Discrete Probability
Theory & Application

Augustin Chaintreau
45 rue d'Ulm
75005 Paris FRANCE
augustin.chaintreau@ens.fr (or chaintreau@cmi.ac.in)

October 4, 2001

Abstract

Note on Lecture given in the Chennai Mathematical Institute, August 2001.

The teaching is made of nine (possibly ten) sets, on five weeks. Each sets is taking an hour and a half.

Mathematical notions : Probability paradigm (sample space, event) — Probability axioms (in the discrete case) — Independence, conditionning and their applications — Random variables and their distribution — Expectation and Moments — Moment generating functions — Introduction to Convergence and Limit Theorems

Applications treated : Numbering and application — Random Walk — Application to Biology — Branching process

Asymptotic methods and results : The Arc Sine law — Bernouilli approximations — Poisson approximation — Law of large numbers — Normal approximation.

Warning :

- The english language is not my mother tongue, and wasn't the language I used to learn mathematics. Incorrect use of English can occur in this document, especially when it comes to mathematical use of word. Remarks and correction are welcomed.
- Mathematical background : No preliminaries are needed to be able to read this lecture notes, out of some set theory and numbering. In particular, the theory of measure is not presented but required results are given as their necessity appear.
- A remark is given to students. Lots of notions developped in this lectures may seem very intuitive, and are very attractive to manipulate. Do not underestimate problem occuring in their using, most notions (independence, conditionning) can become harder and harder to handle as complexity of the problem increases. In this case, only a good understanding of these difficulties and the way they are formally addressed can help you. In particular, writing the theory of probability in the theory of sets and in the theory of measure and integration, is by no means peculiar.

Contents

Chapter 0

What do we mean by "Random" ?

Summary : Sample space, event — examples

Mathematical prelimary : Set Theory

Sketch THE PROBABILITY LIES ON THE WRITING OF "THE POSSIBLE" IN THE LANGUAGE OF SET THEORY.

—

Probability appears when one needs to take into account non deterministic behavior of a system and the aim to achieve is to deal with assertion on what happens, that may come to be true or not.

Example :

- a die is tossed, number that shows on top is observed.
- a coin is tossed 2 times, we note which face appears in order.
- a coin is tossed 2 times, we observe the number of head obtained.
- an airplane wing is assembled with a certain number of rivets, and the number of defective rivets is counted.
- From an urn containing n balls with different colors, three balls is chosen and their color noted.
- a light bulb used in a lamp is turned on, how many time before it is broken.

Assertions can be for the first example "we observed a pair number", "we have more than one head", "the three balls have different colors", or "at time t the light bulb is still working, or has shutted more than three minutes ago"

Sample space Possible is reduced to the choice of one element in a collection. This collection is a set denoted by Ω and called the *Sample Space*.

Example :

- $\{0, 1, 2, 3, 4, 5, 6\}$ for the Die.
- $\{HH, HT, TH, TT\}$ for two consecutive tossing of a coin.
- $\{0, 1, 2\}$ for the number of heads in two tossing of a coin.
- $\{0, 1, \ldots, N\}$ where N is the total number of rivets.
- {subsets made of three elements, of the ball sets}.
- $\mathbb{R}$ for the life time of the bulb.

Events (temporary definition) An event can occur or not, this is an assertion on what happens. As the possible is reduced to the choice of one element in a collection, events will be sub collection of this collection. *Events* are then subsets of Ω.

It is now possible to fully handle intuitive notion about "the possible" in the writing of set theory.

Events easy to understand

- Ω is the event of all possibility (always chosen by "random").
- $\emptyset$ is the event of no possibility at all (never chosen by "random").

Operations on sets standing for operations on events

- Complement stands for the negation of the event.
- Intersection of two events stands for "both events occurs".
- Union of two events stands for "one occurs".
- Difference, defined for one event included in the other, stands for the exclusion of an event in another one.

Relation on sets standing for assertion on events

- Inclusion stands for the implication.
- Being disjoint stands for the incompatibility.

Notation : We will in the following use the notation $A+B$ to denote the union of A and B when these two events are incompatible. Same for $\sum_{n\geq 0} A_n$ (standing for the union of sets from the sequence $(A_n)_{n\geq 0}$ as seen below, where these sets are supposed to be all disjoint).

Events constructed on infinite number of events $(A_n)_{n\geq 0}$ is a sequence of events, we can consider :

- The event $\bigcup_{n\geq 0} A_n$ "At least one event of sequence is happening"
- The event $\bigcap_{n\geq 0} A_n$ "Every event of the sequence is happening"
- The event $\limsup A_n$ "An infinite number of events of the sequence is happening"
- The event $\liminf A_n$ "All events of the sequence is happening except for a finite number"

Properties of $\liminf$ **and** $\limsup$ For a sequence of events $(A_n)_{n\geq 0}$

1. These limits inferior and superior can be written using $\cup$ and $\cap$.

$$\liminf A_n = \bigcup_{n\geq 0} \bigcap_{m\geq n} A_n \quad \text{and} \quad \limsup A_n = \bigcap_{n\geq 0} \bigcup_{m\geq n} A_n$$

2. In general we have $\liminf A_n \subseteq \limsup A_n$, when this two events are equal, the sequence is said to have a limite denoted by $\lim A_n$.
3. In particular if the sequence is non-decreasing (resp. non increasing) we have the previous equalty and $\lim A_n = \bigcup_{n\geq 0} A_n$(resp. $\bigcap_{n\geq 0} A_n$).

—

All of this is very natural as this is a usual way to reason in set theory to pick a point in a set and make supposition on this point.

Part I

Probability

Chapter 1

Providing a Probability on a Sample Space

Summary : σ field (just introduced) — Probability axioms — Probability general properties — additivity seen as continuity

Sketch A PROBABILITY IS CHARACTERIZED IN GENERAL BY ONE CALCULUS RULES, THE σ_ADDITIVITY. WE LOOK AT WHICH DEFINITION CAN BE CHOOSEN AND WHAT ARE THE DIRECT CONSEQUENCES.

a What is behind probability ?

Probability should put a positive number of every event, corresponding to their proportion to be more likely to happen, with the convention that when something is sure, this number is equal to 1. A calculus rules we would like to ask is the σ_additivity : the Number corresponding to the union of disjoint events, possibly in infinite numbers, is equal to the sum of the number corresponding to each one.

Probability (temporary definition) A probability should be an application defined on the events into the interval $[0;1]$ with $P(\Omega) = 1$ and the following rule :

$$P(\sum_{n\geq 0} A_n) = \sum_{n\geq 0} P(A_n)$$

THIS FORMAL DEFINITION CANNOT BE CHOSEN as probability won't be easy to build. One can show that no probability can exist in that way when the sample space is neither finite nor countable.

One solution is to restrict the additivity property to finite family of disjoint sets. Another one is to consider finite or countable Sample space where Probability can exist according to this definition. Another one is to

restrict events to a certain class of subsets of Ω. The last solution provide the most promissing framework. Considering the operation we already defined on events in chapter 0, such a class of events should at least verify the three following properties (to be a so called σ_field).

Definition :A σ_field is a class of subsets verifying :

1. Ω is in the class
2. if an event is in the class, so is its complement.
3. for a sequence made of subset in the class $(A_n)_{n\geq 0}$, their union $\bigcup_{n\geq 0} A_n$ is still in the class.

It is possible to show that this restriction is enough to be able to consider great varieties of probability on sample space (in particular the real line), but this result is a difficult fact coming from the theory of measure.

AS A CONSEQUENCE, ONE SHOULD KEEP IN MIND, before thinking of the probability of a given subset, to first check that this subset is in the class of events on which the probability is defined. This may seem strenuous to obtain, **BUT** that won't be an issue because of the two following facts :

1. In the case of finite or countable sample space, one can take the entire class of subsets of Ω to be the class of events, and be able to define probability.
2. Most of the subsets we can build from events use operation $\cup$ and $\cap$, possibly an infinite countable number of times, so that we still obtain an event at the end of our writing.

Final Definition $(\Omega, \mathcal{F}, P)$ is a probability space (where $\mathcal{F}$ is a σ_field), if P provides each events (elements of $\mathcal{F}$) with a positive real number with respect to :

- $P(\Omega) = 1$
- $P(\sum_{n\geq 0} A_n) = \sum_{n\geq 0} P(A_n)$

Remark: recall that writing $\sum_{n\geq 0} A_n$ supposes $(A_n)_{n\geq 0}$ to be disjoint.

b Direct Consequences of the additivity

Extremal value $P(\Omega) = 1$ is given as a convention, but $P(\emptyset) = 0$ is a consequence.

