale d'espérance nulle et de variance $\frac{1}{a^2}$. Ainsi $E(\mathfrak{E}) = 0$ est une conséquence des hypothèses.

Une conséquence surprenante de la loi faible: le théorème d'approximation de Weierstrass.

Sergueï Natanovitch Bernstein (1880-1968), un mathématicien ukrainien, a fait la découverte étonnante suivante (1912/13): en jouant à pile ou face avec une pièce qui peut être biaisée, la preuve du théorème 4 permet de démontrer facilement l'un des grands théorèmes de l'analyse, à savoir l'approximation uniforme, sur un intervalle compact, d'une fonction continue par une suite de polynômes. Cette approche fournit même explicitement les polynômes. La vitesse de convergence en norme uniforme est toutefois plus mauvaise que celle obtenue à l'aide de polynômes tels que ceux de Lagrange, d'Hermite ou de Legendre.

Théorème de Weierstrass

Toute fonction réelle continue sur un intervalle compact $[a, b] \subset \mathbb{R}$ est limite uniforme d'une suite de polynômes.

Remarque

Il est légitime de se demander pourquoi l'énoncé se limite aux intervalles compacts. Considérons donc une suite de polynômes $(P_n)_{n\geq 1}$ qui converge uniformément sur un intervalle non-borné I. La convergence uniforme entraîne que $(P_n)_{n\geq 1}$ est une suite de Cauchy uniforme et il existe donc $n_0 \geq 1$ tel que pour tout $n > n_0$ et tout $x \in I$:

$$|P_n(x) - P_{n_0}(x)| < 1.$$

Or $P_n - P_{n_0}$ est un polynôme qui est borné sur I non-borné. Ceci est possible seulement si ce polynôme est constant. Ainsi, pour tout $n > n_0$ et tout $x \in I$, $P_n(x) - P_{n_0}(x) = C = P_n(0) - P_{n_0}(0)$. Si f désigne la limite de $(P_n)_{n\geq 1}$ lorsque $n \to +\infty$, alors:

$$\forall \ \ x \in I, f(x) = P_{n_0}(x) + f(0) - P_{n_0}(0)$$

et ainsi f est encore un polynôme. Ce résultat est donc inintéressant.

Notons que pour $n > n_0, P_n(x) = P_n(0) - P_{n_0}(0) + P_{n_0}(x)$. Donc asymptotiquement les polynômes de la suite diffèrent seulement par une constante additive. Cette configuration est donc peu intéressante.

Démonstration probabiliste du théorème de Weierstrass

Nous démontrons d'abord le théorème sur l'intervalle $[0,1]$ et considérons une suite $(X_n)_{n\geq 1}$ de v.a. i.i.d. de Bernoulli de paramètre p. Ainsi $S_n = \sum_{i=1}^{n} X_i$ est une v.a. binômiale de paramètres (n,p) et:

$$P(\frac{S_n}{n} = \frac{k}{n}) = P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, 0 \leq k \leq n.$$

Soit f une fonction réelle continue définie sur $[0,1]$. Un théorème classique nous permet d'écrire l'espérance comme:

$$E(f(\frac{S_n}{n})) = \sum_{k=0}^{n} f(\frac{k}{n}) \binom{n}{k} p^k (1-p)^{n-k} =: B_n(p).$$

Lêxtension aux cas $p = 0$ et $p = 1$ se fait de façon continue avec $B_n(0) = f(0)$ et $B_n(1) = f(1)$. La fonction $B_n(p)$ est un polynôme de degré au plus n appelé polynôme de Bernstein. L'idée de l'approche repose sur la loi faible des grands nombres, qui, dans cette situation nous garantit que:

$$\frac{S_n}{n} \xrightarrow{P} p \text{ lorsque } n \to +\infty.$$

Il faut donc s'attendre à ce que les valeurs de $\frac{S_n}{n}$ aient tendance à se concentrer sur les $\frac{k}{n}$ proches de p.

La fonction f étant continue sur un compact, elle est bornée, i.e. $\exists \ 0 \leq M < +\infty$ tel que $|f| \leq M$ sur $[0,1]$. La fonction f est aussi uniformément continue, c'est-à-dire:

$$\forall\, \varepsilon > 0, \exists\, \delta = \delta(\varepsilon) > 0 \text{ tel que } |f(x) - f(y)| < \varepsilon \text{ si } |x-y| \leq \delta,\ x, y \in [0,1].$$

Ainsi pour tout $p \in [0,1]$, en utilisant l'égalité $\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$:

$$|f(p) - B_n(p)| = | \sum_{k=0}^{n} (f(p) - f(\frac{k}{n})) \binom{n}{k} p^k (1-p)^{n-k} |$$

$$\leq \sum_{|\frac{k}{n} - p| \leq \delta} |(f(p) - f(\frac{k}{n})| \binom{n}{k} p^k (1-p)^{n-k} + \sum_{|\frac{k}{n} - p| > \delta} \text{idem.}$$

