Mathématiques et sciences humaines
Mathematics and social sciences

190 | 2010
190, Mathématiques discrètes : théories et usages. Numéro en hommage à Bruno Leclerc

Des distributions de probabilité singulières

Strange probability distribution

Marc Barbut

Édition électronique
URL : http://msh.revues.org/11702
DOI : 10.4000/msh.11702
ISSN : 1950-6821

Éditeur
Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée
Date de publication : 10 mars 2010
Pagination : 11-18
ISSN : 0987-6936

Référence électronique
Marc Barbut, « Des distributions de probabilité singulières », *Mathématiques et sciences humaines* [En ligne], 190 | Eté 2010, mis en ligne le 16 octobre 2010, consulté le 30 septembre 2016. URL : http://msh.revues.org/11702 ; DOI : 10.4000/msh.11702

Ce document est un fac-similé de l'édition imprimée.

© École des hautes études en sciences sociales

Math. Sci. hum / Mathematics and Social Sciences (48e année, n° 190, *2010*(2), p. 11-18)

DES DISTRIBUTIONS DE PROBABILITÉ SINGULIÈRES

Marc BARBUT[1]

RÉSUMÉ – *Ce texte n'a rien d'original. Son objectif est seulement pédagogique. On montre comment peuvent se construire des fonctions de répartition continues non dérivables et des fonctions de sauts de support partout dense sur l'intervalle de définition.*

MOTS-CLÉS – Cantor, Continu, Dense, Discret, Fonction de répartition, Lévy, Saut

SUMMARY – Strange probability distributions

This paper is in no way original. Its aim is purely pedagogical. We show how to construct continuous cumulative distribution functions without density function and discontinuous ones with a set of discontinuities that are dense over their interval of definition.

KEYWORDS – Cantor, Continuous, Cumulative distribution function, Dense, Discrete, Lévy, Saltus

Le premier livre publié par Bruno Leclerc est *Distributions statistiques et lois de probabilité*, paru en 1972 aux éditions Gauthier-Villars et Mouton, dans la collection « Mathématiques et Sciences de l'Homme », fasc. XV, série « Cahiers mathématiques » (fasc. IV).

Dans cet ouvrage tout à fait remarquable, ne serait-ce que du point de vue pédagogique, B. Leclerc présente une bonne quinzaine de distributions de probabilité (la Normale, bien sûr, mais aussi celles de Fisher, Pareto, Halphen, Polya, lois gamma, binomiale négative, etc…), dont certaines peu usuelles.

Et pour chacune, il fournit plusieurs exemples d'applications statistiques, avec références précises aux articles originaux.

Ainsi, ce livre a été un outil irremplaçable pour de nombreux chercheurs ou ingénieurs lorsqu'ils rencontraient des problèmes statistiques nouveaux pour lesquels des méthodes originales de résolution devaient être trouvées.

Ici, mon objectif est de souligner que la définition toute simple de la fonction de répartition d'une variable aléatoire réelle :

Fonction F à valeurs réelles, monotone non décroissante, continue à droite et telle que :

$$F(-\infty) = 0 \text{ et } F(+\infty) = 1$$

[1] Centre d'Analyse et de Mathématique Sociales, EHESS, 54 bd Raspail 75270 Paris cedex 06, mbarbut@ehess.fr

que cette définition, donc, est compatible avec des cas singuliers de distributions – de peu d'intérêt, il est vrai, pour les applications à la statistique mais de grand intérêt pour comprendre l'extrême diversité des lois de probabilité possibles.

LA DÉCOMPOSITION DE PAUL LÉVY

Au début des « années 1920 », Paul Lévy a montré que toute fonction de répartition F était la somme de trois termes, soit :

(1) $\quad F = F_{ac} + F_s + F_c$

Dans cette expression, F_{ac} est proportionnelle à une fonction de répartition « absolument continue », c'est-à-dire qu'elle est de la forme $\int_{-\infty}^{x} f(t)\,dt$, où f est positive et intégrable ; f_x est appelée sa *densité*. La plupart des distributions utilisées en statistique sont de ce type.