Monotony : As a direct consequence of the positivity ot P, the inclusion $A \subseteq B$ implies $P(A) \leq P(B)$.

Partioning the sample space We have $P(A)+P(A^c) = 1$, and moreover $P(A_1) + \ldots + P(A_N) = 1$ for each partition $A_1, \ldots, A_N$ of Ω.

When events meet For two events A and B we have $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. This rule can be systematized in the following, attributed to Poincaré:

$$P(A_1 \cup \ldots \cup A_N) = \sum_{k=1\ldots N} (-1)^{k-1} \sum_{1 \leq i_1 \leq \ldots \leq i_k \leq N} P(A_1, \ldots, A_n)$$

—

All this were using finite version of the σ_additivity, two useful results use the additivity in the case of an infinite sequence of events. First an inegality consequence of the mononoty of P, and a much more fine property giving behavior of P regarding to the limit of an event sequence as given in chapter 0.

Union and Sums The σ_addivity gives value of the probability of any union of infinite sequence of disjoint events, when they possibly meets, an inegality still holds, attributed to Boole.

$$P(\bigcup_{n \geq 0} A_n) \leq \sum_{n \geq 0} P(A_n)$$

Note that this inegality may not be of any help, as the right term might be higher than 1, or even infinite.

Additivity seen as continuity Recall the definition we gave in chapter 0 for the limit of a sequence of events, the σ_additivity can be interpreted as a "weak continuity" of the application P for this notion of limits. We have in particular :

$$P(\lim_{n \geq 0} A_n) = \lim_{n \geq 0} P(A_n)$$

when the sequence of event $(A_n)_{n \geq 0}$ is monotone.

Chapter 2

Probability in the discrete case

Finite case When the Sample set is finite, the class of events chosen is made of all the subsets of Ω and a probability is given by its value in each singleton. These values are all positive, with sum equal to 1.

Uniform probability in the finite case In the case where there is no reason for any particular sample point to be chosen among the others, the probability is the same on each singleton, equal to one divided by the cardinal of the sample space. And all the probability calculus is reduced to numbering.

Numbering allows one to deduce probability from a uniform probability (usually written on a product space). $n(n-1)\dots(n-r))$.

—

Mathematical prelimary : Series of real number. Let's summarize results we will use in a few proposition,

Convergence of series

- For a sequence of real number $(u_n)_{n\geq 0}$, the series is said to be *convergente* if the partial sums $S_n = u_0 + \ldots + u_1$ converge to a finite value S, that we then denote by $\sum_{n\geq 0} u_n$.
- If the term of the series are positive (possibly infinite), the sequence of partial sums in non decreasing, and thus has always a limit (finite or equal to $+\infty$) denoted by $\sum_{n\geq 0} u_n$

- A series has an *absolute convergence* if $\sum_{n\geq 0} |u_n| < +\infty$, such a series converges (by completness of $\mathbb{R}$).

Playing with the terms of a series In the case of absolute convergence, or if the series have positive terms (possibly infinite), the two following operations are allowed that do not affect the existence and value of the infinite sums :

- **Reordering :** for bijective map $v : \mathbb{N} \to \mathbb{N}$, $\sum_{n\geq 0} u_{v(n)} = \sum_{n\geq 0} u_n$
- **Packet sum :** for $(N_i)_{i\geq 0}$ a partition of $\mathbb{N}$,

$$v_i = \sum_{n\in N_i} u_n \text{ exists for all } i, \text{ and we have } \sum_{i\geq 0} v_i = \sum_{n\geq 0} u_n$$

Remark : These result do not hold in the general case of a convergente series.

—

Mass functions On a countable sample space, a probability is characterized by its value on singleton. All of these values are positive, and their infinite sums is equal to 1. We define the *mass function* in order to describe the probability, as seen below.

Definition : for a probability P, the mass function f_P is given by :

$$\forall \omega \in \Omega,\ f_P(\omega) = P(\{\omega\}),\ \text{such that for any event } A,\ P(A) = \sum_{\omega\in A} f_P(\omega)$$

Chapter 3

Independence and conditionnal probability at first glance

Summary : Independence of two events, of a family of events — Bernouilli trials & Binomial distribution — Non-Incrementality — Conditionning with respect to an event — Conditioning and Partitioning

Mathematical prelimary : none

Sketch HERE WE INTRODUCE THE INTUITION BEHIND THESE TWO KEYS CONCEPTS DEALING WITH A REPRESENTATION OF CAUSALITY AND ITS ABSENCE. DIRECT CONSEQUENCES FROM DEFINITION AND CALCULUS FORMULAS ARE GIVEN.

Warning Being able to deal with independence and conditionnal probability is a real addition of probability to measure and integration theory. In order to manipulate these two very attractive notions, one does not need to keep backing up to the formalization, made in the Lebesgue's theory of integration. However an effort is necessary to fully write intuition in our framework, and cannot be avoid to dissolve further confusion.

a let's first talk about independence

Causality in the randomness can be interpreted in the framework of event as relation between events. If one event occur, then the other one is more likely to happen. We already seen two cases : "positive" causality as inclusion, "negative" causality as incompatibility.

Independence is the most perfect state of non causality between two events. What happens with an event, whether it is occuring or not, does not have any implication on what is happening to the other one. It leads to the following definition :

Definition : Two events are said to be *Independent* if we have

$$P(A \cap B) = P(A)P(B)$$

Events' family $(A_1, \ldots, A_N)$ is said to be *independent* if

$$P(A_{i_1} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \ldots P(A_{i_k})$$

where $i_1, \ldots, i_k$ are chosen distinct in $1, \ldots, N$

Example :

- Typically are independent two successive experience with no relation (two successive toss of a coin). (very common to meet some of these independence in sample space made of product).
- Let's toss a dice, what would you say of this two events concerning the number :
  - { can be divided by 2 } and { can be divided by 3 }.
  - { the number can be divided by 3 } and { it is less or equal than 3 }
  - { the number can be divided by 3 } and { it is 1 }
  - { the number is 1 } and { the number is 8 }
  - { the number can be divided by 2 } and { it is less or equal than 3 }

Practice of the Theory :

- Show that complementation does not have any impact on independence.
- Independence and extremalites :
  - We call *extremal events* events with probability equal to 0 or 1. What can you they about them ?
  - What about an event independent from itself ? More generally show that two events A and B independent, where A implies B, leads to $P(A) = 0$ or $P(B) = 1$.
  - Is the independent relation transitive ?
  - What do you think of : A is independent from $B \cup C$, $B \cap C$, but A is not independent from B and from C.

independent

Non Incrementality Imagine we'd like to add one event to an independent family of events and keep the family independent, checking independence between this event and every elements of the family is not enough ! Intuitively it comes from the fact "available information" increase drastically as the number of events known get higher.

Example : On the probability space $\Omega = \{1, \ldots, 4\}$ provided with uniform probability, every singleton has probability equal to $\frac{1}{4}$. Consider the family made of the two events "less than 2" $A_1 = \{1, 2\}$, and "can be divided by 2" $A_2 = \{2, 4\}$. The event $B = \{2, 3\}$ cannot be add to the family - a way to understand it is to observe that being in the intersection $A_1 \cap A_2$ implies being in B - but it is independent with A_1 and independent with A_2.

—

Practice of the Theory : **Deformation and Independence**
In three successive tossings of a coin we are looking at the events :
$A = \{$Head appears in the first tossing.$\}$
$B = \{$There is at least two Heads appearing.$\}$

- Are these two events independent, when looking at the uniform probability on the sample space ?
- Can you give a probability on the sample space, where these two events are independent ? Same questions but it is now required that A and B are not extremal events (i.e. one with probability 0 or 1) in the new probability ?
- Give all the probability, that makes A and B independent, when it is also required that in the new probability different tossings are independent, as the contrary may seem absurd ?

Remark : Independence strictly speaking depends on the probability put on the sample space, but one can see here that these two events, very unlikely to be independent for intuitive reason, actually won't be independent in any case. It is very unlikely that independence will happen because of a weird choice of probability.

b conditionning with respect to an event

The operation of conditionning refers to make the supposition that one event really occured, and that randomness is now restrict to the case where this events is true. It leads to a new probability on Ω, and it is a good way to make reasonment and calculus.

