HAL
open science

A Perfect Random generator : II

René Blacher

To cite this version:

René Blacher. A Perfect Random generator : II. 2009. hal-00443576

HAL Id: hal-00443576

https://hal.science/hal-00443576v1

Preprint submitted on 30 Dec 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Perfect Random Number Generator : II

René BLACHER

Laboratory LJK
Université Joseph Fourier
Grenoble
France

Summary : In this report one explicates the new method to generate random numbers whose the randomness is *proved*. One transforms data resulting from electronic files or provided by machines or software methods. This method can be applied directly in computers in the same way that the function "random". It can be also applied with the machines and the chips or software methods. In this report, one shows that ont one can use only the Fibonnacci functions. Moreover, one obtains new results about two other methods already obtained in a previous report.

Key Words : Central limit theorem, Or exclusive, Fibonacci sequence, Random numbers, Random noise, Higher order correlation coefficients.

NOTICE

This report has to be read in relation to the first report which we go published on this subject "A perfect random generator I " : cf [18]. Indeed the first report is very long. As a matter of fact, this new report is also a summary of the results of this first report with a simpler presentation and some new results.

Then, in order to read well these two reports, the best way is to read this report, and when it is necessary, to refer to the corresponding part of report I.

Contents

1 Introduction **5**
1.1 General presentation of the matter 5
1.1.1 Presentation of the result 5
1.1.2 Summary of the method 6
1.1.3 Definition of randomness 7
1.2 Presentation of the solution 7
1.2.1 Fundamental properties 8
1.3 Other Methods of construction 11
1.4 Conclusion 11

2 Quality of obtained sequences **13**
2.1 Criteria of randomness 13
2.1.1 Mathematical definitions 13
2.1.2 Statistical definitions 14
2.1.3 Use of random variables 15
2.1.4 Empirical properties 17
2.2 Comparison with the current generators 19
2.2.1 Various current techniques 19
2.2.2 Comparisons 21
2.3 Uses of these results 22

3 Cd-Rom of Marsaglia **24**
3.1 Theoretical study 24
3.1.1 Case of 2-dependence 24
3.1.2 Transformation of datas 25
3.1.3 Independence induced by the data 25
3.1.4 Conclusion 26

4 Basic properties **27**
4.1 Some properties 27

5 Dependence induced by linear congruences **30**
5.1 Theoretical study 30
5.1.1 Notations 30

5.1.2 Theorems . . . 31
5.2 Proof of theorem 1 . . . 33

6 Randomization by the functions of Fibonacci 39
6.1 Study of the problem . . . 39
6.1.1 Some Notations . . . 39
6.1.2 Sequence of real numbers regarded as IID . . . 40
6.1.3 Randomization of Y_n . . . 43
6.2 Empirical Probability . . . 44
6.3 Theoretical probability . . . 46
6.3.1 Two dimensional case . . . 48
6.3.2 General case . . . 51
6.3.3 Continuous density . . . 54
6.3.4 General numerical results . . . 58
6.3.5 Other congruences . . . 59
6.3.6 Remarks . . . 59
6.4 Study of models . . . 60
6.4.1 Continuous densities . . . 60
6.4.2 Another group of models . . . 61
6.4.3 General case . . . 61

7 Limit Theorems 64
7.1 Central Limit Theorem . . . 64
7.2 XOR Limit Theorem . . . 66
7.2.1 Presentation . . . 66
7.3 Examples . . . 67
7.3.1 Example using datas of this paper . . . 69
7.3.2 Other theoretical study . . . 69
7.3.3 Numerical study . . . 69
7.3.4 Rate of convergence in the XORLT . . . 72
7.3.5 Limit theorems for conditional probabilities . . . 78

8 Empirical Theorems 80
8.1 Empirical Theorems . . . 80
8.1.1 First theorems . . . 80
8.1.2 Applications . . . 82
8.1.3 Proof of theorem 8 . . . 82

9 Study of some files 87
9.1 Introduction . . . 87
9.2 Existence of satisfactory datas . . . 87
9.2.1 Definition . . . 87
9.2.2 Objections . . . 88
9.2.3 A finite random sequence . . . 88
9.2.4 Consequence 1 . . . 89
9.2.5 Consequence 2 . . . 89

9.3 Practical example . . . 90
9.3.1 Use of text . . . 90
9.3.2 Other data . . . 91
9.3.3 Conclusion . . . 91

10 Building of IID sequences : 1 **92**
10.1 General method . . . 92
10.1.1 Choice of data . . . 92
10.1.2 Description of the method . . . 92
10.1.3 Properties . . . 93
10.1.4 Example . . . 95
10.1.5 Continuous case . . . 97
10.1.6 Conclusion . . . 98

11 Building of an IID sequence : II **99**
11.1 General method . . . 99
11.1.1 Description of the method . . . 99
11.1.2 Explanation of the conditions about q_0 and r_0 . . . 101
11.2 Example . . . 102
11.2.1 Choice of random datas . . . 102
11.2.2 Building of a random sequence $b^1(n')$ . . . 102
11.2.3 Properties of $B^1(n')$ . . . 102
11.2.4 Tests . . . 103
11.2.5 Conclusion . . . 104
11.3 Continuous case . . . 104

12 Building of IID sequences : III **107**
12.1 Third method . . . 107
12.1.1 Method of construction of the sequence . . . 107
12.1.2 Properties . . . 108
12.1.3 Permutations and associated transformations . . . 108
12.1.4 Example . . . 109
12.1.5 Conclusion . . . 110
12.1.6 Comparison of methods II and III . . . 110

A Continous case in dimension 2 **111**

Chapter 1

Introduction

1.1 General presentation of the matter

In this report, we present a new method to obtain IID sequences x_n of random numbers [1]. This method can be used as well with machines as directly on a computer alone.

1.1.1 Presentation of the result

To have random number two methods exists :
1) Use of pseudo-random generators
2) Use of random noise.

These two methods have different defects.

1) For the best of them, the pseudo-random generators seem nondeterminist only during a certain time. This can be long enough for the cryptographic generators, but it is with the current means of calculations. Moreover in simulation, the pseudo-random generators must be tested for each application : cf [2] page 151.

2) If random noises are used, bias and dependences can appear : cf [3]. One tries to remove them by mathematical transformations. But these methods have defects. They remove bias and the linear correlation, but not necessarily the dependence.
On the other hand, these random noises can be produced by machines or chips. In this case, that thus require additional material which can suffer from malfunctions extremely difficult to detect : cf [1] page 3.

Now, for some applications, a maximum quality is essential (Nuclear power, medical, cryptography). It is thus necessary to have generators without defects.

[1] By abuse of language, we will call "IID sequence" (Independent Identically Distributed) the sequences of random numbers.

But, up to now *no completely reliable solution had been proposed* .

To set straight this situation, Marsaglia has created a Cd-Rom of random numbers by using sequences of numbers provided by Rap music. However, it does not have proved that the sequence obtained is really random.

However, there exists simple means of obtaining random sequences whose the quality is sure.

One can obtain perfect generators by using random noises, for example those produced by the machines or by software-based generators. In this case, one transforms these noises in a more effective way. Indeed, one uses assumptions much weaker than those of the current methods

One can also obtain perfect generators usable directly on computer (without the use of machines). In this case, one uses the electronic files as random noises (like Marsaglia uses Rap music). Then they are transformed by the same method that we use for the machines.

Then, *our technique can be applied with all the current methods.*

One can thus obtain sequences of real numbers which are **proved** random, which is a completely new result.

1.1.2 Summary of the method

Currently, when one uses random noise, bias and dependences are removed. In this aim, one supposes that theses noises check some assumptions. But, generally, those are not checked. Moreover, for each samples x_n there exists many possibles models X_n such that $x_n = X_n(\omega)$. That can be problematic.

Our method consists to transform random noises under very weak hypotheses : we assume only that theses noises are not completely deterministic.

Moreover our results are true for *all* logical models possible. That suppress the problem of the model. That allows also to satisfy the mathematical definitions of random numbers (these definitions are very difficult to etablish cf [1]).

Then, the obtained sequences will be always IID.

Now one can apply this method to many noises. So texts can be regarded as noises which satisfy these assumptions. It is also the case for numerous softwares which are recorded on computers : systems software for example.

Therefore, one can obtain directly IID sequences by transforming the files of computers. In this case, it is not necessary to use machines in order to have true random numbers.

On the contrary some electronics files can be studied logically. Then obtained numbers are surer than thoses obtained by machines which can have also malfunctions.

One can apply also our methods to noises furnished by machines, by chips, by mouse or by keyboard: of course, our results are much surer than those of

current methods.

1.1.3 Definition of randomness

To produce a really random sequence, it is thus necessary to have a definition of the randomness. It is a subject which was studied much. But, it is extremely complex. Philosophical questions are even involved. A summary of this study is in the book of Knuth [1] pages 149-183. One reminds some definitions in section 2.1. In fact, one will understand that no current mathematical definition is really satisfactory.

Though, one can think to define randomness by the following way.

Definition 1.1.1 *: One notes the approximation by $\approx$: for $x, y \in \mathbb{R}$, one sets $x \approx y$ if numerically x is nearly equal to y.*

Let L be the Lebesgue measure. A sequence $x_n \in [0,1]$ is said random if, for all Borel set Bo, for all n+1, if the past $x_1, x_2, , x_n$ is given, one cannot predict the place of x_{n+1} with a probability very different from that of the uniform distribution : $P_e\{x_{n+1} \in Bo|x_1, ..., x_n\} \approx L(Bo)$, where $P_e\{x_{n+1} \in Bo|x_1, ..., x_n\}$ is the empirical conditional probability of Bo when the past is given.

This type of definition is that which one wishes. Unfortunately, it has a defect : one does not have specified enough the approximation. On the one hand, the definition of $\approx$ is very undetermined mathematically. On the other hand, one would like a definition closer to the statistics definitions. But it is difficult to obtain such a definition : cf section 2.1.1.

But these questions of mathematical definition will not obstruct us because we will circumvent this problem by using sequences which are really samples of sequences of random variables X_n.

Unfortunately, an infinity of models X_n corresponds to the sequence x_n. Then, there is the problem of the choice of the model X_n. We will avoid this problem by proving that x_n behave as an IID sequence for *all the logical possible models.*

1.2 Presentation of the solution

Our method rests on a simple idea: to transform random noises by adapted transformations.

Like random noises, one can use those provided by the machines. It is what Vazirani, Neumann, Elias and others (cf [33] [4], [8]) wanted to do, but with too restrictive assumptions.

One can also use some electronic files. It is what Marsaglia did with the Rap music (cf [1], [20]). But he has transformed these data in a too elementary way

(cf chapter 3).

Then, one has sequences of random noises y_n : one can always assume $y_n = Y_n(\omega)$ with the following rule.

Notations 1.2.1 *When one has a sequence of real numbers which one can regard as one realization of a sequence of random variables, one will always note with small letters the data and with CAPITAL LETTERS the random variables which one will suppose defined on a probability space* (Ω, A, P).

When the y_n's mean random noises, to consider that $y_n = Y_n(\omega)$ is a traditional and normal assumption. It is also true for the y_n extracted from certain electronic files.

1.2.1 Fundamental properties

Into this section, we introduce the properties which are at the heart of our study. We will use the following notations.

Notations 1.2.2 *: The notation Ob(.) is that of the classical "O(.)" with the additional condition* $|Ob(1)| \leq 1$.

The sequences $j_1, j_2, ..., j_p$, $p \in \mathbb{N}^*$, *mean alway finite injective sequences* $j_s \in \mathbb{Z}$, *such that* $j_1 = 0$. *On the other hand, the sequences* $j'_1, j'_2, ..., j'_p$ *satisfy moreover* $0 = j'_1 < j'_2 < ..... < j'_p$.

The notation $P\big\{X_n \in Bo\big|x_2, ..., x_p\big\}$ *means always the conditional probability that the random variable* X_n *belongs to the Borel Set Bo given* $X_{n+j_2} = x_2$,.....,$X_{n+j_p} = x_p$.

Let $m \in \mathbb{N}^*$. *We set* $F(m) = \{0/m, 1/m, ..., (m-1)/m\}$ *and* $F^*(m) = \{0, 1, ..., m-1\}$. *We note by* μ_m *and* μ^*_m *the uniform measures on F(m) and* $F^*(m)$ *, respectively :* $\mu_m(k/m) = 1/m$.

Let X_G *be a random variable which has the distribution N(0,1) :* $X_G \sim N(0,1)$. *For all* $b > 0$, *we set* $\Gamma(b) = P\{|X_G| \geq b\}$.

Transformation of Fibonacci

Definition 1.2.3 *Let* fi_n *be the Fibonacci sequence :* $fi_1 = fi_2 = 1$, $fi_{n+2} = fi_{n+1} + fi_n$. *Let T be a congruence* $T(x) \equiv ax$ *modulo m such that there exists* $n_0 > 3$ *satisfying* $a = fi_{n_0}$ *and* $m = fi_{n_0+1}$. *Then T is said a Fibonacci's congruence with parameters a and m (or more simply m),*

Notations 1.2.4 *Let* $h \in \mathbb{Z}$ *and* $m \in \mathbb{N}^*$. *We define* $\overline{h}^m$ *by the following way*
1) $\overline{h}^m \equiv h$ *modulo m.*
2) $0 \leq \overline{h}^m < m$.

If the choice of m is obvious, we simplify $\overline{h}^m$ *into* $\overline{h}$.

Let $h \in F(m)$. *We define* $\overline{h}^1$ *by* $\overline{h}^1 = \overline{mh}^m/m$. *Often we simplify* $\overline{h}^1$ *into* $\overline{h}$.

In the same way, if the choice of m is obvious, and if T is a congruence : $T(x) \equiv ax + c$ modulo m, we set $\overline{T}(x) = \overline{T(x)}^m$.

The reduction of Fibonacci congruences to their first bits will be very useful for our study.

Definition 1.2.5 *Let $q, d \in \mathbb{N}^*$. Let T be the congruence of Fibonacci modulo m.*

We define the function of Fibonacci $T_q^d : F(m) \to F(d^q)$ by $T_q^d = Pr_q^d \circ \widehat{T}$ where
1) $\widehat{T}(x) = \overline{T}(mx)/m$
2) $Pr_q^d(z) = \overline{0, d_1 d_2 .... d_q}$ where $z = \overline{\overline{0, d_1 d_2 ...}}$ is the writing of z in base d.

If d=2, we simplify T_q^d in T_q and Pr_q^d in Pr_q.

To make IID by the functions of Fibonnacci

These functions T_q make independent sequences of random variables $Y_n \in F(m)$. Moreover, they make uniform their marginal distributions.

Now because Y_n is discrete, one can always regard $y_n \in F(m)$ as the realization of a sequence of random variables $Y_n : y_n = Y_n(\omega)$ such that Y_n has a differentiable density with respect to $\mu_m \otimes .... \otimes \mu_m$.

Moreover, assume that this density have a Lipschitz coefficient K_0 which is not too large. That is a logical assumption. As a matter of fact, that is an assumption which most mathematicians admit: that is especially clear when they estimate the densities (which they suppose to exist) when $N << m$ where N is the size of sample.

Now, the conditional probabilities $P\{Y_n | y_{n+j_2} = y_2, ..., Y_{n+j_p} = y_p\}$ have also a continuous density with a coefficient Lipschitz K_0^{cp} which is not too great. Then, one will prove that, for all interval I,

$$P\big\{T_q(Y_n) \in I \mid Y_{n+j_2} = y_2, ..., Y_{n+j_p} = y_p\big\} = L(I)\Big[1 + \frac{O(1)K_0^{cp}}{N(I)}\Big] \qquad (1.1)$$

where $N(I) = card\big\{k/m \;\big|\; k/m \in I,\ k \in \mathbb{N}\big\}$. For example, if $m \geq 2^{100}$, d=2, $q = 50$, $K_0^{cp} \leq 10$, then $P\{T_q(Y_n) \in I \mid Y_{n+j_2} = y_2, ..., Y_{n+j_p} = y_p\} = L(I)\Big[1 + \frac{O(1)10}{2^{50}}\Big]$.

We set $X_n = T_q(Y_n)$. A good choice of the parameters N, m, q will imply that

$$P\big\{X_n \in Bo \big| x_2, ..., x_p\big\} = L(Bo)[1 + Ob(1)\epsilon] \ ,$$

where $X_n = T_q(Y_n)$ and where $\epsilon \approx 0$, not only for intervals I, but also for all Borel sets Bo :

If ϵ is small enough with respect to N, the size of sample, X_n cannot be differentiated from an IID sequence : cf section 2.1.4.

Because the assumption that the Lipschitz coefficient K_0 is not too large is correct, we deduce that the sequence X_n behaves really as an IID sequence.

Now, we are considering a new situation : we are considering the set of all the possible probabilities for sequences Y_n, n=1,2,...,N. We provide it with a uniform probability, i.e. we want to know what occurs when the probabilities are randomly chosen.

Then, we shall prove in chapter 6 that, for all intervals I_s, s=,..,p, with a probability larger than $1 - 2p\Gamma(b)$ approximately, in this set of probabilities,

$$P\Big\{\{X_{n+j_1} \in I_1\} \cap .... \cap \{X_{n+j_p} \in I_p\}\Big\} = \frac{\prod_{r=1}^{p} N(I_r)}{m^p}\Big[1 + \frac{O(1).pb}{\sqrt{Inf_s\{N_{I_s}\}}}\Big]$$

For example suppose $m \geq 2^{100}$, $q = 50$, d=2, b=40. Moreover, we can assume $p \leq Log_2(N)$. Indeed, the following remark is used.

Remark 1.2.1 *One imposes $p \leq Log_2(N)$ because that does not have any meaning to consider the empirical dependence if $p > Log_2(N)$, e.g., if p=5, and if one has a sample of size 10, that has not meaning to study its dependence in $32 = 2^5$ cubes of width 1/2.*

Then, $1 - 2p\Gamma(b) \approx 1 - \frac{2p}{10^{340}}$ and $Inf_s\{N_{I_s}\} \approx 2^{50}$. Moreover, because $\frac{O(1).pb}{\sqrt{Inf_s\{N_{I_s}\}}} \approx \frac{O(1).pb}{2^{25}}$, one will not can differentiate X_n with an IID sequence. Therefore, X_n is IID with a probability larger than $1 - \frac{2Log(N)}{10^{340}}$ in the sets of all the possible models for the sequence y_n. Let us remark also that, if necessary, one can choose b much greater.

Let us notice that the result is thus true even for a very large number of bad model Y_n associated to y_n. It is pointed out that with a sample, one can always associate a certain number of correct models. The other models will be bad: for example a model AR(1) is a bad model for an IID sample. However our result is true with a probability of $1 - \frac{2p}{10^{340}}$. It is thus true even for an infinity of bad model of y_n: it is thus a strong result.

One can still refine this result : in many cases, for example if y_n is obtained from texts, the previous equation holds for all the logical models.

Therefore, X_n is IID for all the logical models of Y_n and even for an infinity of bad models.

That means that

1) X_n is IID in almost all the cases.

2) One avoids the least error in the estimate of K_0^{cp}. It is especially useful if one knows nothing a priori about y_n.

3) The functions T_q are functions which make a sequence IID with a great power.

4) One is sure that X_n is IID: there is no risk of error.

Also let us notice that thus one answers the problem of the definition of a random sequence : for every correct model or for almost all the models even bad Y_n, $P\{X_n \in Bo|x_2, ..., x_p\} = L(Bo)[1 + Ob(1)\epsilon]$, i.e. X_n cannot be differentiated with an IID sequence.

In conclusion there is really a method to obtain IID sequences X_n and this result is proved.

1.3 Other Methods of construction

The first method applies to data having Lipschitz coefficient which are not too large. It is thus better to use only data which check surely this assumption, for example text files. Then one uses the CLT which smoothes the probability very quickly and thus decreases the Lipschitz coefficient. As a matter of fact it is better to summon these data modulo m, which corresponds to new a theorem limit, the XORLT (XOR Limit Theorem) which produces a smoothing even faster.

There are then transformed data which have Lipschitz coefficients not too large. One thus applies the functions of Fibonnacci T_q to them. The choice of the parameters is carried out according to the results quoted previously.

The second method consists in standardizing data and then to apply a transformation to them which has characteristics rather similar to those of a permutation. After, the XORLT is used.

Indeed, in this case one is sure that one can apply results to the rate of convergence of the XORLT which one obtains in section 7.3.4. Indeed, that one is extremely fast. For example one obtains $P\{X_n \in Bo|x_2, ..., x_p\} = L(Bo)[1 + Ob(1)\epsilon$ where $\epsilon = 0(1/2^{50000})$!

We have concretely built IID sequences of real by using the method described here. This sequence can be asked to rene.blacher@imag.fr. Soon one will be able to obtain it in a website.

We carried out the traditional tests of Diehard with these sequences. All were checked : cf section 11.2.4.

1.4 Conclusion

The advantages compared to the current methods are clear:

1) It was **proven** that the numbers obtained are random.

2) There is not to test these numbers, especially in simulation where it had to be done for each new practical application.

3) The method is applicable directly on the computers: it is as easy as to use a function "random". Moreover, there does not need to add a machine or

an additional chip to the computer.

4) If one uses the random noises (Machines, chips, software programs), one removes all the dependence, which generally the current methods do not do. Moreover that can remove certain dysfunctions of the machines.

A more detailed comparison with the current methods is carried out in section 2.2.2.

Chapter 2

Quality of obtained sequences

2.1 Criteria of randomness

2.1.1 Mathematical definitions

To determine the quality of a generator, one needs a definition of the randomness of a sequence of real numbers x_n. Many studies were made to have reasonable definitions: there is a good summary of these studies in chapter 3-5 of Knuth : cf [1]. In this section 2.1.1, we summarize the study of Knuth.

The common wish when one tries to obtain random sequences, it is to obtain a sequences of real numbers x_n which can be regarded as a sample of an IID sequence of random variables X_n. Then, one could propose the following definition.

Definition 2.1.1 *: Let x_n, n=1,2,....,N, be a sequence of real numbers in [0,1]. Then, x_n is random if there exists an IID sequence of random variables $X_n \in [0,1]$ defined on a probability space (Ω, A, P) such that $x_n = X_n(\omega)$ where $\omega \in \Omega$.*

But there is a problem with this definition : for example, x_n can be increasing. Of course, it is possible only with a negligible probability. But it is possible. Then, Franklin proposed another definition.

Definition 2.1.2 *: Let x_n, n=1,2,....,N, be a sequence of real numbers in [0,1]. Then, x_n is random if it has each property that is shared by all samples of an IID sequence of random variables from uniform distribution.*

This definition is not precise and one could even deduce from it that no really random sequence exists (cf [1], Knuth page 149).

One must thus define differently what is a random sequence (or IID sequence). Also, the following definitions were introduced.

Definition 2.1.3 *: For all finite sequence of intervals $I_s \subset [0,1]$, we denote by P_e the empirical probability : $P_e = (1/N_4)\sum_{n=1}^{N_4} 1_{I_1}(x_n)1_{I_2}(x_{n+1}).....1_{I_p}(x_{n+p})$ where $N_4 = N - p$.*

The sequence $\{x_n\}$ is said p-distributed if $|P_e - L(I)| \leq N_4^{-1/2}$ for all $I = I_1 \otimes I_2 \otimes ... \otimes I_p$.

Definition 2.1.4 *The sequence x_n is random if it is p-distributed for all $p \leq Log_2(N_4)$.*

Unfortunately, this definition does not take into account the randomness of subsequences $x_{t_1}, x_{t_2}, ...x_{t_m}$. However, it is known that one cannot extend this definition to all the transformations $s \to t_s$ which define these subsequences : for example, this definition cannot be satisfied by the sequences x_{t_s} increasing. It is necessary thus that the application $s \to t_s$ is too not complicated. Also Knuth proposes the following definition.

Definition 2.1.5 *: The sequence x_n is random with respect to a set of algorithms A, if for all sequence $x_{t_1}, x_{t_2}, ...x_{t_m}$, determined by A, it is p-distributed for all $p \leq Log_2(N)$.*

These definitions summarize those given by Knuth, [1] page 108. In fact he has especially studied the infinite case. But because in practice, there are always samples of finite size, we are limited to this case.

This type of definition was the subject of many studies. In 1966, Knuth had thought that definition 3 defines the randomness perfectly: cf [1] page 163. It seems that he changed opinion since. In any case, none of these definitions is fully satisfactory. Knuth speaks philosophical debate on this subject. Thus, he points out that, according to certain principles, all the finite sequences can be regarded like determinist (cf pages 167-168 [1]).

2.1.2 Statistical definitions

Now, the definitions above are not satisfactory statistically. Indeed, it is known that if x_n is really an IID sample, $\frac{N^{1/2}(P_e - L(I))}{\sigma}$ has approximatively the normal distribution where σ is the variance associated to P_e. That means that it is possible that the event $|P_e - L(I)| > N^{-1/2}$ occurs. This property is thus different from the definition 2.1.5 de Knuth.

Therefore, one specifies statistically the definitions 2.1.4 and 2.1.5 in the following way.

Definition 2.1.6 *: Let $P'_e = (1/N_4)\sum_{n=1}^{N_4} 1_{Bo_1}(x_n)1_{Bo_2}(x_{n+1}).....1_{Bo_p}(x_{n+p})$ where the Bo_i's are Borel sets.*

It is said that x_n is random if, for all the sequences x_{t_s} defined by a set of algorithms A, for all suitable p, for all $Bo = Bo_1 \otimes ..... \otimes Bo_p$, it checks all the tests associated to $N^{1/2}|P'_e - L(Bo)|$ with the same frequency as a really IID sequence would do it.

By "same frequency", we understand that a really IID sequence will not check all the associated tests, but will check them only with a certain probability. For example, if all the tests to 1 percent was checked that would be abnormal.

Moreover in this definition, we use Borel sets rather than interval, because if not, there is an important gap (cf page 24 of [18]).

Remark that it is known that one will always find Borel sets which does not check tests of randomness even for a really IID sequence. This fact is not annoying: this case is envisaged by the use of the terms "with the same frequency ".

On the other hand, it is not obvious that one has forgets not any dependence in the previous definitions. Also, to avoid gaps, we introduced definition 1.1.1 which we remind now.

Definition 2.1.7 *: It is said that x_n is random if $P_e\{x_{n+1} \in Bo|x_1, ..., x_n\} \approx L(Bo)$, where $P_e\{x_{n+1} \in Bo|x_1, ..., x_n\}$.*

This definition seems a priori a good definition of the randomness. Indeed, it says that, knowing the past, one cannot predict the future with a probability too different from that of the uniform distribution. Intuitively, it is understood well that it is well the independence of the X_n which one defines thus.

Besides, it is this condition which one wishes for the random sequences in much books. However, in these books, one does not adopt this definition. Indeed, the definition 2.1.7 is imprecise : one does not have specified the approximation.

In fact, it is also the case in the definition 2.1.6 where one does not have specified the frequency. However, that will pose problems as for the definition of Franklin. It would thus be necessary to specify our definitions and to make a theoretical study.

However that will not be necessary because we have avoided this problem by using sequences which are really samples of random variables.

2.1.3 Use of random variables

Then, we use really random variables. It is this technique used with machines by Von Neumann, Vazirani, and others ones : cf [4], [33], [8]. They assume that x_n is the realization of a sequence (not IID) of random variables and they transforms $\{X_n\}$. But they obtain often the randomness under assumptions whose one is not sure that they are checked : in this case, x_n is provided by physical phenomena in machines generating random numbers. Unfortunately, the intruments of measurements distort the physical phenomenon and induce bias and dependences.

As a matter of fact, it is necessary to choose a correct model for X_n. But it is difficult.

At first, we know that one cannot choose definition 2.1.1 as definition of randomness. For example, an increasing sequence x_n can check $x_n = X_n(\omega)$

where X_n is IID (with a very negligible probability). As a matter of fact, this definition is not a problem solely when it is known a priori that the sequence X_n is IID like the case of a mechanical roulette or a mechanical lotto. In this case, one starts from a machine and one extracts a sample from it.

But this technique is not thus appropriate inevitably when one starts from a real sequence y_n (cf counterexample of increasing sequences).

To a sequence of real, it corresponds an infinity of models. Even if x_n can be regarded as a sample of an IID sequence X_n, it can be also logically regarded as the realization of an infinity of other models X_n^a (thus not IID).

The question thus should be asked: if one associates a model to a sequence x_n which criteria make that it possible to be sure that this model is correct? Generally, the following facts are admitted:

1) There never exists single model: a model is always related so that one wants to make of it.

2) Even when the goal is fixed, there are always several possible models, which all can be as valid the ones as the others.

Then how to be sure that a model is the good? That seems impossible.

A solution of the problem

In order to resolve the problem of the definition of the random sequence, one can transform them : $\{X_n^T\} = \mathcal{F}(\{X_n\})$ as Von Neumann, Vazirani, etc.

But we use transformations which have good properties on a *group of models*. Indeed, one can admit that some sets of model contain always a correct model.

For example, when all the x_n are different, one can admit that X_n has a differentiable density with a Lipschitz coefficient K_0 not too large. This hypothesis is usually admitted by the statisticians especially those which use functional estimate.

Under this assumption, the transformations defined by Fibonacci functions T_q have good properties. Indeed, if a sequence y_n has models with continuous density and a Lipschitz coefficient K_0, it will check

$$P\{T_q(Y_n) \in I \mid Y_{n+j_2} = y_2, ..., Y_{n+j_p} = y_p\} = L(I)\Big[1 + \frac{O(1)K_0}{N(I)}\Big] \ .$$

Then, one is sure that $T_q(Z_n)$ could be regarded reasonably as an IID sequence if K_0 is small enough and q not too large (i.e. N(I) great).

The problem of other models

Then a question is asked : if a model is correct and does not belong to the models with K_0 rather small, is what it will produce the same properties? If it produces another one, it will be a contradiction. There will be two possible

logical conclusions. It seems impossible. However, it is not obvious [1]. The problem of the choice of the definitions is found again.

A total answer

Now, by using the Fibonacci functions, one avoids the problem. In section 6.4.3, one proves mathematically that, *for almost all the models*, $T_q(X_n)$ behaves as an IID sequence . Indeed, one has

$$P\Big\{\{T_q(X_{n+j_1}) \in I_1\} \cap .... \cap \{T_q(X_{n+j_p}) \in I_p\}\Big\} = \frac{\prod_{r=1}^{p} N(I_r)}{m^p}\Big[1 + \frac{O(1).pb}{\sqrt{Inf_s\{N_{I_s}\}}}\Big]$$

It is a very satisfactory result. Indeed, it is wellknown that if one uses all the possible models without a priori, there will be an majority of bad models. Here, we find of it only a negligible number : it is already extraordinary.

Moreover, there is another result. One indeed finds that for some data, for example those resulting from texts, ALL the logical models associated with y_n will check (cf chapter 6.4)

$$P\{T_q(Y_n) \in I \mid Y_{n+j_2} = y_2, ..., Y_{n+j_p} = y_p\} = L(I)\big[1 + \epsilon\big] \ .$$

One could better wish with difficulty like results. It is a very strong result which resolves the problem of definition.

2.1.4 Empirical properties

We remark that in the previous equation, there remain ϵ. We will see now that it is not annoying if it is rather small with respect to N, the size of sample.

Choice of the parameters

One thus chooses the parameters q, m and N according to the sample size. In this paragraph, we will clarify this point.

Let us suppose that we have a really IID sequence with uniform distribution on [0,1/2] and [1/2,0] and with a probability such as $P\{[0, 1/2]\} = 0,501$. Then, this sequence has not the uniform distribution on [0,1]. However, if we have a sample with size 10, we will absoluetely not understand it. To understand this difference, one will need samples with size larger than 1000.

One will thus solve the problem of the choice of ϵ in the same way: according to N, the wished size of the sample, one will choose ϵ and thus T_q. Let us translate that mathematically.

[1] For example to say that, it would be to affirm that there is no characteristic of the English language such as there is a connection between the units $T^{-1}(I_k)$ and English texts : cf section 6.4

Let us note by P_e the empirical probability of an interval I associated with a sequence $x_n^* = X_n^*(\omega)$, n=1,2,....,N. Then, if X_n^* is a sequence of IID random variables with uniform distribution, if N is big enough,

$$P\{N^{1/2}|P_e - L(I)| > \sigma b\} \approx \Gamma(b) \ ,$$

where $\sigma^2 = L(I)[1 - L(I)]$.

Now, if X_n^* checks only $P\{X_n^* \in I|x_2^*, ..., x_p^*\} = L(I) + Ob(1)\epsilon$, one can prove that

$$P\{N^{1/2}|P_e - L(I)| > \sigma b\} \leq \Gamma\big\{b[1 - \eta(\epsilon)]\big\},$$

where $\eta(\epsilon) \geq 0$ and $\eta(\epsilon) \to 0$ as $\epsilon \to 0$.

For example, let us suppose that we built T_q so that $\eta(\epsilon) = 0.1$. In this case, for b=1,5

$$P\{N^{1/2}|P_e - L(I)| > \sigma.1.5\} \leq 0,134 \ \ under\ IID\ hypothsesis,$$

$$P\{N^{1/2}|P_e - L(I)| > \sigma.1.5\} \leq 0,148 \ \ if\ P\{X_n \in I|x_2, ..., x_p\} = L(I) + Ob(1)\epsilon.$$

However, it is known that if there is a really IID sequence, P_e is close to L(I) with a certain probability: it is completely possible that P_e is enough different from L(I), but the probability that occurs is weak.