La première somme est majorée par ε à cause de la continuité uniforme et de la densité de probabilité. La seconde est majorée par:

$$2M \sum_{|\frac{k}{n}-p|>\delta} \binom{n}{k} p^k (1-p)^{n-k} = 2MP(|\ \frac{S_n}{n} - p\ |> \delta).$$

Pour $p, q \in [0,1], p+q=1$ on a $pq \leq \frac{1}{4}, E(\frac{S_n}{n}) = p$ et $Var(\frac{S_n}{n}) = \frac{pq}{n}$, l'inégalité de Chebyshev fournit:

$$P(|\frac{S_n}{n} - p| > \delta) \leq \frac{Var(\frac{S_n}{n})}{\delta^2} = \frac{pq}{n\delta^2} \leq \frac{1}{4n\delta^2}.$$

Ainsi:

$$2MP(|\ \frac{S_n}{n} - p\ |> \delta) \leq \frac{M}{2n\delta^2},$$

et finalement:

$$|f(p) - B_n(p)| \leq \varepsilon + \frac{M}{2n\delta^2}.$$

Cette dernière grandeur sera inférieure à 2ε, uniformément en p, pour n suffisamment grand.

L'extension à un intervalle quelconque $t \in [a,b] \mapsto f(t)$ peut se faire par changement de variable. En posant $t = a + p(b-a) = t(p)$ on obtient, par composition, une fonction continue sur $[0,1]$:

$$p \in [0,1] \mapsto f(t(p)) = f(a + p(b-a)).$$

Pour $\varepsilon > 0$, il existe donc un polynôme $h(p)$ tel que $\max_{p\in[0,1]} |f(t(p)) - h(p)| < \varepsilon$. En remplaçant p par $p(t) = \frac{t-a}{b-a}$, nous obtenons $f(t(p(t))) = f(t)$ et $h(p(t)) = h(\frac{t-a}{b-a})$ qui est également un polynôme comme fonction de t et vérifie de plus la condition souhaitée:

$$\max_{t\in[a,b]} |f(t) - h(\frac{t-a}{b-a})| < \varepsilon.$$

Nous terminerons avec cette citation de Ekström [3]:

If a probability cannot be estimated through Huygen's formula, the expected value operator, then Bernoulli proposes a method of empirical estimation through repeated experiments and the use of the arithmetic mean. This

unassuming proposal constitutes one of the most important contributions to the evolution of modern science. It is in arguing the merits of this proposed method that Bernoulli states and proves his famous limit theorem, which Poisson (1837) termed the law of large numbers.

Petite bibliographie

1. J. Bernoulli, Wahrscheinlichkeitsrechnung (Ars conjectandi), dritter und vierter Theil, Ostwald's Klassiker der exakten Wissenschaften Nr. 108, Wilhelm Engelmann, 1899.

2. M. Henry, La démonstration par Jacques Bernoulli de son théorème, IREM de Besançon, Université de Franche-Comté, Google.

3. J. Ekström, Jakob Bernoulli's theory of inference, Google.

4. P. Picard, Hasard et probabilités, Histoire, théorie et application des probabilités, Vuibert, 2007.

5. R. Durrett, Probability: Theory and Examples, Wadsworth & Brooks/Cole, 1991.

6. N. Meusnier, Argumentation et démonstration: à quoi sert la démonstration de la "Loi des grands nombres" de Jacques Bernoulli (1654-1705), La démonstration mathématique dans l'histoire, pp 81-97, Actes du 7ième colloque inter-IREM Epistémologie et Histoire des Mathématiques, Besançon, 1989.

7. G. Shafer, The Significance of Jacob Bernoulli's *Ars Conjectandi* for the Philosophy of probability Today, Google

8. Jakob Bernoulli, On the law of large numbers, Translated into english by Oscar Sheynin, Berlin, 2005, Google

9. H. Bauer, Wahrscheinlichkeitsrechnung, de Gruyter, 2002.

10. L. Breiman, Statistics: With a View Toward Applications, Houghton Mifflin, 1973.

11. W. Feller, An introduction to Probability Theory an its Applications, Vol II, Wiley (1971).

12. H. Bauer, Wahrscheinlichkeitstheorie und Grundzüge der Masstheoriem de Gruyter (1978).

13. M. Kuczma, An Introduction to the Theory of Functional Differential Equations and Inequalities, Panstwowe wydawnictwo Naukowe (1985).

14. Histoire de la loi normale, Google.

15. M. Woodroofe, Probability with Applications, McGraw-Hill (1975).