*Definition :*The *Conditionnal probability* given the event A, that we will write $P(.|A)$ has value on any event B given by :

$$P(B|A) = \frac{1}{P(A)} P(A \cap B)$$

- $P(A)$ is here supposed to be strictly positive.
- $P(A \cap B) = P(B|A)P(A)$ holds, even if $P(A) = 0$.

last observation can be generalized in the following :

$$\begin{aligned} P(A_1 \cap \ldots \cap A_N) &= P(A_N|A_1 \cap \ldots \cap A_{N-1})\, P(A_{N-1}|A_1 \cap \ldots \cap A_{N-2}) \ldots \\ &\quad \ldots P(A_2|A_1)P(A_1) \end{aligned}$$

Conditioning with A is "assuming A" : Doing so ...

- Events incompatible with A cannot occur with positive probability.
- Probability on the subsets of A just suffers a $P(A)$ division.
- Independent events are not affected (their probability are the same).

Conditioning and Partitioning Imagine we are given an event A and a partition of the sample space $B_1, \ldots, B_N, \ldots$ finite or event countable.

- It is possible to calcul probability of A with respect to the assumption that there is exactly one B_i occuring (Bayle's rule):

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

- It is also possible, assuming that A occurs, to know which one of the B_i is occuring with which probability (usually referred by "probability of the causes").

$$P(B_I|A) = \frac{1}{\sum_i P(A|B_i)P(B_i)} P(A|B_I)P(B_I)$$

All of these formulas are useful methods to compute probability of an event, with respect to causality that were in the initial writing of the system.

—

Practice of the Theory : **Conditionnal independence**
A and B are said *independent conditionnally* to the event C if they are independent in the probability obtained by conditionning from respect to C.

- (Downstream) What do you think of two events A and B, independent for P, seen after conditionning by C ?
- (Upstream) What do you think of two events A and B conditionnally independent with respect to C and to C^c, seen in in the initial probability ?
- Give, if necessary, the additionnal assumption(s) needed to have transmission of the independence.

Part II

Random Variable

Chapter 4

Random variable and its distribution

Summary : Distribution of a random variable — composition with a map — Joint distribution of several random variables — independence

A random variable (temporary definition) is a some value that is not a constant but depends on randomness of the system. As the impact of random has been captured in the choice of one sample point, a random variable assign a value to each sample point, then follows the definition.

Definition : In the discrete case, that we are looking here, a *Random Variable* is a map from the sample set to a set a values.

Example : typically a random variable can be the gain of a gambler in a game, or the time of ruin of this same gambler.

a Distribution function of a random variable

The importance of the inverse function For each possible value of the random variable $x_1, x_2, \ldots$ one can consider the collection of sample points that are associated with each of these values $X^{-1}(x_1), X^{-1}(x_2), \ldots$, all are subsets of Ω. we can consider there probabilities, it gives us the probability for X to be equal to these different values.

$$P(X = x_i) = P(X^{-1}(\{x_1\}))$$

Distribution function we can consider the following function defined on $(x_1, x_2, \ldots)$ possible values of X, that we called the *distribution function* of X, given by

$$f_X(x_i) = P(X^{-1}(\{x_i\}) \text{ (also referred as } P(X = x_i))$$

One can see that this function is a mass function, that defines a unique probability on $(x_1, x_2, \ldots)$, referred as the *distribution* of X. The distribution of X may also be called the law of X, this is a probability on the new sample space $\{x_1, x_2, \ldots\}$ made of values of X.

The law of X gives us a behavior of this map, and characterize for example which values of X are chosen more often by the random, but it does not characterize at all the map.

Any probability on a sample space Ω can be written using random variable, whose law is given. This procedure avoids reference to abstract sample spaces, and simplifies the theory in many ways. However two things should be kept in mind.

- This gives much more opacity to the probability background
- Random variable are not entirely given by their ditribution, their "functional nature" is not peculiar to the theory of probability.

a.1 the case of the image of a random variable

When looking at a random variable, and given a map g defined on values of X, we can consider the random variable given by compound of these two maps, and the distribution function associated.

The distribution of $g(X)$ can be entirely deduced from the distribution of X and from the map g, with the following rules. For all value y taken by $g(X)$, we have :

$$f_{g(X)}(y) = \sum_{x \in g^{-1}(\{y\})} f_X(x) = f_X(x_{i_1}) + f_X(x_{i_2}) + \ldots$$

where $(x_{i_1}, x_{i_2}, \ldots)$ are the solution of $g(x) = y$

As a consequence : If one random variable X describe entirely the system - i.e. all values taken into account in the problem can be written as $g(X)$ -, giving the distribution of X is enough to characterize all the distributions in the system, and thus all the results that may be observed.

b The case of several random variables

We are now considering two random variable X and Y defined ON THE SAME SAMPLE SPACE Ω, taking values $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$.

b.1 joint distribution

One can consider the aggregate of sample points in which the two conditions $X = x_i$ and $Y = y_j$ are satisfied, and its probability. The function p defined on couples of values by

$$p(x_i, y_j) = P(X = x_i, Y = y_j)$$

will be called the *joint distribution function* of X and Y.

It is a mass function on the couple of values of X and Y, defining a unique probability on this space, called *joint probability* of X and Y.

Joint and Marginal distribution One can retrieve distribution function of X and Y from the function p. In fact the joint function satisfies

$$\text{for all } i \text{ we have : } \; p(x_i, y_1) + p(x_i, y_2) + p(x_i, y_3) + \ldots = f_X(x_i)$$

$$\text{for all } j \text{ we have : } \; p(x_1, y_j) + p(x_2, y_j) + p(x_3, y_j) + \ldots = f_Y(y_j)$$

BUT given that these two properties are satisfied, the joint function can have take any in the general case. In particular, given the distribution function f_X and f_Y, we cannot know how the joint random variable (X, Y) is distributed.

Conditionnal probability We can consider the conditionnal distribution of X given that $Y = y_j$ (assuming $P(Y = y_j) > 0$), it is given by the distribution function :

$$f_{X|Y=y_j}(x_i) = \frac{p(x_i, y_j)}{f_Y(y_j)}$$

This is a mass function giving the law of X given that $Y = y_j$.

b.2 Independence of random variables

Two Random variables will be independent if the value taken by one has no influence on the value taken by the other. So that it can be written by $f_{X|Y=y_j} = f_X$ for all y_j value of Y, or equivalently

$$P(X = x_i, Y = Y_j) = P(X = x_i)P(Y = y_j)$$

It corresponds to a joint function $p(x_i, y_j)$ with a product form, where all column and lines are proportionnal. In particular, if X and Y are independent, with known distributions, the joint distribution is also known.

Independent family Similarly, the family $(X, Y, \ldots, W)$ will be named *independent* if for any $x_i, y_j, \ldots, w_k$ we have :

$$P(X = x_i, Y = y_j, \ldots, W = w_k) = P(X = x_i)P(Y = y_j) \ldots P(W = w_k)$$

An infinite family of random variable will be named *independent* if any finite sub family is.

Composition with applications Considering a family $X, Y, \ldots W$ of independent random variables, and maps $f, g, \ldots, h$, the random variables $f(X), g(Y), \ldots, h(W)$ are independent.

Chapter 5

Expectation

Summary : Expectation — existence property — other properties — deviation theory — Moments and Variance — Correlation

a Definition and direct consequences

We define the expectation, or mean value, of a random value as the barycenter of the different value of X, each one with mass equal to its probability of occuring. Note that with this definition, the expected value of X just depends on its distribution.

Expectation is given by the following formulas, where the series is assumed to converge absolutely (else by convention $\mathbb{E}(X)$ is said infinite)

$$\mathbb{E}(X) = \sum_{x_i} f_X(x_i) x_i = \sum_{x_i} P(X = x_i)\ x_i$$

To assure existence and finiteness of $\mathbb{E}(X)$

- $\mathbb{E}(X)$ is finite iff $\mathbb{E}(|X|)$ is finite
  and in this case we have always $\mathbb{E}(X) \leq \mathbb{E}(|X|)$.
- $\mathbb{E}(X)$ is finite if we have $|X| \leq Y$ and $\mathbb{E}(Y)$ finite.
- $\mathbb{E}(X)$ is finite if $-\infty < a \leq X \leq b < +\infty$
  and in this case we have $a \leq \mathbb{E}(X) \leq b$.