Now, if $P\{X_n \in I|x_2, ..., x_p\} = L(I) + Ob(1)\epsilon$, it is also possible that P_e is enough different from L(I), but that is not likely much more to occur than in really IID case.

With such a result, it will be thus difficult to differentiate the x_n^* from a really IID sample.

Of course, if it is necessary, one can impose $\eta(\epsilon)$ smaller : for example, $\eta(\epsilon) = 0.01$. In this case,

$$P\{N^{1/2}|P_e - L(I)| > \sigma.1.5\} \leq 0,135 \ \ if\ P\{X_n \in I|x_2, ..., x_p\} = L(I) + Ob(1)\epsilon.$$

This type of result holds again for $I_1 \otimes .... \otimes I_p$ where the I_i's are intervals. Moreover, one obtains a similar result for the empirical conditional probability $P_e^C = P_e\{x_n \in I|x_2, ..., x_p\}$:

$$P\big\{N^{1/2}\big|P_e^C - L(I)\big| > b.\sigma_p^C\big\} \leq \Gamma\big(b[1 - \eta'(\epsilon)]\big),$$

where $\eta'(\epsilon) \to 0$ as $\epsilon \to 0$.

Case of Borel sets

Then, we can obtain $P\{X_n \in I|x_1, ..., x_p\} = L(I) + Ob(1)\epsilon$ for the intervals I, by using the properties of T_q. Then, for all Borel set Bo,

$$P\{X_n \in Bo|x_2, ..., x_p\} = L(Bo) + Ob(1)2^q\epsilon \ .$$

It is thus enough to choose q not too large and ϵ enough small so that $2^q\epsilon$ is also enough small.

Relations about $B^0(n')$

The previous results being true for all Borel sets, one deduced equivalents results about the bits $b^0(n')$ provided by the writing of x(n) bases 2 :

$$P\{B^0(n') = b|b_2, ..., b_p\} = 1/2 + Ob(1)\epsilon' \ .$$

In practice $\epsilon' = \frac{Ob(1)\alpha}{\sqrt{qN}}$ will be chosen where $\alpha \leq 0.02$ and where qN is the size of sequence $b^0(n')$.

Checking of definitions

It is thus proven that, that the model $B^0(n')$ built from ANY logical models of the data a(j) - or except maybe for a negligible minority (according to the case) - cannot be differentiated from a sequence of IID random variables.

In particular, it satisfies the properties

$$P\{B^0(n) = b \mid B^0(n + j_s) = b_s, s = 2, .., p\} = 1/2 + \frac{Ob(1)\alpha}{\sqrt{Nq}} \ ,$$

$$P\{N^{1/2}|P_e^C - 1/2)| > \sigma_p^C x\} \leq \Gamma\big(x[1 - \eta]\big) \ ,$$

which correspond theoretically and empirically to the definition 2.1.7 of the randomness.

It satisfies also

$$P\{N^{1/2}|P_e - 1/2^p| > \sigma_p x\} \leq \Gamma\big(x[1 - \eta]\big) \ ,$$

which corresponds empirically to the definition 2.1.6 of the randomness.

Then all definitions of a random sequences are satisfied : the sequence $b^0(n')$ cannot be differentiated from a sample of IID random variables.

2.2 Comparison with the current generators

2.2.1 Various current techniques

Generators using algorithms

1) Pseudo Random generator for simulation .
2) Pseudo Random generator for cryptography.
3) Irrational numbers : for example π and e.

For these generators, it is admited that it will never provide really random sequence : "Any one who considers arithmetical methods of producing random digits is, of course, in a state of sin" : John Von Neumann (1951).

These generators will certainly not check the definitions of randomness given by Knuth. One cannot thus mathematically regard them as random sequences. Then, the pseudo-random generators must be tested for each application : cf [2] page 151.

Generators using random noises

1) Hardware-based random bit generators : they exploit the randomness which occurs in some physical phenomena :e.g. quantum phenomena. They use machine or chips.
2) Processes upon which software random bit generators may be based include 1) the system clock; 2) elapsed time between keystrokes or mouse movement; 3) content of input/output buffers.

A true random bit generator requires a naturally occuring sources of randomness. Designing a hardware device or software program to exploit this randomness and produce a bit sequence that is free of biases and correlation is a difficult task. Moreover, random bit generators based on natural sources of randomness are subject to influence by external factors, and also to malfunctions. It is imperative that such devices be tested periodically (cf [3]). Moreover, it is impossible to reproduce calculations exactly a second time when cheking out a program.

The major defect of all these systems is that there can be correlations and bias in the generated sequences. The underlying physical process can be random. But there are many measuring devices between the digital part of the computer and the physical device. These intruments can thus introduce bias and correlations (cf [5] ch 17.14 Bias and Correlation).

One removes these bias and these linear correlations by various mathematical transformations like that of Neuman or Vazirani ([4] [33]). One can also use hash functions (cf 17.14, [5]).

However, the linear correlation are only one of the possible correlation. There exists correlations of higher order (quadratic cubic, etc : cf [10]). If there are bits, the correlation of higher order are the multilinear correlation between 3,4,5,..... bits. If one does not remove the correlation of higher order, one will not have independence.

Therefore, a priori the sequences of numbers built by the current methods to remove the linear correlation are not IID. It is a serious defect of the hardware device or software.

Tables of random numbers

One can obtain such tables by mechanical processes like the lotto or the roulette. They are the alone tables having results which are guaranteed IID.

But most of the time, these tables are obtained by the previous methods. They thus have the defects of them

In any case, these tables have a major defect: they are limited by their size.

A particular case is the CD-Rom of Marsaglia. Indeed, the random bits of this CD-ROM were made by combining music rap with sources of electronic white noise and the output of deterministic random number generators.

But, the randomness of the obtained sequence was not proved mathematically. In this report, we want to know logically if this sequence were random: cf section 3. This study shows that to have more certainty, it is necessary that these sequences are built by a certain way

That led us to take up the idea of Marsaglia: to regard certain electronic files as random noises, but to apply transformations a little more complex to them. That thus makes it possible to obtain true random numbers with a computer alone and to do without machines and chips. But, one can also use the numbers produced by machines

Conclusion

On none of the current generators there is certainty that the obtained sequences are random: that which approaches more this result is the Cd-Rom of Marsaglia.

However, much of users think that the provided generators are completely reliable and use them without precaution. All this already led to some scientific errors (cf [1] page 32). Thus, it is necesary to obtain a more reliable solution.

2.2.2 Comparisons

At first, it is **proved** that the obtained numbers by our method are really random. That had been obtained in no other method. Moreover,

A) Comparison with the pseudo-random generators

Our method thus brings obvious concrete advantages. In particular, in cryptography, there is no risks that the system can be broken. In simulation, there is not to test the numbers obtained.

Remind also that the usual opinion was that no generator built on computer is random. It is understood that it is an error : the truth is that no generator built by algorithm is random.

B) Comparison with generators based on natural sources of randomness

B-1) When one directly uses the program on a computer.

1) There does not need to add an additional machine to the computer.

2) There are no possible malfunctions as on the machines. Therefore, there is not to regularly test them like those.

3) The sequence obtained starting from the electronic files can be reproduced (it is useful for the checking of calculations).

B-2) When one uses the program on a source of random noises

1) That removes all the dependences, and maybe even certain effects of the malfunctions.

2) One can have very long sequences quickly (contrary to the methods using software).

C) Comparison with the CD-Rom of Marsaglia.

1) The results are proved.

2) The CD-ROM has a limited size.

3) A priori, it is possible that the sequence of the CD-Rom have defects.

2.3 Uses of these results

Direct programming on computer It is enough to transform certain files recorded on the computers. It is as simple to use as the function "random". Moreover, our generators are perfect. It is thus a method quite superior to the current generators.

Application to hardware devices One applies our transformations to data provided by machines or chips. That offers several advantages.

On the one hand any dependence is removed, (and not only linear correlations).

On the other hand, our method can be applied as soon as the data are not completely deterministic. It is certainly the case for the data provided by the machines maybe even if they have malfunctions.

It is thus a new method which one proposes to transform the noises provided by these machines. The advantage, it is that it needs extremely weak assumptions to be applied.

Application to software methods One can choose as data those provided by the software methods. As for hardware devices, our method can be applied under wery weak assumptions. However, it is simpler to use text files than the system clock for example.

Use of files of IID sequences By using our method, one can develops files of numbers which are proved IID.

They could thus be placed for public use in the form of files to download, of files recorded on hard disk, of DVD or of CD-Rom as it is the case for the CD-Rom of Marsaglia (cf Internet site [20]).

Transformations of $b^0(n')$ From sequences $b^0(n')$ which are proved random, one can obtain a multitude of others by using any sequences y_n provided by generators which are pseudo-random. Indeed, $\overline{b^0(n') + y_{n'}}$ modulo 2 is also IID (cf theorem 6).

Software for data external to the computer One can build softwares allowing to transform the majority of data external to the computer in random numbers, for example, texts.

Complete construction It is the matter to completely use the method of programming defined in this report with new data and choice of new parameters.

This method can be used when one wants, for various reasons, to obtain new sequences x_n completely reliable.

Combination of several methods If one wants to avoid any risk of human error, of machine's error, of computer's error or other ones, one can build several sequences $b^0(n')$ as described above in section 2.3. Indeed, if one summons modulo 2 : $b_n = \overline{\sum_{s=1}^{I} b_n^s}$, it is enough that only one sequence b_n^s is random so that b_n is it.

One will thus build them with different data. In this case, one can also use machines, even different machines. One can even employ the files of random numbers which exist over the world. That will reduce infinitely the probability of any potential error, human or different.

Chapter 3

Cd-Rom of Marsaglia

In this chapter, one will study the method which Marsaglia employed to create its CD-ROM. Marsaglia mixed digital tracks from rap and classical music selections. Then the random bits were made by combining three sources of electronic white noise with the output from a pseudo-random number generator. "They seem to pass all tests I have put to them – and I have some very stringent tests," Marsaglia says. Then, Marsaglia has studied his CD-Rom by using tests. But, it is possible to study it by logical reasoning.

In this section we give examples of such reasoning. In order to simplify we study only the case $my'_n = \overline{g_n + my_n} \in F^*(m)$ where g_n is a pseudo random sequence and where the y_n's derive from a text [1].

3.1 Theoretical study

3.1.1 Case of 2-dependence

In section 11.2, we understand that the data d(j) which we use can be regarded as 2 dependent. Then, we study now the case where y_n is 2-dependent. Suppose $j'_{s+1} > j'_s$ and $j'_{s_0+1} - j'_{s_0} \geq 2$. Then,

$$(y'_{n+j'_1}, y'_{n+j'_2}, ..., y'_{n+j'_p}) = \big((y'_{n+j'_1}, y'_{n+j'_2}, ..., y'_{n+j'_{s_0}})(y'_{n+j'_{s_0+1}}, ..., y'_{n+j'_p})\big),$$

where $(y_{n+j'_1}, y_{n+j'_2}, ..., y_{n+j'_{s_0}})$ and $(y_{n+j'_{s_0+1}}, ..., y_{n+j'_p})$ are independent.

Therefore, in order to study the dependence of y'_n, it is sufficient to study the case $j'_s = s - 1$, i.e. the $(y'_n, y'_{n+1}, ..., y'_{n+p-1})$. Now, one supposes $p \leq log(N)/log(2)$: cf remark 1.2.1. For example suppose $p \leq 22$. In this case, if the g_n's produce sequences where $(g_n, g_{n+1}, ..., g_{n+22})$ are independent, the

[1] Marsaglia has not used texts but Rap music. It is no important. We want only to study logically the method of Marsaglia. Then, we use texts because we studied them in a detailed way cf [18].

$(y'_n, y'_{n+1}, y'_{n+2}, ..., y'_{n+22})$ are independent. Therefore, one can consider that $(y'_{n+j'_1}, y'_{n+j'_2}, ..., y'_{n+j'_p})$ are also independent.

Then, it is enough to choose pseudo-random generators such that $(g_n, g_{n+1}, ..., g_{n+22})$ are independent. In this case, to suppose that the y'_n are independent will be a reasonable assumption.

Therefore the method of Marsaglia can be sufficient to obtain IID sequences if the parameters and the type of data have been suitably chosen.

However it remains to be checked that the marginal distributions are quite uniform: the tests of uniformity of the g_n means that some tests are checked, for example for intervals. But is this case for all Borel sets? It is similar for independence : they are independence for some hypercubes of the g_n: what is it for the others?

3.1.2 Transformation of datas

Now, one can use transformed data. It is what we do for sequence $c(j) \in F^*(32)$ defined in section 10.1.2 : we set $d(j) = \sum_{r=1}^{r_0} c(r_0(j-1)+r)32^{r-1}$.

Size of r_0 and conditional dependence

We choose again data resulting from texts. Then, if one finds a ". ", there is a strong probability so that it is followed by a "space character".

Therefore, it is possible that it has there some strong dependences between c(j) and c(j+e) (where c(j) are the letters modulo 32) especially for e=1. But this dependence decreases very quickly if e increases.

That will mean that the possible concentrations of d(j+1) given $d(j) = \sum_{r=1}^{r_0} c(r_0(j-1)+r)32^{r-1}$ will be less strong if r_0 increases. Indeed, suppose that $d(j-1)$ means a piece of text ending in a "." . Then, d(j) belongs to the set of the part of texts starting with a "space character".

Then, the behavior of y'_n depends on the choice of transformation and of parameters.

3.1.3 Independence induced by the data

We use again the example of ".". We note by p_o their numerical value. Let $z_t \in \{1, 2, ..., m\}$ be the value of successive "n" such as $y_n = p_o$, i.e. $y_{z_t} = p_o$. This sequence z_t is random : one can write $z_t = Z_t(\omega_5)$ where Z_t is a sequence of variable increasing in a random way, defined on a probability space $(\Omega_5, \mathcal{A}_5, Proba_5)$. Then, in order to obtain $my'_n = \overline{g_n + my_n}$, we add g_{z_t} to $y_{z_t} = p_o$.

In practice, we understand that $Z_{t+1} - Z_t$ is close to an IID sequence (not necessarily with uniform distribution). It is enough to make some numerical simulations to realize that (cf section 3.1.4 of [16].

This result means that $my'_{z_t} = \overline{g_{z_t} + my_{z_t}} = \overline{g_{z_t} + mp_o}$, has a behavior close to an IID sequence because g_{z_t} can be regarded as chosen randomly.

3.1.4 Conclusion

The previous study shows that one can improve the result by choosing better the parameters.

If they are well chosen, there is many reasons to think that y_n is IID. But we have not a certainty : that is difficult to specify mathematically. Maybe a thorough study would allow to arrive at certainties.

But it is simpler to use transformations T_q whose properties are appropriate well for the construction of an IID sequence and can be studied more easily. It is the aim of this report.

Chapter 4

Basic properties

4.1 Some properties

Let $X_n \in F(m)$ be a sequence of random variables. In this section we study some properties of conditional probabilites when $P\big\{X_n \in Bo|x_2, ..., x_p\big\} = L(Bo) + Ob(1)\epsilon$ for all Borel set Bo.

Proposition 4.1.1 *Let $Bo = Bo_1 \otimes Bo_2 \otimes ..... \otimes Bo_p$ be a Borel set of $F(m)^p$. Assume that, for all $s \in \{1, 2, ..., p\}$, for all sequence x_t, t=2,3,..,p, and for all $n \in \mathbb{N}^*$, $P\big\{X_n \in Bo_s|x_2, ..., x_p\big\} = L(Bo_s) + Ob(1)\epsilon$.*

Then, for all injective sequence $j_s \in \mathbb{Z}$ such that $j_1 = 0$,

$$P\Big\{\{X_{n+j_1} \in Bo_1\} \cap ..... \cap \{X_{n+j_p} \in Bo_p\}\Big\} = \Big[L(Bo_1) + Ob(1)\epsilon\Big]..... \Big[L(Bo_p) + Ob(1)\epsilon\Big] \ .$$

In order to prove this proposition the following lemma is needed

Lemma 4.1.1 *Let $Y_s \in F(m)$, s=1,2,...,N be a sequence of random variables defined over a probability space (Ω, A, P). Let $f \in L^1$ be a measurable function defined over $Y^-(\Omega)$ where $Y^- = (Y_1, Y_2, ..., Y_{n-1}, Y_{n+1}, ..., Y_N)$ and $n \in \{1, 2, ..., N\}$. Let Bo_1 be a Borel set of F(m).*

Assume $P\big\{Y_n \in Bo_1|y_1, ..., y_{n-1}, y_{n+1}, ..., y_N\big\} = L(Bo_1) + Ob(1)\epsilon$ for all $(y_1, ..., y_{n-1}, y_{n+1}, ..., y_N)$. Then,

$$E\big\{1_{Bo_1}(Y_n)f(Y^-)\big\} = L(Bo_1)E\big\{f(Y^-)\big\} + Ob(1)\epsilon E\big\{|f(Y^-)|\big\} \ .$$

Proof Let Q be the distribution of $(Y_1, Y_2, ..., Y_N)$ and let Q^- be the distribution of $(Y_1, Y_2, ..., Y_{n-1}, Y_{n+1}, ..., Y_N)$. Let $Q(.|y_1, ..., y_{n-1}, y_{n+1}, ..., y_N)$ be the distribution of Y_n given $Y_s = y_s$, for s=1,2,...,n-1,n+1,...,N. Let $y^- = (y_1, ..., y_{n-1}, y_{n+1}, ..., y_N)$. Then,

$$E\big\{1_{Bo_1}(Y_n)f(Y^-)\big\} = \int 1_{Bo_1}(y_n)f(y^-)Q(dy)$$

$$= \int \Big(\int 1_{Bo_1}(y_n)Q(dy_n|y_1, ..., y_{n-1}, y_{n+1}, ..., y_N)\Big)f(y^-)Q^-(dy^-)$$

$$= \int P\big\{Y_n \in Bo_1 | y_1, ..., y_{n-1}, y_{n+1}, ..., y_N)\big\} f(y^-) Q^-(dy^-)$$

$$= L(Bo_1) \int f(y^-) Q^-(dy^-) + \int Ob(1)\epsilon\big(y^-)\big) f(y^-) Q^-(dy^-),$$

where $|\epsilon\big(y^-)\big)| \leq \epsilon$. We deduce the lemma.

Proof 4.1.2 *We prove the proposition 4.1.1*

We use the lemma 4.1.1 with N=p, $X_{n+j_s} = Y_s$. Moreover, we choose $f(Y^-) = 1_{Bo_2}(Y_{n+j_2}).....1_{Bo_p}(Y_{n+j_p})$. Then,

$$P\Big\{\{X_{n+j_1} \in Bo_1\} \cap ..... \cap \{X_{n+j_p} \in Bo_p\}\Big\}$$

$$= \big(L(Bo_1) + Ob(1)\epsilon\big) P\Big\{\{X_{n+j_2} \in Bo_2\} \cap ..... \cap \{X_{n+j_p} \in Bo_p\}\Big\} .$$

Then, we prove the proposition by recurence. ■

Now, one obtains a similar resul about conditional probability.

Proposition 4.1.2 *Let Bo be a Borel set of $F(m)^p$, $Bo = Bo_1 \otimes ..... \otimes Bo_p$. Assume that $P\big\{X_n \in Bo_1 | x_2, ..., x_p\big\} = L(Bo_1) + Ob(1)\epsilon$. Then,*

$$P\Big\{X_n \in Bo_1 \Big| \{X_{n+j_2} \in Bo_2\} \cap ..... \cap \{X_{n+j_p} \in Bo_p\}\Big\} = L(Bo_1) + Ob(1)\epsilon.$$

Proof By using the proof 4.1.2 ,

$$P\Big\{X_{n+j_1} \in Bo_1 \Big| \{X_{n+j_2} \in Bo_2\} \cap ..... \cap \{X_{n+j_p} \in Bo_p\}\Big\}$$

$$= \frac{P\Big\{\{X_{n+j_1} \in Bo_1\} \cap \{X_{n+j_2} \in Bo_2\} \cap ..... \cap \{X_{n+j_p} \in Bo_p\}\Big\}}{P\Big\{\{X_{n+j_2} \in Bo_2\} \cap ..... \cap \{X_{n+j_p} \in Bo_p\}\Big\}}$$

$$= \frac{\Big(L(Bo_1) + Ob(1)\epsilon\Big) P\Big\{\{X_{n+j_2} \in Bo_2\} \cap ..... \cap \{X_{n+j_p} \in Bo_p\}\Big\}}{P\Big\{\{X_{n+j_2} \in Bo_2\} \cap ..... \cap \{X_{n+j_p} \in Bo_p\}\Big\}}$$

$$= L(Bo_1) + Ob(1)\epsilon. \blacksquare$$

The proof of the following theorem is a consequence of proposition 4.1.1.

Proposition 4.1.3 *The sequence X_n, n=1,2,.....,N, is IID if and only if, for all $p \in \{1, 2, ..., N-1\}$, for all sequence j_s, for all $n \in \mathbb{N}^*$, for all Borel set Bo, for all sequence x_s, s=1,...,p*

$$P\Big\{X_n \in Bo \Big| X_{n+j_2} = x_2, ..., X_{n+j_p} = x_p\Big\} = L(Bo) \ .$$

Chapter 5

Dependence induced by linear congruences

We study in this chapter the dependence induced by $\left(T^n(x_0), T^{n+1}(x_0)\right)$ when T a congruence $T(x) \equiv ax$ mod(m) with $0 < a < m$ and where a and m are fixed. Then, we study the set $E_2 = \left\{\ell, \overline{T(\ell)} \mid \ell \in \{0, 1, ..., m-1\}\right\}$.

5.1 Theoretical study

5.1.1 Notations

We will understand that this dependence depends on the continued fraction $\frac{m}{a}$, i.e. it depends on sequences r_n and h_n defined in the following way.

Notations 5.1.1 *Let $r_0 = m$, $r_1 = a$. One denotes by r_n the sequence defined by $r_n = h_{n+1}r_{n+1} + r_{n+2}$ the Euclidean division of r_n by r_{n+1} when $r_{n+1} \neq 0$.*

One denotes by d the smallest integer such as $r_{d+1} = 0$. One sets $r_{d+2} = 0$.

Therefore, $h_n \geq 1$ for all n=1,2,...,d and $r_{d-1} = h_d r_d + r_{d+1} = h_d r_d + 0 = h_d r_d$.

The full sequence r_n is thus the sequence $r_0 = m$, $r_1 = a$, ..., $r_{d+1} = 0$, $r_{d+2} = 0$. Then, it is easy to prove the following result.

Proposition 5.1.1 *The congruence $T(x) \equiv a$ mod(m) is a Fibonacci conguence if $h_n = 1$ for n=1,2,...,d, $h_d = 2$ and $r_d = 1$*

In this case, r_n is the Fibonacci sequence fi_n, except for the last terms. In addition, one considers also the following sequences.

Notations 5.1.2 *One sets $k_0 = 0$, $k_1 = 1$ and $k_{n+2} = h_{n+1}k_{n+1} + k_n$ if $n + 1 \leq d$.*

Remark that if $h_n = 1$ for n=1,2,...,d-1, k_n is also the Fibonacci sequence for n=1,2,...,d.

5.1.2 Theorems

One will understand that dependence depends on the h_i: more they are small , more the dependence is weak. As $h_i \geq 1$, the best congruence will satisfy $h_i = 1$ and $h_d = 2$. It will be thus the congruence of Fibonacci.

Theorem 1 *Let $n \in \{2, 3...., d\}$. Then*

If n is even , $E_2 \cap \big\{[0, k_n[\otimes [0, r_{n-2}[\big\} = \big\{(k_{n-1}\ell\ ,\ r_{n-1}\ell) \big| \ell = 0, 1, 2, ..., h_{n-1}\big\}$. Moreover the points $(k_{n-1}\ell\ ,\ r_{n-1}\ell)$ are lined up.

If n is odd,
$E_2 \cap \big\{]0, k_n] \otimes]0, r_{n-2}]\big\} = \big\{(k_{n-2} + k_{n-1}\ell\ ,\ r_{n-2} - r_{n-1}\ell) \big| \ell = 0, 1, 2, ..., h_{n-1}\big\}$. *Moreover, the points $(k_{n-2} + k_{n-1}\ell\ ,\ r_{n-2} - r_{n-1}\ell)$ are lined up.*

That means that the rectangle $[0, k_n/2] \otimes [r_{n-2}/2, r_{n-2}[$ does not contain points of E_2 if n is even : $E_2 \cap \big\{[0, k_n/2] \otimes [r_{n-2}/2, r_{n-2}[\big\} = \emptyset$. If h_{n-1} is large, that will mean that an important rectangle of $\mathbb{R}^2$ is empty of points of E_2: that will mark a breakdown of independence.

Of course, one has equivalent results for rectangles modulo m : $R_0 = \overline{R^0}$ where $R^0 = \big\{[x_0, x_0 + k_n] \otimes [y_0, y_0 + r_{n-2}[\big\}$.

For example suppose n=2. Then, one has a wellknown result. Indeed, $m = r_0$, $r_1 = a$, $k_1 = 1$ and $k_2 = h_1 = \lfloor m/a \rfloor$ ou $\lfloor x \rfloor$ means the integer part of x. Thus, the rectangle $Rect_2 = [0, m/(2a)] \otimes [m/2, m[$ will not contain any point of E_2. However, this rectangle has its surface equal to $m^2/(4a)$. Thus if "a" is not sufficiently large, i.e if h_1 is too large, there is breakdown of independence.

Principal theorem

Now, one takes in account the number of points of E_2 contained in rectangles of the type $R_{ect} = [x, x + L[\otimes [y, y + L'[$.

Theorem 2 *It is supposed that T is invertible. Let R_{ect} be a rectangle of $F^*(m)^2$, length $L_{on} \geq 1$, width $L_{ar} \geq 1$. Let $N(R_{ect})$ be the number of points of E_2 which belong to R_{ect} and let $S_{R_{ect}}$ be its surface. One denotes by Log the Neperian logarithm : Log(e)=1. Then,*

$$\big|N(R_{ect}) - \frac{S_{R_{ect}}}{m}\big| \leq (p^o + 1)[sup(h_i) + 1]\ ,$$

where p^o is a function of (L_{on}, L_{ar}) satisfying $2.0782.Log(m_{in}) + 2.00005 \geq p^o$ where $m_{in} = Min(L_{on}, L_{ar})$.

The proof is page 135 of [18] (cf also [13]).

This theorem shows that if $sup(h_i)$ is small, the only rectangles where there is maybe breakdown of independence are the rectangles of the type $R^0 = [x, x + k_n[\otimes[y, y + r_{n-2}[$: cf page 135-140 of [18] .

Then, these rectangles do not contain enough points to make tests if h_i is small. If $y = \overline{T}(x)$ the breakdown with independence is proved by theorem 1 : there is $h_{n-1} + 1$ lined up points. If h_i is small, it is easy to understand that it is not important.

Thus, in the case of the Fibonacci sequence, all rectangles satisfy the test of normality. In fact, it is even statistically too. It is not important. We do not make use of it like sample of independent couple.

Numerical examples

We confirm by graphs the previous conclusion. We suppose m=21. If a = 13, we have a Fibonacci congruence : cf figure 5.1. If one chooses a=10, $sup(h_i) = 20$: cf figure 5.2 . If one chooses a=5, $sup(h_i) = 5$: cf figure 5.3.

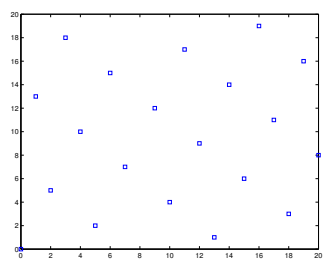

Figure 5.1: $sup(h_i) = 1$

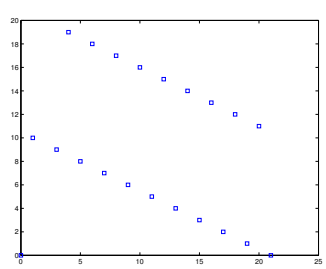

Figure 5.2: $sup(h_i) = 20$

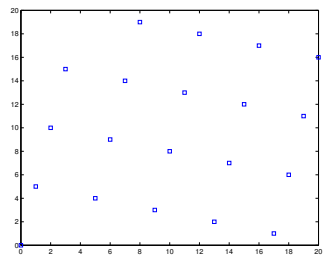

Figure 5.3: $sup(h_i) = 5$ Fig

Conclusion

To avoid any dependence, it is necessary that $sup(h_i)$ is small. In the case of the Fibonacci congruence, independence is checked on all rectangles R_{ect}.

Remark 5.1.1 *For the Fibonacci congruence $T^2 = \pm Id$ where Id is the identity (cf page 141 of [18]) One cannot thus apply it to create directly a pseudo-random sequence.*

5.2 Proof of theorem 1

In this section, the congruences are conguences modulo m. Now the first lemma is obvious.

Lemma 5.2.1 *For n=3,4,...,d+1, $k_{n+1} > k_n > k_{n-1}$. Moreover $k_{n+2} = h_{n+1}k_{n+1} + k_n$ is the Euclidean division of k_{n+2} by k_{n+1}.*

Now, we prove the following results.

Lemma 5.2.2 *Let n=0,1,2,...,d. If n is even, $\overline{k_n a} = m - r_n$. If n is odd, $\overline{k_n a} = r_n$.*

Proof : We prove this lemma by recurrence. For n=0, $\overline{k_n a} = \overline{0} = 0 = m - m = m - r_0$. For n=1, $\overline{k_n a} = \overline{a} = a = r_1$.

We suppose that it is true for n.
One supposes n even. Then, $k_{n+1}a \equiv ah_nk_n + ak_{n-1} \equiv -h_nr_n + r_{n-1} = r_{n+1}$.
One supposes n odd. Then, $k_{n+1}a \equiv ah_nk_n + ak_{n-1} \equiv h_nr_n - r_{n-1} = -r_{n+1} \equiv m - r_{n+1}$. Therefore, $\overline{k_{n+1}a} = m - r_{n+1}$. ■

Lemma 5.2.3 *Let n=2,3,...,d+1. Let $t \in \{1, 2, ..., k_n - 1\}$. If $n \geq 2$ is even, $r_{n-1} \leq \overline{at} < m - r_n$. If $n \geq 3$ is odd, $m - r_{n-1} \geq \overline{at} > r_n$.*
Moreover, if $n \geq 2$ is even, $\overline{k_n a} = m - r_n$. If $n \geq 3$ is odd, $\overline{k_n a} = r_n$.

Proof : The second assertion is lemma 5.2.2. Now, we prove the first assertion by recurrence.

One supposes n=2. Then, $m = r_0 = h_1 r_1 + r_2 = h_1 a + r_2$. Moreover, $k_2 = h_1$. If $1 \leq t < h_1 = k_2$, $r_1 = a \leq at < h_1 a = m - r_2$.

If $h_1 = k_2 = 1$, $\{1, 2, ..., k_2 - 1\} = \emptyset$. In this case, we study $t \in \{1, 2, ..., k_3 - 1\}$ where $k_3 = h_2 k_2 + k_1 = h_2 + 1$. Then, $1 \leq t \leq h_2$. Then, $at \equiv tak_2 \equiv -tr_2$.

Moreover, $m - r_2 \geq m - tr_2 \geq m - h_2 r_2 = r_0 - h_2 r_2 = r_0 - (r_1 - r_3) = r_3 + (r_0 - r_1) > r_3$.
Therefore, because $at \equiv m - tr_2$, $\overline{at} = m - tr_2$.
Therefore, $m - r_2 \geq \overline{at} > r_3$.

One supposes that the first assertion is true for n where $2 \leq n \leq d$.

Let $0 < t' < k_{n+1}$. Let $t' = fk_n + e$ be the Euclidean division of t' by k_n : $e < k_n$.

Then, $f \leq h_n$. If not, $t' \geq (h_n + 1)k_n + e \geq h_n k_n + k_{n-1} = k_{n+1}$.

One supposes n even.
In this case, $r_{n-1} \leq \overline{at} < m - r_n$ for $t \in \{1, 2, ..., k_n - 1\}$.
Moreover, $at' \equiv fak_n + ae \equiv f(m - r_n) + ae \equiv -fr_n + ae$.

First, one supposes $e = 0$. Then, $f \geq 1$.
Moreover, because $n \geq 2$, $m - r_n \geq m - fr_n \geq m - h_n r_n = m - (r_{n-1} - r_{n+1}) = r_0 - r_{n-1} + r_{n+1} \geq r_0 - r_1 + r_{n+1} > r_{n+1}$.
Therefore, because $at' \equiv -fr_n$, $\overline{at'} = m - fr_n$.
Therefore, $m - r_n \geq \overline{at'} > r_{n+1}$.