Le terme F_s de (1) est une fonction de sauts, telles que les classiques loi de Poisson ou loi de Pascal.

Ces fonctions de saut sont discontinues et le nombre des sauts est fini ou dénombrable: en effet, ils sont positifs et leur somme vaut au plus 1; donc, quel que soit l'entier naturel n, il n'y en a qu'un nombre fini d'amplitude supérieure ou égale à $1/n$.

Dans les deux exemples évoqués ci-dessus (Poisson et Pascal) le support de la distribution est discret ; c'est même l'ensemble des entiers positifs ou nul. Nous verrons que les choses peuvent ne pas être aussi simples.

Quant au troisième terme F_c de l'expression (1), c'est une fonction de répartition *continue* partout *mais non dérivable* en une infinité de ses points.

DISTRIBUTION UNIFORME ET ÉCRITURE DES NOMBRES

Partons de la banale distribution uniforme sur l'intervalle [0,1] (cf. Figure 1). Pour tout x de cet intervalle, on a :

(2) $\quad F(x) = x$

Pour ceux des x qui sont décimaux, on a ainsi, par exemple :

$$F(0{,}5) = \frac{1}{2} \qquad F(0{,}673) = 67{,}3\ \%, \text{ etc…}$$

L'ordre des décimaux est l'ordre lexicographique (i.e., alphabétique), l'alphabet étant constitué des 10 chiffres, par exemple :

$$0{,}673 < 0{,}6731 < 0{,}6735$$

C'est un ordre *dense en soi* : entre deux décimaux on peut toujours en intercaler un autre, et donc une infinité d'autres.

Pour les non-décimaux, l'écriture décimale se ferait avec une infinité de chiffres après la virgule.

Revenons aux décimaux ; il y a une ambiguïté. Chacun peut s'écrire de deux façons dont l'une avec une infinité de chiffres. Par exemple :

$$0{,}673 = 0{,}\,672999\ldots$$

C'est cette ambiguïté qu'on va essayer de lever.

Pour simplifier l'exposé, nous supposons désormais que les nombres sont exprimés en écriture *binaire*, les deux chiffres étant 0 et 1, avec l'ordre « alphabétique » :

$$0 < 1$$

Notre problème est de faire correspondre à chaque *écriture* d'un nombre au moyen d'une suite de chiffres 0 et 1, un *point x* et un seul du segment de droite unitaire [0,1] (noté AB sur les figures).

– L'*écriture* peut être *figurée* par une arborescence. À chaque bifurcation, le chiffre 0 envoie à gauche et le chiffre 1 à droite.

  Ainsi, la Figure 1 représente les 8 nombres binaires à 3 chiffres, engendrés par ce procédé et rangés dans l'ordre lexicographique, c'est-à-dire alphabétique, comme pour les mots d'un dictionnaire ou d'un annuaire.

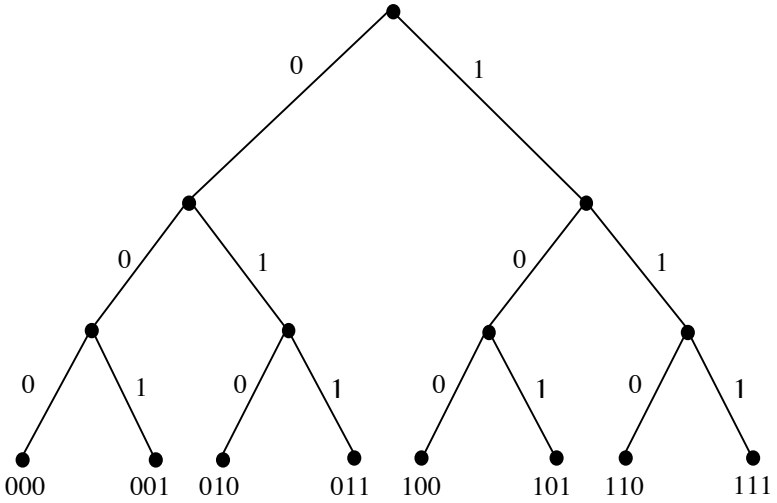


FIGURE 1

– La *position* sur le segment unitaire du point correspondant à chaque écriture va être déterminée par une suite de partitions de ce segment emboîtées les unes dans les autres, selon la correspondance :

  Chiffre 0 → segment de gauche
  Chiffre 1 → segment de droite.