—

Expectation of the Image of a random variable let X be a R.V., we consider its image by a map g. The series defining the expectation of $g(X)$ can be derived from the serie defining the expectation of X.

$$\mathbb{E}(g(X)) = \sum_{x_i} P(X = x_i)g(x_i) = \sum_{y_j} \left(\sum_{x_i | g(x_i) = y_j} P(X = x_i) \right) y_j$$

Expectation with intervention of several random variables It is very important to undestand what is the meaning of $\mathbb{E}(XY)$, $\mathbb{E}(X + Y)$, $\mathbb{E}(\sqrt{X} \ln Y)$, etc. All this writing mean the expectation of a random variable $E(W)$, that can be written $W = g(X, Y)$ where g is a map.

In particular, being the expectation of the image of the joint random variable (X, Y), in the general case it depends on the joint distribution of (X, Y), and in many general case, giving only the distribution of X and Y is not enough, there is two case where things simplify :

- when X and Y are independent, the joint distribution is given once marginal distributions [1] are fixed, and thus $\mathbb{E}(g(X, Y))$.
- when we look at the function "sum" of two random variables (as seen in the next paragraph).

—

Others consequence from the definition

X and Y are here supposed to have a finite expectation

- **Linearity** $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$, also $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X)$
- **Monotony** $X \geq 0$ implies $\mathbb{E}(X) \geq 0$,
  as a consequence, $X \leq Y$ implies $\mathbb{E}(X) \leq \mathbb{E}(Y)$.
- **Independence** when X and Y are independent

$$\mathbb{E}(X\ Y) = \mathbb{E}(X)\mathbb{E}(Y)$$

—

[1] recall we denote *marginal distributions* the distribution of X and Y

Almost sure properties One property that may come to be true or not depending on ω, is *almost sure* if it is true with probability equal to 1. (resp. ω such that the condition does not hold are contained in a set of zero probability).

Two random variables equal a.s. have the same expectation, then to characterize the expectation, properties on random variables need only to be almost sure.

—

Practice of the Theory : **Sum with independent number of terms**
let $X_1, X_2, \ldots$ be a sequence of random variable with the same distribution, let N be a random variable valued in the integers, such that the family $N, X_1, X_2, \ldots$ is also independent. Show that we have :

$$\mathbb{E}(X_1 + \ldots + X_N) = \mathbb{E}(N)\mathbb{E}(X)$$

The first thing to do is of course to understand the writing $X_1 + \ldots + X_N$ as the definition of a random variable.

Practice of the Theory : **Conditionnal expectation**
Remember that conditionning with respect to an event A si changing the probability on the sample space P into $P(.|A)$. Similarly we define :

$$\mathbb{E}(X|A) = \sum_{x_i} P(X = x_i|A)\, x_i$$

(conditionnal expectation of X with respect to A)

- Giving $(A_0, A_1, A_2, \ldots)$ gives a partition of the sample space, show that, similar to Baye's formula seen on the chapter about conditionning, we have :

$$\mathbb{E}(X) = \sum_{i \geq 0} P(A_i)\mathbb{E}(X|A_i)$$

Practice of the Theory : **Expectation and Extremal Values**
Warning The following results may seem silly but, as surprising as it may be, they are quite powerful in the formalism. We consider a random value taking positive reals values, possibly infinite, in $\{x_1, x_2, \ldots\} \cup \{+\infty\}$. Definition of the expectation can be given in this case as an infinite sum of positive real number with $\mathbb{E}(X) = +\infty$ if the series diverges.

- Show that $\mathbb{E}(X) < +\infty$ implies $P(X < +\infty) = 0$
  What do you think of the converse implication ?
- Take X any real valued R.V., possibly infinite,
  Show that $\mathbb{E}(|X|) = 0$ implies $X = 0$ a.s.
  What can you say if $\mathbb{E}(X^2) = 0$?

Deviation theory When X is positive with finite mean, we have the following inegality, attributed to Markov :

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

Proof : looking at the random variables $\frac{X}{a}$, and dividing it in two part we have by linearity :

$$\frac{1}{a}\mathbb{E}(X) = \mathbb{E}(\frac{X}{a}\mathbb{I}_{\{X<a\}}) + \mathbb{E}(\frac{X}{a}\mathbb{I}_{\{X\geq a\}})$$

The first one is greater than 0 and the second one, as $X\mathbb{I}_{\{X\geq a\}} \geq a\mathbb{I}_{\{X\geq a\}}$, is greater than $P(X \geq a)$

b Variance and Moments

Moments for an integer r, and a random variable X, if the expectation of random variable X^r exists, then it is called the r_th moment of X. Note that as $|X|^{r-1} \leq |X|^r + 1$, existence of r_th moment implies existence of all preceding ones.

Variance As long as the second moment of a random variable exists, we can define its variance (usually denoted by σ^2).

$$Var(X) = \mathbb{E}((X-\mu)^2) = \mathbb{E}(X^2) - \mu^2 \text{ where } \mu = \mathbb{E}(X)$$

Deviation theory (cont'd) : The markovs inegality leads directly to the **Chebishev's inegality**, true for any random variable X with a finite second moment.

$$P(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2} \text{ where } \sigma = \sqrt{Var(X)}$$

This inegality is very useful and handy because of its generality, but in most of the applications, this is far from being the optimal comparison we can make. Consider it as a theoretical tool.

Covariance For two random variables X and Y, defined on the same sample space, with finite second moments. we can then define :

$$Cov(X,Y) = \mathbb{E}((X-\mu_X)(Y-\mu_Y)) = \mathbb{E}(XY) - \mu_X\mu_Y$$

and the normalized correlation coefficient : $\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$.

In particular we have for any sum of n random variables :

$$Var(X_1 + X_2 + \ldots + X_n) = \sum_{k=1..n} \sigma^2_{X_k} + \sum_{i\neq j} Cov(X_i, X_j)$$

Extremal values of correlation We always have $|\rho(X,Y)| \leq 1$

- When $\rho(X,Y) = 0$ the random variables X and Y are said to be *non correlated*. In particular Two independent random variables are non correlated, but this not at all necessary.
- $|\rho(X,Y)| = 1$ only if we can write Y=aX+b a.s.

Non correlation does not imply independence. In fact it is even possible to have non correlation and a functionnal dependence : Taking X equal to $\{-2,-1,+1,2\}$ with uniform probability and $Y = X^2$. correlation is zero, but Y depends only on X. For all of these reasons, correlation cannot be interpreted as a general measure of dependence between X and Y.

However, some results are not using all the aspects of independence of random variables, and non correlation, much less stronger than independence, may sometimes be sufficient (cf. remark on law of large numbers).

Practice of the Theory : **Weakness of non correlation :**
Consider X and Y random variables with equal variance. Show that $X+Y$ and $X-Y$ are non correlated.

—

Practice of the Theory : **Manipulating the framework**
Warning : This exercice has non complexity neither mathematical interest, but recapitulate different notions and serve as a checking point.
For an event A of the sample space we define the associated indicating random variable I_A taking value in $\{0;1\}$ where $\{I_A = 1\} = A$.
All usual operations on events and their relations can be written clearly using this functions : $\cup$, $\cap$, ^{c}, independence, etc. And we have $\mathbb{E}(I_A) = P(A)$
Show directly from the definition that :

- **Covariance :** $Cov(I_A, I_B) = P(A \cap B) - P(A)P(B)$, $Cov(I_A^c, I_B) = -Cov(I_A, I_B)$.
  What can you say when of two events are *non correlated* (i.e. their indicating variables are) ?
- **Variance :** give general formula giving $Var(I_A)$.x
- **Correlation :** Give the simplest criteria on the event for $\rho(I_A, I_B)$ to be defined. What can you say on A and B when correlation take extremal values $+1$ and -1 ?

—

*Practice of the Theory :***Queue repartition Method** Simple observation presented here can be indeed very useful in the applications.

The method Taking an integer valued, possibly infinite, random variable X (see exercice *Expectation and extremal values* for more details on that). Show that expectation are given by the repartition of the queue :

$$\mathbb{E}(X) = \sum_{k \geq 0} P(X > k) = \sum_{k \geq 1} P(X \geq k)$$

Some applications :

- You are left with n keys in front of a door, you have been told that k keys can open the door. You are trying each key, stopping when the door is open.