Now, one supposes $f < h_n$ and $e > 0$.
By recurrence, $m - r_n \geq \overline{ae} \geq \overline{ae} - fr_n \geq r_{n-1} - fr_n \geq r_{n-1} - (h_n - 1)r_n = r_n + r_{n+1} > r_{n+1}$.
Therefore, because $at' \equiv -fr_n + ae$, $\overline{at'} = \overline{ae} - fr_n$.
Therefore, $m - r_n \geq \overline{at'} > r_{n+1}$.

One supposes $f = h_n$, $e \neq k_{n-1}$ and $e > 0$.
If $e \neq k_{n-1}$, $\overline{ae} \neq \overline{k_{n-1}a}$. Indeed, if not, $\overline{a(e - k_{n-1})} = 0$. For example, if $e - k_{n-1} > 0$, $k_n > e - k_{n-1} > 0$. Then, because our recurence, $\overline{a(e - k_{n-1})} > r_{n-1} > 0$: it is impossible.

Now, if $n = 2$, $\overline{k_{n-1}a} = \overline{k_1 a} = \overline{a} = r_1 = r_{n-1}$.
Moreover, if $n > 2$, $n \geq 4$. Then, by recurence $\overline{k_{n-1}a} = r_{n-1}$.
Then, if $e \neq k_{n-1}$, $\overline{ae} \neq \overline{k_{n-1}a} = r_{n-1}$. Then, $\overline{ae} > r_{n-1}$.

Moreover, $m - r_n \geq \overline{ae} \geq \overline{ae} - fr_n > r_{n-1} - fr_n \geq r_{n-1} - h_n r_n = r_{n+1}$.
Therefore, because $at' \equiv -fr_n + ae$, $\overline{at'} = \overline{ae} - fr_n$.

Therefore, $m - r_n \geq \overline{at'} > r_{n+1}$.

One supposes $f = h_n$ and $e = k_{n-1}$. Then, $t' = h_n k_n + k_{n-1} = k_{n+1}$. It is opposite to the assumption.

Then, in all the cases, for $t' \in \{1, 2, ..., k_{n+1} - 1\}$, $m - r_n \geq \overline{at'} > r_{n+1}$. Therefore, the lemma is true for n+1 if n is even. Then, it is also true for n+1=3.

One supposes n odd with $n \geq 3$. One proves the recurrence by the same way as if n is even. Then the lemma is true for n+1. ■

Lemma 5.2.4 *The following inequalities holds :* $k_{d+1} \leq m$.

Proof If $t \in \{1, 2, ..., k_{d+1} - 1\}$, by lemma 5.2.3, $r_d \leq \overline{at} < m - r_{d+1}$ or $m - r_d \geq \overline{at} > r_{d+1}$, i.e. $r_d \leq \overline{at} < m$ or $m - r_d \geq \overline{at} > 0$ where $r_d > 0$. Then, $0 < \overline{at} < m$ or $m > \overline{at} > 0$.

Then, if $k_{d+1} > m$, there exists $t_0 \in \{1, 2, ..., k_{d+1} - 1\}$ such that $t_0 = m$, i.e. $\overline{at_0} = \overline{am} = 0$. It is impossible. ■

Lemma 5.2.5 *Let* $t,\ t' \in \{1, 2, ..., k_{d+1} - 1\}$ *such that* $\overline{at} = \overline{at'}$. *Then, t=t'.*

Proof Suppose $t > t'$. Then, $a(t - t') \equiv 0$ and $\overline{a(t - t')} = 0$. Then, by lemma 5.2.3, $r_d \leq \overline{a(t - t')} < m - r_{d+1}$ or $m - r_d \geq \overline{a(t - t')} > r_{d+1} = 0$ where $r_d > 0$. Then, $0 < \overline{a(t - t')}$. It is a contradiction. ■

Lemma 5.2.6 *Let n=1,2,...,d. Let* $H_n = h_1 k_1 + h_2 k_2 + h_3 k_3 + ... + h_n k_n$. *Then,* $H_n = k_{n+1} + k_n - 1$

The proof is basic.

Lemma 5.2.7 *Let n=1,2,3,...,d-1 . Let* $L_n = \left\{t \middle| t = 0, 1, 2, ..., H_n\right\}$. *Then, for all* $n \geq 1$, $L_{n+1} = \left\{t = l + gk_{n+1} \middle| l \in L_n, g \leq h_{n+1}\right\}$.

Proof Let $l \in L_n$, $l \leq H_n$. Let $g \leq h_{n+1}$. Therefore, if $t = l + gk_{n+1}$, $t \leq H_n + h_{n+1} k_{n+1} = H_{n+1}$. Therefore, $\left\{t = l + gk_{n+1} \middle| l \in L_n, g \leq h_{n+1}\right\} \subset L_{n+1}$.

Reciprocally, let $t \in L_{n+1}$ and let $t = fk_{n+1} + e$, $e < k_{n+1}$ be the Euclidean division of t by k_{n+1}.

We know that $H_n = k_{n+1} + k_n - 1 \geq k_{n+1}$. Therefore, $e \leq H_n$. Therefore, $e \in L_n$.

Therefore, if $f \leq h_{n+1}$, $t = fk_{n+1} + e \in \left\{t = l + gk_{n+1} \middle| l \in L_n, g \leq h_{n+1}\right\}$.

Moreover, if $f > h_{n+1} + 1$, $t = fk_{n+1} + e \geq (h_{n+1} + 2)k_{n+1} + e \geq h_{n+1}k_{n+1} + 2k_{n+1} = H_{n+1} - H_n + 2k_{n+1} = H_{n+1} - k_{n+1} - k_n + 1 + 2k_{n+1} = H_{n+1} + k_{n+1} - k_n + 1 \geq H_{n+1} + 1$. Therefore, $t \notin L_{n+1}$.

Then, suppose $f = h_{n+1} + 1$. Then, $t = fk_{n+1} + e = (h_{n+1} + 1)k_{n+1} + e = h_{n+1}k_{n+1} + k_{n+1} + e = H_{n+1} - H_n + k_{n+1} + e = H_{n+1} - k_{n+1} - k_n + 1 + k_{n+1} + e = H_{n+1} - k_n + 1 + e$.
Because $t \in L_{n+1}$ and $t = H_{n+1} - k_n + 1 + e$, $e + 1 - k_n \leq 0$. Therefore, $e \leq k_n - 1$.
Therefore, $t = fk_{n+1} + e = h_{n+1}k_{n+1} + k_{n+1} + e$,
where $k_{n+1} + e \leq k_{n+1} + k_n - 1 = H_n$
Therefore, $t = h_{n+1}k_{n+1} + e'$ where $e' \leq H_n$.
Therefore, $t \in \left\{t = l + gk_{n+1} \middle| l \in L_n, g \leq h_{n+1}\right\}$.
Therefore, $L_{n+1} \subset \left\{t = l + gk_{n+1} \middle| l \in L_n, g \leq h_{n+1}\right\}$.

Therefore, $L_{n+1} = \left\{t = l + gk_{n+1} \middle| l \in L_n, g \leq h_{n+1}\right\}$. ■.

Lemma 5.2.8 *Let* $F_n = \left\{\overline{at} \middle| t = 0, 1, 2, ..., H_n\right\}$.
Let $E_n = \left\{\overline{at} + km \middle| t = 0, 1, 2, ..., H_n, k \in \mathbb{Z}\right\}$. *We set* $E_n = \{o_s^n | s \in \mathbb{Z}\}$ *where* $o_0^n = 0$ *et* $o_{s+1}^n > o_s^n$ *for all* $s \in \mathbb{Z}$.
Then, for all $s \in \mathbb{Z}$, $o_{s+1}^n - o_s^n = r_n$ *or* $o_{s+1}^n - o_s^n = r_{n+1}$.

Proof We prove this lemma by recurrence.
Suppose n=1. Then, $r_1 = a$, $H_1 = h_1k_1 = k_2 = h_1$. Therefore,
$F_1 = \left\{\overline{at} \middle| t = 0, 1, 2, ..., h_1\right\} = \left\{0, a, 2a, ..., h_1a\right\} = \left\{0, r_1, 2r_1, ..., h_1r_1 = m - r_2\right\}$.
Therefore, the lemma is true for n=1.

Suppose that the lemma is true for n.
Then, $E_{n+1} = \left\{\overline{at} + km \middle| t = 0, 1, 2, ..., H_{n+1}, k \in \mathbb{Z}\right\}$, where $H_{n+1} = h_1k_1 + h_2k_2 + h_3k_3 + ... + h_{n+1}k_{n+1} = H_n + h_{n+1}k_{n+1}$.

Because $t \in \{0, 1, 2, ..., H_{n+1}\}$, $t \in L_{n+1}$. By lemma 5.2.7, si $t \in L_{n+1}$, $t = l + gk_{n+1}$ where $g \leq h_{n+1}$. By lemma 5.2.2, $\overline{at} \equiv \overline{a(l + gk_{n+1})} \equiv \overline{al} + (-1)^{n+2}gr_{n+1} \equiv \overline{al} + (-1)^n gr_{n+1}$.
Therefore,

$$E_{n+1} = \left\{\overline{at} + km \middle| t \in L_{n+1}, k \in \mathbb{Z}\right\}$$

$$= \left\{\overline{at} + km \middle| t = l + gk_{n+1}, l \in L_n, g \leq h_{n+1}, k \in \mathbb{Z}\right\}$$

$$= \left\{\overline{al} + (-1)^n gr_{n+1} + km \middle| l \in L_n, g \leq h_{n+1}, k \in \mathbb{Z}\right\}$$

$$= \{f + (-1)^n gr_{n+1} + km \big| f \in F_n, g \leq h_{n+1}, k \in \mathbb{Z}\}$$
$$= \{o_s^n + (-1)^n gr_{n+1} + km \big| s \in Z, g \leq h_{n+1}, k \in \mathbb{Z}\}.$$

Suppose that n is even.
Then, $o_s^n + (-1)^n gr_{n+1} = o_s^n + gr_{n+1} \leq o_s^n + r_n - r_{n+2}$ because $gr_{n+1} \leq h_{n+1}r_{n+1} = r_n - r_{n+2}$.

Use the recurrence. Suppose $o_{s+1}^n - o_s^n = r_n$. Then, $o_s^n + (-1)^n gr_{n+1} \leq o_s^n + r_n - r_{n+2} = o_{s+1}^n - r_{n+2}$.
Therefore,
$\{o_t^{n+1} \mid o_s^n \leq o_t^{n+1} < o_{s+1}^n\} = \{o_s^n < o_s^n + r_{n+1} < .... < o_s^n + h_{n+1}r_{n+1} < o_{s+1}^n\}$.
Therefore, $o_{t+1}^{n+1} - o_t^{n+1} = r_{n+1}$ or r_{n+2} if $o_s^n \leq o_t^{n+1} < o_{t+1}^{n+1} \leq o_{s+1}^n$.

Suppose $o_{s+1}^n - o_s^n = r_{n+1}$. Then, s is fixed .
Let $T = min\{t = 0, 1, ..., \ |o_{s+t+1}^n - o_{s+t}^n = r_n\}$. Therefore, $o_{s+T+1}^n - o_{s+T}^n = r_n$.
Let $O = \cup_{t=0}^T \{o_{s+t}^n + gr_{n+1} \mid 0 \leq g \leq h_{n+1}\}$.
Then, $O = \{o_s^n, o_{s+1}^n, ..., o_{s+T-1}^n\} \cup \{o_{s+T}^n + gr_{n+1} | \ 0 \leq g \leq h_{n+1}\}$.
Therefore, $O = \{o'_s, o'_{s+1}, ..., o'_{s+K}\}$ where $o'_{s'+1} - o'_{s'} = r_{n+1}$. Moreover, $o_{s+T+1}^n - o'_{s+K} = r_n - h_{n+1}r_{n+1} = r_{n+2}$.
Therefore, if $o_{t'}^{n+1}$ and $o_{t'+1}^{n+1} \in \{o_t^{n+1} \mid o_s^n \leq o_t^{n+1} \leq o_{s+T+1}^n\}$, $o_{t'+1}^{n+1} - o_{t'}^{n+1} = r_{n+1}$ or r_{n+2}.

Suppose that n is odd. One proves this result by the same way as when n is even. ■

Proof 5.2.9 *Now one proves theorem 1.*

Suppose that n is even.
Then, $\overline{k_{n-1}a} = r_{n-1}$, $2\overline{k_{n-1}a} = 2r_{n-1}$, ...$h_{n-1}\overline{k_{n-1}a} = h_{n-1}r_{n-1} = r_n - r_{n-2}$.
Now, $\overline{ak_{n-1}\ell} = \overline{\ell r_{n-1}} = \ell r_{n-1}$ for $\ell = 0, 1, 2, ..., h_{n-1}$.
Therefore,
$\{(k_{n-1}\ell, r_{n-1}\ell) \big| \ell = 0, 1, 2, ..., h_{n-1}\} = \{(k_{n-1}\ell, \overline{ak_{n-1}\ell}) \big| \ell = 0, 1, 2, ..., h_{n-1}\} \subset E_2$.
Moreover, $r_{n-2} = h_{n-1}r_{n-1} + r_n$. On the other hand, by lemma 5.2.8 , all the points of $E_2 = (t, \overline{at})$, $t \leq H_{n-1}$, have ordinates distant of r_n or r_{n-1}.
Therefore, if there is other points of $E_2 \cap \{[0, H_{n-1}] \otimes [0, r_{n-2}[\}$ that the points $\{(k_{n-1}\ell, r_{n-1}\ell) \big| \ell = 0, 1, 2, ..., h_{n-1}\}$, there exists $\ell_0 \in \{1, 2, ..., h_{n-1}\}$ and $(x_1, y_1) \in E_2 \cap \{[0, H_{n-1}] \otimes [0, r_{n-2}]\}$ such that $r_{n-1}\ell_0 - y_1 = r_n$.

Because $H_{n-1} = k_n + k_{n-1} - 1 < k_{n+1} \leq k_{d+1}$, by lemma 5.2.5, there exists an only $t \in \{1, ..., H_{n-1}\}$, such that $\overline{at} = y_1 : t = x_1$. Because $y_1 \neq 0$, there exists an only $t \in \{0, 1, ..., H_{n-1}\}$, such that $\overline{at} = y_1$.

Now, $r_{n-1}\ell_0 - y_1 = \overline{a\ell_0 k_{n-1}} - \overline{at} = r_n = \overline{-ak_n}$. Then, $\overline{a\ell_0 k_{n-1}} - \overline{-ak_n} = \overline{at}$. Then, $\overline{a(\ell_0 k_{n-1} + k_n)} = \overline{at}$.

Because $r_{d-1} = h_d r_d$ with $r_{d-1} > r_d$, $h_d \geq 2$. Moreover, $d \geq n \geq 2$. Then, $d - 1 > 0$. Then, $k_{d-1} > 0$.

Then, by lemma 5.2.4, $2k_n - k_{n-2} \leq 2k_d < 2k_d + k_{d-1} \leq h_d k_d + k_{d-1} = k_{d+1} \leq m$. Then, $0 < \ell_0 k_{n-1} + k_n < k_{d+1}$.

Then, by lemma 5.2.4, $0 < k_{n-1} + k_n \leq \ell_0 k_{n-1} + k_n \leq h_{n-1}k_{n-1} + k_n \leq k_n - k_{n-2} + k_n = 2k_n - k_{n-2} \leq 2k_d < 2k_d + k_{d-1} \leq h_d k_d + k_{d-1} = k_{d+1} \leq m$. Then, $0 < \ell_0 k_{n-1} + k_n < k_{d+1}$.

Now $0 < t \leq H_{n-1} = k_n + k_{n-1} - 1 < k_d + k_{d-1} \leq k_{d+1}$. Moreover, $0 < \ell_0 k_{n-1} + k_n < k_{d+1}$.

Then, because $\overline{a(\ell_0 k_{n-1} + k_n)} = \overline{at}$, by lemma 5.2.5, $t = \ell_0 k_{n-1} + k_n$.

Then, $t = \ell_0 k_{n-1} + k_n \geq k_{n-1} + k_n > H_{n-1}$. It is a contradiction.

Therefore, there is not other points of $E_2 \cap \big\{[0, H_{n-1}] \otimes [0, r_{n-2}[\big\}$ that $\big\{(k_{n-1}\ell, r_{n-1}\ell)\big|\ell = 0, 1, 2, ..., h_{n-1}\big\}$.

Therefore, there is not other points of $E_2 \cap \big\{[0, k_n[\otimes[0, r_{n-2}[\big\}$ that the points $\big\{(k_{n-1}\ell, r_{n-1}\ell)\big|\ell = 0, 1, 2, ..., h_{n-1}\big\}$.

Therefore,

$$E_2 \cap \big\{[0, k_n[\otimes[0, r_{n-2}[\big\} = \big\{(k_{n-1}\ell, r_{n-1}\ell)\big|\ell = 0, 1, 2, ..., h_{n-1}\big\} \ .$$

According to what precedes,

$$\big\{(k_{n-1}\ell, \overline{ak_{n-1}\ell})\big|\ell = 0, 1, 2, ..., h_{n-1}\big\} = \big\{(k_{n-1}\ell, r_{n-1}\ell)\big|\ell = 0, 1, 2, ..., h_{n-1}\big\}$$

is located on the straight line $y = (r_{n-1}/k_{n-1})x$ if n is even.

Suppose that n is odd. One proves this result by the same way as when n is even. ■

Chapter 6

Randomization by the functions of Fibonacci

6.1 Study of the problem

In this section, we assume that our data y_n are provided by texts. Then, one will understand how one can make IID the y_n thanks to the Fibonacci functions $T_q^d = Pr_q^d \circ \widehat{T}$ (cf Definition 1.2.5) (if d=2, one simplifies T_q^d in T_q).

6.1.1 Some Notations

In this chapter, the following notations are used.

Notations 6.1.1 *In this chapter 6, $q, d \in \mathbb{N}^*$. Moreover, m is an element of the Fibonacci sequence : $m = fi_{n_0}$. Then, we set $m = d^Q$ where $Q \in \mathbb{R}_+$. Moreover, $Y_n \in F(m)$ is a sequence of random variables defined on a probability space $(\Omega, \mathcal{A}, P)$ and $X_n = T_q^d(Y_n)$.*

Notations 6.1.2 *Let $k \in \{0, 1, ..., d^q - 1\}$. We set $I_k = [k/d^q, (k+1)/d^q[$. We define the interval $[c_k/m, c'_k/m[$ with $c_k, c'_k \in F^*(m)$ by $[c_k/m, c'_k/m[\cap F(m) = [k/d^q, (k+1)/d^q[\cap F(m)$. We set $N(I_k) = c'_k - c_k$.*

More generally, we denote by I the intervals $I = [k/d^q, k'/d^q[$. Then, we define $[c/m, c'/m[$ with $c, c' \in F^*(m)$ by $[c/m, c'/m[\cap F(m) = [k/d^q, k'/d^q[\cap F(m)$.

Then, $N(I_k)$ is the number of $t/m \in F(m)$ such that $k/d^q \leq t/m < (k+1)/d^q$.

Notations 6.1.3 *Let $x_s \in F(m)$. We set $p_{x_s} = P\big\{\overline{T}(mY_n)/m) = x_s\big\}$.*

Of course, $P\big\{X_n = k/d^q\big\} = P\big\{\overline{T}(mY_n)/m \in [c_k/m, c'_k/m[\big\} = \sum_{x_s \in [c_k/m, c'_k/m[} p_{x_s}$. Moreover, the following lemma holds.

Lemma 6.1.1 *With the previous notations, $(c_k - 1)/m < k/d^q \leq c_k/m$ and $(c'_k - 1)/m < (k+1)/d^q \leq c'_k/m$.*

Lemma 6.1.2 *Let $1/d^q = h_0/m + r$ where $0 \leq r < 1/m$ and $h_0 \in \mathbb{N}$. Then, $N(I_k) = h_0$ or $N(I_k) = h_0 + 1$. Moreover, $m/d^q = h_0 + e$ where $0 \leq e < 1$.*

6.1.2 Sequence of real numbers regarded as IID

We show now that about any sequence of real numbers can be regarded as the permutation of an IID sequence.

Proposition 6.1.1 *Let z_n , $n = 1, 2, ..., n_0$ be a sequence of integers $z_n \in F^*(m)$ such that all the z_n's are different.*

Then, there exists a permutation ϕ such that $z'_n = z_{\phi(n)}$, $n = 1, 2, ..., n_0$, can be regarded as an IID sample having a distribution M_Z .

Proof One can associate to z_n a continuous distribution function F which is the smoothest possible and which have a density function which is the smoothest possible.

Let $x_n = x(n)$ be an IID sample with the distribution associated to F. For any function f, $f(x_n)$ is a priori an IID sample. But it is necessary to be careful: it is better than f is not too complicated. For example $f(x_n)$ can be increasing.

To avoid it, we denote by r_x and r_z the number of order of $x(n)$ and $z(n) = z_n$, respectively : $r_x(n)$ and $r_z(n)$ are the permutations of $\{1, 2, ..., n_0\}$ such that $x_{r_x^{-1}(1)} < x_{r_x^{-1}(2)} < .... < x_{r_x^{-1}(n_0)}$ and $z_{r_z^{-1}(1)} < z_{r_z^{-1}(2)} < .... < z_{r_z^{-1}(n_0)}$.

Then, there exists a continuous function f such that $f(x_{r_x^{-1}(n)}) = z_{r_z^{-1}(n)}$ for $n = 1, 2, ..., n_0$. One can force this function to be smoothest possible with a Lipschitz coefficient not too large. Moreover, if it is not smooth enough, one can also use another IID sample $\{x_n^1\}$.

Then, the following conjecture is applied.

Conjecture 6.1.2 *Let x_n be an IID sample. One suppose that, for all function which is smooth enough with a Lipschitz coefficient not too large, $f(x_n)$ can be regarded as an IID sample.*

For this function f , $f(x_n)$ can be regarded as an IID sample which has the same law as $f(X_1)$.

Now, $\{f(x_n) \mid n = 1, 2, ..., n_0\} = \{z_n \mid n = 1, 2, ..., n_0\}$. Then, there exists a permutation ϕ such that $z_{\phi(n)} = f(x_n)$ for $n = 1, 2, ..., n_0$. Then, $z_{\phi(n)}$ can be regarded as an IID sample. ■.

Remark 6.1.3 *Conjecture 6.1.2 is very likely. However it implies to choose a function according to the sample in order to deduce from it that the transformed sample is IID : it is always delicate. In fact it would be necessary to explicate all that. But it would be too long in this report.*

But one understands well the meaning of this conjecture and one understands that it is very reasonable. Moreover, we have carries out many simulations which all, confirm it.

Now what means this result? It means that the sample $z'_n = f(x_n)$ behaves like a sample IID of law M_Z. However if z'_n is an IID sample, that means that, for almost all the permutations ψ, $z'_{\psi(n)}$ is a priori IID. Thus for almost all the permutations ψ', $z_{\psi'(n)}$ is IID.

It is thus enough to check that it is indeed the case for various permutations chosen randomly. One thus tests the independence of the obtained sample.

For example, we choose a sample of the sequence not satisfying the CLT of Ibragimov-Linnik page 384 of [21]. Then, we have estimated multilinear correlation coefficients for various dimentions p : with a empirical variance equal to 39.57, the following values have been obtained :

p=2	- 0.0019	0.0032	-0.0009	-0.0062	0.0041	0.0036	0.0102
p=3	0.0014	0.0005	-0.0094	-0.0052	0.0034	-0.0087	-0.0027
p=4	-0.0057	-0.0041	-0.0013	0.0124	-0.0074	0.0033	0.0051

We have also used classical Diehard tests (cf [1] , [2]). All confirm independence. Moreover, if one takes subsamples, those are also independent and have the same law M_Z.

Now, one would like to apply the CLT. For studying this problem, we will apply proposition 6.1.1 to the sums.

Corollary 6.1.4 *Let $\sum_{n\in F} z_n$ where $F \subset \{1, 2, ..., n_0\}$. Then $\sum_{n\in F} z_n = \sum_{n\in F'} z'_n$, where $z'_n = z_{\psi(n)}$ is an IID sample which has the distribution M_Z and where $F' = \psi(F)$ and $\psi = \phi^{-1}$.*

Corollary 6.1.5 *For all sets F" $= \psi'(F')$, except a negligible minority, $\sum_{n\in F"} z'_n$ behaves as the sum of an IID sample which has this same distribution M_Z.*

Corollary 6.1.6 *Let H be a measurable function. For all sets F" $= \psi'(F')$, except a negligible minority, $\sum_{n\in F"} H(z'_n)$ behaves as the sum of IID samples which have a same distribution M_{HZ}.*

Corollary 6.1.7 *For almost all the sets F, $\frac{1}{card(F)} \sum_{n\in F} z_n \approx L$ where L does not depend on F.*

Now, let us suppose that F is chosen randomly. Then, one can admit that F' is also chosen randomly.

Thus each time one has a sum over a set chosen randomly, one carries out a sum of a sample of an IID sequence of random variable Z'_n. Then, $\frac{1}{card(F)} \sum_{n\in F} Z'_n \to E\{Z'_1\}$.

Let us notice immediately that these results does not means that all the CLT can be regarded as a sum of IID sample. In no case, the sum $Z_1 + ... + Z_n$ is necessarly close to a normal law.

Now suppose that one uses the sample $f(x(n)) = z(\phi(n))$. One separates it into several subsequences with the same size and one sums these subsequences. Then, the empirical distribution of those sums will be close to a normal law.

In figure 6.1, we have transformed by a permutation MATLAB chosen randomly a sample of size 900000 of the sequence not satisfying the CLT (cf page 384 of [21]). Then, one has it separated in 9000 successive subsequences which one has summed. It is understood that the distribution is close to a normal law.

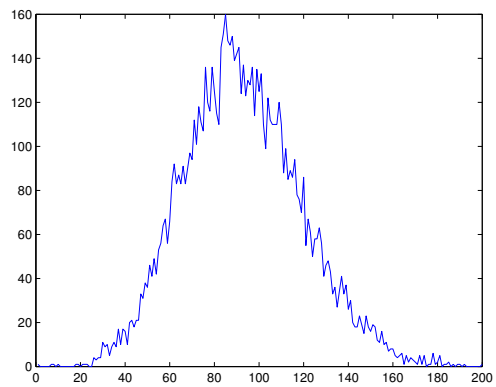

Figure 6.1: Sums of subsamples chosen randomly

In the same way, if sums of the z_n are chosen randomly, the sample consituted by these sums shows that one has a distribution close to a normal law.

But these results do not mean inevitably that the exact distribution of a sum chosen randomly will be close to the normal law. By example, let us choose independent samples of the type $z_{\psi(1)} + ... + z_{\psi(n_1)}$, $n_1 < n_0$, where the subsamples $z_{\psi(1)}, ..., z_{\psi(n_1)}$ are all built with the same permutation ψ. Suppose that the sequence Z_n has the distribution of the example not checking the CLT (page 384 of [21]).

Then these samples of sums will behave like a nonnormal samples. An estimate of the law obtained by these sample is in figure 6.2.

These results can appear strange. They thus should be explained.
1) If a sequence Z_n does not satisfy the CLT, it is anyway possible that samples extracted of a sequence $Z_{k+n_2}, Z_{2k+n_2}, ..., Z_{nk+n_2}$ do not satisfies the CLT, i.e. $Z_{k+n_2} + Z_{2k+n_2} + ... + Z_{nk+n_2}$ has not a distribution close to the normal law.

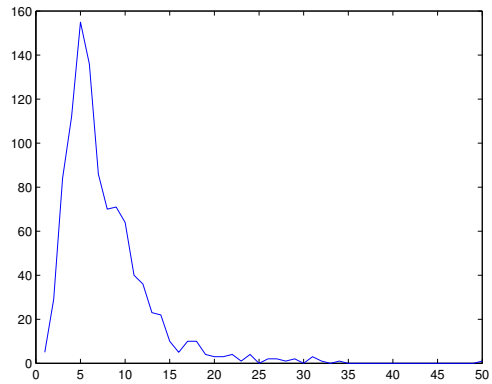

Figure 6.2: Law of a subsample chosen randomly

In this case, it is plausible that samples of type $z_{\psi(1)+n_2}, z_{\psi(2)+n_2} ... z_{\psi(n)+n_2}$ do not satisfy the CLT (where ψ is a permutation, for example of $\{1, 2, ..., kn\}$). Then, $Z_{\psi(1)+n_2} + Z_{\psi(2)+n_2} + ... + Z_{\psi(n)+n_2}$ has not a distribution close to the normal law.
2) But, by proposition 6.1.1, we know that if $n_0 = kn_1$, the sample $z'_1 + ... + z'_{n_1}$, $z'_{n_1+1} + ... + z'_{2n_1}$, ... , $z'_{(k-1)n_1+1} + ... + z'_{kn_1}$ can be regarded as a sample with a distribution close to the normal law.

Indeed, it is possible that $Z'_1 + ... + Z'_{n_1}$ has not a normal distribution. But the $Z'_{n_1+1} + ... + Z'_{2n_1}$, ... , $Z'_{(k-1)n_1+1} + ... + Z'_{kn_1}$ have not the same distribution. Because that, they behave indeed as samples with normal law.

Then, when one uses sums $z_{\psi(1)+n_2}, z_{\psi(2)+n_2} ... z_{\psi(n)+n_2}$, they can regarded by two different ways : we have to choose which is the good one.

In subsection 6.1.3 below, what we want, it is to use $P\{X_n = k/d^q\} = \sum_{x_s \in [c_k/m, c'_k/m[} p_{x_s}$ where the sets $\overline{T}^{-1}([c_k, c'_k[)$ can be regarded as randomly chosen. Then, it is about sum randomly chosen : they behave as approximately normal.

6.1.3 Randomization of Y_n

We have $P\{X_n = k/d^q\} = \sum_{x_s \in [c_k/m, c'_k/m[} p_{x_s}$. Now, there is no logical connection between text and the distribution of the points of $\{a_1, a_2, ...\} = \widehat{T}^{-1}(I)$ where I is an interval. These two events are logically independent. Indeed, the sequence $\{a_1, a_2, ...\} = \widehat{T}^{-1}(I)$ is built by a specific and relatively simple mathematical application whereas the data y_n are the realization of a sequence of random variables Y_n and thus unpredictable in an exact way. Moreover the sequence $\{a_1, a_2, ...\}$ is well distributed in F(m). It is reasonable to think that this set is independent of sequences obtained starting from text. One can thus

regard this set as randomly chosen.

That means that $\sum_{x_s \in [c_k/m, c'_k/m[} p_{x_s}$ can be regarded as a sum $\sum_{s \in F} p_{x_s}$ where the set F is a Borel set chosen randomly. According to the corollary 6.1.7, that means that, for all k, $(d^q/m) \sum_{x_s \in [k/d^q, (k+1)/d^q[} p_{x_s}$ converges to the same limit L.

One is all the more sure of this result that only a negligible minority of the possible sets F will not check this property: because there is only d^q "k" possible, it is thus enough to choose m enough large compared to d^q.

At last, $\sum_k \sum_{x_s \in [k/d^q, (k+1)/d^q[} p_{x_s} = 1$. Therefore, $\sum_{x_s \in [k/d^q, (k+1)/d^q[} p_{x_s} \approx 1/d^q$.

Example In figure 6.3, one supposes $card\big([c, c'[\cap F^*(m)\big) = 10$. One understands that $\{a'_1, a'_2, ..., a'_{10}\}$ is about uniformly distributed in [-4,4]. We obtain, $P\{X_n = k/d^q\} \approx 1/d^q$.

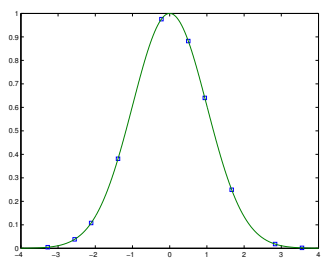

Figure 6.3: Example : Normal curve

6.2 Empirical Probability

Unidimensional case

Let us be interested with a sample $x^*_n = T_q(y_n)$, $n = 1, 2, ..., n_0$, where all the y_n are distinct. Let $pe_{x_s} = P_e\big\{\overline{T}(mY_n)/m = x_s\big\}$ where P_e is the empirical probability. Then, one has

$$P_e\big\{T_q(Y_n) \in [c/m, c'/m[\big\} = \sum_{x_s \in [c/m, c'/m[} pe_{x_s} \ .$$

There is no logical connection between the set $\mathcal{J} =$

[Newton's theory]
[of gravitation was]
[soon accepted wit]

[hout question, and]
[it remained unques]
[tioned until the begin]
[ning of this century.]
.....

and the sets $\mathcal{H}_s =$

[whgkudf ly cuqhjg]
[aamxgusdggbxckmp]
[x;cbkutcc ze xycyc x]
[qtdxucdzlcxy yx vyxy]
[ory of Relativity in 190]
[xwtex pez! i yi qy yqhfg]
.....

Thus a priori, the probability that $[Newton'stheory] \in \mathcal{H}$ is approximatively of $card(\mathcal{H})/32^{18}$ if $y_n \in \{0, 1, ..., 31\}$ [1].