La Figure 2 représente ainsi la partition en 4 sous-segments correspondant aux deux premiers chiffres : à un nombre commençant par 00 correspond un point du

segment AA_1 ; de même, si le nombre commence par 10, le point appartient au segment MB_1 ; etc…

e façon à avoir une *partition* du segment unitaire, il faut préciser à quel intervalle correspondent les points de division. Ici, tous nos intervalles seront fermés à gauche et ouverts à droite.

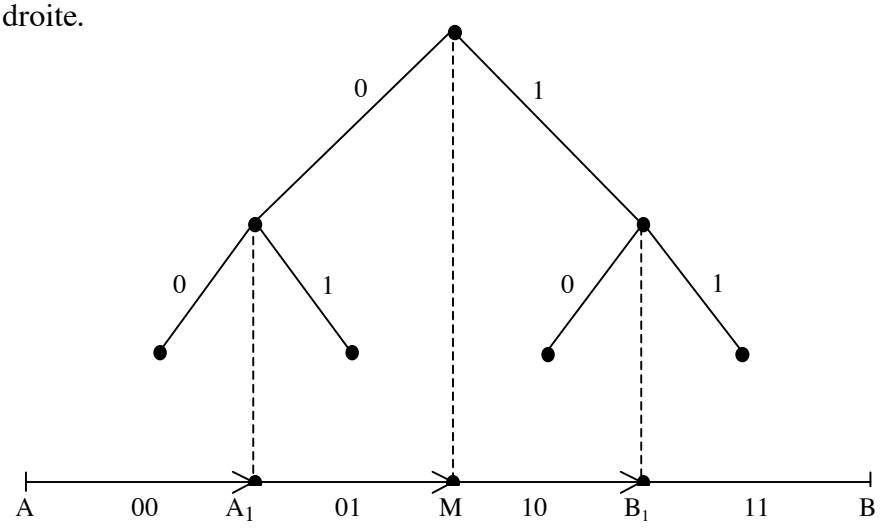


FIGURE 2

Il est clair qu'à chaque nouveau chiffre de l'écriture correspond un intervalle emboîté dans le précédent et que l'intersection de cette suite d'intervalles emboîtés, prolongée indéfiniment, converge vers un point *x* et un seul du segment unitaire.

Dans le cas des points de division, qui représentent les nombres *binaires* (i.e. ceux qui peuvent s'écrire avec un nombre fini de chiffres), il y a deux écritures possibles. Par exemple, pour le point B_1, on écrira soit : « 11 » (sous-entendu « 11000… »), soit « 10111… » (avec ici, explicitement, une suite infinie de 1 après les deux premiers chiffres « 10 »), comme illustré par la Figure 3.

Ainsi, pour ces points, il y a ambiguïté : à chacun correspondent deux écritures, et non une seule, comme pour les points images de nombres non-décimaux.