  Give the expectation of the number of trials you will have to perform ?

  Without using queue distribution method, find the result for the case $k = 1$.
- Same questions when you cannot differentiate the key after trying one, so that you may try it again another time.

  Same remark, another simple method should give you the result using Binomial distribution.
- **A better example** One urn contains N balls with number $1, 2, \ldots, N$. We are taking n times a balls from the urn, and putted in back just after, denoting by X the greater number we have obtained.
  - Give expectation of X.
  - Propose a method to evaluate N when it is not known.
  - Should we better keep the ball after taking it from the urn ?

Part III

Introduction to Limit Theorems and Generating Functions

Chapter 6

As an introduction to limit theorems : The law of large Number

Warning : We are in this chapter as in all of this lecture notes in the case of discrete probability. Sample space we can consider need to be finite or countable.

Here we are looking at an arbitrary numbers of bernouilli trials. We cannot provide the infinite number of trials with a sample space because the set of sequence $\varepsilon_1, \varepsilon_2, \ldots$ where each elements is 0 or 1 is not countable.

In this context, limit theorems can just be enunciated as : all probability are compute in a case where n is fixed, then we look at the behavior of this probability when n goes to infinity. In particular we will look at :

for n natural number, we can consider $X_1, X_2, \ldots, X_n$ random variables on the sample space, and look at the average $S_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$.

S_n will be said to *converge stochastically* to a constant μ if we have :

$$\text{for any } \varepsilon > 0,\ P(|S_n - \mu| \geq \varepsilon) \to 0 \text{ as } n \to +\infty$$

but note that this is not a convergence of a sequence of elements of the same space to an element of this space.

In the continuous case It will be possible to define probability on non countable sample space, and in this context, this convergence will interpreted as a convergence of elements in one space.

a From the Thebychev inegality to the first law of large numbers

Nature of the problem The aim of this chapter is to prove a convergence result on sequence of random variables. We are looking at the sequence

$X_1, X_2, \ldots$ and as we supposed that it is different trials of the same experience of estimating the values of X, this variables will be supposed to have the same distribution law. Then one think we would like to have is that results we obtain, looking on the average of a large numbers of trials, converge to the expectation. Which mean

$$S_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \quad \text{should be close to } \mu = \mathbb{E}(X)$$

Note that for any n, S_n has expectation equal to μ, so what do we need to show is that with increase of n the distribution of S_n reduced to value close to its expectation, or if you prefer that the sums of the error: $\frac{(X_1-\mu)+(X_2-\mu)+\ldots+(X_n-\mu)}{n}$ will compensate and converge to zero.

Chebichev's method For some reason it is not possible to use a deviation result of order 1, as in this case the errors sum additionnaly in a linear scheme, are normalized by division by n, but the order of magnitude seems to stay constant, and no convergence to zero can be extracted.

But looking at the second order, which mean considering the variance of $(X_n)_{n\geq 1}$ and the variance of S_n, we observe that the variance of $X_1 + X_2 + \ldots + X_n$ can be written as the sum of two terms:

- $Var(X_1) + Var(X_2) + \ldots + Var(X_n)$ that increases linearly with n. This is indeed very slow as we will then divide the sum by n and thus its variance by n^2. This term will vanish when $n \to +\infty$.
- the second term $\sum_{i\neq j} Cov(X_i, X_j)$ depends heavily on the relation between the variables. It can in certain case grows as n^2, but it can vanish, in particular when the variables considered are independent (pairwise independent - or uncorrelated - being also sufficient).

As by Chebichev's inegality,

$$P(|S_n - \mu| \geq \varepsilon) \leq \frac{Var(S_n)}{\varepsilon^2}$$

We have the following results:

LAW LARGE NUMBER iid $\mathbb{L}^2$ For a sequence of random variables $X_1, X_2, \ldots$ independent, with the same distribution, that admits a second moment with expectation μ, we have :

$$(S_n)_n \text{ converges stochastically to } \mu \text{ where } S_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

—

The following of this chapter propose to see different comments and extension made on this first result.

b Is this result robust to distribution perturbation of the sequence ?

Imagine that the sequence of results observed $X_1, X_2, \ldots, X_n$ that should be independent with same law suffer in its process some error, for example some experience have different conditions, and so slightly different law, or independence of the entire family cannot be obtained.

First concerning relation on the sequence, pairwise uncorrelation is the weakest qualitative assumption we can make for the moment. Others that can be regarded are bounds given on correlations of different variables, that we will not present here.

Second on the "law fluctuation" one can see that everything still holds while assuming the two following fact :

- *Mean can vary but should converge :*

$$\mu_n \to \mu \text{ as } n \to +\infty \text{ where } \mu_n = \mathbb{E}(X_n).$$

- *Variance should not grow too much :* keeping

$$Var(X_1) + Var(X_2) + \ldots + Var(X_n) = o(n^2) \text{ as } n \to +\infty.$$

c Getting out of the $\mathbb{L}^2$ context

For the case where random variables admits expectation but no second moment, a truncation technique allows us to show the same result. We are giving the fundamental case where the law does not vary.

Method we would like to show

$$\forall \delta \ \forall \eta, \ P(|S_n - \mu| \geq \delta) \leq \eta, \text{ for } n \geq N$$

we are now fixing the value of n, fixing one real number ε that will tends to be small, and considering the variables $X_1, \ldots, X_n$ that we divide in two parts

$$X_k = A_k + B_k, \text{ where } A_k = X_k \mathbb{I}_{|X_k| \leq n\varepsilon} \text{ and } B_k = X_k \mathbb{I}_{|X_k| > n\varepsilon}$$

Bounded part of the sequence Having $A_1, A_2, \ldots, A_n$ bounded, with same law of mean μ_n, we have

$$P(|\frac{A_1 + \ldots + A_n}{n} - \mu_n| \geq \frac{\delta}{2}) \leq \frac{4}{\delta^2} \frac{Var(A)}{n}$$

One mistake has to be avoided here, as $(A_k)_k$ has been build fixing the value of n, we have no reason to be able to make $\frac{Var(A)}{n}$ going small.

Fortunately, making the best use of the bound given on $(A_k)_k$:

$$Var(A) \leq \mathbb{E}(A^2) \leq n\varepsilon\mathbb{E}(|A|) \text{ such that}$$

$$P(|\frac{A_1 + \ldots + A_n}{n} - \mu_n| \geq \frac{\delta}{2}) \leq \frac{4\varepsilon\mathbb{E}(|X|)}{\delta^2}$$

that is more likely to become small with ϵ.

Adding that we suppose n bigger than a threshold N_1 so that $\mu_n - \mu \leq \frac{\delta}{2}$, we found

$$P(|\frac{A_1 + \ldots + A_n}{n} - \mu| \geq \delta) \leq \frac{4\varepsilon\mathbb{E}(|X|)}{\delta^2}$$

Event reasonment : X will be different from A just in the case where B is different than zero. So that probability that X far from μ by δ is smaller that the sum:

$$P(|\frac{A_1 + \ldots + A_n}{n} - \mu| \geq \delta) + P(B_1 \neq \text{ or } \ldots \text{ or } B_n \neq 0)$$

looking at the second term, we found as $(B_k)_k$ have the same law that it is less than $nP(|X| \geq n\varepsilon)$ that we need to show small.

Large values of X Markov is not enough, as the bound provided will diverge for large value of n that we need to be able to consider, but fortunately, we can make better use of the situation :

$$\mathbb{E}(\mathbb{I}_{|X_k|>n\varepsilon}) \leq \frac{1}{n\varepsilon}\mathbb{E}(|X_k|\mathbb{I}_{|X_k|>n\varepsilon})$$

last term is converging, for any value of ε fixed, to zero as $n \to +\infty$, by dominated convergence.

Conclusion Choosing an ϵ such that the first part has probability smaller than $\frac{\eta}{2}$, that is not depending on n, we can then consider big value of n and for a threshold, the second probability will be arbitrary smaller. We have then shown the ...