Now, there is no logical connection between the y_n and the set $A = \{a_1, a_2, ...\}$. Then, the set $\{a_1, ..., a_{c'-c}\}$ is well chosen randomly. Then, by corollary 6.1.7, $\frac{1}{n_0}\sum_{n=1}^{n_0} 1_{\{a_1,.....,a_{c'-c}\}}(y_n) \rightarrow \frac{N(A)}{m}$ where $N_A = c' - c$.
Finally, $P_e\big\{X_n \in [k/d^q, (k+1)/d^q[\big\} \approx 1/d^q$ for any k.

One can also understand this result by a more classical way : cf section 6.2 of [18]

Now, because the subsample of the y_n such that $y_n \in A$ can be regarded as IID (cf corollary 6.1.6), $\frac{1}{\sqrt{n_0}\sigma_{N_A}}\sum_{n=1}^{n_0}\left[1_{\{a_1,.....,a_{c'-c}\}}(y_n) - N_A/m\right]$ has asymptotically a standard normal distribution where $\sigma^2_{N_A} = N_A/m - (N_A/m)^2$.

One has checked these results by testing them with the the sequence d(j), j=1,2,... : cf section 10.1 . All the tests conclude to the uniformity
In figure 6.4 , we have the histogram for the sequence d(j). The figure 6.5 which represents the histogram for a pseudo-random sequence of uniform distribution.

Multidimensional case

For t=1,2,...,p, we set $I^t = \left[\frac{c_t}{m}, \frac{c'_t}{m}\right[$ where c_t, $c'_t \in F^*(m)$. We set $A^t = \widehat{T}^{-1}(I^t) = \{a^t_1, ..., a^t_{c'_t - c_t}\}$ for t=1,2,...,p.

[1] 32 corresponds to 26 letters, punctuation, signs and space

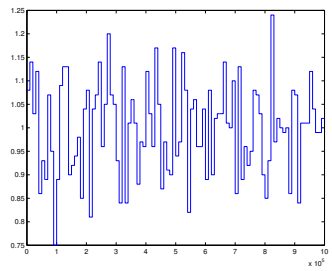

Figure 6.4: Histogram of f(1,j)

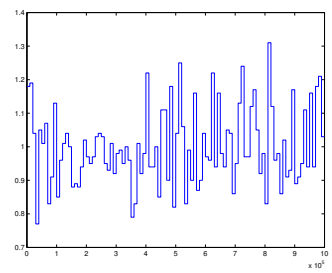

Figure 6.5: Histogram of uniform data

There is no logical connection between the sets $\mathcal{J}$ and $\mathcal{H}_s$. Then, the probability that {[Newton's theory], [of gravitation was], [soon accepted wit]} $\subset \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_3$ is approximatively equal to $\prod_{t=1}^{3} \left[card(\mathcal{H}_t)/32^{18}\right]$. This result can be understood by simulation.

Because there is always no connection between parts of texts $(y_{n+j_1}, ..., y_{n+j_p})$ and the sets $A^1 \otimes .... \otimes A^p$, it is thus logical that sums on the various possible sets $A^1 \otimes .... \otimes A^p$ (where $p \leq Log(n_0)/log(2)$), behave as sums over sets chosen randomly, i.e.

$$P_e\Big\{\big\{X_{n+j_1} \in I^1\big\} \cap .... \cap \big\{X_{n+j_p} \in I^p\big\}\Big\} = \sum_{x^1_{s_1} \in I^1} .... \sum_{x^p_{s_p} \in I^p} pe_{x^1_{s_1},...,x^p_{s_p}} \approx \prod_{t=1}^{p} L(I^t)\,,$$

where $pe_{x^1_{s_1},...,x^p_{s_p}} = P_e\big\{\{\overline{T}(mY_{n+j_1})/m = x^1_{s_1}\} \cap .... \cap \{\overline{T}(mY_{n+j_p})/m = x^p_{s_p}\}\big\}$.

6.3 Theoretical probability

Let us choose a random vector $(X_{n+j_1}, X_{n+j_2}, ..., X_{n+j_p})$. We set $p_{x^1_{s_1},...,x^p_{s_p}} = P\big\{(\overline{T}(mY_{n+j_1})/m = x^1_{s_1}) \cap ... \cap (\overline{T}(mY_{n+j_p})/m = x^p_{s_p})\big\}$.

Probabilities chosen randomly As for empirical probabilities $P\Big\{\big\{X_{n+j_1} \in I^1\big\}\cap....\cap\big\{X_{n+j_p} \in I^p\big\}\Big\}$ is equal to a sum of $p_{x^1_{s_1},...,x^p_{s_p}}$. Because y_n means texts and is independent with sets $T_q^{-1}(I^1) \otimes .... \otimes T_q^{-1}(I^p)$, the sums are about sets randomly chosen. Therefore the $p_{x^1_{s_1},...,x^p_{s_p}}$'s can be regarded as an IID sample of random variables which have a distribution M_Z.

Model Because $n_0 << m$, there are many possibles models associated to the sample y_n, $n = 1, 2, .., n_0$. As a matter of fact there in an infinity of them. Then, because there are several possible correct models meaning this text, we will study the various possible probabilities $p_{x^1_{s_1},...,x^p_{s_p}}$.

One will provide the set of the $p_{x^1_{s_1},...,x^p_{s_p}}$ of a measure which is a probability : i.e. the set of the possible probabilities $p_{x^1_{s_1},...,x^p_{s_p}}$ is itself the realization of a probability space.

For example, one can choose $p_{x^1_{s_1},...,x^p_{s_p}} = \frac{p'_{x^1_{s_1},...,x^p_{s_p}}}{\sum_{i_1=1}^m .... \sum_{i_p=1}^m p'_{i_1/m,...,i_p/m}}$ where the $p'_{x^1_{s_1},...,x^p_{s_p}}$'s are a sample of a sequence of IID random variables $P'_{x^1_{s_1},...,x^p_{s_p}}$ defined on a probability space $(\Omega_1^p, \mathcal{A}_1^p, Proba_1^p)$ and which have a distribution M.

With a such model, we have proved in [18] section 8.4.2, that with a probability very close to 1,

$$P\Big\{\big\{X_1 \in I_1\big\} \cap ..... \cap \big\{X_p \in I_p\big\}\Big\} \approx \frac{\prod_s (c'_s - c_s)}{m^p} \left[1 + \frac{Ob(1).b\sigma_M}{E_M\sqrt{\prod_s N(I_s)}}\right].$$

Remark 6.3.1 *This approximation is not the same as* $\frac{1}{n_0}\sum_{n=1}^{n_0} 1_{\{a_1,..,a_{c'-c}\}}(y_n) = N_A/m + \frac{Ob(1).b\sigma_{N_A}}{\sqrt{n_0}}$ *because in this last one, the empirical approximation is involved.*

As a matter of fact with this probability $P'_{x^1_{s_1},...,x^p_{s_p}}$, p is fixed. If p is changed, one changes space of measure.

The problem of marginals distributions

If the model defined on spaces $(\Omega_1^p, \mathcal{A}_1^p, Proba_1^p)$ is chosen (cf [18] page 217 section 8.4.2), there is a problem : one does not take account of the marginal probabilities.

If p=2, in space $(\Omega_1^2, \mathcal{A}_1^2, Proba_1^2)$ those sums will be already sums of probabilities taken randomly. That means that, with this measure, the marginal probabilities $p_{x^1_{s_1}} = \sum_{x^2_{s_2}} p_{x^1_{s_1},x^2_{s_2}}$ in their vast majority will have a priori uniform distribution. One thus does not take in account that the $p_{x^1_{s_1}}$ are probabilities in two dimensions with marginal laws, i.e. with constraints.

It is thus a result which seems not to correspond to reality, for example if $P'_{x^1_{s_1},...,x^p_{s_p}}$ has continuous densities.

Anyway, this is not very important: measures of spaces $(\Omega^p_1, \mathcal{A}^p_1, Proba^p_1)$ are only measures giving an idea of the numbers of models close to a sequence IID. Moreover, that does not change anything with the ultimate result : only a negligible minority of models does not check

$$P\Big\{\big\{X_1 \in I_1\big\} \cap ..... \cap \big\{X_p \in I_p\big\}\Big\} \approx \frac{\prod_s (c'_s - c_s)}{m^p} \Big[1 + Ob(1)\epsilon\Big]$$

where $|\epsilon| << 1$.

6.3.1 Two dimensional case

If one wants to take account of the marginal laws, it is necessary to consider the probabilities of each Y_{n+j_t}, i.e the probabilities of the marginal laws. Then it should be considered that the sums are taken randomly: for example $\sum_{x_{s_1}=x^0_{s_1}, x_{s_2}\in A_2} p_{x_{s_1},x_{s_2}} \approx p_{x^0_{s_1}}$.

Now, it is necessary to define associated probability spaces. One thus chooses probability spaces $(\Omega_{x_{s_1}}, \mathcal{A}_{x_{s_1}}, Proba_{x_{s_1}})$ for each x_{s_1}: one uses product space

$$(\Omega^2, \mathcal{A}^2, Proba^2) = \Big(\prod_{x_{s_1}} \Omega_{x_{s_1}}, \mathcal{T}\Big(\prod_{x_{s_1}} \mathcal{A}_{x_{s_1}}\Big), \prod_{x_{s_1}} Proba_{x_{s_1}}\Big)$$

where $\mathcal{T}\Big(\prod_{x_{s_1}} \mathcal{A}_{x_{s_1}}\Big)$ is the σ-algebra generated by $\prod_{x_{s_1}} \mathcal{A}_{x_{s_1}}$.

One takes also into account the probability space $(\Omega^1, \mathcal{A}^1, Proba^1)$ associated with the first marginal law and finally one uses product space $(\Omega, \mathcal{A}, Proba) = (\Omega^1, \mathcal{A}^1, Proba^1) \otimes (\Omega^2, \mathcal{A}^2, Proba^2)$.

Let us notice that it poses the problem then to know if one chooses to take the sums $\sum_{x_{s_1}=x^0_{s_1}, x_{s_2}\in A_2} p_{x_{s_1},x_{s_2}}$ or $\sum_{x_{s_1}\in A_1, x_{s_2}=x^0_{s_2}} p_{x_{s_1},x_{s_2}}$. As a matter of fact for the results which we try to obtain, we will understand that does not have importance.

Hypothesis 6.3.1 *Suppose that*

$$p_{x^1_{s_1},x^2_{s_2}} = \frac{p'_{x^1_{s_1}} \frac{p'_{x^1_{s_1},x^2_{s_2}}}{\sum_{i_2=1}^m p'_{x^1_{s_1},i_2/m}}}{\sum_{i_1=1}^m p'_{i_1/m}} .$$

We assume that the $p'_{x^1_{s_1}}$'s are a sample of an IID sequence of random variables $P'_{x^1_{s_1}}$. We assume also that, for each $x^1_{s_1}$, the $p'_{x^1_{s_1},x^2_{s_2}}$'s are a sample of an IID sequence of random variables $P'_{x^1_{s_1},x^2_{s_2}}$. Then, $p'_{x^1_{s_1},x^2_{s_2}} = P'_{x^1_{s_1},x^2_{s_2}}(\omega)$ and $p'_{x^1_{s_1}} = P'_{x^1_{s_1}}(\omega)$ where $\omega \in \Omega$.

One supposes that $P'_{x^1_{s_1}}$ and, for each $x^1_{s_1}$, $P'_{x^1_{s_1},x^2_{s_2}}$ have the distribution M. Let E_M and σ^2_M be the associated expectation and variance.

Hypothesis 6.3.2 *We assume that $I_t = \big[k_t/d^q, (k_t+1)/d^q\big[$. Let $N(I_t)$ be the number of $r/m \in F(m)$ such that $k_t/d^q \leq r/m < (k_t+1)/d^q$. Let $c_t, c'_t \in F^*(m)$ such that $I_t \cap F(m) = [c_t/m, c'_t/m[\cap F(m)$.*

We suppose m enough large compared to d^q and to h_0. We suppose $d^q >> 1$. We suppose that b and σ_M are not too large and that E_M is not too small.

We shall need the following notations.

Notations 6.3.1 *Let $b > 0$. Let $\Gamma'_1(b) = Max_{n \geq h_0}\Big(Proba\big\{|S_n| \geq b\big\}\Big)$ where $S_n = \frac{\sum_{i_1=1}^n P'_{i_1/m}}{\sigma_M\sqrt{n}}$.*

Remark that $\Gamma'_1(b) = Max_{n \geq h_0}\Big(Proba\big\{|S'_n| \geq b\big\}\Big)$ where $S'_n = \frac{\sum_{i_1=1}^n P'_{x^1_{s_1},i_2/m}}{\sigma_M\sqrt{n}}$ and $x^1_{s_1} \in F(m)$. Moreover, $\Gamma'_1(b) \approx \Gamma(b)$ because m/d^q is large enough.

Then, one has the following proposition.

Proposition 6.3.1 *Under the hypotheses 6.3.1 and 6.3.2, with a probability larger than $1 - 4\Gamma'_1(b)$,*

$$P\Big\{\{X_1 \in I_1\} \cap \{X_2 \in I_2\}\Big\} = \frac{N(I_1)N(I_2)}{m^2}[1 + Ob(1).\epsilon'_{(2)}] \ ,$$

where $\epsilon'_{(2)} \approx \frac{2b\sigma_M}{E_M\sqrt{h_0}}$.

Proof At First,

$$\sum_{x^1_{s_1} \in I_1}\sum_{x^2_{s_2} \in I_2} p_{x^1_{s_1},x^2_{s_2}} = \sum_{x^1_{s_1} \in I_1}\sum_{x^2_{s_2} \in I_2} \frac{p'_{x^1_{s_1}} \frac{p'_{x^1_{s_1},x^2_{s_2}}}{\sum_{i_2=1}^m p'_{x^1_{s_1},i_2/m}}}{\sum_{i_1=1}^m p'_{i_1/m}} = \sum_{x^1_{s_1} \in I_1} \frac{p'_{x^1_{s_1}} \frac{\sum_{x^2_{s_2} \in I_2} p'_{x^1_{s_1},x^2_{s_2}}}{\sum_{i_2=1}^m p'_{x^1_{s_1},i_2/m}}}{\sum_{i_1=1}^m p'_{i_1/m}} \ .$$

We use the CLT. Then, with a probability larger than $1 - \Gamma'_1(b)$,

$$\frac{1}{N(I_2)}\sum_{x^2_{s_2} \in I_2} p'_{x^1_{s_1},x^2_{s_2}} = E_M + \frac{Ob(1).b\sigma_M}{\sqrt{N(I_2)}} \ .$$

Moreover, with a probability larger than $1 - \Gamma'_1(b)$,

$$\frac{1}{m}\sum_{i_2=1}^m p'_{x^1_{s_1},i_2/m} = E_M + \frac{Ob(1).b\sigma_M}{\sqrt{m}} \ .$$

Then, with a probability larger than $1-2\Gamma_1'(b)$,

$$p'_{x^1_{s_1}} \frac{\sum_{x^2_{s_2}\in I_2} p'_{x^1_{s_1},x^2_{s_2}}}{\sum_{i_2=1}^m p'_{x^1_{s_1},i_2/m}} = p'_{x^1_{s_1}} \frac{N(I_2)}{m} \frac{E_M + \frac{Ob(1).b\sigma_M}{\sqrt{N(I_2)}}}{E_M + \frac{Ob(1).b\sigma_M}{\sqrt{m}}}$$

$$= \frac{N(I_2)p'_{x^1_{s_1}}}{m} \frac{1+\frac{Ob(1).b\sigma_M}{E_M\sqrt{N(I_2)}}}{1+\frac{Ob(1).b\sigma_M}{E_M\sqrt{m}}} = \frac{N(I_2)p'_{x^1_{s_1}}}{m}[1+Ob(1).\epsilon_2] \ ,$$

where $\epsilon_2 \approx \frac{b\sigma_M}{E_M\sqrt{N(I_2)}} + \frac{b\sigma_M}{E_M\sqrt{m}}$.

Moreover, with a probability larger than $1-2\Gamma_1'(b)$,

$$\frac{1}{N(I_1)} \sum_{x^1_{s_1}\in I_1} p'_{x^1_{s_1}} = E_M + \frac{Ob(1).b\sigma_M}{\sqrt{N(I_1)}} \ ,$$

$$\frac{1}{m}\sum_{i_1=1}^m p'_{i_1/m} = E_M + \frac{Ob(1).b\sigma_M}{\sqrt{m}} \ .$$

Then,

$$\frac{\sum_{x^1_{s_1}\in I_1} p'_{x^1_{s_1}}}{\sum_{i_1=1}^m p'_{i_1/m}} = \frac{N(I_1)}{m} \frac{E_M + \frac{Ob(1).b\sigma_M}{\sqrt{N(I_1)}}}{E_M + \frac{Ob(1).b\sigma_M}{\sqrt{m}}} = \frac{N(I_1)}{m}[1+Ob(1).\epsilon_1] \ ,$$

where $\epsilon_1 \approx \frac{b\sigma_M}{E_M\sqrt{N(I_1)}} + \frac{b\sigma_M}{E_M\sqrt{m}}$.

Moreover,

$$\frac{1}{N(I_1)} \sum_{x^1_{s_1}\in I_1} Ob(1)p'_{x^1_{s_1}} = \frac{Ob(1)}{N(I_1)} \sum_{x^1_{s_1}\in I_1} p'_{x^1_{s_1}} = Ob(1)\Big[E_M + \frac{Ob(1).b\sigma_M}{\sqrt{N(I_1)}}\Big] \ .$$

Then, with a probability larger than $1-2\Gamma_1'(b)$,

$$\frac{\sum_{x^1_{s_1}\in I_1} Ob(1)p'_{x^1_{s_1}}}{\sum_{i_1=1}^m p'_{i_1/m}} = \frac{Ob(1)N(I_1)}{m} \frac{E_M + \frac{Ob(1).b\sigma_M}{\sqrt{N(I_1)}}}{E_M + \frac{Ob(1).b\sigma_M}{\sqrt{m}}} = \frac{Ob(1)N(I_1)}{m}[1+Ob(1).\epsilon_1] \ .$$

Then, with a probability larger than $1-4\Gamma_1'(b)$,

$$\sum_{x^1_{s_1}\in I_1}\sum_{x^2_{s_2}\in I_2} p_{x^1_{s_1},x^2_{s_2}} = \sum_{x^1_{s_1}\in I_1} \frac{p'_{x^1_{s_1}} \frac{\sum_{x^2_{s_2}\in I_2} p'_{x^1_{s_1},x^2_{s_2}}}{\sum_{i_2=1}^m p'_{x^1_{s_1},i_2/m}}}{\sum_{i_1=1}^m p'_{i_1/m}} = \sum_{x^1_{s_1}\in I_1} \frac{\frac{N(I_2)p'_{x^1_{s_1}}}{m}[1+Ob(1).\epsilon_2]}{\sum_{i_1=1}^m p'_{i_1/m}}$$

$$= \frac{N(I_1)N(I_2)}{m^2}[1+Ob(1).\epsilon_1][1+Ob(1).\epsilon_2] = \frac{N(I_1)N(I_2)}{m^2}[1+Ob(1).\epsilon'_{(2)}]$$

where $\epsilon'_{(2)} \approx \frac{b\sigma_M}{E_M\sqrt{N(I_1)}} + \frac{b\sigma_M}{E_M\sqrt{m}} + \frac{b\sigma_M}{E_M\sqrt{N(I_2)}} + \frac{b\sigma_M}{E_M\sqrt{m}} \approx \frac{2b\sigma_M}{E_M\sqrt{h_0}}$. ■

The form of this result resolve the problem of knowing if one chooses initially the sums $\sum_{x_{s_1}=x^0_{s_1},x_{s_2}\in A_2} p_{x_{s_1},x_{s_2}}$ or $\sum_{x_{s_1}\in A_1,x_{s_2}=x^0_{s_2}} p_{x_{s_1},x_{s_2}}$. Whatever the chosen sum, the result remains the same one : $\epsilon'_{(2)} \approx \frac{2b\sigma_M}{E_M\sqrt{h_0}}$.

We have regarded the probabilities associated with (X_{n+j_1}, X_{n+j_2}) when n and the sequence j_s are fixed. But we will need also a definite probability space for any n and all j_2. Then, we shall use the following assumptions.

Hypothesis 6.3.3 *We generalize by natural way the notations of hypothesis 6.3.1. For each n and each j_2, we replace the notation of the probability space $(\Omega, \mathcal{A}, Proba)$ by $(\Omega_{(n,j)}, \mathcal{A}_{(n,j)}, Proba_{(n,j)})$.*

In this case, we denote by $(\Omega, \mathcal{A}, Proba)$ the probability space $(\Omega, \mathcal{A}, Proba) = \prod_{n,j}(\Omega_{(n,j)}, \mathcal{A}_{(n,j)}, Proba_{(n,j)})$.

6.3.2 General case

Hypothesis 6.3.4 *We generalize by natural way the notations of hypothesis 6.3.1. with probabilities $p_{x^1_{s_1},x^2_{s_2},\ldots,x^p_{s_p}}$. Then we generalize the $p'_{x^1_{s_1}}$ and $p'_{x^1_{s_1},x^2_{s_2}}$ in $p'_{x^1_{s_1},x^2_{s_2},\ldots,x^p_{s_p}}$ which are samples of sequences of IID random variables defined on the probability spaces $(\Omega_{(n,j)}, \mathcal{A}_{(n,j)}, Proba_{(n,j)})$ and which have the distribution M. We denote by $(\Omega, \mathcal{A}, Proba)$ the probability space $(\Omega, \mathcal{A}, Proba) = \prod_{n,j}(\Omega_{(n,j)}, \mathcal{A}_{(n,j)}, Proba_{(n,j)})$.*

Then, one generalize easily proposition 6.3.1.

Proposition 6.3.2 *Under the hypotheses 6.3.4 and 6.3.2 , with a probability larger than $1-2p\Gamma'_1(b)$,*

$$P\Big\{\{X_1 \in I_1\} \cap .. \cap \{X_p \in I_p\}\Big\} = \frac{\prod_{r=1}^p N(I_r)}{m^p}[1+Ob(1).\epsilon'_{(p)}] \ ,$$

where $\epsilon'_{(p)} \approx \frac{pb\sigma_M}{E_M\sqrt{h_0}}$.

We deduce the following proposition.

Proposition 6.3.3 *Under the hypotheses 6.3.4 and 6.3.2, with a probability larger than $1-2p\Gamma'_1(b)$,*

$$P\big\{X_1 = k_1/d^q, \ldots.., X_p = k_p/d^q\big\} \approx \frac{1}{d^{pq}}[1+Ob(1).\epsilon'_{(p)}] \ .$$

Proof We have $P\big\{X_1 = k_1/d^q, ..., X_p = k_p/d^q\big\}$
$= P\Big\{\{X_1 \in I_1\} \cap .. \cap \{X_p \in I_p\}\Big\} = \frac{\prod_s (c'_{k_s} - c_{k_s})}{m^p}\ [1 + Ob(1).\epsilon'_{(p)}]$.
Moreover, $\prod \frac{c'_{k_s} - c_{k_s}}{m} = \frac{1}{d^{pq}}[1+\epsilon_a]$, where $\epsilon_a = \frac{p.d^q Ob(1)}{m} + \frac{p(p-1)d^{2q}Ob(1)}{2m^2} + ... << \epsilon'_{(p)}$ by hypothesis 6.3.2. Then,

$$P\big\{X_1 = k_1/d^q, ..., X_p = k_p/d^q\big\} = \frac{1}{d^{pq}}[1+\epsilon_a][1+Ob(1).\epsilon'_{(p)}] \approx \frac{1}{d^{pq}}[1+Ob(1).\epsilon'_{(p)}] \ . \blacksquare$$

Now, we study the Borel sets of $F(d^q)^{\otimes p}$: $Bo = \cup_{(k_1,...,k_p)\in\Theta}\{(k_1/d^q, ..., k_p/d^q)\} = \cup_{(k_1,...,k_p)\in\Theta}\{I_{k_1} \otimes .... \otimes I_{k_p}\}$ where $I_{k_t} = [k_t/d^q, (k_t+1)/d^q[$. Then, $L(Bo) = \frac{K_\Theta}{d^{pq}}$.

We have the following proposition.

Proposition 6.3.4 . *Assume that, for all* $(k_1, ..., k_p)$, $P\big\{X_1 = k_1/d^q, ..., X_p = k_p/d^q\big\} = \frac{1}{d^{pq}}[1 + Ob(1)\epsilon'_{(p)}]$.
Then, $P\big\{(X_{n+j_1}, ..., X_{n+j_p}) \in Bo\big\} = L(Bo)[1 + Ob(1)\epsilon'_{(p)}]$.

Proof We can write

$$P\big\{(X_{n+j_1}, ..., X_{n+j_p}) \in Bo\big\} = \sum_{(k_1,...,k_p)\in\Theta} \frac{1}{d^{pq}}[1+Ob(1)\epsilon'_{(p)}] = \frac{K_\Theta}{d^{pq}}[1+Ob(1)\epsilon'_{(p)}] \ . \blacksquare$$

Now, one can prove the following result.

Proposition 6.3.5 *We assume that the the hypotheses 6.3.4 and 6.3.2 hold. Let* $1 \le n \le n_0$. *Let* $P_{X_n}(Bo) = P\big\{(X_{n+j_1}, ..., X_{n+j_p}) \in Bo\big\}$. *One supposes that M is the uniform distribution. Then,*

$$Proba\Bigg\{\bigcap_{n+j_t, Bo} \Big\{\big|P_{X_n}(Bo) - L(Bo)\big| \le \frac{2b.pL(Bo)}{\sqrt{3.d^{Q-q}}}\Big\}\Bigg\} \ge 1 - 2p.n_0^p d^{pq}\Gamma'_1(b) \ . \tag{6.1}$$

Proof Let $\{k\} = \{k_1\} \otimes .... \otimes \{k_p\}$. Clearly, $\frac{|Ob(1)|.pb\sigma_M}{E_M\sqrt{h_0}} \approx |\epsilon'_{(p)}| \le \frac{2pb\sigma_M}{E_M\sqrt{d^{Q-q}}}$. Then, by the proof of proposition 6.3.3, with a probability larger than $1 - 2p\Gamma'_1(b)$,

$$P_{X_n}(\{k\}) = L(\{k\})\Big[1 + \frac{Ob(1).2p.b\sigma_M}{E_M.\sqrt{d^{Q-q}}}\Big] \ .$$

Then, because $\sigma_M^2 = 1/12$, $E_M = 1/2$,

$$Proba\Big\{\big|P_{X_n}(\{k\}) - L(\{k\})\big| > \ L(\{k\})\Big[\frac{2b.p}{\sqrt{3.d^{Q-q}}}\Big]\Big\} \le 2p\Gamma'_1(b) \ .$$

By proposition 6.3.4,

$$\left\{ \bigcap_{n+j_t,Bo} \left\{ \left| P_{X_n}(Bo) - L(Bo) \right| \leq L(Bo) \frac{2pb\sigma_M}{E_M \sqrt{d^{Q-q}}} \right\} \right\}$$

$$\supset \left\{ \bigcap_{n+j_t,\{k\}} \left\{ \left| P_{X_n}(\{k\}) - L(\{k\}) \right| \leq L(\{k\}) \frac{2pb\sigma_M}{E_M \sqrt{d^{Q-q}}} \right\} \right\} .$$

There are d^{pq} sets $\{k\}$. Moreover, there is at the maximum $(n_0)^p$ "$n + j_t$" possible. Then,

$$Proba\left\{ \bigcap_{n+j_t,Bo} \left\{ \left| P_{X_n}(Bo) - L(Bo) \right| \leq \ L(Bo) \ \frac{2b.p}{\sqrt{3.d^{Q-q}}} \right\} \right\}$$

$$\geq Proba\left\{ \bigcap_{n+j_t,\{k\}} \left\{ \left| P_{X_n}(\{k\}) - L(\{k\}) \right| \leq \ L(\{k\}) \ \frac{2b.p}{\sqrt{3.d^{Q-q}}} \right\} \right\}$$

$$= 1 - Proba\left\{ \complement \bigcap_{n+j_t,\{k\}} \left\{ \left| P_{X_n}(\{k\}) - L(\{k\}) \right| \leq \ L(\{k\}) \ \frac{2b.p}{\sqrt{3.d^{Q-q}}} \right\} \right\}$$

$$= 1 - Proba\left\{ \bigcup_{n+j_t,\{k\}} \left\{ \left| P_{X_n}(\{k\}) - L(\{k\}) \right| > \ L(\{k\}) \ \frac{2b.p}{\sqrt{3.d^{Q-q}}} \right\} \right\}$$

$$\geq 1 - \sum_{n+j_t,\{k\}} Proba\left\{ \left| P_{X_n}(\{k\}) - L(\{k\}) \right| > \ L(\{k\}) \ \frac{2b.p}{\sqrt{3.d^{Q-q}}} \right\}$$

$$= 1 - \sum_{n+j_t,\{k\}} 2p\Gamma_1'(b)$$

$$= 1 - 2p.n_0^p d^{pq} \Gamma_1'(b) \ . \ \blacksquare$$

Now, $\Gamma_1'(b) \approx \Gamma(b) \approx \frac{\sqrt{2}}{\sqrt{\pi}b} \ e^{-b^2/2}$ when b is big (cf (28)' page 56 [31]). Then, if d=2, $2p(n_0)^p 2^{pq} \Gamma_1'(b) \approx \frac{2\sqrt{2}.p(n_0)^p 2^{pq} e^{-b^2/2}}{\sqrt{\pi}b} \approx \frac{2\sqrt{2}.pe^{Log(n_0)p} e^{log(2)pq} e^{-b^2/2}}{\sqrt{\pi}b}$. Then, we have to impose $b \geq \sqrt{2[Log(n_0)p + Log(2)pq]}$ in order that the inequality 6.1 is useful.

We deduce the following properties.

Property 6.3.2 *Assume d=2. Then, in order that the inequality 6.1 is useful, we can impose* $b = \sqrt{3p[Log(n_0) + q]}$.

Proof Indeed, in this case,

$$2p(n_0)^p 2^{pq}\Gamma_1'(b) \approx \frac{2\sqrt{2}.pe^{Log(n_0)p}e^{Log(2)pq}e^{-b^2/2}}{\sqrt{\pi}b}$$

$$= \frac{2\sqrt{2}.pe^{Log(n_0)p}e^{Log(2)pq}e^{-1.5.p[Log(n_0)+q]}}{\sqrt{3\pi p[Log(n_0)+q]}}$$

$$\leq \frac{2\sqrt{2}.p.e^{-Log(n_0)p/2}e^{-pq/2}}{\sqrt{3\pi p[Log(n_0)+q]}} \ . \ \blacksquare$$

Now, b has to be not too large.

Property 6.3.3 *Assume d=2. Then, in order that the inequality 6.1 is useful, we can impose* $\epsilon_3 = \frac{[Log(n_0)+q]p^3 2^q}{m} << 1$.

Proof If one chooses $b = \sqrt{3p[Log(n_0)+q]}$

$$\frac{2bp}{\sqrt{3*2^{Q-q}}} \leq \frac{2p\sqrt{3p[Log(n_0)+q]}}{\sqrt{3*2^{Q-q}}} = \frac{2\sqrt{p^3[Log(n_0)+q]}}{\sqrt{2^{Q-q}}} \leq 2\sqrt{\frac{[Log(n_0)+q]p^3 2^q}{m}} \ .$$

If $\epsilon_3 << 1$,

$$Proba\left\{\bigcap_{n+j_t,Bo}\left\{\left|P_{X_n}(Bo) - L(Bo)\right| \leq \ 2\sqrt{\epsilon_3}L(Bo)\right\}\right. \approx 1 \ . \ \blacksquare$$

Property 6.3.4 *If d=2, in order that the inequality 6.1 is useful, we can impose*
1) if one regards all the possible "p", $\frac{[Log(n_0)+q](n_0)^3 2^q}{m} << 1$,
2) if one regards all the $p \leq p_m = \lfloor Log(n_0)/log(2) \rfloor$, $\frac{[Log(n_0)+q](p_m)^3 2^q}{m} << 1$.

6.3.3 Continuous density

Consider the vector $(X_{n+j_1}, ..., X_{n+j_p})$. We know that

$$\sum_{x_{s_1}^1 \in \left[\frac{c_1}{m}, \frac{c_1'}{m}\right[} ..... \sum_{x_{s_p}^p \in \left[\frac{c_p}{m}, \frac{c_p'}{m}\right[} p_{x_{s_1}^1, ..., x_{s_p}^p}$$

$$= \sum_{(y_{n+j_1}, ..., y_{n+j_p}) \in A^1 \otimes .... \otimes A^p} \mathbb{E}\left\{1_{A^1 \otimes .... \otimes A^p}(y_{n+j_1}, ..., y_{n+j_p})\right\} \ .$$

Let us suppose that $(Y_{n+j_1}, ..., Y_{n+j_p})$ has a density function with a Lipschitz coefficient K_0 not too large. Suppose that the $c'_t - c_t$ are large enough. We know that the sets $A^t = \overline{T}^{-1}(mI_t)/m = \overline{T}^{-1}([c_t, c'_t[)/m$ are well distributed in F(m), i.e. the sets A^t have a distribution close to that of $\{r/N(I_t)|r = 0, 1, ..., N(I_t)\}$ (see below). Then, by applying the traditional methods of integration, it is clear that

$$\sum_{x^1_{s_1} \in I_1} ..... \sum_{x^p_{s_p} \in I_p} p_{x^1_{s_1}, ..., x^p_{s_p}} \approx \prod_{t=1}^{p} L(I_t) \ .$$

Under this assumption, one thus obtains easily IID sequences. Let us recall that, when $n_0 << d^p$, one often accepts like model a model with continuous density and a Lipschitz coefficient K_0 not too large. That shows that the functions T_q are a good tool to obtain IID sequences.