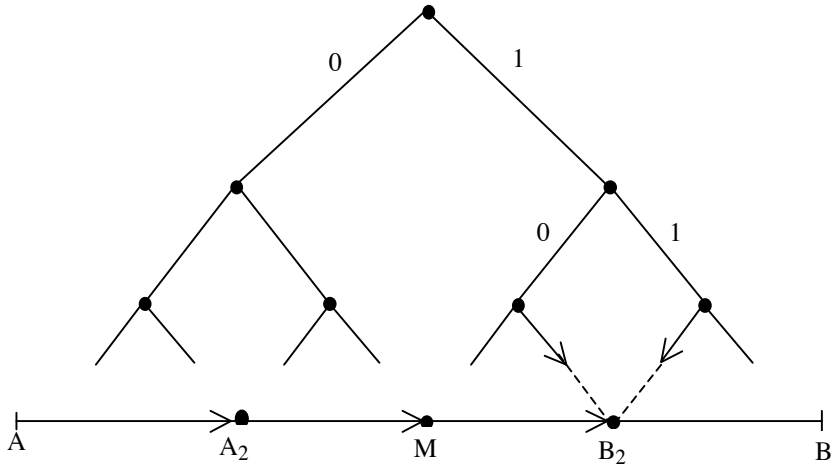


FIGURE 3

LA CONSTRUCTION DE CANTOR

Georg Cantor a imaginé une façon simple de lever cette ambiguïté : à chaque étape du processus dichotomique, enlever du segment unitaire un intervalle *ouvert* dont chacune des extrémités correspond à l'une des deux écritures. Par exemple (Figure 4), à l'écriture « 10111… » correspond l'extrémité gauches B1 de l'intervalle ouvert $B_1B'_1$ et à l'écriture « 11000… » l'extrémité droite B'_1.

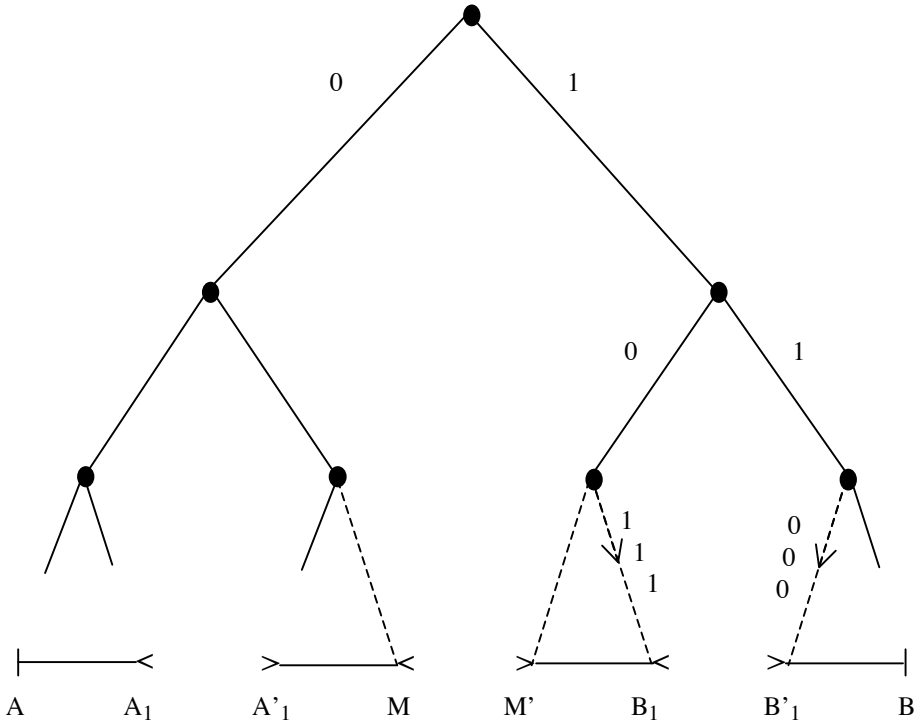


FIGURE 4

L'ensemble *fermé* des points subsistant après une infinité d'itérations, souvent appelé *ensemble ternaire* (également dénommé *triadique*) de *Cantor*, est en correspondance bi-univoque avec les écritures. Il a donc le même cardinal $\aleph_1$ (on dit encore « la puissance du continu ») que le segment unitaire AB de départ.

LA FONCTION DE RÉPARTITION ASSOCIÉE

Comment le procédé de Cantor modifie-t-il la distribution uniforme sur le segment [0,1] ?

Supposons pour simplifier qu'à chaque itération la $k^{\text{ième}}$ par exemple, chacun des 2^k *nouveaux « trous »* soit centré au milieu de l'intervalle troué et mesure le tiers de sa longueur.

Ainsi, à la première étape, le « trou » MM' (Figure 5(a)), qui est un intervalle ouvert, est centré au point d'abscisse 1/2 et a pour longueur 1/3.