LAW LARGE NUMBER iid $\mathbb{L}^1$ For a sequence of random variables $X_1, X_2, \ldots$ independent, with the same distribution, that admits an expectation μ, we have :

$$(S_n)_n \text{ converges stochastically to } \mu \text{ where } S_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

remark Note that, if this result holds, we cannot though here ignore the method originally written in the context $\mathbb{L}^2$, that properly creates here the convergence.

d Increasing the convergence : convergence in mean

We would like now to increase our knowledge of the convergence, and in particular show that S_n can be replaced by μ in integral form when looking at the limit. This require a $\mathbb{L}$ convergence.

The sequence $(S_n)_n$ converge to a constant μ as a $\mathbb{L}^r$ *convergence* ($r = 1, 2, \ldots$) if we have $\mathbb{E}(|S_n - \mu|^r) \to 0$ as $n \to +\infty$.

remark To be able to speak of $\mathbb{L}^r$_convergence, it is implicitly assumed that random variables of the sequence admits all a r_moment.

The case $\mathbb{L}^2$ is already treated by Chebychev's method. The writing the $\mathbb{L}^2$_convergence exactly corresponds to $Var(S_n) \to 0$ that we have shown.

The $\mathbb{L}^1$ case is not treated in this lectures note. A general technique that will be presented to you in the general case of continuous random variable, allows one to deduce the $\mathbb{L}^1$_convergence from Stochastic convergence, provided that the random variables of the sequence are uniformly distributed.

The question to know whether it holds in the case of the sequence S_n made on the sequence $X_1, X_2, \ldots$ of independent and identically distributed random variables, is not answered here.

Chapter 7

Method of Generating functions

Considering generating function of integer valued random variables is very similar to taking the logarithm of positive real numbers. In both case we introduce this transformation, that might be inverted in a sense to precise, to be able to change one operation into another one. Here We'd like to change the sum of independent random variables into product.

Definition : We define, for any Random variables X taking values in the integers, the *generating function* on the real numbers given by :

$$\text{gen}_X : s \to \sum_{k \geq 0} P(X = k) s^k$$

- Note that the series defining this function, as coefficients are no greater than 1 in module, with finite sum, converges absolutely for $-1 \leq s \leq 1$. Also note that $\text{gen}_X(1) = 1$
- Also Note that the generating function depends only on the distribution of the random variable.

Derivative and Expectation Generating function contains in particular information about the expectation, written in its derivative (that is always well defined on $]-1;1[$).

- If $\mathbb{E}(X)$ exists and is finite, then the derivative converge in 1 to the value of the expectation. This also means that the generating function admit a derivative in 1 with same value.
- Else, the derivative diverge in 1.

This can be written : $\text{gen}'_X(1) = \mathbb{E}(X)$

Variance Similarly Variance can be derived from the generating function, and we found :

$$Var(X) = \text{gen}''_X(1) + \text{gen}'_X(1) - (\text{gen}'_X(1))^2$$

when variance is finite, else we have divergence of gen''_X in the value 1.

Inversion As two power series with same value on $]-1;1[$ have same coefficient, a generating function is characterizing the distribution of the random variable.

Example : Binomial case

—

But the practicality of calculus for expectation and variance is not the reason why we are interested in the generating function. As we have said in introduction, as the logarithm on real number, the one to one correspondence between law and these functions converts some of the usual operations on random variables.

Operation Conversion : Sums $\rightarrow$ Product If X and Y are independent random variables, The sum $X+Y$ verify :

$$\text{for any real number } s,\ \text{gen}_{X+Y}(s) = \text{gen}_X(s)\,\text{gen}_Y(s)$$

—

Operations Conversion : Compound Sums Another remarkable property of the generating functions, not as famous and used as the first one is the following. Considering $X_1, X_2, \ldots$ independent random variables with the same law, whose generating function is f, we are regarding the sum :

$$S = X_1 + \ldots + X_N$$

where N is a positive integer valued random variables, whose law has g for distribution function.

Then the distribution function of S is given by composition :

$$\text{for all real number } s,\ \text{gen}_S(s) = g(f(s))$$

Before Ending the lectures

We have been able to show one limit theorems, Taking a random variable which take a countable number of values on the real line, we showed that when we sample independently n times the random variables, and take the average of the results. Then we will end with the expectation value.

From now, one can go into more details, or try to generalize the context in which this result hold :

Fluctuation in a random walk

The final results of a random walk (final gain for one player) is given by his number of success. When n is increasing, how are this value distribted among the possible values (from 0 to n). When looking at symmetrical random walk, The expectation will be $n/2$ and the law of large number tells us that $\frac{S_n}{n}$ will be close to this expectation with probability 1. So that deviation $S_n - \frac{n}{2}$ will be $o(n)$.

One question will be to have a much more precise evaluation of the deviation $S_n - \frac{n}{2}$ than an upper bound, and be able to quantify it. The answer is using the notion of *normal law*, and the results is that when dividing the deviation $S_n - \frac{n}{2}$ by $\sqrt{n}$, we end with some convergence to a normal law.

Getting further from the independence property

Markov's paragdigm

Looking at a sequence of results $X_1, X_2, \ldots$. We are no more assuming that this is independent result, but that the next result X_{n+1} depends only on the value of X_n : once the value of X_n is known, the precedent value $X_1, X_2, \ldots, X_{n-1}$ have no impact on the future of the sequence. This that can be written :

$$\mathbb{E}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1)$$

$$= \mathbb{E}(X_{n+1} = j | X_n = i)$$

If we assume additionnaly that this does not depend on the place we are in the sequence. This is equal to the transition probability from i to j, that we denote $p_{i,j}$.

The behavior of the sequence can be known studying the transition matrix $(p_{i,j})_{i,j}$.

In particular, it is possible to extract from this study an invariant probability, one such that if X_n has this probability, X_{n+1} has also this probability a distribution law.

Adding assumption, we can sometimes show a limit theorem : starting from any places, the sequence will converge, in a sense that we need to precise, to a random variables with this "fixed point" probability.

Martingales

Martingale stands in probability equivalently to constant sequence of real number. Submaringale are by comparison the equivalent of non decreasing sequence of real number.

Martingales can be seen as better evaluation of a random variable, with greater and greater information. M_{n+1} is evaluation of a r.v. W with a set of information $\mathcal{F}_{n+1}$ greater that $\mathcal{F}_n$. At each step approximation is the better we can find, such that $M_{n+1} - M_n$ cannot differ when information at step n is only considered.

This can be explicitly written in probability framework, using a new conditionnal method on random variable. Conditionnal expectation related to class of set (equivalent of information), denoted $\mathbb{E}(M|\mathcal{F})$.

The condition written above is : $\mathbb{E}(M_{n+1} - M_n|\mathcal{F}_n) = 0$.

For a submartingale we will have $\mathbb{E}(M_{n+1} - M_n|\mathcal{F}_n) \geq 0$

Such that increment of submartingale $(M_{n+1} - M_n)_{n\geq 0}$ can be related to non decreasing sequence, or a series of positive terms. We then have both operations on the series allowed, and convergence results, provided bounded assumption.

A first step in the application of martingale is the Doob's theorem, stating that the expectation of a martingale do not vary with n, $\mathbb{E}(M_1) = \mathbb{E}(M_2) = \ldots$, and this is true even if we look at the

The particular operation valid on martingales' increment allows one to make sums with respect to the increment of a martingale, leading to the definition of stochastic integration, that is used in many applications of probability frm EDP study, statistical physic, to financial mathematics.

Bibliographical notes

A very good validation and introduction to concept of probability is Khinchin's book [Khin]. This book was intended for a large audience, and can help the

student in understanding the building of the probability framework. Independence, and conditionning, are in particular very well explained.

William Feller's book [Fell] contains definition and results of the theory, as well as lot of applications and treated case. Subject of this lecture notes are treated in the first volume. Another book with lot of examples and applications, as well as a presentation of definitions and theorems is M. Loeve's two volumes on Probability.

Another book by Paul Andre Meyer [Meyer] present notions of probability with lot of treated case and applications. Grimett Strizaker contains lot of applications and examples. Neveu's book has been translated to french, and contains very clear presentation of probability definition and result, as well as fundamental applications.

Three references are given that is not at all suited to follow this course, but can be of used for later purpose: The two Patrick Billingsley's books on Probability and Weak convergence, presenting limit results and for the second one a very precise description of random process topology. Shiryayev's book is a lot of help.