Theoretical study

For example if p=1, the following property holds.

Property 6.3.5 *Let $m >> 1$. Let h_N be the probability density function of $Y \in F(m)$ with respect to μ_m : $\int_0^1 h_N(u)\mu_m(du) = 1$. Let $h'_N = (1/c_0)h_N$ be the probability density function such that $\int_0^1 h'_N(u)du = 1$.*

Let $K_0 \in \mathbb{R}_+$ such that $|h_N(r) - h_N(r')| \leq K_0|r' - r|$ and $|h'_N(r) - h'_N(r')| \leq K_0|r' - r|$ when $r, r' \in [0, 1]$.

Then, the following equality holds : $P\{\overline{T}(mY)/m \in I\} = L(I)\Big[1 + \frac{O(1)K_0}{N(I)}\Big]$, where $N(I) \leq m/2$.

Proof We need the following lemmas.

Lemma 6.3.6 *The following equality holds :*

$$c_0 = 1 + \frac{O(1)K_0}{m} \ .$$

Proof The following equalities hold :

$$1 = \sum_t \int_{t/m}^{(t+1)/m} h'_N(u)du = \sum_t \int_{t/m}^{(t+1)/m} \big[h'_N(t/m) + Ob(1)K_0/m\big]du$$

$$= \frac{1}{m}\sum_t h'_N(t/m) + \frac{Ob(1)K_0}{m} = \int_0^1 h'_N(u)\mu_m(du) + \frac{Ob(1)K_0}{m} \ .$$

Then, $\int_0^1 h'_N(u)\mu_m(du) = 1 + \frac{Ob(1)K_0}{m}$. Therefore,

$$1 = \int_0^1 h_N(u)\mu_m(du) = c_0 \int_0^1 h'_N(u)\mu_m(du) = c_0\Big[1 + \frac{Ob(1)K_0}{m}\Big] \ . \ \blacksquare$$

Lemma 6.3.7 *The following equality holds :* $\frac{1}{N(I)}\sum_r h_N(r/N(I)) = 1+\frac{2Ob(1)K_0}{N(I)}$.

Proof The following equalities hold :

$$1 = \sum_r \int_{r/N(I)}^{(r+1)/N(I)} h'_N(u)du = \sum_r \int_{r/N(I)}^{(r+1)/N(I)} \big[h'_N(r/N(I))+Ob(1)K_0/N(I)\big]du$$

$$= \frac{1}{N(I)}\sum_r h'_N(r/N(I)) + \frac{Ob(1)K_0}{N(I)} \ .$$

Therefore $c_0 = \frac{1}{N(I)}\sum_r h_N(r/N(I)) + \frac{Ob(1)c_0K_0}{N(I)}$.

Therefore, by lemma 6.3.6,

$$c_0 = 1 + \frac{O(1)K_0}{m} = \frac{1}{N(I)}\sum_r h_N(r/N(I)) + \frac{Ob(1)[1+\frac{O(1)K_0}{m}]K_0}{N(I)} \ .$$

Because $m >> 1$ and $N(I) \leq m/2$, we deduce the lemma. ■

Then, the following property holds.

Property 6.3.8 *Let* $I = [c/m, c'/m[$. *Let* $g_N(k) = h_N\big(\overline{T}^{-1}(k)/m\big)$. *Assume again that* T *is a Fibonacci congruence. The following approximation holds*

$$\frac{1}{N(I)}\sum_{k=c}^{c'-1} g_N(k) = 1 + \frac{6Ob(1)K_0}{N(I)} \ .$$

Proof Let k^n, n=1,2,..,c'-c, be a permutation of $I \cap F(m) = \{c/m, (c+1)/m, ..., (c'-1)/m\}$ such that $\overline{T}^{-1}(k^1) < \overline{T}^{-1}(k^2) < \overline{T}^{-1}(k^3) < ..... < \overline{T}^{-1}(k^{c'-c})$. Then, for all numerical simulations which we executed, one has always obtained

$$|\overline{T}^{-1}(k^r)/m - r/N(I)| \leq 4/N(I) \ .$$

We deduce that $|g_N(k^r) - h_N(r/N(I))| \leq 4K_0/N(I)$.

Therefore, by lemma 6.3.7,

$$\frac{1}{N(I)}\sum_{k=c}^{c'-1} g_N(k) = \frac{1}{N(I)}\sum_r g_N(k^r)$$

$$= \frac{1}{N(I)} \sum_r h_N(r/N(I)) + \frac{1}{N(I)} \sum_r \left[g_N(k^r) - h_N(r/N(I)) \right]$$

$$= \frac{1}{N(I)} \sum_r h_N(r/N(I)) + \frac{4Ob(1)K_0}{N(I)} = 1 + \frac{2Ob(1)K_0}{N(I)} + \frac{4Ob(1)K_0}{N(I)} \ . \blacksquare$$

Remark 6.3.9 *The only result which is not proven mathematically is*

$$|T^{-1}(k^r)/m - r/N(I)| \leq 4/N(I) \ .$$

It is enough to prove this result in order that property 6.3.5 is fully proven. We point out that, by our numerical study, this result seems sure.

Proof of property 6.3.5 By the previous equalities,

$$P\{\overline{T}(Y)/m \in I\} = \frac{1}{m} \sum_k g_N(k) = \frac{N(I)}{m} \Big[1 + \frac{6Ob(1)K_0}{N(I)} \Big]$$

$$= L(I) \Big[1 + \frac{Ob(1)}{m} \Big] \Big[1 + \frac{6Ob(1)K_0}{N(I)} \Big] = L(I) \Big[1 + \frac{O(1)K_0}{N(I)} \Big] \ . \blacksquare$$

Remark 6.3.10 *One can easily generalize the proof of property 6.3.5 to the two-dimensional case. For example, if p=2, by proposition A.0.1,*

$$P\big\{ \big(\overline{T}(mY_1)/m, \overline{T}(mY_2)/m \big) \in I_1 \otimes I_2 \big\} = L(I_1)L(I_2) \Big[1 + \frac{O(1)K_0}{Inf_s[N(I_s)]} \Big] \ .$$

Remark 6.3.11 *The previous results can be proved in another manner. In this case, there is a less fine approximation : cf property 7.1.21 of [18].*

Numerical results

All the results which we have obtained confirm the previous result. For example, when $h_N(y) \approx \frac{10.e^{-[10(y-0.5)]^2/(2\sigma^2)}}{2\pi\sigma^2}$, we have obtained the following tables where m= 2178309 and I = [541231,1905574[, $\epsilon = N(I)/m - P\{X \in I\}$.

σ^2	ϵ
1/4	0.0000042
1/2	0.0000028
1	0.0000016
1/16	0.0000225
1/128	0.0001511
1/1024	0.0003016

Many other numerical results are in [18] section 7.1.2. All the results which we have obtained confirm the property 6.3.5.

6.3.4 General numerical results

In [18] , we have studied the case where the probability density function of Y with respect to μ_m is written in a form : $h(y)\big[1+\frac{\eta(y)}{co}\big]$, where $\eta(y)$ is a sample of a white noise independent of h and where $\int h(y)\mu_m(dy) \approx 1$. We have obtained equivalent result. Here we recall some results when h is a normal or uniform density.

Normal case

We assume that $h(y) \approx \frac{10.e^{-[10(y-0.5)]^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$, $co \geq 10$. Then, we have proved that $P\Big\{\frac{m\sqrt{\sigma}|\epsilon_I^G|}{0.0485\sqrt{N(I)}} \geq b\Big\} \approx \Gamma(b)$ where $\epsilon_I^G = N(I)/m - P\{X \in I\}$. This result thus gives us a probable increase of $|\epsilon_I^G|$.

Example 6.3.12 *Suppose that we do not have more than* 10^6 *possible intervals I. We know that* $\Gamma(6) \leq 10^{-9}$*.Then, if N(I) is not too small, one can assume*

$$|\epsilon_I^G| \leq \frac{0.291\sqrt{N(I)}}{m\sqrt{\sigma}} \ . \tag{6.2}$$

For example, suppose m= 267914296 , a= 165580141. We choose intervals I length $L(I) = (1/5)10^{-j}$ for j =1,...,6 . Choose standard deviations $\sigma = 1/2$, 1/4, 1/8, 1/40.

For each j, one calculated each ϵ_I^G for 50 intervals I_s, s=1,2,..,50 length $(1/5)10^{-j}$. Then, one obtains the following table of $Max_s\{0.5*10^6|\epsilon_I^G| \mid I_s,\ \sigma\}$ on these 50 terms.

L(I) $\setminus\ \sigma$	1/2	1/4	1/8	1/40
$(1/5)10^{-1}$	0.0708	0.0837	-0.0321	0.5361
$(1/5)10^{-2}$	0.1096	-0.0114	-0.1507	-0.1077
$(1/5)10^{-3}$	0.0067	0.0328	0.0097	-0.1834
$(1/5)10^{-4}$	-0.0008	0.0004	0.0046	0.0083
$(1/5)10^{-5}$	-0.0008	-0.0014	0.0025	-0.0152
$(1/5)10^{-6}$	-0.0000	-0.0010	0.0010	0.0013
$(1/5)10^{-7}$	-0.0006	0.0002	-0.0032	-0.0044

We have other various numerical results : a more complete study has been done in section 7.1 of [18]

Uniform distribution

Here we study the model $P_Y\{Y = k/m\} = \frac{1}{m}\Big[1 + u_k\Big]$ where u_k is a sample of an IID sequence U_k with variance σ_U^2.

Let I be an interval. We set $\epsilon_I = N(I)/m - P\{X \in I\}$. Let $N_{Iel} = Sup_k\Big|card\Big[F(m) \cap [k/2^q, (k+1)/2^q[\Big]\Big| = \lfloor m/2^q \rfloor + 1$. Then, for all interval I_k, generally,

$$P\Big\{\frac{m|\epsilon_{I_k}|}{\sigma_U\sqrt{N_{Iel}}} \geq b_q\Big\} \leq 4^{-q} \ .$$

Because there are only 2^q intervals $I_k = [k/2^q, (k+1)/2^q[$, one can admit

$$|\epsilon_{I_k}| \leq \frac{b_q\sigma_U\sqrt{N_{Iel}}}{m} \ . \tag{6.3}$$

For example for datas $h(n)$ used section 11.1.1, if $m \geq 1.4 * 10^{31}$, $\sigma_U \leq 1$ and if q=84, one choose $b_q = 15$, $N_{Iel} \approx 7520$. Then, one can suppose

$$|\epsilon_{I_k}| \leq \frac{9.3}{10^{29}} \ .$$

A more complete study has been done in section 7.2 of [18]

6.3.5 Other congruences

Similar results are obtained with other congruences. But the approximations are less good. For example we proved the following result in section 4.1.1 of [18].

Proposition 6.3.6 *Let $(d,p) \in \{2,3,....,\}^2$. Let $T(x) \equiv d^p x \ mod \ m = d^{2p} - 1$.*

Let $Z \in F(m)$ be a random variable. Let f be the density of Z with respect to μ_m. Let $K_0 > 0$ such that, for all $z, z' \in F(m)$, $|f(z) - f(z')| \leq K_0|z - z'|$. Then,

$$P\big\{\overline{T}(Z)/m \in I\big\} = \frac{N(I)}{m} + \frac{O(6K_0)}{d^p} \ .$$

6.3.6 Remarks

Remark 6.3.13 *The previous results were obtained by considering that one chose randomly a measure in the set of the possible probabilities. But, for that , one needs that the probabilities of the X_n are not concentrated in a small number of points. If not, the majority of the p_{x_s} will be equal to 0 and could not thus be regarded as chosen randomly. Of course, it is one of the exception envisaged in section 5.4 of [18]. But it is better to remove this case.*

At first, it is necessary that one has a sample of y_n which all are different. In this aim, it is enough that m is large with respect to n_0.

Moreover, it is better that a priori *all the possible values of $F(m)^p$ can exist in a sample. For p=1, it is reasonably the case when one adds modulo m a*

pseudo-random sequence g_n of period m : $my'_n = \overline{g_n + my_n}$. For p=2, one can use two generators $g^1_{2n'}$ and $g^2_{2n'+1}$: if n=2n', $my'_{2n'} = \overline{g^1_{2n'} + my_{2n'}}$

Remark 6.3.14 *We can make this study without supposing that T is the congruence of Fibonacci. But, the conclusion does not hold when T is the identity and that the curve of the probabilities have the shape of a normal curve. It is natural : contrary to the congruence of Fibonacci, there is a dependence between T and text : if y_n means an extract of texts, $T(y_n)$ means the same extract of text. It is the same when $T(x) \equiv d^p x$ modulo $d^{2p} - 1$ (cf proposition 6.3.6).*

Remark 6.3.15 *There exists another method to prove our conclusions: it should be considered that the probabilities are fixed and the a_i are taken randomly : cf section 8.3 of [18].*

6.4 Study of models

A more detailed study of models is in chapter 13 of [18].

6.4.1 Continuous densities

Let us suppose that one has a sample resulting from texts, y_n, $n = 1, 2, ..., n_0$, $y_n \in F(m)$ where $n_0 << m$. We suppose $y_n \neq y_{n'}$ if $n \neq n'$. Generally, that occurs always if m is great enough with respect to n_0. This assumption involves that, for all subsequence $y_{t(n)}$ and for all p, $(y_{t(n)}, ..., y_{t(n+p-1)}) \neq (y_{t(n')}, ..., y_{t(n'+p-1)})$ if $n \neq n'$.

One can always regard $y_n \in F(m)$ as the realization of a sequence of random variables Y_n : $y_n = Y_n(\omega)$ such that Y_n has a a differentiable density with respect to $\mu_m \otimes .... \otimes \mu_m$. One assume also that this density have a Lipschitz coefficient K_0 which is not too large.

It is a logical assumption. In fact it is an assumption which most mathematicians admit when $N << m$: that is especially clear when they estimate the densities.

We have studied numerous examples which corroborate this hypothesis : cf [18].

Now, in the case where densities are continuous, the conditional densities are also continuous. Then, the conditional probabilities $P\{Y_n | y_{n+j_2} = y_2, ..., Y_{n+j_p} = y_p\}$ has a continuous density $f_{y_2,...,y_p}$ with a coefficient Lipschitz K_0^{cp} which is not too great. Then, by the same technique as for property 6.3.5, one obtains

$$P\{T_q(Y_n) \in I \mid Y_{n+j_2} = y_2, ..., Y_{n+j_p} = y_p\} = L(I)\Big[1 + \frac{O(1)K_0^{cp}}{N(I)}\Big] ,$$

where h_0 is chosen big enough : $N(I) \geq h_0$.

6.4.2 Another group of models

In order to prove the previous equation, we used K_0. But it is enough to read the proofs of property 6.3.5 in order to understand that it would be possible to use the coefficients of Lipschitz K^r associated with each interval $[r/N(I), (r+1)/N(I)[$ to obtain the same type of results.

In this case, it would be enough that $\sum_r K^r$ is not too large. It is felt well intuitively that this kind of conditions is satisfied by our models.

Admittedly, it is easier to understand for the classical densities of $(Y_1, ..., Y_N)$. But what interests us, are the conditional probabilities. Then to understand that the property "$\sum_r K^r$ not too large" is checked for the conditional densities, simplest way is to remember that the conditional density $f_{y_2,,y_p}(y_1)$ is equal to the product of the marginal densities $f_2(y_2, y_3, ., y_p)$ and of the density of dependence (cf [10] or proposition 14.3.2 of [18]) :

$$f_{y_2,...,y_p}(y_1) = f^{dep}(y_1; y_2, y_3, ..., y_p) f_2(y_2, y_3, ..., y_p) \ .$$

Then, one confirms this assumption with simulations by using this equality.

Therefore, this kind of conditions on $\sum_r K^r$ seems checked by our models. That makes our end result even surer.

6.4.3 General case

The use of the previous models is interesting because that y_n behaves well like a sample of one of these possible models Y_n. We thus do not make any error while putting to us under these assumptions. Our calculations are thus right. That implies that the bits $b^4(n')$ obtained by our construction in section 10.1.4 behave well like IID sequences.

But it is probable that there do not need even to suppose to be under the assumptions of these models: it is probable that, for *any logical model*, one will still obtain

$$P\{T_q(Y_n) \in I \mid Y_{n+j_2} = y_2, ..., Y_{n+j_p} = y_p\} \approx L(I) \ .$$

A very strong result

Indeed, we understood that if one provides the set of possible probabilities with the measure defined in section 6.3, our results are checked for almost all the possible probabilities.

There is thus a slight restriction which is normal: in the set of ALL the models, there will be an infinity of them which will not be appropriate. However, it is already extraordinary that the result is true for *almost all* the possible models.

In order to understand it, let us take for example a sample really IID y_n. One wants to associate with y_n a model Y_n. Clearly, among all the possible

models, there is an infinity of good models and an infinity of bad ones. One can even think that there is much more bad models than goods.

In section 6.3, it is the opposite: in the set of all the possible models, almost all will be good. It is a very strong result.

A result checked by all the logical models

As a matter of fact, one can remove the bad models : one can admit that $P\{T_q(Y_n) \in I \mid Y_{n+j_2} = y_2, ..., Y_{n+j_p} = y_p\} \approx L(I)$ will be checked for all the *logical* possible models.

Indeed, there is no connection between the $\overline{T}^{-1}(mI_k)$ and texts. Therefore, if a model was bad, that would mean that there is a logical connection between the $T_q^{-1}(I_k)$ and text. One can thus a priori exclude a such model.

Thus our result holds with all the possible logical models, those where there is no connection between text and the $T_q^{-1}(I)$.

Now, it is obtained that $P\big\{Y_1 \in T_q^{-1}(I)\big\} \approx \frac{(c'-c)}{m}\left[1 + \frac{Ob(1).b}{\sqrt{3N(I)}}\right]$ for all the logical models. Then the question is put: which value to choose for b?

In order to know that, it is necessary to go back to the themself texts : i.e. it is necessary to study the associated empirical probabilities.

We thus estimated b for various texts and for various $T_q^{-1}(I)$.

If p=1, all the numerical studies that we have made show that, for intervals I of the same length, the sets $T_q^{-1}(I)$ contains about the same number of possible texts : cf [18] section 13.1.2.

Finally, it is found that one can admit - and by far - in all the cases

$$P\big\{Y_1 \in T_q^{-1}(I)\big\} = \frac{(c'-c)}{m}\left[1 + \frac{Ob(1).20}{\sqrt{3N(I)}}\right]. \tag{6.4}$$

This increase (b=20) is not astonishing. Indeed, according to proposition 6.3.2, it occurs with a probability larger than $1 - 2p\Gamma_1'(b)$. Now, if b=20, $\Gamma_1'(b) \approx 1.12/10^{88}$. Then, a priori, in order to find a case where that is not true, it would be necessary to use a such large number of texts that it is impossible to realize.

In any case it is even not sure that one can find cases where this equality is not checked empirically. Indeed, one does not use texts representing samples which have a fixed law. What one uses, it is, on the one hand, sequences which have the logic of the English language, and on the other hand, sets which have simple mathematical properties. Anyway, we never have encounter such a case. It is thus possible logically that it has no text not checking the equation 6.4.

If p=2, sets $T_q^{-1}(I_1)$ and $T_q^{-1}(I_2)$ behave like randomly selected compared to the text.

If $p > 2$, we have obtained results equivalent for $p \leq \frac{log(n_0)}{log(2)}$:

$$P\Big\{\big\{Y_1 \in T_q^{-1}(I_1)\big\}\cap.....\cap\big\{Y_p \in T_q^{-1}(I_p)\big\}\Big\} \approx \frac{\prod_s (c'_s - c_s)}{m^p}\left[1+\frac{Ob(1).20.p}{\sqrt{3.Inf\{N(I_s)\}}}\right].$$

Anyway a such value of b is not important because the equation

$$P\{X_n \in I \mid X_{n+j_2} = x_2, ..., X_{n+j_p} = x_p\} = L(I)[1 + Ob(1)\epsilon]$$

is too strong. Indeed it is enough that

$$P\{X_n \in I \mid X_{n+j_2} = x_2, ..., X_{n+j_p} = x_p\} = L(I) + Ob(1)\epsilon$$

where $\epsilon = \frac{\alpha}{\sqrt{N}}$ when N is the size of sample and $0 < \alpha \leq 0.02$. Indeed, in this case one cannot differentiate X_n with an IID sequence : cf section 2.1.4 or example in section 11.2.3.

Then, the equation

$$P\{X_n \in I \mid X_{n+j_2} = x_2, ..., X_{n+j_p} = x_p\} = L(I)[1 + Ob(1)\epsilon]$$

is too strong.

Therefore a very strong connection between text and the sets $T^{-1}(I_k)$ would be necessary in order not to obtain this kind of equation. It is not therefore advisable to worry about the value of b.

Conditional probabilities

The fact that there is no connection between text and the sets $T_q^{-1}(k/d^q) = \{a_1, ..., a_{c'-c}\}$ applies to the conditional probabilities. Indeed, there is always no logical connection between text and the sets $T_q^{-1}(I) = \{a_1, ..., a_{c'-c}\}$ in the conditional probabilities:

$$P\{X_n \in I \mid Y_{n+j_2} = y_2, ..., Y_{n+j_p} = y_p\}$$

$$= \frac{P\Big\{\ \{Y_n \in T_q^{-1}(I)\} \cap \{Y_{n+j_2} = y_2\} \cap .... \cap \{Y_{n+j_p} = y_p\}\ \Big\}}{P\Big\{\ \cap \{Y_{n+j_2} = y_2\} \cap .... \cap \{Y_{n+j_p} = y_p\}\ \Big\}}\ .$$

Therefore, the conditional probabilities behave well as sums on sets taken randomly, i.e.

$$P\{T_q(Y_n) \in I \mid Y_{n+j_2} = y_2, ..., Y_{n+j_p} = y_p\}$$

$$= \sum_s P\{Y_n = a_s \mid Y_{n+j_2} = y_2, ..., Y_{n+j_p} = y_p\} \approx L(I)\ .$$

Chapter 7

Limit Theorems

7.1 Central Limit Theorem

The Central Limit Theorem (CLT) produces the limit distribution of $(X_1 + ...X_n)/\sigma$ when X_n is a sequence of random variables such that $\mathbb{E}\{X_n\} = 0$ and σ^2 is the variance.

It has been proved under various hypotheses of asymptotical independence., in particular under the strong mixing condition or under martingale assumptions : cf [21] and [28]. Now, these condition are too strong for most of datas. Then, some authors have introduce weaker hypotheses : Versik Ornstein ([22], [23]), Cogburn [25] Rosenblatt [26], Pinskers [7], Doukhan-Louhichi [27]. But theses assumptions are still strong in order to be used with data.

Fortunately, another look is possible : in [11] , one can use higher order correlation coefficients (cf Lancaster [9], Blacher [10]). Then, in [11] we have turned the convergence of moments into an equivalent conditions on these coefficients. For example we have proved the following theorem.

Theorem 3 *Assume that the X_n have the same distribution with variance σ^2 and that there exists $bo > 0$ such that $|X_n| \leq bo$. Assume that*

$$\frac{\sum_{s=1}^{n} \sum_{r \neq s} \left[\mathbb{E}\{(X_s)^2 (X_r)^2\} - \mathbb{E}\{(X_s)^2\}\mathbb{E}\{(X_r)^2\}\right]}{n^2} \to 0 \ .$$

Let $\mu_p = \mathbb{E}\{(X_G)^p\}$ where $X_G \sim N(0,1)$. Then, for all $p \in \mathbb{N}^*$,

$$\mathbb{E}\Big\{\Big(\frac{X_1 + X_2 + ... + X_n}{\sqrt{(N_2 + \sigma^2)n}}\Big)^p\Big\} \to \mu_p \quad as \ n \to \infty$$

if and only if, for all $p \in \mathbb{N}^*$,

$$p! \frac{\sum_{t_1=1}^{n} \sum_{t_2=t_1+1}^{n} .... \sum_{t_p=t_{p-1}+1}^{n} \mathbb{E}\{X_{t_1} X_{t_2} ..... X_{t_p}\}}{n^{p/2}} \to (N_2)^p \mu_p \ .$$

From this type of results we have deduced CLT with minimal assumptions whose the conditions are close to strong mixing assumptions.

In this aim, one decomposes $X_1+X_2+.....+X_n$ in $X_1+X_2+.....+X_u$, $X_{u+1}+X_{u+2}+.....+X_{u+t}$ and $X_{u+t+1}+X_{u+t+2}+.....+X_{u+t+u}$ where u=u(n) and t=t(n).

Notations 7.1.1 *We denote by $\kappa(n)\in\mathbb{N}$, an increasing sequence such that $\kappa(1)=0$, $\kappa(n)\leq n$ and $\kappa(n)/n\to 0$. We define the sequences u(n) and t(n) by : u(1)=1, $u(n)=max\big\{m\in\mathbb{N}^*\big|2m+\kappa(m)\leq n\big\}$, and t(1)=0, t(n) = n-2u(n) if $n\geq 2$.*

Notations 7.1.2 *Let $\sigma(u)^2=\mathbb{E}\{(X_1+X_2+.....+X_u)^2\}$. One sets*

$$S_u=\frac{X_1+X_2+.....+X_u}{\sigma(u)}\ ,\ \xi_u=\frac{X_{u+1}+X_{u+2}+.....+X_{u+t}}{\sigma(u)}$$

$$and\ \ S'_u=\frac{X_{u+t+1}+X_{u+t+2}+.....+X_{u+t+u}}{\sigma(u)}\ .$$

Then, one can define almost minimal assumptions for the convergence of moments.

Notations 7.1.3 *: Let $k\in\mathbb{N}^*$. We define conditions $H_{mS}(k)$ and $H_{mI}(k)$ by the following way :*
$H_{mS}(k)$: $\forall p\in\mathbb{N}$, $p<k+1$, $E\big\{(S_u)^p\big\}-E\big\{(S'_u)^p\big\}\to 0$ as $n\to\infty$.
$H_{mI}(k)$: $\forall(p,q)\in(\mathbb{N}^*)^2$, $p+q<k+1$,

$$E\big\{(S_u)^p(S'_u)^q\big\}-E\big\{(S_u)^p\big\}E\big\{(S'_u)^q\big\}\to 0$$

as $n\to\infty$.

Equivalent conditions can be defined for the convergence in distribution.

Notations 7.1.4 *: We define condition H_S and H_I by the following way.*

$$H_S:\forall k\in\mathbb{N},\forall j\in\mathbb{N},P\{A_{k,j}\}-P\{B_{k,j}\}\to 0\ \ as\ \ n\to\infty\ ,$$
$$H_I:\forall k\in\mathbb{N},\forall(j,j')\in\mathbb{N}^2,P\{A_{k,j}\cap B_{k,j'}\}-P\{A_{k,j}\}P\{B_{k,j'}\}\to 0$$

as $n\to\infty$, where $A_{k,j}=\big\{S_u\in I_{k,j}\big\}$ and $B_{k,j}=\big\{S'_u\in I_{k,j}\big\}$ with $I_{k,j}=\big[j.4^{-k},(j+1)4^{-k}\big[\big\}$.

Then the following CLT holds : cf [14] [15].

Theorem 4 *We assume that $H_{mS}(\infty)$ and $H_{mI}(\infty)$ hold. We assume also that, for all $p\in\mathbb{N}^*$, $\mathbb{E}\{(\xi_u)^p\}\to 0$ as $n\to\infty$. Then, $S_n\overset{D}{\to}N(0,1)$.*

Theorem 5 *We assume that H_S , H_I , $H_{mS}(4)$ and $H_{mI}(4)$ hold. We assume also that $\mathbb{E}\{(S_u)^2\}-\mathbb{E}\{(S'_u)^2\}\to 0$ and $\mathbb{E}\{\xi_u^2\}\to 0$ as $n\to\infty$. Then, $S_n\overset{D}{\to}N(0,1)$.*

It is this CLT that we use with our datas (cf chapter 9).

7.2 XOR Limit Theorem

Our secund limit theorem is based on the property of XOR. But it holds also modulo m. Then, by misusing of language, we call this this result "XOR Limit Theorem" (XORLT). The XORLT is much more general than the CLT. Then, one can use with many type of datas, in particular with the most part of electronic files.

7.2.1 Presentation

The XORLT relates to sums $\alpha(n)(X_1 + X_2 + ... + X_n)/\sigma(n)$, in particular $\overline{X_1 + X_2 + ... + X_n}$ (in this section, $\overline{h} \equiv h$ modulo 1).

Definition 7.2.1 *Let $(X_n^1, X_n^2, ...X_n^p) \in \mathbb{R}^{\mathbf{p}}$ be a sequence of random vectors. For s=1,...,p, let $\sigma_s(n)^2 = \mathbb{E}\{(X_1^s + ... + X_n^s)^2\}$. Then, we set*

$$S_n^s = \frac{X_1^s + ... + X_n^s}{\sigma_s(n)} \ .$$

The XOR limit theorem holds for $(X_n^1, X_n^2, ...X_n^p)$ if there exists p sequences $\alpha_s(n) \to \infty$ as $n \to \infty$, such that $(\overline{\alpha_1(n)S_n^1}, ..., \overline{\alpha_p(n)S_n^p})$ has asymptotically the uniform distribution on $[0,1[^p$.

As a matter of fact, we have always obtained that $\overline{X_1 + X_2 + ... + X_n}$ has asymptotically the uniform distribution on $[0,1[$. In order to understand that, we recall the following theorem (cf theorem 4 [18]).

Theorem 6 *Let X and Y be two independent random vectors, $X,\ Y \in F^*(m)^p$. Assume that X has the uniform distribution. Then, $\overline{X+Y} \in F^*(m)^p$ has also the uniform distribution.*

For example assume that $X_1 \in [0,1[$ has the uniform distribution and that X_1 is independent of $(X_2, ..., X_n)$. Then $\overline{X_1 + X_2 + ... + X_n}$ has the uniform distribution on [0,1[.

This result is general : if the CLT is satisfied, then the XORLT is satisfied.

Proposition 7.2.1 *: Let X_n be a sequence of random variables such that $\mathbb{E}\{X_n\} = 0$ and $S_n = \frac{X_1+....+X_n}{\sigma(n)} \xrightarrow{D} S$ with $\mathbb{E}\{S^2\} = 1$. Assume that S has a probability density function f with respect to the Lebesgue measure such that $|f(x) - f(x')| \leq K_0|x - x'|$.*

Then, there exists a sequence $\alpha(n) \to \infty$ as $n \to \infty$ such that, for all $0 \leq t \leq 1$, $P\{\overline{\alpha(n)S_n} \in [0,t[\} \to t$ as $n \to \infty$.

The proof is in proposition 5.2.3 of [18]. Remark that analog results hold also for random vectors $(S_1^n, S_2^n, ..., S_p^n) \in \mathbb{R}^p$.

In general, $\alpha(n)\sigma(n) = 1$, i.e. $\overline{(X_1 + ...X_n)}$ has asymptotically the uniform distribution on $[0,1[$. Indeed let $\mu_n'(\frac{k}{m\sigma(n)}) = \frac{1}{m\sigma(n)}$ for all $k \in \mathbb{Z}$. Then,

$\mu'_n([0,1]) \to 1$ as $n \to \infty$. Now the following XORLT holds (cf proposition 5.2.4 of [18]).

Proposition 7.2.2 *Let $(S_1^n, S_2^n, ..., S_p^n) \in \mathbb{R}^{\mathbf{p}}$ be a random vector such that $E\{(S_n^s)^2\} = 1$ for s=1,2,...,p.*

Let μ^A be a measure on $\mathbb{R}$: one assumes that $\mu^A = \mu^1 \otimes .... \otimes \mu^p$ where $\mu^s = \mu$ the Lebesgue measure for s=1,...,p or where $\mu^s = \mu'_n$ for s=1,..,p. Assume that $(S_1^n, S_2^n, ..., S_p^n)$ has a probability density function f_n with respect to μ^A such that $|f_n(x_1, ..., x_p) - f_n(x'_1, .., x'_p)| \leq K_0 max(|x_s - x'_s|)$.

Let $\alpha(n)$ be a sequence such that $\alpha(n) \to \infty$ as $n \to \infty$.

Then $\overline{\alpha(n)(S_1^n, S_2^n, ..., S_p^n)}$ has asymptotically the uniform distribution over $[0,1[^p$.

Now, the condition " $\exists K_0 \ : \ |f_n(x) - f_n(x')| \leq K_0|x - x'| \ \ \forall \ n \in \mathbb{N}^*$ " is not a necessary condition of the CLT. Then, in some cases, the hypotheses of proposition 7.2.2 seem stronger than those of the CLT. But the reciprocal assertion is true too : e.g. the XORLT holds if $X_s = X_1$ for all s : $\overline{X_1 + ... + X_n} = \overline{nX_1}$.