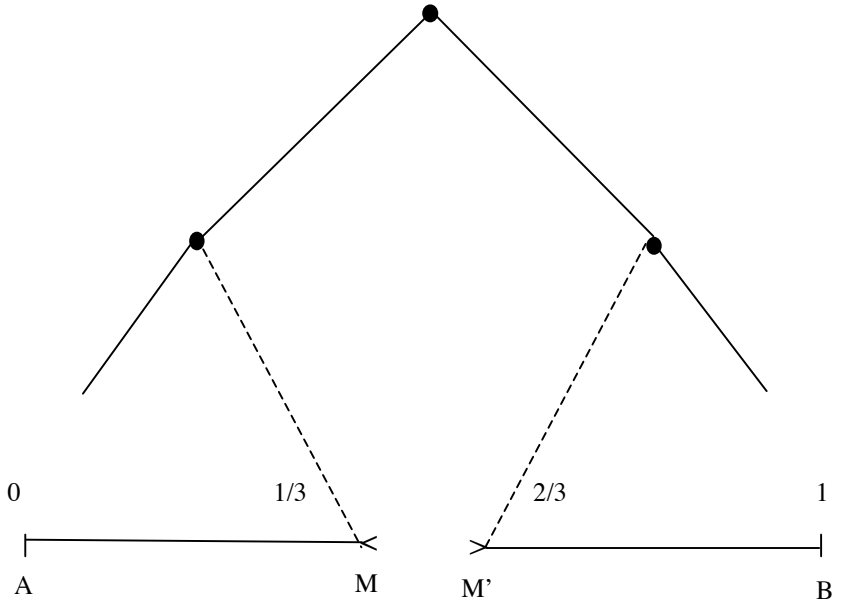


FIGURE 5(a)

Or, il y a (distribution uniforme) une même probabilité 1/2 pour qu'un point se trouve soit à gauche, soit à droite du milieu de l'intervalle unitaire [0,1].

Donc la fonction de répartition correspondante a un « plateau » d'ordonnée 1/2 et de longueur 1/3 sur l'intervalle MM'. On devra ensuite compléter la définition de la fonction de répartition F_1 à cette première étape par interpolation linéaire (distribution uniforme) sur chacun des deux intervalles AM et M'B (Figure 5(b)).

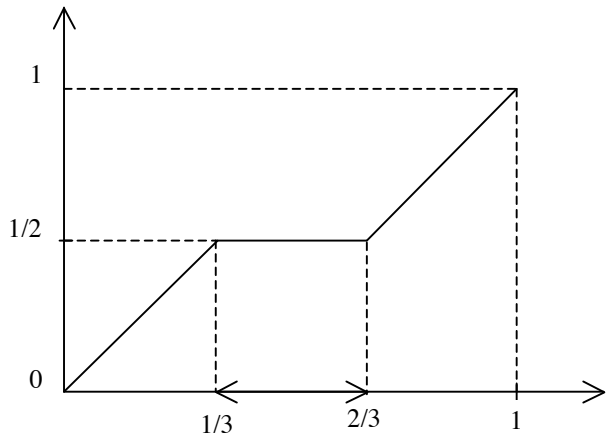


FIGURE 5(b). La fonction de répartition F_1

De même, à la deuxième itération, chacun des deux intervalles AM et M'B est amputé d'un « trou » de longueur 1/9, centrés aux points d'abscisse 1/6 pour le premier $A_1A'_1$, 5/6 pour le second $B_1B'_1$ (cf. Figure 4). À chacun de ces deux nouveaux trous

correspondent des « plateaux » d'ordonnées 1/4 et 3/4 respectivement de la fonction de répartition F_2, complétée par interpolation linéaire (Figure 6).

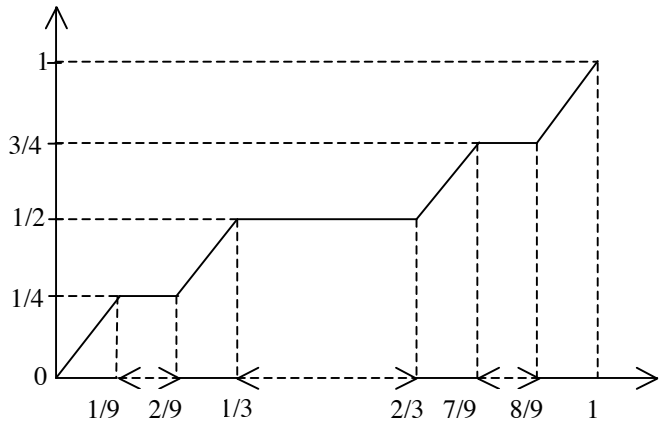


FIGURE 6. La fonction de répartition F_2

On remarquera d'ailleurs que l'on a *uniformément* l'inégalité :