Bibliography

[Khin] A.Y. Khinchin and B.V. Gredenko, *An elementary introduction to the theory of probability*, Dover Publication.

[Fell] W. Feller, *A Introduction to Probability theory and its applications (Volume 1 and 2)*, Wiley (1957).

[Loev] M. Loeve Probability theory (I and II), Springer.

[Meyer] Paul Andre Meyer, *Introductory Probability and statistical applications*, Addison Wesley, Oxford IBH publications

[GrimStriz] G. R. Grimett and D. R. Strizaker, *Probability and Random Processes* (1982).

[Neveu] J.Neveu, Mathematical foundation of probability

[Bill] P. Billingsey, *Probability and Measure* (1978).

[BillMeas] P. Billingsley, *Convergence of probability measure* Wiley.

[Shir] A.N. Shiryayev *Probability*, Springer

Applied Probability, Set Number 1

Augustin Chaintreau

October 3, 2001

Abstract

This is a combination of treated case, example of application of probability to problem solving. This set, that takes place after chapter 2 of the lectures notes, contains consequences of numbering results to probability applications.

Most of this are using the following method :

- Consider a product sample space, with some uniform probability.
- Build another sample space (made of the real "results of our experience"), and provide it with probability, made out of the first one using numbering.

a Numbering and Meeting

Problem number 1 : Birthday In a group made of N people, what is the probability to have at least two people sharing the same birthday ?

Problem number 2 : Accident during a week In a city, seven accident are occuring in a week, give the probability that two accidents happen in the same day.

b Numbering and Indistinguishability

b.1 Simultaneous tossing of several coins, or several dices

N coins are tossed together, these coins cannot be distinguish, give the probability we should provide on the sample space.

Give the probability that tossing twelve dice in the same time, we observe each number $1, 2, 3, 4, 5, 6$ given twice.

b.2 Occupancy problem

Accidents during a week (cont'd) : Gives the probability that there is two days with two accidents, and three with one accident

b.3 Sampling a property

In a set made of n particles, $p \leq n$ are "mutant". We observed r particle. Our goal is to understand how this proportion of mutant can be represented in the sample.
Give the law of the number m equal to the number of mutant particles in the sample ?
Try to locate which value of m is the most probable ? Does that give you an idea of a method using sampling to quantify the proportion of mutant in a population.

—

A limit results : We intuitively feel that there is roughly for each particles of the sample a probability $\frac{p}{n}$ to be mutant.

- Why is this false ? (Give the differences between the case which is described for hypergeometric sampling, and the case of a set of r particle, each one having a probability $\frac{p}{n}$ to be mutant.)
- Make it True ! (In other words, which asymptotics can make the second model a good approximation ? Show the result.)

c A first step in random walk

We consider a particle moving in a space, starting from (0,0), at each step, the particle is going one unit on the right, then choose to go one unit up or down.

We denote the successive position of the particle by $(0,0),(1,s_1),\ldots,(n,s_n)$.

- $\varepsilon_i = s_i - s_{i-1}$ for $i = 1..n$ are successive choices made by the particle.
- When looking at n step, we define $p = \#\{i = 1\ldots n \mid \varepsilon_i = +1\}$ and $q = \#\{i = 1\ldots n \mid \varepsilon_i = +1\}$ one can easily observe :

$$p + q = n \text{ and } s_n = p - q$$

For a given path $(0,0),(1,s_1),\ldots,(n,s_n)$, we say that at time n there occur :

- a return to the origin if $s_n = 0$
- a first return to the origin if $s_1 \neq 0, s_2 \neq 0, \dots s_{n-1} \neq 0, s_n = 0$
- a first passage through $r > 0$ if $s_1 < r, s_2 < r, \dots, s_{n-1} < r, s_n = r$

Framework 1 : Fixing the final position of the random walk We are doing the assumption that the final position is fixed (as length is also fixed, it is the same as suppose that p is given).
What we are looking at is the different paths going from $(0,0)$ to the position $(n, p-q)$. (an example of such a situation is the count vote of an election.)
Probability : Do all paths have the same probability to occur ? How many are there (denoted it by N_{n,s_n}) ?

- The answer is YES, all paths are equivalently probable.

  To fully understand that, one can number each votes contained initially in the urn by $(1, \dots, p, p+1, \dots, n)$ (the p first are "positive" votes, then they all are "negative").

  What is then a vote-counting ? It is exactly a *reordering* of this whole sets. It is clear that two different reordering are equivalently probable. The question now is "Do one ordering refers to exactly one path ?", that is not true as to compute the path, one do not look at the number of the vote, but just whether this vote is positive or negative, BUT for one particular path, there is exactly $p!q!$ *reordering* corresponding to this path (once fixed which votes of the reordering are "positive" or "negative", you just need an order on the "positive" votes ($p!$ choices possible), and an order on the negative votes ($q!$ choices possible).

  So that each path can be equally chosen (corresponding each one to the same number of reordering).

- There is in total $\binom{n}{p} = \binom{n}{\frac{n+s_n}{2}}$ paths leading from $(0,0)$ to (n, s_n).

Framework 2 : Final position is not given We look at all the paths of length n starting from $(0,0)$. At each time, there is probability $\frac{1}{2}$ to go up or down.
Probability : Do all paths have the same probability to occur ? How many are there ?

- The answer is again affirmative. Each path corresponds to n successive choices whether it goes up or down, each choice being probably equivalent.
- There is in total 2^n paths starting from (0,0) and with length n.

Sympathic property : Imagine a property depending on the k first steps of the random walk, would the value of n have any impact ? Is this true with the framework 1 ?

—

We have now provided sample space and probability framework for two different situations dealing with choices of "paths of a random walk". In both of this model, a path is equally equivalent to another, all the probability calculus is then reduced to NUMBERING.

c.1 Problem of arrangments

We are in this section in the framework number 1.

Reflection principle Let's first show a geometrical result, that will help the numbering : Let $A = (a, \alpha)$ and $B = (b, \beta)$ are two points with $a < b$ above the x_axis (α and β strictly positive).

The number of path from A to B that cross the x_axis, is equal to the number of path from $A' = (a, -\alpha)$ to B.

We can now prove the **Ballot theorem** : for $s_n > 0$, the number of paths from $(0, 0)$ to (n, s_n) that do not touch the x_axis after departure, is given by

$$\frac{s_n}{n} N_{n,s_n} = N_{n-1,s_n-1} - N_{n-1,s_n+1}$$

Application Give in a election between two candidates, the probability that one candidate is leading all the time during counting vote. What do you need for that ?

c.2 Long leads, Arc Sine law

One may think that when n increase, the random walk will have some repetition of the same behavior. One win for some time, then the other. We will observe that most of our intuitive understanding of the evolution of the random walk is wrong.

We will consider path with length $2n$ and will consider the last return of 0 in the path. Let's denote $\texttt{last-Ret}_{2k}^{2n}$ proportion of paths with last egality in $2k$. Our goal is to be able to give the value of $\texttt{last-Ret}_{2k}^{2n}$.

Step number 1 We show

$$\texttt{last-Ret}_{2k}^{2n} = \underbrace{P(S_{2k} = 0)}_{\texttt{Ret}_{2k}} \times \underbrace{P(S_1 \neq 0, \ldots, S_{2(n-k)} \neq 0)}_{\texttt{No-Zero}_{2(n-k)}}$$

($\texttt{Ret}_{2k}$ for "return in $2k$", $\texttt{No-Zero}_{2(n-k)}$ for "no zero during $2(n-k)$")

Step number 2 We observe some symetry, actually we have

$$\texttt{Ret}_{2k} = \texttt{No-Zero}_{2k}$$

Such that last return in $2n$ can be with equal probability in $2k$ and in $2(n-k)$. We take a look a some consequence.

Evaluation and asymptotics It is easy to see $\texttt{Ret}_{2k} = \frac{1}{2^{2k}}\binom{2k}{k} = \frac{1}{2^{2k}}\frac{(2k)!}{k!^2}$
Stirling's equivalent of $n!$ gives an asymptotic estimation for $n \to +\infty$

—

The preceding fact, concerning last return to the origin provided us with non intuitive behavior of the system, we would like to confirm the ability of this random walk to choose one of the opponent.

let's consider the proportion of path, with length $2n$, where a time equal to $2k$ has been passed on the positive side ($2(n-k)$ on the negative side).