Moreover, proposition 7.2.2 suggests that if the CLT holds, then, $\overline{X_1 + ... + X_n}$ has asymptotically the uniform distribution. Anyway, under the assumptions of datas studied in this report, we have never found a single case where it is not verified.

7.3 Examples

In this section, we compare limit distributions. In these examples we shall note the strength of the XORLT.

Let $S_n^2 \in \mathbb{R}^2$ such that $S_n^2 \xrightarrow{D} S_0^2$ where $S_0^2 \sim N_2(0, C)$ when C is a covariance matrix. One knows that $g(S_n^2) \xrightarrow{D} g(S_0^2)$ if g is continuous with $P_{S_0^2}$ probability 1 (cf [29] page 24). Then, $\overline{S_n^2} \xrightarrow{D} \overline{S_0^2}$. Moreover, we shall note that the dependence of S_0^2 does not exist any more for $\overline{S_0^2}$. We shall deduce the XORLT for $\sigma(n)\overline{S_n^2}$.

Example 7.3.1 *Let X and Y be two independent random variable with distribution N(0,1). Let $Z = \frac{X+aY}{\sqrt{1+a^2}}$ where $a \in \mathbf{R}$.*

Test of the linear correlation coefficient Under the previous hypotheses, Z has the N(0,1) distribution. Moreover the linear correlation coefficient of X and Z is $\rho = (1 + a^2)^{-1/2}$. For example, $\rho = 0.701$ si a=1.

let ρ_n be the empirical linear correlation coefficient associated to a sample (X_s, Z_s). Let ρ_n^U be the empirical linear correlation coefficient associated to the sample $(\overline{X_s}, \overline{Z_s})$.

Then, ρ_n et ρ_n^U allow us to estimate the linear correlation coefficients of (X_0, Z_0) and $(\overline{X_0}, \overline{Z_0})$.

Let N be the size of the sample. The following results have been obtained

ρ	N	ρ_n	ρ_n^U	ρ_n	ρ_n^U
0.7071	1000	0.7063	-0.0607	0.6941	0.0367
0.4472	1000	0.4597	0.0017	0.4488	-0.0260
0.2425	1000	0.2472	0.0054	0.2167	-0.0252
0.7071	5000	0.7034	-0.0294	0.6996	-0.0012
0.4472	5000	0.4536	0.0002	0.4436	-0.0270
0.2425	5000	0.2351	0.0075	0.2290	0.0216
0.7071	10000	0.7108	0.0061	0.7107	0.0010
0.4472	10000	0.4469	-0.0020	0.4454	-0.0049
0.2425	10000	0.2675	0.0099	0.2478	0.0101
0.7071	100000	0.7074	-0.0011	0.7056	-0.0007
0.4472	100000	0.4433	-0.0013	0.4467	0.0002
0.2425	100000	0.2466	-0.0037	0.2445	-0.0015

Then ρ_n^U is smaller than ρ_n. As a matter of fact, if we do tests, we can even consider that ρ_n^U is the estimate of the correlation coefficient equal to 0.

Chi squared independence test We test the independence of $\overline{X_n}$ and $\overline{Z_n}$ by the chi squared independence test.

Assume that the linear correlation coefficient is equal to 0.7071. We use a partition (15,15). The chi-squared statistics has asymptotically the distribution N(0,1) (cf [1] page 44) : $\sqrt{2\chi_2} - \sqrt{2d-1}$ where d is the degree of freedom : (15-1)(15-1). With this statistics, we obtained the following numerical results.

1.1256	-0.1246	2.0030	-0.8977	-0.7952	0.6594	-0.7758
-0.3079	0.5618	-0.3380	-1.2630	-0.5369	-1.0617	0.9458
-1.6506	-0.3484	0.9821	-0.8853	-0.1215	-0.5373	1.0599

As a matter of fact $(\overline{X}, \overline{Z})$ is enough close to an independent vector.

Conclusion Under the previous hypotheses,

$$(\overline{X1 + ... + X_n}, \overline{Z_1 + ... + Z_n}) \to \overline{\sigma(n)\overline{(X, Z)}} \ .$$

Now $\overline{(X, Z)}$ is already close to an independent vector. Then, it will thus be even more true for $\overline{\sigma(n)\overline{(X, Z)}}$ because the multiplication by $\sigma(n)$ modulo 1 makes uniform the distribution as soon as $\sigma(n)$ is enough big.

In conclusion, the fact that (X, Z) is already almost independent shows the rate of the convergence of the XORLT.

7.3.1 Example using datas of this paper

In section 5.3.3 of [16] we have studied an example using the datas $G(j)$ and H(j) defined in section 11. We note that the estimated density are close to the normal or uniform density.

As a matter of fact, we studied numerically various examples using data of the type "text", "computer programs", "mathematical reports", etc., we always found that $\overline{X_1 + X_2 + ... + X_n}$ has asymptotically the uniform distribution.

We obtained results similar in several dimensions: for the data used in this report, we always found that $(\overline{X_{1,1} + X_{2,1} + ... + X_{n,1}}, \overline{X_{1,2} + X_{2,2} + ... + X_{n,2}})$ has asymptotically the uniform distribution on $[0,1]^p$ for p=2. We obtained similar results for p=3,4,5,6.

7.3.2 Other theoretical study

One can confirm that it is more practical to use the XORLT than the CLT by another theoretical study : one can compare the the conditional densities of the sequences G(j) and H(j) (cf section 11.1.1). Indeed, in corollary 5.6.2 of [18] we have proved the following result.

Proposition 7.3.1 *Let $f^*_{g_2,g_3,..}(g/(mS))$ be the conditional density of $G_i/(mS) = g/(mS)$ given $G_{i+j_s} = g_s$ and let $f^*_{h_2,h_3,..}\{h/m\}$ be the conditional density of $H_i/m = h/m$ given $H_{i+j_s} = h_s$.*

Let $K_{f_{G/(Sm)}}$ and $K_{f_{H/m}}$ be the Lipschitz coefficients associated to $f^*_{g_2,g_3,..}$ and $f^*_{h_2,h_3,..}$.

Then, $K_{f_{H/m}} \leq \frac{K_{f_{G/(Sm)}}}{S}$.

7.3.3 Numerical study

In [18], we have studied several examples of the rate of convergence de $X_1 + X_2 + ... + X_n$ and $\overline{X_1 + X_2 + ... + X_n}$ when $X_s \in \{0, 1, ..., q\}$. It is true as soon as, $n \geq 7$ or if q is large enough $q \geq 20$. For example, in figures 7.1 and 7.2, we obtain curves close to those of the normal or uniform distributions for non -independent vectors $(X_1, X_2, ..., X_8)$. Then, in [18], we notice that the graphs are about the ones of a normal distribution or a uniform distribution except when the probabilities are concentrated near a small number of points : cf figures 7.3 and 7.4.

We have studied numerically the rate of convergence of the XORLT when the CLT is satisfied. Then, we assume $Y = \frac{X_1 + ... + X_n}{\sigma\sqrt{n}} \sim N(0,1)$. Then, $X_1 + ... + X_n = \sigma\sqrt{n}Y \sim N(0, n\sigma^2)$.

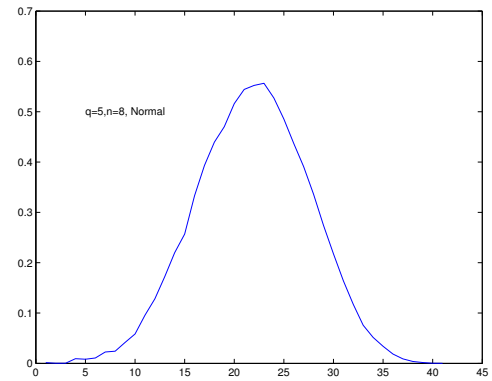

Figure 7.1: Normal convergence

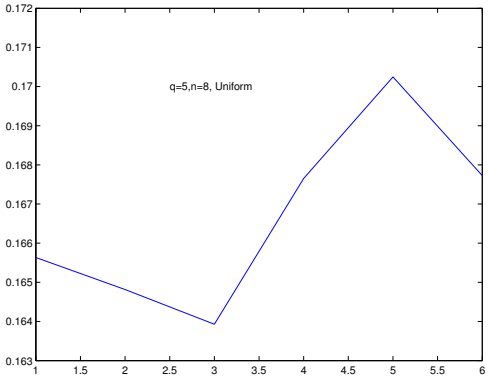

Figure 7.2: Uniform convergence

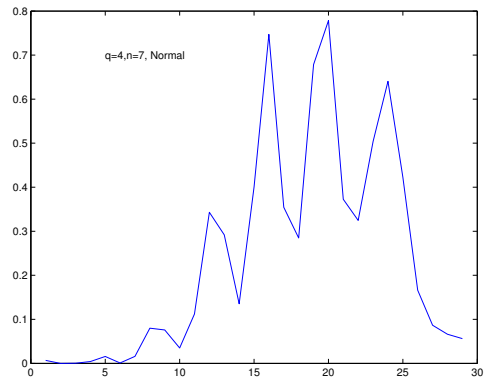

Figure 7.3: Normal convergence

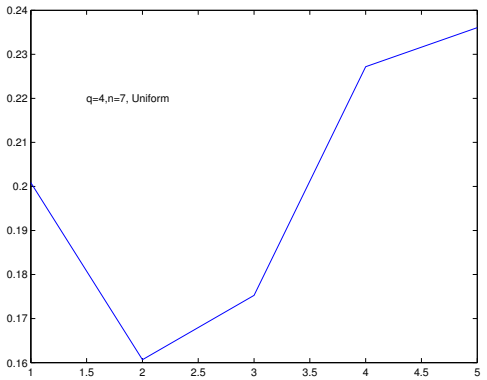

Figure 7.4: Uniform convergence

Here we study the distribution of $\overline{X_1 + ... + X_n}$ when n=10 with the variances 1/50, 1/200 : cf figure 7.5, 7.6. We understand that we are enough near of the uniform distribution if σ^2 is not too big.

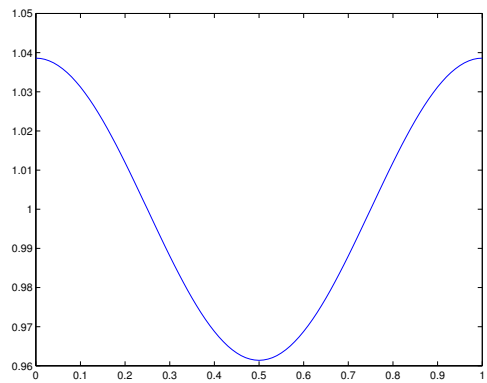

Figure 7.5: n=10, $\sigma^2 = 50$

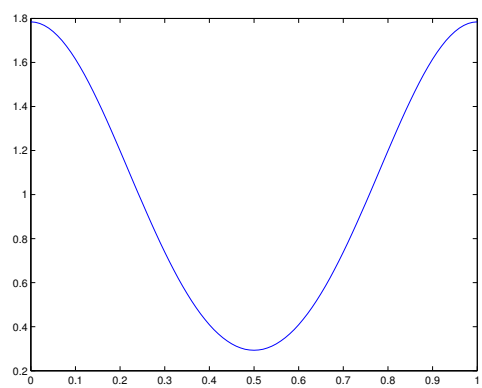

Figure 7.6: n=10, $\sigma^2 = 200$

For the data used in the construction of $b^1(n')$ in section 11.2, we can think that a sum of 10 terms is sufficient so that our hypotheses are satisfied.

7.3.4 Rate of convergence in the XORLT

In this section we understand that, in some cases, the convergence to the uniform distribution can be very fast.

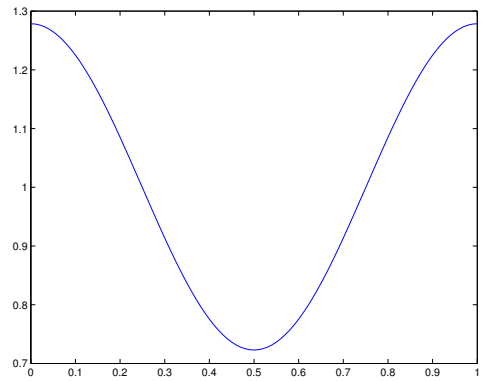

Figure 7.7: n=10, $\sigma^2 = 200$

Notations 7.3.1 *Let X_i, i=1,2,...,S, be a sequence of independent random variables with values in $\{0, 1, ..., N-1\}$. For all $s \in \{1, 2, ..., S\}$, we set $p"^{s}_{x_n^s} = P\{X_s = x_n^s\}$.*

Hypotheses

We assume that $p"^{i}_{x_n} = P"^{i}_{x_n}(\omega_7)$ where $p"^{i}_{x_n} = (1/N)[1 + r_N^i(v^i_{x_n} - v^i_N)]$ and where $v^i_{x_n} = V^i_{x_n}(\omega_7)$ is a realization of an IID sequence defined by the following way.

Hypothesis 7.3.1 *For all $i \in \{1, 2, ..., S\}$, we assume that $v^i_{x_n}$ is a realization of an IID sequence of random variables $V^i_{x_n}$ defined on a probability space $(\Omega_7, \mathcal{A}_7, Proba_7)$ such that $-1 \leq V^i_{x_n} \leq N-1$, $\mathbb{E}\{V^i_{x_n}\} = 0$, and all the $V^t_{x_{n_t}}$'s, t=1,..,S, $n_t = 1, .., N$, are independent.*

Then, we set $v^i_N = (1/N)\sum_{x_s} v^i_{x_s}$ et $V^i_N = (1/N)\sum_{x_s} V^i_{x_s}$.

Then, the following results holds

Lemma 7.3.2 *There exists a sequence of random variables $0 < R^i_N \leq 1$ such that $-1 \leq R^i_N(V^i_{x_n} - V^i_n) \leq N-1$ and $R^i_N \xrightarrow{P} 1$ as $N \to \infty$.*

Proof : One can write $-1 - e \leq V^i_{x_n} - V^i_N \leq N - 1 + e$ where $e > 0$. Then, one can write $-1 \leq R^i_N(V^i_{x_n} - V^i_N) \leq N-1$ where $0 < R^i_N \leq 1$. By the CLT, $V^i_N \xrightarrow{P} 0$. Therefore, $R^i_N \xrightarrow{P} 1$. ■

Then, we can define probabilities.

Proposition 7.3.2 *For all $x_n \in F^*(N)$, we set $P"^{i}_{x_n} = (1/N)[1 + R^i_N(V^i_{x_n} - V^i_N)]$. Then, $0 \leq P"^{i}_{x_n} \leq 1$ and $\sum_{x_n} P"^{i}_{x_n} = 1$*

Proof : We have $\sum_{x_n} P"^i_{x_n} = \sum_{x_n}(1/N) + (R^i_N/N)\sum_{x_n}(V^i_{x_n} - V^i_N)] = 1$. ■

Then, we assume that the following hypothesis holds.

Hypothesis 7.3.2 *For all $i \in \{1, 2, ..., S\}$, we assume that $p"^i_{x_n}$ is the realization of the sequence of random variables $P"^i_{x_n}$ defined over $(\Omega_7, \mathcal{A}_7, Proba_7)$ by $p"^i_{x_n} = P"^i_{x_n}(\omega_7)$.*

Then, we have the rate of convergence of XORLT.

Theorem 7 *Assume that, for all $s \in \{1, 2, ..., S\}$, the variance of V^s_1 is $\sigma^2_{V_s}$. Then, with a probability greater than $1 - \Gamma(b)$ approximately,*

$$P\{\overline{X_1 + ... + X_S} = y\} = \frac{1}{N}\left[1 + \frac{b.Ob(1)\sigma_{V_1}....\sigma_{V_S}}{\sqrt{N^{S-1}}}\right].$$

Remark 7.3.3 *If P^i_x has a distribution similar to that of $P^i_x = \frac{P'^i_x}{\sum_{t=1}^N P'^i_t}$ when P'^i_x has the uniform distribution, then $\sigma^2_{V_r} = O(1)$. For example, one can choose $\sigma^2_{V_r} \leq 1$.*

Proof of theorem 7

At first, the following proposition holds.

Lemma 7.3.4 *The following equality holds :*
$P\{\overline{X_1 + ... + X_S} = y\} = \frac{1}{N} + \frac{r^1_N.....r^S_N}{N^S}\sum_{\overline{x_1+...+x_S}=y}(v^1_{x_1} - v^1_N).....(v^S_{x_S} - v^S_N)$.

Proof At first, $P\{\overline{X_1 + X_2 + ... + X_S} = y\} = \sum_{\overline{x_1+...+x_S}=y} p"^1_{x_1}.....p"^S_{x_S} = (1/N^S)\sum_{\overline{x_1+...+x_S}=y}[1 + r^1_N(v^1_{x_1} - v^1_N)].....[1 + r^S_N(v^S_{x_1} - v^S_N)]$.

Now, $[1 + r^1_N(v^1_{x_1} - v^1_N)].....[1 + r^S_N(v^S_{x_S} - v^S_N)]$

$= 1 + \left[r^1_N(v^1_{x_1} - v^1_N) + ... + r^S_N(v^S_{x_n} - v^S_N)\right]$
$+.....$
$+\sum_{i_1<i_2<....<i_q} r^{i_1}_N(v^{i_1}_{x_{i_1}} - v^{i_1}_N)r^{i_2}_N(v^{i_2}_{x_{i_2}} - v^{i_2}_N).....r^{i_q}_N(v^{i_q}_{x_{i_q}} - v^{i_q}_N)$
$+.....$
$+r^1_N(v^1_{x_1} - v^1_N)r^2_N(v^2_{x_2} - v^2_N).....r^S_N(v^S_{x_S} - v^S_N)$.

We deduce the proposition by using the following lemma (7.3.5). ■

Lemma 7.3.5 *Suppose $q < S$. Then,*
$\sum_{\overline{x_1+....+x_S}=y}\left[\sum_{i_1<i_2<....<i_q} r^{i_1}_N(v^{i_1}_{x_{i_1}} - v^{i_1}_N).....r^{i_q}_N(v^{i_q}_{x_{i_q}} - v^{i_q}_N)\right] = 0$.

Proof We have
$\sum_{\overline{x_1+....+x_S}=y}\left[\sum_{i_1<i_2<....<i_q} r_N^{i_1}(v_{x_{i_1}}^{i_1}-v_N^{i_1}).....r_N^{i_q}(v_{x_{i_q}}^{i_q}-v_N^{i_q})\right]=0$
$=\sum_{i_1<i_2<....<i_q}\left[\sum_{\overline{x_1+....+x_S}=y} r_N^{i_1}(v_{x_{i_1}}^{i_1}-v_N^{i_1}).....r_N^{i_q}(v_{x_{i_q}}^{i_q}-v_N^{i_q})\right]$.

For example, if $i_q<S$,
$\sum_{\overline{x_1+....+x_S}=y}(v_{x_{i_1}}^{i_1}-v_N^{i_1}).....(v_{x_{i_q}}^{i_q}-v_N^{i_q})$
$=\sum_{x_{i_1}}\sum_{x_{i_2}}....\sum_{x_{i_{S-1}}}\sum_{x_S=\overline{y-x_1+....-x_{S-1}}}(v_{x_{i_1}}^{i_1}-v_N^{i_1}).....(v_{x_{i_q}}^{i_q}-v_N^{i_q})$
$=\sum_{x_{i_1}}\sum_{x_{i_2}}....\sum_{x_{i_{S-1}}}(v_{x_{i_1}}^{i_1}-v_N^{i_1}).....(v_{x_{i_q}}^{i_q}-v_N^{i_q})$
$=\sum\left[\left[\sum_{x_{i_1}}(v_{x_{i_1}}^{i_1}-v_N^{i_1})\right].....\left[\sum_{x_{i_q}}(v_{x_{i_q}}^{i_q}-v_N^{i_q})\right]\right]$
$=0$ because $\sum_{x_{i_1}} v_{x_{i_1}}^{i_1}=Nv_N^{i_1}$. ■

Proposition 7.3.3 *Under the hypothesis 7.3.1,* $\frac{\sum_{\overline{x_{i_1}+....+x_{i_S}}=y} V_{x_{i_1}}^1 V_{x_{i_2}}^2....V_{x_{i_S}}^S}{N^{(S-1)/2}}$ *has asymptotically a distribution* $N(0,\sigma_{V_1}^2.....\sigma_{V_S}^2)$.

Proof We apply theorem 3 with $X_{t_s}=V_{i_1}^1....V_{i_{S-1}}^{S-1}V_{\overline{y-i_1-.....-i_{S-1}}}^S$ and $n=N^{S-1}$.

The first relation of theorem 3 is obvious. For example, if S=3, this relation is equivalent to the convergence of $(1/N^4)\sum_{r\neq s}\left[\mathbb{E}\{(X_s)^2(X_r)^2\}-\mathbb{E}\{(X_s)^2\}\mathbb{E}\{(X_r)^2\}\right]$, which is equivalent to the convergence of
$\frac{\left|\sum_{(i,j)\neq(i',j')}\left[\mathbb{E}\{(V_i^1V_j^2V_{\overline{y-i-j}}^3)^2(V_{i'}^1V_{j'}^2V_{\overline{y-i'-j'}}^3)^2\}-\mathbb{E}\{(V_i^1V_j^2V_{\overline{y-i-j}}^3)^2\}\mathbb{E}\{(V_{i'}^1V_{j'}^2V_{\overline{y-i'-j'}}^3)^2\}\right]\right|}{N^4}$.

Now, in order that $\mathbb{E}\{(V_i^1V_j^2V_{\overline{y-i-j}}^3)^2(V_{i'}^1V_{j'}^2V_{\overline{y-i'-j'}}^3)^2\}\neq\mathbb{E}\{(V_i^1V_j^2V_{\overline{y-i-j}}^3)^2\}$ $\mathbb{E}\{(V_{i'}^1V_{j'}^2V_{\overline{y-i'-j'}}^3)^2\}$, it is necessary that $i=i'$ or $j=j'$. Therefore, at the maximum, there is $2N^3$ such $V_i^1V_j^2V_{\overline{y-i-j}}^3V_{i'}^1V_{j'}^2V_{\overline{y-i'-j'}}^3$. Then, there exists a constant C_3^2 such that $\frac{2C_3^2}{N}$ is greater than
$\frac{\left|\sum_{(i,j)\neq(i',j')}\left[\mathbb{E}\{(V_i^1V_j^2V_{\overline{y-i-j}}^3)^2(V_{i'}^1V_{j'}^2V_{\overline{y-i'-j'}}^3)^2\}-\mathbb{E}\{(V_i^1V_j^2V_{\overline{y-i-j}}^3)^2\}\mathbb{E}\{(V_{i'}^1V_{j'}^2V_{\overline{y-i'-j'}}^3)^2\}\right]\right|}{N^4}$.

Now we study the condition $p!\ \frac{\sum_{t_1<t_2<.....<t_p}\mathbb{E}\{X_{t_1}X_{t_2}.....X_{t_p}\}}{(N^{S-1})^{p/2}}\to(N_2)^p\mu_p$.

First, assume S=2 : in this case, $X_{t_1}=V_{x_n^1}^1V_{\overline{y-x_n^1}}^2$. Then, $\mathbb{E}\{X_{t_1}.....X_{t_p}\}=\mathbb{E}\{V_{x_{n_1}^1}^1V_{\overline{y-x_{n_1}^1}}^2.....V_{x_{n_p}^1}^1V_{\overline{y-x_{n_p}^1}}^2\}=\mathbb{E}\{V_{x_{n_1}^1}^1\}.....\mathbb{E}\{V_{x_{n_p}^1}^1\}\mathbb{E}\{V_{\overline{y-x_{n_1}^1}}^2.....V_{\overline{y-x_{n_p}^1}}^2\}=0$ because the x_n^1 are all dissimilar.

Assume S=3 : in this case, $X_{t_1} = V^1_{x^1_{n_1}} V^2_{x^2_{n_2}} V^3_{\overline{y-x^1_{n_1}-x^2_{n_2}}}$. Then, one can write

$$\mathbb{E}\{X_{t_1}X_{t_2}.....X_{t_p}\}$$
$$= \mathbb{E}\{V^1_{x_{n_1}}.....V^1_{x_{n_p}}\}\mathbb{E}\{V^2_{x'_{n_1}}.....V^2_{x'_{n_p}}\}\mathbb{E}\{V^3_{\overline{y-x_{n_1}-x'_{n_1}}}.....V^3_{\overline{y-x_{n_p}-x'_{n_p}}}\} \ .$$

If p=2, $\mathbb{E}\{X_{t_1}X_{t_2}\} = \mathbb{E}\{V^1_{x_{n_1}} V^1_{x_{n_2}}\}\mathbb{E}\{V^2_{x'_{n_1}} V^2_{x'_{n_2}}\}\mathbb{E}\{V^3_{\overline{y-x_{n_1}-x'_{n_1}}} V^3_{\overline{y-x_{n_p}-x'_{n_p}}}\}$. Because $t_1 < t_2$, $x_{n_1} \neq x_{n_2}$ or $x'_{n_1} \neq x'_{n_2}$. Then, $\mathbb{E}\{X_{t_1}X_{t_2}\} = 0$. If p=3, we have the same conclusion.

If p=4,

$$\mathbb{E}\{X_{t_1}X_{t_2}X_{t_3}X_{t_4}\}$$
$$= \mathbb{E}\{V^1_{x_{n_1}} V^1_{x_{n_2}} V^1_{x_{n_3}} V^1_{x_{n_4}}\}\mathbb{E}\{V^2_{x'_{n_1}} V^2_{x'_{n_2}} V^2_{x'_{n_3}} V^2_{x'_{n_4}}\}$$
$$\mathbb{E}\{V^3_{\overline{y-x_{n_1}-x'_{n_1}}} V^3_{\overline{y-x_{n_2}-x'_{n_2}}} V^3_{\overline{y-x_{n_3}-x'_{n_3}}} V^3_{\overline{y-x_{n_4}-x'_{n_4}}}\} \ .$$

In order that $\mathbb{E}\{X_{t_1}X_{t_2}X_{t_3}X_{t_4}\} \neq 0$, it is necessary that, for example, $x_{n_1} = x_{n_2}$, $x_{n_3} = x_{n_4}$, $x'_{n_1} = x'_{n_3}$ and $x'_{n_2} = x'_{n_4}$. *Then, we assume that these relations hold.*

Then, in order that

$$\mathbb{E}\{V^1_{x_{n_1}} V^1_{x_{n_2}} V^1_{x_{n_3}} V^1_{x_{n_4}}\}\mathbb{E}\{V^2_{x'_{n_1}} V^2_{x'_{n_2}} V^2_{x'_{n_3}} V^2_{x'_{n_4}}\}$$
$$\mathbb{E}\{V^3_{\overline{y-x_{n_1}-x'_{n_1}}} V^3_{\overline{y-x_{n_2}-x'_{n_2}}} V^3_{\overline{y-x_{n_3}-x'_{n_3}}} V^3_{\overline{y-x_{n_4}-x'_{n_4}}}\} \neq 0 \ ,$$

it is necessary that

OR $\overline{y-x_{n_1}-x'_{n_1}} = \overline{y-x_{n_2}-x'_{n_2}}$ and $\overline{y-x_{n_3}-x'_{n_3}} = \overline{y-x_{n_4}-x'_{n_4}}$. Therefore, $x'_{n_1} = x'_{n_2}$. Then, $X_{t_1} = V^1_{x_{n_1}} V^2_{x'_{n_1}} V^3_{\overline{y-x_{n_1}-x'_{n_1}}} = V^1_{x_{n_2}} V^2_{x'_{n_2}} V^3_{\overline{y-x_{n_2}-x'_{n_2}}} = X_{t_2}$: it is impossible.

OR $\overline{y-x_{n_1}-x'_{n_1}} = \overline{y-x_{n_3}-x'_{n_3}}$ and $\overline{y-x_{n_2}-x'_{n_2}} = \overline{y-x_{n_4}-x'_{n_4}}$. Then, $x_{n_1} = x_{n_3}$: it is impossible.

OR $\overline{y-x_{n_1}-x'_{n_1}} = \overline{y-x_{n_4}-x'_{n_4}}$ and $\overline{y-x_{n_2}-x'_{n_2}} = \overline{y-x_{n_3}-x'_{n_3}}$. Then, $x_{n_1}+x'_{n_1} \equiv x_{n_3}+x'_{n_2}$ and $x_{n_1}+x'_{n_2} \equiv x_{n_3}+x'_{n_1}$. Therefore, $2(x'_{n_1}-x'_{n_2}) \equiv 0$ and $2(x_{n_1}-x_{n_3}) \equiv 0$. If N is odd, it is impossible.

If N is even, $x'_{n_1} - x'_{n_2} = \delta_1(N/2)$ and $x_{n_1} - x_{n_3} = \delta_2(N/2)$ where $\delta_s = 0$, -1 or 1. Therefore, there are $\frac{C'_0 N^4}{N^2}$ possible variables $X_{t_1} = V^1_{x_{n_1}} V^2_{x'_{n_1}} V^3_{\overline{y-x_{n_1}-x'_{n_1}}}$, $X_{t_2} = V^1_{x_{n_2}} V^2_{x'_{n_2}} V^3_{\overline{y-x_{n_2}-x'_{n_2}}}$, $X_{t_3} = V^1_{x_{n_3}} V^2_{x'_{n_3}} V^3_{\overline{y-x_{n_3}-x'_{n_3}}}$, $X_{t_4} = V^1_{x_{n_4}} V^2_{x'_{n_4}} V^3_{\overline{y-x_{n_4}-x'_{n_4}}}$ such that $\mathbb{E}\{X_{t_1}X_{t_2}X_{t_3}X_{t_4}\} \neq 0$. Therefore,

$$\frac{\sum_{t_1<t_2<t_3<t_4} \mathbb{E}\{X_{t_1}X_{t_2}X_{t_3}X_{t_4}\}}{(N^2)^2} < \frac{C'_0 N^2}{N^4} \to 0 \ .$$

One prove the general case by the same way.

Then all conditions of theorem 3 hold. Then, $\frac{\sum_{i=1}^{n_0} X_n}{\sqrt{N}} \xrightarrow{D} N(0, \sigma^2_{V_1} .... \sigma^2_{V_S})$ because $\mathbb{E}\{X^2_{t_r}\} = \mathbb{E}\Big\{\Big(V^1_{x^1_{n_r}} V^2_{x^2_{n_r}} ... V^{S-1}_{x^{S-1}_{n_r}} V^S_{y-x^1_{n_r}-...-x^{S-1}_{n_r}}\Big)^2\Big\} = \prod_{r=1}^S \mathbb{E}\Big\{(V^r_1)^2\Big\}$. ■

Now, one can assume that $\sum_{x^s_n} v^s_{x^s_n} = 0$.

Lemma 7.3.6 *Assume that the assumptions of proposition 7.3.3 hold. Then,*

$$\frac{1}{\sqrt{N^{S-1}}} \sum_{x^1_n, x^2_n, ..., x^{S-1}_n} \Big[\prod_{t=1}^{S-1} R^t_N (V^t_{x^t_n} - V^t_N)\Big] R^S_N (V^S_{y-x^1_n-....-x^{S-1}_n} - V^S_N)$$

has asymptotically the distribution $N(0, \sigma^2_{V_1} .... \sigma^2_{V_S})$.

Proof Assume S=2. Then,

$$\frac{1}{\sqrt{N}} \sum_{x^1_n} R^1_N (V^1_{x^1_n} - V^1_N) R^2_N (V^2_{y-x^1_n} - V^2_N)$$

$$= \frac{R^1_N R^2_N}{\sqrt{N}} \sum_{x^1_n} V^1_{x^1_n} [V^2_{y-x^1_n} - V^2_N] - \frac{R^1_N R^2_N V^1_N}{\sqrt{N}} \sum_{x^2_n} [V^2_{x^2_n} - V^2_N]$$

$$= \frac{R^1_N R^2_N}{\sqrt{N}} \sum_{x^1_n} V^1_{x^1_n} [V^2_{y-x^1_n} - V^2_N]$$

$$= \frac{R^1_N R^2_N}{\sqrt{N}} \sum_{x^1_n} V^1_{x^1_n} V^2_{y-x^1_n} - \frac{R^1_N R^2_N}{\sqrt{N}} V^2_N \sum_{x^1_n} V^1_{x^1_n} \ ,$$

where $\frac{R^1_N R^2_N}{\sqrt{N}} \sum_{x^1_n} V^1_{x^1_n} V^2_{y-x^1_n}$ and $\frac{1}{\sqrt{N}} \sum_{x^1_n} V^1_{x^1_n}$ have asymptotically a normal distribution (cf proposition 7.3.3). Moreover, V^2_N converges in probability to 0.

Then, $\frac{1}{\sqrt{N}} \sum_{x^1_n} R^1_N (V^1_{x^1_n} - V^1_N) R^2_N (V^2_{y-x^1_n} - V^2_N)$ has asymptotically the distribution $N(0, \sigma^2_{V_1} \sigma^2_{V_2})$.