$$(1) \qquad |F_2 - F_1| \leq \frac{1}{4}$$

D'une façon générale, à la *k*ième itération, on aura pour la fonction de répartition F_k, 2^k nouveaux « plateaux », chacun de longueur $\frac{1}{3^k}$ et d'ordonnées $\frac{2n-1}{2^k}$, avec $n = 1,2,3,\ldots, 2^{k-1}$. Par construction, on a uniformément :

$$(2) \qquad |F_k - F_{k-1}| \leq \frac{1}{2^k}$$

d'où, pour tout couple d'entiers k et h :

$$|F_{k+h} - F_k| < \sum_{i=1}^{h} |F_{k+i} - F_{k+i-1}| \leq \sum_{i=1}^{h} \frac{1}{2^{k+i}} = \frac{1}{2^k} \sum_{i=1}^{h} \frac{1}{2^i} < \frac{1}{2^k} \sum_{i=1}^{\infty} \frac{1}{2^i}$$

Comme $\sum_{i=1}^{\infty} \frac{1}{2^i} = 1$, nous avons finalement, quels que soient les entiers positifs k et h :

$$(3) \qquad |F_{k+h} - F_k| < \frac{1}{2^k}$$

De (3), il résulte que la suite des fonctions de répartition F_m converge uniformément vers une fonction de répartition limite F ; comme chacune des F_m est *continue*, leur limite uniforme l'est également.

Nous avons ainsi l'exemple d'une fonction de répartition continue, presque partout dérivable de dérivée nulle, mais non dérivable sur son support (qui est de mesure linéaire nulle) : la pente des segments de droite d'interpolation tend évidemment vers l'infini lors des itérations successives.

N.B. On peut généraliser. À chaque itération, les 2^k nouveaux « trous » peuvent être de longueur arbitraire et centrés où l'on veut, pourvu qu'ils soient tous ouverts et strictement intérieurs à l'intervalle duquel ils sont ôtés.

De même, les ordonnées des nouveaux « plateaux » correspondants de la fonction de répartition F_k peuvent être à des ordonnées arbitraires mais distinctes de celles des plateaux de F_{k-1} et intercalées entre celles-ci. Dans ce dernier cas, la fonction de répartition limite continue F ne procède plus de la distribution uniforme, comme dans la construction qui a été détaillée *supra*.

DE SINGULIÈRES FONCTIONS DE SAUTS

Si maintenant nous inversons la fonction de répartition F, en échangeant le rôle des abscisses et des ordonnées, qu'obtenons-nous ? Encore un objet bien singulier : une fonction Φ de sauts, mais dont les sauts, d'amplitudes $\frac{1}{3^k}$ (k entier positif quelconque) forment un ensemble dénombrable mais partout dense sur l'intervalle [0,1].

On est bien loin de la représentation habituelle de fonctions de sauts dont le support, comme dans la distribution de Poisson ou celle de Pascal, est un ensemble dénombrable *discret*.

N.B. Ici aussi, on peut généraliser.

PS. Les rapporteurs anonymes sur ce texte, outre plusieurs remarques fort judicieuses dont ils sont remerciés, suggèrent d'indiquer des références bibliographiques pour le lecteur qui désirerait en savoir plus. Ils proposent :

BILLINGSLEY P., (1995), *Probability and Measure*, John Wiley and Sons, chapitre 31.

LEBESGUE H., (1904), *Leçons sur l'intégration et la recherche des fonctions primitives*, Paris, Gauthiers-Villars, [2e édition 1928, chap. IV, p. 56-57].

LÉVY P., (1937), *Théorie de l'addition des variables aléatoires*, Paris, Gauthier-Villars, chapitre III, §.12.