Step Number 1 : Introducing proportion $\texttt{Fst-Ret}_{2k}$ of first return in $2k$, we observe :

$$\texttt{Pos}_{2k}^{2n} = \sum_{r=1\ldots n} \frac{1}{2}(\texttt{Fst-Ret}_{2r} \times \texttt{Pos}_{2(k-r)}^{2(n-k)} \; + \; \texttt{Fst-Ret}_{2r} \times \texttt{Pos}_{2k}^{2(n-k)})$$

Besides we have for "extremal values of $2k$" :

$$\texttt{Pos}_{2n}^{2n} = \texttt{Ret}_{2n} = \texttt{Pos}_{0}^{2n}$$

Step Number 2 : Finding the solution including the value $\texttt{Ret}_{2k}$ and $\texttt{Ret}_{2(n-k)}$ seems natural, a study of relation between $(\texttt{Ret}_{2k})_k$ and $(\texttt{Fst-Ret}_{2k})_k$ allows us to check that $\texttt{Pos}_{2k}^{2n} = \texttt{Ret}_{2k} \times \texttt{Ret}_{2(n-k)}$ gives a solution satysfying all observation made in *Step Number 1.* We can then conclude by induction, and show explicitely that this is the probability we are looking for.

Evaluation, asymptotic result As in the previous study, it is possible to expand the value of the probability using an equivalent given by Stirling's formulas. We find an asymptotic results including the Arcsine fonctions.

—

GENERAL CONCLUSION : Looking at the framework of probability where the only analytical method used is NUMBERING, we have been able to study some particular non deterministic system, and observe some non intuitive asymptotic behavior. A first step has been made to understand how probability theory made simple can provide one with new understanding.

Applied Probability, Set Number 2

Augustin Chaintreau

October 3, 2001

Abstract

This is a combination of treated case, example of application of probability to problem solving. This set follows chapter 3, and is made of application of independence and conditioning.

a Bernouilli trials, Binomial distribution + an asymptotic result

We are regarding a test (result being whether success or fail) where success is supposed to occur with probability p (fail with probability $q = (1-p)$. (Bernouilli trials).

Doing n successive test. Give the probability framework (sample space and probability), when test are assumed independent.

- Sample space are given by n successive choice whether success or fail, $\Omega = \{S, F\}^n$.
- an element $(\varepsilon_1, \ldots, \varepsilon_n)$, is being given the probability

$$p^{\#\{i=1..n|\varepsilon_i=S\}} q^{\#\{i=1..n|\varepsilon_i=F\}}$$

  usually written $p^k q^{n-k}$ (k being the number of success observed)

Imagine now we are doing simultaneously n tests, make the same work.

- Sample space is $\{1, \ldots, n\}$ (giving the number of success observed).
- Probability of results of n successive tests are given by their number of success, there is $\binom{n}{k}$ results of successive tests with k success, we then have :

$$P(\{k\}) = \binom{n}{k} p^k q^{n-k}$$

a.1 Binomial as limit of hypergeometric

Recall the expression of the hypergeometric probability :

$$P(\{m\}) = \frac{\binom{p}{m}\binom{n-p}{r-m}}{\binom{n}{r}} = \binom{r}{m}\frac{p\dots(p-m)\times(n-p)\dots(n-p-(r-m))}{n\dots(n-m)\times(n-(m+1))\dots(n-r)}$$

(sample of r elements, in a population of n, p of them being mutant).

Asymptotic : Can you give asymptotic to make hypergeometric a good approximation of the binomial case.

b Suspecting your ennemy

You play with a coins with one personn, he keeps winning for n times, what can you say assuming that it is not possible to cheat.

—

Now you are beginning to doubt from his sincerity. Being very pessimistic, and assuming that a cheater certainly win and never stop to cheat when he's playing, can you quantify his honnesty ? Is this possible to say anything without giving an abstract probability p that you are in front of a cheater ?

—

Finally you make the following experience, you come in a casino where the cheater are supposed to be one on a hundred people, you play with someone that keep winning, when should you stop the game, and call for the director ?

c Simple applications

Problem number 1 : Family case You are visiting one family where you know there is two children. What is the probability that this is two girls ?

- when you know that there is at least a girl ?
- when you know that the elder is a girl ?

Complementary questions : You are visiting one of your friend, knowing she has a brother and sister. As you do not know if she is the elder or the younger, can you conclude that all your information is that at least one children of the family is a girl ?

Problem number 2 : Prisoner's paradox Three revolutionnary are arrested and put into jail. The dictator is taking well care of his image, need to show strength as killing two personns, and to show some mercy in releasing one of the man
One of the prisoner asked to the guardian to give him the name of one that should die, the guardian, that do not want to tell him what will be his fate, as it's forbidden whether it is a good news or bad, tell the name of one of the other prisonner.

—

What is his chance to live ?

Problem number 3 : Several tossing and independence You are tossing three coins successively, we are considering the two following events.

- A is the event that head appears in the first tossing.
- B is the event that at least two heads appears.

We would like to deal with independence of event A and B :

- If we take the uniform probability on the sample space, what can you say ?
- Can you find ONE probability on the sample space that makes these events independent ? is this possible to have $0 < P(A) < 1$ and $0 < P(B) < 1$?
- We consider now that all probability we can put on the sample space should respect that different tossing are independent, the reason is that else it is considered to be absurd.

  Can you give in this case all the possible probability making A and B independent ?

Discrete Probability : Problems given

Augustin Chaintreau : augustin.chaintreau@ens.fr

First Semester (2001-2002), third year undergraduate, CMI Chennai

These problems should be given back to Mr. Sripathy on September, Tuesday 4th. Students will then come an take it. I will send you results and comments by mail, as soon as I can.

a Validate formulas on events probability : a method

$A_1, \ldots, A_n$ are events, $c_1, \ldots, c_n$ real numbers, and we define for any probability P :

$$I(P) = \sum_{k=1\ldots n} c_k P(A_k)$$

Show that following assertions are equivalent :

(i) $I(P) \geq 0$ for any probability P.

(ii) $I(P) \geq 0$ for any probability P with $P(A_k) \in \{0, 1\}$ for any k.

Hint : You are recommended to treat the case of disjoint events, or the case $n = 2$, prior to the general case.

Application : We want to prove the formula :

$$a = \sum_{k=1\ldots n} c_k P(A_k)$$

where $a, c_1, \ldots, c_n$ are real numbers, $(A_k)_k$ events, P a probability. Show that we can assume that $P(A_k)$ is 0 or 1 for each k.

Comment : It is possible to show that this still holds even if I is defined using not $(A_k)_k$ but any sequence of event $B_1, \ldots, B_m$ where each B_k is written from $A_1, \ldots, A_n$ using operations $\cap, \cup, ^c$. This extended result can be used in may ways, validating lot of different formulas (Poincaré, and much more).

As an example : the probability $\mathtt{Exac}_n^{(r)}$ that exactly r events in the family $A_1, A_2, \ldots, A_n$ occur is given by :

$$\texttt{Exac}_n^{(r)} = \sum_{k=0}^{n-r} (-1)^k \binom{r+k}{k} S_n^{(k)} \text{ where } S_n^{(k)} = \sum_{1<i_1<\ldots<i_k<n} P(A_{i_1} \cap \ldots \cap A_{i_k})$$

b Retrieving your coat after a meal

n personns are getting out of a restaurant, asking for their coat, as the waiter did not recognize anybody, he gave randomly anycoat to anybody.

- Provide this situation with correct sample space and probability.
- Let's S_n be the number of persons who retrieve their coat, give its expectation and variance.
- Show that the probability that S_n is at least 11 is smaller than 0.01, whatever may be the value of n ($n \geq 11$ will be of course assume in this question).

  remark Deviation theory is highly recommended for the last question.

c Applications of generating functions

c.1 Deformation ?

Do you think it is possible to create two fake dices (a dice where probability for the different numbers are not the same), that can be different, such that when we toss these two dices in the same time, the sum of the two numbers have a uniform distribution in $\{2, 3, 4, \ldots, 12\}$?

c.2 The mushroom picker

The number of mushrooms taken by one mushroom picker when he goes to the forest is N. This is a random variable whose distribution function is G.

Besides, every mushroom has a probability p that it can be eaten.

Show properly, with all the reasonnable independence assumption on the problem, that the probability that every mushroom picked can be eaten is $G(p)$.