In the general case, we prove this proposition by the same way . ■

Proof 7.3.7 *Now we prove theorem 7*

By proposition 7.3.6,

$$R_N^1.....R_N^S \ \frac{\sum_{\overline{x_1+...+x_S}=y}(V_{x_1}^1 - V_N^1).....(V_{x_S}^S - V_N^S)}{\sqrt{N^{S-1}}}$$

has asymptotically the distribution $N(0, \sigma_{V_1}^2.....\sigma_{V_S}^2)$. We deduce (cf proposition 7.3.4) that, with a probability greater than $1 - \Gamma(b)$ approximately,

$$r_N^{i_1}.....r_N^{i_S} \ \frac{\sum_{\overline{x_1+...+x_S}=y}(v_{x_1}^1 - v_N^1).....(v_{x_S}^S - v_N^S)}{N^S} = \frac{b.Ob(1)\sigma_{V_1}.....\sigma_{V_S}}{\sqrt{N^{S+1}}} \ . \ \blacksquare$$

Problem in some cases

Theorem 7 is only a mathematical theorem with a measure on the set of the probabilities chosen a priori. This measure is not thus inevitably adapted to certain assumptions.

It is not difficult to understand that theorem 7 has absurd consequences in the case of continuous density : cf pages 118-121 of [18].

To avoid this problem, one can transform the random variables : for example one can multiply each X_t by a suitably chosen number α_t modulo 1: $X_t' = \overline{\alpha_t X_t}$. For example, one can transform some lines by various Fibonacci congruences or various Fibonacci functions T_q : cf [18] pages 118-121.

As a matter of fact, the multiplication by α_t modulo 1 defines a permutation if α_t is suitably selected. But in this case, one has again the problem of the choice of the permutations: the permutations too simple are not appropriate. Is this case here? This problem is not so simple. On the one hand, Knuth ([1]) explains why one cannot use permutations built by algorithm (cf also section 2.1.1). On the other hand, one understands in chapter 6 that the multiplication corresponding to a Fibonacci congruence is a good permutation.

Now a simpler solution is to use transformation which have the same characteristic as a permutation really random. It is what we do in section 12.

7.3.5 Limit theorems for conditional probabilities

Here, we study $G(j) = \sum_{i=1}^S F(i,j)$ where the rows F (i,.) are independent : cf section 11. In that case, the distribution of the sums admitting for probabilities the conditional probabilities is that one of a sum of independent variables.

Proposition 7.3.4 *Let $X_{i,j}$, i=1,...,I, j=1,...,p, be a sequence of random variables. We assume that the rows $X_{i,.} \in F(m)^p$ are independent. Then,*

$$P\Big\{\overline{X_{1,1} + ... + X_{I,1}} \in Bo \ \Big| \ \overline{X_{1,2} + ... + X_{I,2}} = y_2, ..., \overline{X_{1,p} + ... + X_{I,p}} = y_p\Big\}$$

$$= \sum_{x_{i,j}:\forall j, \overline{x_{1,j}+...+x_{I,j}}=y_j} \eta'_{\{x_{i,j}\}} P\Big\{\overline{X_{1,1} + ... + X_{I,1}} \in Bo \ \Big| \ X_{i,j} = x_{i,j}, i = 1, .., I, j = 2, .., p\Big\},$$

where $\sum_{x_{i,j}:\forall j, \overline{x_{1,j}+...+x_{I,j}}=y_j} \eta'_{\{x_{i,j}\}} = 1$.

The proof is section 5.7 of [18]. These results show that, in many cases, $P\big\{\overline{X_{1,1} + ... + X_{I,1}} \in Bo \,\big|\, \overline{X_{1,2} + ... + X_{I,2}} = y_2, ..., \overline{X_{1,p} + ... + X_{I,p}} = y_p\big\} \rightarrow L(Bo)$. In particular, results obtained in section 7.3.1 show that this limit is checked for all the data used to build the random sequences $b^1(n')$.

Chapter 8

Empirical Theorems

8.1 Empirical Theorems

8.1.1 First theorems

In this section, we use the following notations (cf also property 6.3.3).

Notations 8.1.1 *Let $X_n^0 \in F(m)$, $n \in \mathbb{N}^*$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{A}, P)$.*

Let j_s , s=1,2,...,p, $j_s \in \mathbb{Z}$, be an injective sequence such that $j_1 = 0$. Let $d_0 = \big|min(j_s|s = 1, 2, .., p)\big|$. Then, we set $X_n = X_{n+d_0}^0$.

Notations 8.1.2 *Let $Bo = Bo_1 \otimes Bo_2 \otimes .... \otimes Bo_p \subset F(m)^p$ be a Borel set where $L(Bo_s) \leq 1/2$ for s=1,...,p. We set $L_n = E\Big\{1_{Bo}(X_n)\Big\}$ and $L^N(Bo) = (1/N)\sum_{n=1}^N L_n$ where $1_{Bo}(X_n) = 1_{Bo_1}(X_{n+j_1})1_{Bo_2}(X_{n+j_2})....1_{Bo_p}(X_{n+j_p})$.*

Hypothesis 8.1.1 *One supposes that, for all $p \in \mathbb{N}^*$, for all Borel set $Bo \subset F(m)^p$, for all injective sequence j_s, for all $n \in \mathbb{N}^*$,*

$$\mathbb{E}\big\{1_{Bo}(X_n)\big\} = L(Bo) + Ob(1)L(Bo)\epsilon_{Bo}^p \ ,$$

where $\epsilon_{Bo}^p = 2\sqrt{\frac{[Log(n_0)+q]p^3 2^q}{m}} = \sqrt{\frac{e_0 p^3 2^q}{m}} << 1$.

Notations 8.1.3 *We set $\sigma_1^2 = (1/N)E\Big\{\big[\sum_{n=1}^N \big(1_{Bo}(X_n) - L_n\big)\big]^2\Big\}$. Moreover, if X_n is IID, we write σ_B^2 instead of σ_1^2 .*

For example, if p=1, $\sigma_B^2 = L(Bo)[1 - L(Bo)]$.

Now, we can expound the first empirical theorem.

Theorem 8 *Let $\beta_{1,p} = \frac{\sqrt{N}[L^N(Bo) - L(Bo)]}{\sigma_B}$ and $\gamma'_{1,p} = \frac{N\epsilon_{Bo}^p}{2A(p)}\Big[2^{3/2} + 2L(Bo)\ \Big]$ where $A(p) = 1 - (p^2 - p + 1)2^{-p}$.*

Let $P_e = \frac{1}{N}\sum_{n=1}^{N} 1_{Bo}(X_n)$. *Then, the following inequality holds*

$$P\Big\{\sqrt{N}\big|P_e - L(Bo)\big| \geq \sigma_B x\Big\} \leq K_1\Big(\frac{1-\beta_{1,p}/x}{1+\gamma'_{1,p}}x\Big) \ ,$$

where $K_1(x) = P\Big\{\frac{\sqrt{N}|P_e - L^N(Bo)|}{\sigma_1} \geq x\Big\}$.

Remark that if P_e is asymptotically normal, $\frac{\sqrt{N}[P_e - L^N(Bo)]}{\sigma_1}$ has asymptotically the distribution N(0,1).

Now, one can also obtain similar results to theorem 8 if one replaces hypothesis 8.1.1 by the following way.

Hypothesis 8.1.2 *Let* $\epsilon > 0$. *One supposes that, for all Borel set* $Bo \subset F(m)$, *for all* $p \in \mathbb{N}^*$, *for all sequence* j_s, *for all* $x_2, ..., x_p$, *for all* $n \in \mathbb{N}^*$, *such that* $n > d_0$,

$$P\big\{X_n^0 \in Bo\big|X_{n+j_2}^0 = x_2, ..., X_{n+j_p} = x_p\big\} = L(Bo) + Ob(1)\epsilon \ .$$

Then, one obtains results similar to theorem 8. These results can be specified when X_n is asymptotically independent. In this case, one uses increases of $\sum_d Max_{n\in\mathbb{N}^*}\Big|E\big\{\big(1_{Bo}(X_n) - L_n\big)\big(1_{Bo}(X_{n+d}) - L_{n+d}\big)\big\}\Big|$: cf chapter 8 of [18].

For example, if X_n is q-dependent the following theorem holds.

Theorem 9 *We suppose that* X_n *is q-dependent. We set* $\epsilon_p = (1/2+\epsilon)^p - (1/2)^p$ *and*

$$\gamma_{1,p} = \frac{1}{2A(p)L(Bo)}\Big[(p^2-p+1)\Big(\epsilon_p + 2q\epsilon_{2p} + (1+2q)\big[2^{1-p}\epsilon_p + \epsilon_p^2\big]\Big)\Big] \ .$$

$$Then,\ P\Big\{\sqrt{N}\big|P_e - L(Bo)\big| \geq \sigma_B x\Big\} \leq K_1\Big(\frac{1-\beta_{1,p}/x}{1+\gamma_{1,p}}x\Big) \ ,$$

On the other hand, results similar can be obtained for empirical conditional probabilities : cf theorems 8 and 10 of [18].

Theorem 10 *We suppose that* X_n *is q-dependent. We assume that the hypothesis 8.1.2 holds and that the assumptions of theorem 8 of [18] holds.*

We set $p_e = (1/N)\sum_{n=1}^{N} 1_{Bo_2}(X_{n+j_2})....1_{Bo_p}(X_{n+j_p})$. *Then, if N is great enough, there exists* $K_2 \approx \Gamma$ *such that*

$$P\Big\{\sqrt{N}\Big|\frac{P_e}{p_e} - L(Bo_1)\Big| > \sigma_{cp}\ x\Big\} \leq K_2\Big(\frac{1-\beta_{2,p}/x}{1+\gamma_{2,p}}x\Big)$$

where $\beta_{2,p} \approx 0$ *and* $\gamma_{2,p} \approx 0$ *if* ϵ *is small enough and where* $\sigma_{cp}^2 = N.\mathbb{E}\Big\{\Big[\frac{P_e - L(Bo_1)p_e}{L(Bo_2)...L(Bo_p)}\Big]\Big\}$ *when* X_n *is IID.*

8.1.2 Applications

Now we study how one can apply the previous theorems. We are interested by the sequence of random bits $b^1(n')$ built in section 11.2 : we assume that $B^1(n')$, n'=1,....,N', is a sequence of random bits satisfying

$$P\big\{B^1(n') = b \;\big|\; B^1(n'+j_2) = b_2, ..., B^1(n'+j_p) = b_p\big\} = 1/2 + \frac{Ob(1)\alpha}{\sqrt{N'}} \ .$$

Now, the sequence C(j) defined in section 11.2 is Qd-dependent with Qd=22. Then, $B^1(n')$ is also Q'-dependent. Moreover,
1) $\epsilon_p = (1/2+\epsilon)^p - (1/2)^p \approx \frac{p\epsilon}{2^{p-1}} = \frac{2p\epsilon}{2^p}$,
2) $\beta_{1,p} \leq \frac{\sqrt{N'}\epsilon_p}{\sqrt{A(p)L(Bo)}} \approx \frac{2p\alpha}{A(p)^{1/2}2^{p/2}}$,
3) $\gamma_{1,p} \approx \frac{(p^2-p+1)\alpha}{2A(p)\sqrt{N'}}\Big[2p + \frac{(1+4Q')4p}{2^p}\Big]$.

We want that $B^1(n')$ behaves like an IID sequence. Thus, we use theorem 8 for Q'-dependent sequences and we increase $P\Big\{\frac{\sqrt{N}|P_e - L(Bo)|}{\sigma_B} \geq x\Big\}$. Let H1 be the hypothesis "$B^1(n')$ defined in section 11.2 satisfies theorem 8". Under the assumptions IID and H1, one has the following tables of increases of $P\Big\{\sqrt{N}\big|P_e - L(Bo)\big| \geq \sigma_B x\Big\}$ regarded as function of (x,p).

(x,p)	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,10)	(1,15)	(1,20)
Under IID	0.317	0.317	0.317	0.317	0.317	0.317	0.317	0.317
Under H1	0.334	0.359	0.356	0.357	0.346	0.340	0.361	0.380

(x,p)	(2,1)	(2,2)	.(2,3)	(2,4)	(2,5)	(2,10)	(2,15)	(2,20)
Under IID	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030
Under H1	0.049	0.052	0.051	0.052	0.053	0.050	0.061	0.073

Similar results are obtained for theorem 10 : cf section 9.5.2 of [18] .

8.1.3 Proof of theorem 8

At first, one can prove the following lemma.

Lemma 8.1.1 *For all n, we set* $H(n) = \big\{m \in \mathbb{N}^* \big| \exists (e,e') : X_{n+j(e)} = X_{m+j(e')}\big\}$ *and* $H^*(n) = \big\{m \in \mathbb{N}^* \big| \exists (e,e') : X_{n+j(e)} = X_{m+j(e')},\ m \neq n\big\}$. *Then,* $card\big(H(n)\big) \leq p^2 - p + 1$.

Lemma 8.1.2 *The following inequalities hold :* $\sigma_B^2 \geq A(p)L(Bo) \geq L(Bo)/8$.

Proof One can write

$$\sigma_B^2 = (1/N)\sum_{n=1}^{N}\sum_{m\in H(n)}\Big(E\big\{1_{Bo}(X'_n)1_{Bo}(X'_m)\big\} - L(Bo)^2\Big)$$

$$= (1/N)\sum_{n=1}^{N}\Big(E\big\{1_{Bo}(X'_n)\big\} + \sum_{m\in H^*(n)} E\big\{1_{Bo}(X'_n)1_{Bo}(X'_m)\big\} - \sum_{m\in H(n)} L(Bo)^2\Big)$$

$$\geq (1/N)\sum_{n=1}^{N}\Big(E\big\{1_{Bo}(X'_n)\big\} - (p^2-p+1)L(Bo)^2\Big) = L(Bo)\Big(1-(p^2-p+1)L(Bo)\Big)\ .$$

Now, $(p^2-p+1)L(Bo) \leq (p^2-p+1)2^{-p}$. Moreover,

$$\frac{d(p^2-p+1)2^{-p}}{dp} = (2p-1)2^{-p} - Log(2)(p^2-p+1)2^{-p}$$

which has the roots $p_1 \approx 0.7888$ and $p_2 \approx 3.5423$

Therefore, $(p^2-p+1)2^{-p}$ decreases and converges to 0 if $p \geq 4$. Moreover, $(p^2-p+1)2^{-p} = 3/4$ if p=2, 7/8 if p=3, 13/16 if p=4, 21/32 if p=5. Then, $(p^2-p+1)L(Bo) \leq 7/8$. ■

Lemma 8.1.3 *If $m \notin H(n)$, $\mathbb{E}\big\{1_{Bo}(X_n)1_{Bo}(X_m)\big\} = L(Bo)^2 + Ob(1)\epsilon^3$, where $\epsilon^3 = L(Bo)^2\sqrt{\frac{e_0(2p)^3 2^q}{m}} = L(Bo)^2 2^{3/2}\epsilon_{Bo}^p$.*

If $m \in H(n)$, $\mathbb{E}\big\{1_{Bo}(X_n)1_{Bo}(X_m)\big\} = \mathbb{E}\big\{1_{Bo}(X'_n)1_{Bo}(X'_m)\big\} + Ob(1)\epsilon^4$, where X'_n is an IID sequence and where $\epsilon^4 = L(Bo)\epsilon_{Bo}^{2p} = L(Bo)2^{3/2}\epsilon_{Bo}^p$.

Proof If $m \notin H(n)$, by notation 8.1.1,

$$\mathbb{E}\big\{1_{Bo}(X_n)1_{Bo}(X_m)\big\} = L(Bo\otimes Bo) + L(Bo\otimes Bo)\sqrt{\frac{e_0(2p)^3 2^q}{m}}\ .$$

Assume $m \in H(n)$. Clearly if n=m or if p=1, equation holds by notation 8.1.1.

Suppose $p \geq 2$ and $n \neq m$. One can assume $n < m$. Then, there exists a sequence i_s, s=1,...,p', $p' < 2p$, and a sequence of Borel sets Bo'_s, s=1,...,p', such that $L(Bo'_s) \leq 1/2$ and

$$1_{Bo}(X_n)1_{Bo}(X_m)$$

$$= 1_{Bo_1}(X_n)1_{Bo_2}(X_{n+j_2}).....1_{Bo_p}(X_{n+j_p})1_{Bo_1}(X_m)1_{Bo_2}(X_{m+j_2}).....1_{Bo_p}(X_{n+j_p})$$
$$= 1_{Bo'_1}(X_n)1_{Bo'_2}(X_{n+i_2}).....1_{Bo'_{p'}}(X_{n+i_{p'}})\ . \qquad (8.1)$$

Clearly $p \leq p' < 2p$. Then,

$$\mathbb{E}\{1_{Bo}(X_n)1_{Bo}(X_m)\} = \mathbb{E}\{1_{Bo}(X'_n)1_{Bo}(X'_m)\} + L(Bo)\theta\sqrt{\frac{e_0(p')^3 2^q}{m}} \ ,$$

where $0 \leq \theta \leq 1$. ■

Lemma 8.1.4 *The following equality holds :* $L_n L_m = L(Bo)^2 + Ob(1)\epsilon^5$, *where* $\epsilon^5 \approx 2Ob(1)L(Bo)^2\epsilon^p_{Bo}$.

Proof We have $L_n = L(Bo) + Ob(1)L(Bo)\epsilon^p_{Bo}$. Then,

$$L_n L_m = \Big[L(Bo) + Ob(1)L(Bo)\epsilon^p_{Bo}\Big]\Big[L(Bo) + Ob(1)L(Bo)\epsilon^p_{Bo}\Big]$$

$$= L(Bo)^2 + 2L(Bo)^2 Ob(1)\epsilon^p_{Bo} + Ob(1)L(Bo)^2(\epsilon^p_{Bo})^2$$

$$\approx L(Bo)^2\Big[1 + 2Ob(1)\sqrt{\frac{e_0 p^3 2^q}{m}}\ \Big] \ . \ ■$$

Lemma 8.1.5 *The following equality holds*

$$\sigma_1^2 = \sigma_B^2\big[1 + Ob(1)2\gamma'_{1,p}\big].$$

Proof Let X'_n be an IID sequence with uniform distribution. Then,

$$\sigma_1^2 = (1/N)E\Big\{\big[\sum_{n=1}^{N}\big(1_{Bo}(X_n) - L_n\big)\big]^2\Big\}$$

$$= (1/N)E\Big\{\sum_{n=1}^{N}\sum_{m=1}^{N}\big(1_{Bo}(X_n) - L_n\big)\big(1_{Bo}(X_m) - L_m\big)\Big\}$$

$$= (1/N)E\Big\{\sum_{n=1}^{N}\sum_{m\in H(n)}\big(1_{Bo}(X_n)1_{Bo}(X_m) - L_n L_m\big)\Big\}$$

$$+(1/N)E\Big\{\sum_{n=1}^{N}\sum_{m\notin H(n)}\big(1_{Bo}(X_n)1_{Bo}(X_m) - L_n L_m\big)\Big\}$$

$$\approx (1/N)E\Big\{\sum_{n=1}^{N}\sum_{m\in H(n)}\big(1_{Bo}(X'_n)1_{Bo}(X'_m) + Ob(1)\epsilon^4 - L(Bo)^2 + Ob(1)\epsilon^5\big)\Big\}$$

$$+(1/N)E\Big\{\sum_{n=1}^{N}\sum_{m\notin H(n)}\big(L(Bo)^2+Ob(1)\epsilon^3-L(Bo)^2+Ob(1)\epsilon^5\big)\Big\}$$

$$=\sigma_B^2+(1/N)E\Big\{\sum_{n=1}^{N}\sum_{m\in H(n)}Ob(1)\epsilon^4\Big\}$$

$$+(1/N)E\Big\{\sum_{n=1}^{N}\sum_{m\notin H(n)}Ob(1)\epsilon^3\Big\}+(1/N)E\Big\{\sum_{n=1}^{N}\sum_{m}Ob(1)\epsilon^5\Big\}$$

$$=\sigma_B^2+(1/N)\sum_{n=1}^{N}\sum_{m\in H(n)}Ob(1)2^{3/2}L(Bo)\epsilon_{Bo}^p$$

$$+(1/N)\sum_{n=1}^{N}\sum_{m\notin H(n)}Ob(1)2^{3/2}L(Bo)^2\epsilon_{Bo}^p$$

$$+(1/N)\sum_{n=1}^{N}\sum_{m}Ob(1)2L(Bo)^2\epsilon_{Bo}^p$$

$$=\sigma_B^2+(1/N)\sum_{n=1}^{N}\sum_{m\in H(n)}Ob(1)2^{3/2}L(Bo)\epsilon_{Bo}^p$$

$$+(1/N)\sum_{n=1}^{N}\sum_{m\notin H(n)}Ob(1)2^{3/2}L(Bo)\epsilon_{Bo}^p$$

$$+(1/N)\sum_{n=1}^{N}\sum_{m}Ob(1)2L(Bo)^2\epsilon_{Bo}^p$$

$$=\sigma_B^2+NOb(1)2^{3/2}L(Bo)\epsilon_{Bo}^p+2NOb(1)L(Bo)^2\epsilon_{Bo}^p$$

$$=\sigma_B^2+2A(p)L(Bo)\ \frac{Ob(1)}{2A(p)L(Bo)}\Big[2^{3/2}NL(Bo)\epsilon_{Bo}^p+2NL(Bo)^2\epsilon_{Bo}^p\ \Big]$$

$$= \sigma_B^2 + 2A(p)L(Bo)\ \frac{Ob(1)N\epsilon_{Bo}^p}{2A(p)}\Big[2^{3/2} + 2L(Bo)\ \Big]$$

$$= \sigma_B^2\big[1 + Ob(1)2\gamma'_{1,p}A(p)L(Bo)/\sigma_B^2\big] = \sigma_B^2\big[1 + Ob(1)2\gamma'_{1,p}\big]$$

(by lemma 8.1.2). ■

Now, one proves the following lemma by basic method : cf lemma 9.2.9 of [18]

Lemma 8.1.6 *The following inequality holds :*

$$\sigma_1 \leq (1 + \gamma_{1,p})\sigma_B.$$

Proof 8.1.7 *We prove now the theorem 8*

The following inequalities hold.

$$P\left\{\left|\sqrt{N}\big[P_e - L(Bo)\right| > \sigma_B x\right\}$$

$$\leq P\left\{\left|\sqrt{N}\big[P_e - L^N(Bo)\big]\right| > \sigma_B x - \sqrt{N}|L^N - L(Bo)|\right\}$$

$$\leq P\left\{\left|\sqrt{N}\big[P_e - L^N(Bo)\big]\right| > \sigma_B x[1 - \beta_{1,p}/x]\right\}$$

$$\leq P\left\{\left|\sqrt{N}\big[P_e - L^N(Bo)\big]\right| > \frac{1 - \beta_{1,p}/x}{1 + \gamma"_{1,p}}\sigma_1 x\right\}$$

$$= K_1\Big(\frac{1 - \beta_{1,p}/x}{1 + \gamma"_{1,p}}x\Big)\ .\ \blacksquare$$

Chapter 9

Study of some files

9.1 Introduction

In this chapter, we study the data resulting from certain electronic files, especially from texts. By a study of these data based on logic, we will understand that one will be able to conclude that they behave like asymptotically independent sequences (and even Qd-dependent sequences).

In this section, we use a sequence y_n which one can regard as a realization of a sequence of random variables : $y_n = Y_n(\omega)$ for all n=1,...,N .

We do not impose that the Y_n are independent or identically distributed. But it can be useful that the CLT is satisfied.

9.2 Existence of satisfactory datas

9.2.1 Definition

At first, we had to know when a sequence y_n can be regarded as a realization of a sequence of really random variables : $y_n = Y_n(\omega)$ for all i=1,...,N .

First, any sequences of reals numbers can be regarded as a realization of a sequence of random variable of a certain type (completely deterministic, IID, etc) : this sequence of random variable is the model. But this model is correct with a some probability.

Then, to suppose "$y_n = Y_n(\omega)$" is a traditional scientific assumption if the y_n represents a physical phenomenon. One wants thus to show in an unquestionable way that it is also the case when y_n is resulting from certain electronic files

As a matter of fact a such sequence is simply a not-determinist sequence : that is to say, a sequence such that it is impossible to predicte fully y_{n+p}, when, one knows $y_1, y_2, ..., y_n$.

Now, if we want that the CLT holds, we impose that there is an asymptotic independence. Of course, a such sequence is non-determinist.

9.2.2 Objections

But is what such sequences y_n exist? It is a physical question. It is also a philosophical question. As a matter of fact, some people claimed that there does not exist finite random sequences : e.g. cf [1] page 167.

It is due partly so that any sample of a sequence of random variables can be regarded as fully determinist. Indeed the following proposition is obvious.

Proposition 9.2.1 *Let x_n, n=1,...,N, a sequence of real numbers. Then, there exists a function $g : \{1, 2, ..., N\} \to \mathbb{R}$ such that for all $n \in \mathbb{N}$, $x_n = g(n)$.*

Moreover, there exists p and a function $g : \mathbb{R}^p \to \mathbb{R}$ such that for all $n \in \{1, 2, ..., N - p\}$, $x_{n+p} = f(x_n, x_{n+1}, ..., x_{n-p+1})$.

Moreover, some philosophies claim that all is fixed. For example, meteorology would be fully determined by all data of earth (all temperatures in all point of earth, all the atmospheric pressures, etc).

In the same way, actions of the men would be fully determined by the context in which they live and by the cells of their brains. Then, a book is fully determined before his writing by theses events.

Of course, that involves problems : for example, the quantum theory is rejected. In order to reject this theory, one can call upon various reasons: 1) it is valid only for the infinitely small. 2) It is only a theory 3) It involves inadmissible contradictions for some people (Schrodinger cat).

But, all theses objections are false. In order to prove that, we use a counterexample : one can exhib a finite unpredictable sequence.

9.2.3 A finite random sequence

Let P(x) = $(x - x_1)(x - x_2)....(x - x_{2N})$ where $0 \leq x_1 < x_2 < .... < x_{2N} < 1$, $x_{j+1} - x_j \geq 1/4N$ for j=1,2,...,2N-1. Let $z_1, z_2, ..., z_N$ be a pseudo-random sequence with values in [0,1] obtainded by a good pseudo-random generator. Let $y_i = P(z_i)$ for i=1,2,....,N.

Then, it is no possible to predict y_{n+p} , $n \leq n + p \leq N$ if one knows only $y_1, y_2, ..., y_n$.

Indeed, even if one knew $z_1, z_2, ..., z_{n+p}$, it would not possible because any polynomial Q such that $deg(Q) = 2N$ and $y_s = Q(z_s)$ for s=1,2,...,n is a correct prediction of P. Then, all $y^*_{n+p} = Q(z_{n+p})$ is a correct prediction of y_{n+p}.

Now there exists an infinite number of possible polynomials Q.

Then, it is no possible to predict y_{n+p} even if one knew the sequence z_n and if one had an infinite computing power (cf example 9.2.1 of [18]).

Now there is no reasons that the $Y_n's$ have the same distribution, ($y_n = Y_n(\omega)$). But is is not important because the philosophical objections are that the sequence is not independent.

Anyway, one can build a sequence y'_n where the $Y_n's$ have the uniform distribution : one uses $y'_n = F^{-1}(y_n)$, where F is the distribution function of P(X) when X has the uniform distribution: $F^{-1}(P(X))$ has also the uniform distribution.

There is another reason that it no possible to predict y_{n+p} . In order to estimate P, it would be necessary to compute all the polynomial correlation coefficient of order smaller than 2N (cf [10]) .

It would thus be necessary to calculate the empirical orthogonal polynomials P_j^N of order J smaller than 2N associated with $z_1, z_2, ..., z_N$. However $P_j^N \equiv 0$ if $j > N$: the empirical polynomials of a order larger than the sample size are impossible to estimate.

Moreover, it is not surprising that y_n is unpredictable : indeed P depends on more parameters than N. As matter of fact, many simple functions using more than N parameter z_n can be appropriate to obtain unpredictable sequence. For example if $y_i = Q(k_1, k_2, ..., k_N, k'_1, k'_2, ..., k'_n, z_i)$.

Indeed, in order to estimate the $k'_i s$ and the $k'_i{}'s$, one has to resolve the N equations : $y_i = Q(k_1, k_2, ..., k_N, k'_1, k'_2, ..., k'_n, z_i)$ for i=1,2,...,N, that is there are more parameters than equations.

Then, all sequence $y_1, y_2, ..., y_N$ which depend more parameters than N may be an unpredictable sequence.

9.2.4 Consequence 1

Then, the sequence y_n is random : a sequence whose it is impossible to predict the future, it is inevitably random. It is even an independent sequence.

Then, the philosophy which affirms that there does not exists finite random sequences x_n , n=1,...,n, does not corresponds to reality : a sequence which one cannot predict is obligatorily random. To say the opposite is illogical.

9.2.5 Consequence 2

In order to obtain sequences which satisfy concretely some asymptotical independence assumptions, we shall use data which depend a priori on a number of parameters much many larger than the size of sample.

9.3 Practical example

The sequence $b^1(n')$ which we have built in section 11.2 has been obtained by using texts. Concretely, one has used, in various languages dictionary, Encyclopaedia, Bible, etc. The dictionaries and the encyclopaedias are very good examples: the definitions which are consecutive in a dictionary generally represent independent facts : for example "decibel" is followed by "decide" in some dictionaries. The numbers which correspond to them are thus extracted from independent random sequences.

9.3.1 Use of text

Now, we show that **one can prove by logical reasonings that texts are asymptotically independent**. It is an davantage with respect to sequences furnished by machines for example. Indeed, this asymptotical independence is proved.

In the majority of the sequences obtained from texts, it is reasonable to admit asymptotic independence.

1) The writing of a book depend of a very large number of parameters. Normally, the number of parameters whose the content of the book depend will be always larger than the sample size of the example. One thus finds the argumentation introduced in section 9.2.3.

2) When they write a book many authors do not know what they will write exactly one page later. Concretely they would not predicte exactly what words he will use 100 words later. It will be even more difficult for letters. Then the dependence is weaker between more distant lines. That is, there is asymptotical independence.

3) Of course, it is more difficult to predict the letters used for the people who are not the author of the book.

4) Let us take the example of a novel. In fact if the beginning of a novel is known, there is a very great number of possible alternatives for the continuation of the history. Even for each alternative, there is a very great number of possible texts.

5) Not only, it quasi-impossible to predict about a text. But it is even more difficult to envisage the letters used.

6) To predict logically what is written in a book, it should initially be known that it is written in a certain language. It is not sure that one can arrive at this conclusion. Thus, one is unable to even currently decipher some languages. Could one have deciphered the Egyptian hieroglyphs if the Rosetta Stone had not been written in several languages?

In addition, it has to be known that this text is written with an alphabet of 26 letters for example. If the same book is written in Chinese, one has an alphabet much more important. If this book were written in a rational written form, but not yet invented by men, it would be still other matter. Then, it is not at all certain that, even with means of infinite calculations, it is possible to know that the sequences of numbers obtained has a meaning as a text of English

language.

Then, in most of texts it is very clear that it is many more difficult to predict what words will used 200 words later than 100 words later. That is, there is asymptotical independence (for dictionary or encyclopaedia, there is Qd-dependence).

All these facts mean that logic implies that the files obtained starting from texts are asymptotically independent. One thus obtains a result concerning the first step of our method of construction of the random bits $b^1(n')$. That is logically surer than if one uses random sequences supposed being provided by machines always subjected to possible dysfunctions: if certain electronic files are used, there are certain assumptions which can be admitted because of logical reasoning.

Remark 9.3.1 *Of course, we have tested theses conclusions. All tested texts conclude to asymptotical independence (and even Qd-dependence).*

9.3.2 Other data

One can use other datas in order to obtain the sequences of random numbers : softwares, mathematical texts, musics , etc. Then, it is necessary to study by logical reasonning each type of files in order to the obtained sequences are fully proven random.

Moreover, an important thing is that in conclusion, the XORLT holds. However probably that arrives in much case since it does not require asymptotic independence.

Moreover, the number obtained in chapter 11 satisfies all these tests of randomness.

One can use several files, for example, a dictionary and a software. Those are often completely independent from each other. The sequences of numbers which they provide are thus also independent.

9.3.3 Conclusion

1) For this type of files, one can assume that y_n is a realization of a sequence of random variable Y_n : $y_n = Y_n(\omega)$ where $\omega \in \Omega$ and $Y_n \in \{0, \frac{1}{\kappa}, \frac{2}{\kappa}, ..., \frac{\kappa-1}{\kappa}\}$ where $\kappa = 32$. Moreover, there is asymptotical independence, and the CLT holds : often there is Qd-dependence.
2) By using certain files as sources of noises, there are assumptions much surer than if machines are used.

Chapter 10

Building of IID sequences : 1

10.1 General method

10.1.1 Choice of data

Notations of data

It is supposed that one has a sequence of data a(j) translated in number: a(j), $j = 1, 2, ..., N_3$, $a(j) \in \{0, 1, ..., Ka-1\}$. One supposes that Ka is small enough. If it is not the case, one can break up the a(j)'s in order to have Ka small enough.

It is supposed that a(j) can be regarded as a sample of a sequence of random variables A(j) defined over a probability space (Ω, Δ, P) : $a(j) = A(j)(\omega)$ where $\omega \in \Omega$.

10.1.2 Description of the method

Shortening of the a(j)'s

Let $\kappa \in \mathbb{N}^*$. We set $c(j) = \overline{a(j)} \bmod \kappa$.
Comment One chooses κ in order to obtain a sequence c(j) such as, for all $t \in \{0, 1, ., \kappa - 1\}$, $P'_e\{C(j) = t\} > 0$ where P'_e is the empirical probability associated with $c(1), c(2), ..., c(N_3)$.

Choice of the parameters

At first, we need the following notation.

Notations 10.1.1 *For all $x \geq 2$, we set $m^F(x) = fi_{n_0-1}$ where $fi_{n_0-1} \leq x < fi_{n_0}$ (fi_n : cf definition 1.2.3).*

Then, one chooses now q_1 and $r_1 \in \mathbb{N}^*$ such that
1) $\sqrt{\frac{(n_0)^2(p_m)^3 2^{q_1}}{m}} = \epsilon^1 << 1$
2) $\frac{2\sqrt{[Log(n_0)+q_1](n_0)^3 2^{q_1}}}{\sqrt{m}} = \epsilon^2 << 1$
where $m = m^F(\kappa^{r_1})$ and $n_0 = \lfloor N_3/r_1 \rfloor$.

Building of the sequence

a) We set $d(j) = \sum_{r=1}^{r_1} c(r_1(j-1)+r)\kappa^{r-1}$ for $j = 1, 2, ..., n_0$.
b) We set $e^1(j) = \lfloor d(j)[m/\kappa^{r_1}] \rfloor$ for $j = 1, 2, ..., n_0$.
c) We set $e^2(j) = \overline{e^1(j) + rand_0(j)}$ mod m for $j = 1, 2, ..., n_0$ where $rand_0(j) \in F^*(m)$ is a pseudo-random generator with period m or $k_4.m$, $k_4 \in \mathbb{N}^*$.
d) For $j = 1, 2, ..., n_0$, we set $e^3(j) = T_{q_1}(e^2(j)/m)$.
e) Let $2^{q_1} e^3(j) = \overline{\overline{b_1^j, b_2^j .....b_{q_1}^j}}$, $b_s^j \in \{0, 1\}$, the binary writting of $2^{q_1} e^3(j)$.
f) We set $b'_{q_1 j - r + 1} = b_r^j$ for $j = 1, ..., n_0$, and $r = 1, ..., q_1$: $b'_1 = b_{q_1}^1$, $b'_2 = b_{q_1-1}^1$,....., $b'_{q_1} = b_1^1$, $b'_{q_1+1} = b_{q_1}^2$, $b'_{q_1+2} = b_{q_1-1}^2$,.....
g) The sequence $\{b'_n\}$ is noted $b^3(n')$.

Remark 10.1.1 *Step c) is not albsolutely necessary.*

Study of data

It is supposed that the sequence d(j) is not fully deterministic. That can be checked, for example by logical reasonings as for texts : cf section 9.

One checks that $Min_{j,j' \in \{1,...,n_0\}}(|d(j) - d(j')|)$ is not too small. If not, one can choose r_1 more large.

10.1.3 Properties

Use of proposition 6.3.5

We study a sequence of random variables $E^3(j)$ associated to $e^3(j)$.

We use proposition 6.3.5 and properties 6.3.3 and 6.3.2, in the probability space $(\Omega, \mathcal{A}, Proba)$ with the the uniform distribution M, associated to $E^3(j)$ and defined in hypothesis 6.3.4. Then, for all n, for all p, for all sequence j_t, for all Borel set Bo, with a probability larger than $1 - \frac{2\sqrt{2}e^{-Log(n_0)n_0/2}e^{-n_0 q_1/2}}{\sqrt{3\pi n_0[Log(n_0)+q_1]}}$,

$$\left|P\left\{\left(E^3(j+j_1), ..., E^3(j+j_p)\right) \in Bo\right\} - L(Bo)\right| \leq \frac{2\sqrt{[Log(n_0)+q_1](n_0)^3 2^{q_1}} L(Bo)}{\sqrt{m}} \tag{10.1}$$

Then, because, $\frac{2\sqrt{[Log(n_0)+q_1](n_0)^3 2^{q_1}}L(Bo)}{\sqrt{m}} = \epsilon^2 << 1$, for all the models $E^3(j)$ except for for a very negilgible probability,

$$\big|P\big\{\big(E^3(j+j_1),....,E^3(j+j_p)\big) \in Bo\big\} - L(Bo)\big| \leq \epsilon^2 L(Bo) \ .$$

As a matter of fact, by property 6.3.3, one can even admit that if the parameters q_1 and r_1 are well chosen,

$$\big|P\big\{\big(E^3(j+j_1),....,E^3(j+j_p)\big) \in Bo\big\} - L(Bo)\big| \leq 2\sqrt{\epsilon_3}L(Bo) \ ,$$

for all the logical models $E^3(j)$ where, for example, one can impose $\epsilon^3 = \frac{[Log(n_0)+q_1](n_0)^3 2^{q_1}}{m} \leq 1/10000$: cf section 6.4 ($2\sqrt{\epsilon_3} = \epsilon^2$).

Use of theorem 8

In this case, $p \leq p_m = \lfloor Log(n_0)/Log(2) \rfloor$ is supposed : if not, it doesn't make sens. Now, by lemma 8.1.2, $\sigma_B^2 \geq A(p)L(Bo)$. Then, in theorem 8,

$$\beta_{1,p} = \frac{\sqrt{n_0}[L^N(Bo) - L(Bo)]}{\sigma_B} = \frac{\sqrt{n_0}Ob(1)L(Bo)\epsilon_{Bo}^p}{A(p)L(Bo)}$$

$$= \frac{\sqrt{n_0}Ob(1)\Big[2\frac{\sqrt{[Log(n_0)+q_1](p_m)^3 2^{q_1}}}{\sqrt{m}}\Big]}{\big[1-(p^2-p+1)2^{-p}\big]}$$

$$\leq \frac{\Big[2\frac{\sqrt{[Log(n_0)+q_1]n_0(p_m)^3 2^{q_1}}}{\sqrt{m}}\Big]}{\big[1-(p^2-p+1)2^{-p}\big]} \leq \frac{2\sqrt{[Log(n_0)+q_1]/n_0}}{\big[1-(p^2-p+1)2^{-p}\big]}\ \epsilon^1 \ .$$

Moreover, $\gamma'_{1,p} = \frac{n_0\epsilon_{Bo}^{p_m}}{2A(p)}\Big[2^{3/2} + 2L(Bo)\ \Big]$ where $\epsilon_{Bo}^{p_m} = \sqrt{\frac{e_0(p_m)^3 2^{q_1}}{m}}$. Then

$$\gamma'_{1,p} \leq \frac{1}{2\big[1-(p^2-p+1)2^{-p}\big]}\sqrt{\frac{e_0(n_0)^2(p_m)^3 2^{q_1}}{m}}\ \Big[2^{3/2} + 2L(Bo)\ \Big]$$

$$= \frac{\sqrt{e_0}\big[2^{3/2} + 2L(Bo)\big]}{2\big[1-(p^2-p+1)2^{-p}\big]}\ \epsilon^1 \ .$$

Then,

$$P\Big\{\sqrt{N}\big|P_e - L(Bo)\big| \geq \sigma_B x\Big\} \leq K_1\Big(\frac{1-\beta_{1,p}/x}{1+\gamma'_{1,p}}x\Big) \approx K_1\big(\theta x\big) \ ,$$

where $K_1(x) = P\Big\{\frac{\sqrt{N}|P_e - L^N(Bo)|}{\sigma_1} \geq x\Big\}$, $P_e = \frac{1}{N}\sum_{j=1}^{N} 1_{Bo}(X_n)$ with the notations of theorem 8 when $X_j = E^3(j)$ and where $\theta \leq 1$, $\theta \approx 1$ if $x \geq 0.1$, $\beta_{1,p} << 1$ and $\gamma'_{1,p} << 1$.

Then, it is no possible to differentiate $e^3(j)$ and $b^3(n')$ of IID sequences : cf section 2.1.4. For example we have the following tables of $\Gamma\big(\theta x\big) \approx K_1\big(\theta x\big)$ if n_0 is large enough for x=1 and x=2.

θ	0.8	0.9	0.95	0.98	0.99	0.995	0.9975	1 (case IID)
$\Gamma(\theta)$	0.4237	0.3681	0.3421	0.3271	0.3222	0.3197	0.3185	0.3173

θ	0.8	0.9	0.95	0.98	0.99	0.995	0.9975	1 (case IID)
$\Gamma(2\theta)$	0.1096	0.0719	0.0574	0.0500	0.0477	0.0466	0.0460	0.0455

10.1.4 Example

By using this technique, we have created a real sequence ξ_n. This sequence can be asked to rene.blacher@imag.fr. Soon one will be able to obtain it in a website [1].

This sequence consists of the last ξ_n which one finds in this sequence : 1.444.240 $< n \leq$ 1.508.040. Its size is 66000 : N = 66000= 1.508.040- 1.444.240.

One obtains the sequence of bits b_s by writing in base 2 these ξ_n in the form $\xi_n = \overline{\overline{b_1^n.....b_{50}^n}}$. Then, we denote $b^4(n')$ the bits obtained by joining the b_r^n. Then, one has a sequence of $N = 66000 * 50 = 3.300.000$ bits $b^4(n')$.

In order to obtain $b^4(n')$, we have used a sequence a(j), $j = 1, 2, ..., N_3$ with $N_3 = 2.000.000$ and $1 \leq a(j) \leq 256$ obtained from texts : dictionary, encyclopedia, and Bible.

Then, we transform these sequences of letters in numbers. Now, there are only 26 letters. But it is necessary to add, the capital letters, the ":" , ";" , etc. There will be many of these 256 numbers which will not appear not or little. Also, we will write these numbers in base $\kappa = 32$ so that each number can have a probability reasonable to appear.

We choose $r_1 = 20$ ($32^{20} \approx 1.2677 * 10^{30}$), $n_0 = N_3/r_1 = 10^5$, $m = m^F(1.27 * 10^{30})$, $q_1 = 33$ ($2^{q_1} = 8.5895 * 10^9 \approx 10^{10}$). Then $log(n_0) \approx 11.513$, $log(n_0) + 33 \approx 44.5$. Then, one obtains a sequence of 3.300.000 bits which one denotes by $b^4(n')$. Then,

1) We have $\frac{(n_0)^2(p_m)^3 2^{q_1}}{m} \approx \frac{10^{10}(5.Log(10)/Log(2))^3 10^{10}}{10^{30}} \approx \frac{(16.62)^3}{10^{10}} \approx \frac{4591}{10^{10}} \approx \frac{4.6}{10^7} \approx (\epsilon^1)^2 << 1$. Then, $\epsilon^1 \approx \frac{\sqrt{0.46}}{10^3} \approx \frac{0.68}{10^3}$.

[1] In order to know if this website is created, type the words "Rene Blacher random numbers" in Google for example

Then, $\gamma'_{1,p_m} \leq \frac{\sqrt{e_0}\Big[2^{3/2}+2L(Bo)\Big]}{2\Big[1-(p^2-p+1)2^{-p}\Big]}\ \epsilon^1 = \frac{2\sqrt{44.5}\Big[2^{3/2}+2L(Bo)\Big]}{2\Big[1-(p^2-p+1)2^{-p}\Big]}\ \frac{0.68}{10^3} \approx 6.68\big[2^{3/2} + 2L(Bo)\big]\ \frac{0.68}{10^3} \approx \frac{12.845}{10^3} \approx 0.012$.

2) We have $\beta_{1,p_m} \leq \frac{2\frac{\sqrt{[Log(n_0)+q_1]n_0(p_m)^3 2^q}}{\sqrt{m}}}{1-(p^2-p+1)2^{-p}} \approx \frac{2\epsilon^1\sqrt{[Log(n_0)+q_1]/n_0}}{1} \approx \frac{1.36}{10^3}\sqrt{\frac{44.6}{10^5}} = 0.000029$.

Then, if $x \geq 1$, $K_1\Big(\frac{1-\beta_{1,p}/x}{1+\gamma'_{1,p}}x\Big) \approx \Gamma\big(\theta x\big) \approx \Gamma\Big(\frac{1-0.000029}{1+0.011}x\Big) = \Gamma\Big(0.98x\Big) = 0.3271$ if x=1.

Because in the IID case, $\Gamma(x) = 0.3173$, it is no possible to differentiate the sequence $E^3(j)$ or $B^4(n')$ from an IID sequence.

Remark 10.1.2 *It is not obliged that $\beta_{1,p}$ and $\gamma'_{1,p}$ are very small : cf example 11.2. One can thus moderate these conditions. What is sure, it is that under these assumptions, nothing can distinguish $E^3(j)$ or $B^4(n')$ from an IID sequence : cf section 2.1.4 .*

3) We have $\frac{(n_0)^3 2^{q_1}}{m} = \frac{10^{15}10^{10}}{10^{30}} \approx \frac{1}{10^5} << 1$.
Then $\epsilon_{Bo} = \frac{2\sqrt{44.5}}{\sqrt{10^5}} = 0.0422 = \epsilon^2 << 1$.

Then, for all n, for all p, for all sequence j_s,

$$\big|P\big\{\big(E^3(j+j_1),....,E^3(j+j_p)\big) \in Bo\big\} - L(Bo)\big| \leq 0.0422.L(Bo)\ ,$$

$$\big|P\big\{\big(B'_{n+j_1},....,B'_{n+j_p}\big) = (b_1,....,b_p)\big\} - \frac{1}{2^p}\big| \leq 0.0422\ \frac{1}{2^p}\ .$$

Then, $E^3(j)$ or $B^4(n')$ are very close to IID sequences.

Use of theorem 10

If one uses the other empirical theorems one obtains equivalents results. For example, one can assume that D(j) is Q-dependent cf section 10.4.5 of [18]. Then, one can use theorem 10, i.e. $P\Big\{\sqrt{N}\Big|\frac{P_e}{p_e} - L(Bo_1)\Big| > \sigma_{cp}\ x\Big\} \leq K_2\Big(\frac{1-\beta_{2,p}/x}{1+\gamma_{2,p}}x\Big)$. Then, one obtains the following majorations for $P\Big\{\sqrt{N}\Big|\frac{P_e}{p_e} - L(Bo_1)\Big| > \sigma_{cp}\ x\Big\}$ (cf section 11.2.11 of [18]):

		x=1	x=1.5	x=2	x=2.5	x= 3
Under IID assumption	p=1	0.317	0.133	0.045	0.012	0.0027
Under Q-dependence	p=1	0.322	0.134	0.047	0.011	0.0028
	p=3	0.325	0.136	0.048	0.012	0.0029
	p=5	0.328	0.137	0.050	0.013	0.0030
	p=10	0.3331	0.139	0.052	0.015	0.0032

10.1.5 Continuous case

Now, the relation

$$\left|P\big\{\big(E^3(j+j_1),....,E^3(j+j_p)\big)\in Bo\big\}-L(Bo)\right|\leq\frac{2\sqrt{[Log(n_0)+q_1](n_0)^3}L(Bo)}{\sqrt{h_0}}$$

holds for all model $E^1(j)$ except a tiny minority. Moreover all the logical models are correct cf section 6.4.

As a matter of fact for all the models which we studied we have found approximations still better than those which we have just understood previously. It is the case for the models with continuous density.

Let us choose $E^2(j)$ as a sequence of random variables which has a continous density with a Lipschitz coefficient K_0 not too big (it is equivalent that D(j) has a continous density with a Lipschitz coefficient K_1 not too big). Then, the conditional probabilty of $E^2(j)$ given $E^2(j+j_2)=e_2$, $E^2(j+j_3)=e_3$, ...,$E^2(j+j_p)=e_p$ has also a Lipschitz coefficient K_0' :

$$\left|P\Big\{E^2(j)=e_0\mid E^2(j+j_2)=e_2,...\Big\}-P\Big\{E^2(j)=e_0'\mid E^2(j+j_2)=e_2,...\Big\}\right|\leq K_0'|e_0'-e_0|\ .$$

Now one can apply property 6.3.5 to conditional probabilities. Then, for all interval I,

$$P\big\{E^3(j)\in I\mid E^3(j+j_2)=e_2,...\big\}=L(I)\Big[1+\frac{O(1)K_0'}{N(I)}\Big]\ .$$

It is clear that under this assumption, the approximation with an IID sequence is better: there is a denominator in N(I) instead of $\sqrt{N(I)}$. One thus ensures thus approximations better than those obtained by using the hypothesis 6.3.4. For example

$$P\big\{E^3(j)\in I_k\mid E^3(j+j_2)=e_2,...\big\}=L(I_k)\Big[1+\frac{O(1)K_0 2^{q_1}}{m}\Big]\ .$$

instead of

$$P\big\{E^3(j)\in I_k\mid E^3(j+j_2)=e_2,...\big\}=L(I_k)\Big[1+\frac{Ob(1)e_0'\sqrt{2^{q_1}}}{\sqrt{m}}\Big]\ ,$$

where $e_0'=2\sqrt{[Log(n_0)+q_1](n_0)^3}$.

Therefore under reasonable hypotheses, we can prove that we have an approximation of an IID sequence which is better than that defined in the general case, for example in equation 10.1.

10.1.6 Conclusion

Now, almost all the models D(j) are good models. Moreover all the logical models are correct. On the other hand, the models with continous densities are closer to the IID sequences.

Chapter 11

Building of an IID sequence : II

11.1 General method

The building studied here must be associated with a model where the data have a density admitting a coefficient of Lipschitz not too large (it is known that it is a correct assumption : cf section 6.4.1).

11.1.1 Description of the method

We use again a sequence of data a(j) translated in number: a(j) , $j = 1, 2, ..., N_3$, as in section 10.1.

Choice of the parameters

a) We choose $\alpha \in \mathbb{R}_+$ such that $\alpha \leq 0.02$ according to the quality of the desired approximation [1].
b) One choose S=10.
c) One chooses now $r_0 = r_1$ and $q_0 \in \mathbb{N}^*$ such that :
c-1) q_0/r_0 is maximum
c-2) $m_S/2^{q_0} \geq 1001$,
c-3) $m_S = m^F([m^F(\kappa^{r_0})]^{3/4})$ is sufficiently large but not too (cf remark 11.1.7 of [18])
c-4) $\sqrt{q_0 2^{q_0}}\Gamma^{-1}(a_2^S) \leq \frac{2\alpha\sqrt{S}}{\sqrt{N_3}}\sqrt{r_0 m_S}$, where $a_2^S = \Gamma\Big(\Gamma^{-1}(4^{-q_0})\sqrt{\frac{\lfloor m_S/2^{q_0}\rfloor}{m_S/2^{q_0}} + \frac{2^{q_0}}{m_S}}\Big) \approx 1/4^{q_0}$.

[1] As a matter of fact, in function of $\beta_{1,p}$: cf theorem 8

First transformation

a) We transform the sequence of data a(j), $j = 1, 2, ..., N_3$, into a sequence of random bits $e^2(j)$ by the same way as in section 10.1.
b) We set $e^3_S(j) = mT^m_1(e^2_S(j)/m^1)$, [2] where $m^1 = m^F(\kappa^{r_0})$.

Remark 11.1.1 *One can also use $e^3_S(j) = mT^m_1(e^2_S(j)/m^1)$ only for the first $j \in \{1, 2, ..., \lfloor N_3/r_0 \rfloor\}$: cf Remarks 11.1.1 and 11.1.2 of [18]. Moreover, one can also suppress this step : in this case, one sets $e^3(j) = e^2(j)$.*

It is supposed that sequence $E^3(j)$ has asymptotic independence. One checks this asymptotic independence by logical and numerical studies : e.g. cf chapter 9.

Use of the limit theorems

a) Because $e^3(j)$ depends on S, we write $e^3_S(j)$ instead of $e^3(j)$.

b) We denote by $e^4_S(t)$, $t = 1, 2, ..., N_2$, $N_2 = NS \leq \lfloor N_3/r_0 \rfloor$, a subsequence of $e^3_S(j)$ obtained by suppressing some subsequences $e^3_S(\rho_u), e^3_S(\rho_u + s_{u_1})$, ...,$e^3_S(\rho_u + s_{u_n})$ in order to ensure independence between the lines defined below. If one does not have independent files, this step is not necessary forcing.

c) We set $f_S(i, n) = e^4_S(n + N(i - 1))$ for i=1,...,S , n = 1,...,N.

d) If $i \in 2\mathbb{N}$, we set $f_1(i, n) = f(i, N - n + 1)$ for i=1,...,S, n = 1,...,N [3].

e) We set $g_S(n) = \sum_{i=1}^{S} f_S(i, n)$ mod m_S for n = 1,...,N. This corresponds to use the CLT.

f) We set $h_S(n) = \overline{g_S(n)}$ mod m_S for n = 1,...,N. This corresponds to use the XORLT.

Checking of S

a) One checks by numerical calculations that the curve of the

$$h \mapsto P\{H_S(n) = h \mid H_S(n + j_2) = h_2, ..., H_S(n + j_p) = h_p\}$$

[2] cf definition of T^m_1 : definition 1.2.5

[3] Indeed, in section 9.3, it was understood that, for some files (e.g. texts),

$$P\Big\{F(i, n) = f \;\Big|\; F(i, n - j'_2) = f_2, ..., F(i, n - j'_p) = f_p\Big\} \to P\{F(i, n) = f\}$$

as $j'_2 \to \infty$ when $j'_1 = 0 < j'_2 < .... < j'_p$. In order to have the same result for

$$P\{F(i, n) = f | F(i, n + j'_2) = f_2, ..., F(i, n + j'_p) = f_p\} \to P\{F(i, n) = f\},$$

we invert the even lines.

Therefore, logically, when one will summon the lines $f_1(i, n)$, it is reasonable to think that it will be difficult to predict $\sum_i f_1(i, n)$ knowing elements which are passed **or** future.

is enough close to that of the uniformity: it is necessary that the condition of equation 6.3 is satisfied. In general, it is well the case if S= 10.

If it is not the case, one remakes several times the previous operations with various $S > 10$. One chooses smallest $S \geq 10$ which is appropriate. It is noted S_0.

b) We set $h(n) = h_{S_0}(n)$ for n = 1,...,N.

Use of the Fibonacci function

a) Let $m = m_{S_0} = fi_{n_3+1}$ and $a = fi_{n_3} < m$ where $n_3 \in \mathbb{N}$. Let T_{q_0} be the Fibonacci function with parameters a, m and q_0. We set $x(n) = T_{q_0}\big(h(n)/m\big) = \overline{\overline{0, b_1^n, b_2^n .....b_{q_0}^n}}$ [4] where q_0 was defined previously in 11.1.1.

b) We set $b'_{q_0 n-r+1} = b_r^n$ for n=1,..,N and $r = 1, 2, ..., q_0$ (cf also step "f" section 10.1.2)

c) The sequence $\{b'_n\}$ is noted $b^0(n')$, $n' = 1, 2, ..., Nq_0$.

11.1.2 Explanation of the conditions about q_0 and r_0

Because the various steps of this construction, one can accept the model of the section 6.3.4 : $P\{H(n) = h \mid H(n + j_s) = h_s, s = 2, 3...\} = \frac{1}{m}\Big[1 + u_k\Big]$: cf chapter 7 of [18].

We deduce $P\big\{X(n) = k/2^{q_0}\big|X(n + j_2) = x_2, X(n + j_3) = x_3, ...\big\} = 1/2^{q_0} + Ob(1)\epsilon_{I_k}$, where $\mid \epsilon_{I_k}\mid \leq \epsilon = \frac{\Gamma^{-1}(4^{-q_0})\sqrt{N_{Iel}}}{m}$: cf sections 11.1.3 and 11.1.4 of [18].

We deduce that, for all sequences of bits bi_n, for all finite injective sequence j_s,

$$P\big\{B^0(n') = bi_1\big|B^0(n' + j_2) = bi_2, B^0(n' + j_3) = bi_3, ...\big\} = 1/2 + Ob(1)\epsilon \ ,$$

where $\epsilon = \alpha/\sqrt{q_0 N}$ when Nq_0 is the size of sample $\{b^0(n')\}$: cf section 11.1.3 of [18].

Now apply the theorem 9 : $P\Big\{\sqrt{N}\big|P_e - (1/2)^p\big| \geq \sigma_B x\Big\} \leq \Gamma\big(\big[\frac{1-\beta_{1,p}/x}{1+\gamma_{1,p}}\big]x\big)$, where $\beta_{1,p} \leq \frac{\sqrt{Nq_0}\epsilon_p}{\sqrt{A(p)L(Bo)}} \approx \frac{\sqrt{Nq_0}.2p\epsilon}{A(p)^{1/2}2^{p/2}} = \frac{2p\alpha}{A(p)^{1/2}2^{p/2}}$.

Then, $\beta_{1,p}$ is enough small in order that $P\big\{(B^0(n'), B^0(n'+j_2), ..., B^0(n'+j_p)) = (bi_1, ..., bi_p)\big\}$ is about also close to $(1/2)^p$ that it would be it in case

[4] $\overline{\overline{0, b_1^n, b_2^n .....}}$ is the binary writting of $\overline{T}(h(n))/m$.

IID. One obtains the same type of results for theorem 10. It is not thus finally possible to distinguish the sequence $b^0(n')$ from an IID sample.

11.2 Example

In section 11.2 of [18] we study an example : we obtain a sequence of random bits $b^1(n')$. This sequence can be asked to rene.blacher@imag.fr. Soon one will be able to obtain it in a website [5].

Currently, this sequence $b^1(n')$ is the first part of the sequence of numbers ξ_n : $n \leq 1000000$. Its size is N = 1.000.000.

One finds the sequence of bits b_s by writing these ξ_n in binary system in the form $\xi_n = \overline{\overline{b_1^n b_2^n ....b_{50}^n}}$ [6].

11.2.1 Choice of random datas

We have used a sequence of data a(j) with $N_3 = 298.159.056$ and $1 \leq a(j) \leq 256$. The data result from texts, mathematical texts and file of programming : cf section 9.3.

In the study of data, our numerical results prove that one can consider that the sequence C(j) and D(j) are Qd-dependent with Qd=22 and Qd=2 : cf [18].

We choose $\alpha = 0.02$, S=10, $q_0 = 57$, $r_0 = 28$ and $m^1 = m^F(32^{28})$.

11.2.2 Building of a random sequence $b^1(n')$

We have suppress some sequences $\{e^3(\rho_u), e^3(\rho_u + 1)....., e^3(\rho_u + n_4)\}$ in order that f(i,n) and f(i',n') belongs to different files if $i \neq i'$. Then, the lines are independent.

Then, $h(n) = \overline{\sum_{i=1}^{10} f_1(i,n)}$ for n= 1,...,N where N = 1.000.000 and $x(n) = \overline{\overline{0, b_1^n b_2^n ....b_{57}^n}}$.

We deduce the sequence $b^1(n')$, n'=1,2,...,57000000, where $\{b^1(n')\} = \{b_r^n\}$.

11.2.3 Properties of $B^1(n')$

In section 11.2.10, 11.2.11, 11.2.13 of [18], we study the samples $b^1(\psi(n))$, $n = 1, 2, ..., N_1$, where $N_1 \leq N$ and where $\psi : \{1, 2, ..., N_1\} \rightarrow \{1, 2, ..., N\}$ is an injective function [7].

[5] In order to know if this website is created, type the words "Rene Blacher random numbers" in Google for example

[6] As a matter of fact, we obtain initially $\xi_n = 2^{57}x(n)$ where $x(n) = \overline{\overline{0, b_1^n b_2^n ....b_{57}^n}}$. But, Matlab 2006 does not write numbers which have more 50 bits. Then it is simpler to forget the last bits $b_{51}^n, ..., b_{57}^n$.

[7] The worst approximation occurs for the sample of maximum size: $N_1 = q_0 N$.

Then, we have used the assumption II :

$$P\big\{B^1(n') = b \;\big|\; B^1(n'+j_2) = b_2, ..., B^1(n'+j_p) = b_p\big\} = 1/2 + \frac{Ob(1)\alpha}{\sqrt{Nq_0}} \ ,$$

where $\alpha/(Nq_0)^{1/2} \approx 2.649/10^6$.

For example, we have studied the empirical aspect by using theorem 10 where one can consider that, the sequence $b^1(n')$ is surely Qd_B-dependent with $Qd_B = 57$. Various results are obtained in [18] : e.g. one obtains the following increases for $P\Big\{\sqrt{N_1}\big|P_e^B - (1/2)^p\big| > \sigma_{cp}\ x\Big\}$:

		x=1	x=1.5	x=2	x=2.5	x= 3
Under IID assumption	p=1	0.317	0.133	0.045	0.012	0.0027
Under assumption II	p=1	0.401	0.180	0.065	0.019	0.0045
	p=3	0.337	0.144	0.050	0.014	0.0031
	p=5	0.323	0.137	0.047	0.013	0.0028
	p=10	0.319	0.135	0.046	0.013	0.0028

Then, it is difficult to differentiate the sequence $b^1(n')$ from an IID sample. Indeed, if our data were not IID, that would imply that $\sqrt{Nq_0}\big|P_e^B - (1/2)^p\big|$ would be large. That can certainly occur for some $(bi_1, bi_2, ..., bi_p)$ but, as the previous increases show it, with a probability which is not too different from that of IID case.

In the previous results, one has increase our approximations by using the 2-dependence which exists for the sequence D(j). For the sequence $B^1(n')$, the results are much better because we did everything in our building so that it is identical to a sequence IID.

One could thus have finer increases. For that, it is necessary to calculate γ_p^1 where $\sigma_1^2 = \sigma_B^2[1 + 2\gamma_p^1]$. In [18] we have estimated σ_1^2 and have compared it with the exact value of σ_B^2 for p=1,2,3,4,5.

Remark that one obtains the same type of equations with X(n) as with the $B^1(n')$: cf section 11.2.13 [18].

11.2.4 Tests

In section 11.2.14 of [18], we are have controled the conclusions of this study by making tests. We use the classical Diehard tests ([2], [1]) and the Higher order correlation coefficients tests ([10]). We tested the sequences $b^1(n')$ or x(n).

Results are in accordance with what we waited: for sequences $b^1(n')$ and x(n), the hypothesis "randomness" can be accepted by all these tests.

11.2.5 Conclusion

The inequalities above show that it could be that $b^1(n')$ does not check certain tests of an IID sequence, but that will occur with hardly more probabilities that for a sample of a really IID sequence. It is thus not possible to differentiate the sequence $b^1(n')$ from an IID sample by using these tests.

Thus, it is not possible to differentiate the sequence $b^1(n')$ from an IID sample by using P_e^B and P_e^B/p_e^B. Morever $B^1(n')$ satisfies also the very important additional property : $P\big\{B^1(n') = b\big|B^1(n'+j_2) = b_2, ..., B^1(n'+j_p) = b_p\big\} = 1/2 + Ob(1)\epsilon$.

Then, the sequence $b^1(n')$ satisfy all the conditions which we have indicates in section 2.1 and also this theoretical property. Then, one can admit that $b^1(n')$ is an IID sample.

11.3 Continuous case

Let us notice that the use of the CLT smoothes the probabilities of the sums G(j). One can thus admit that they have a continuous density. It is particularly true if $E^3(j)$ has already a continuous density. Then, this model has to be studied under the assumption as the density of $E^3(j)$ with respect to $\mu_m \otimes ..... \otimes \mu_m$ is continuous and has a coefficient of Lipschitz which is not too large.

It is known that it is a correct assumption : cf section 6.4.1. Thus, this method is completely sure : we are sure that the sequence $b^1(n')$ is IID.

This technique could be used with the machines which produce random numbers very quickly. It remains valid even if there are dysfunctions (provided that the produced numbers are not completely deterministic or very near to a completely deterministic model).

In fact, it is supposed that there is asymptotic independence: one can thus choose sums of S terms $Z_{i,n}$ with S=10, 20 or much if it is necessary. This study thus applies perfectly to the case of machines not having too great dysfunctions and quickly producing random numbers.

Let $G'_n = \sum_i \frac{Z_{i,n} - \mathbb{E}\{Z_{i,n}\}}{\sigma_n}$ where $\sigma_n^2 = \mathbb{E}\big\{\big([Z_{1,n} - \mathbb{E}\{Z_{1,n}\}] + ... + [Z_{S,n} - \mathbb{E}\{Z_{S,n}\}]\big)^2\big\}$. By the CLT, the conditional distribution of G'_n knowing $G'_{n+j_2} = g_2$,.....,$G'_{n+j_p} = g'_p$ has a distribution close to a normal one. This normal law has a small variance if the linear correlation coefficients are rather close to ± 1.

Now it is supposed that, for any n, any p, any sequence j_s, s=1,2,, p, $P\big\{G'_{n+j_1} = g \mid G'_{n+j_2} = g_2, ..., G'_{n+j_p} = g'_p\big\}$ is too not concentrated nearly one only point. Because the CLT is used, that means effectively that the linear correlation coefficients are not too close to ± 1.

As a matter of fact, what interests us is the conditional probability of the $H_n = \overline{Z_{1,n} + ... + Z_{S,n}}$: it is not wanted that it is concentrated nearly one only point. It is thus supposed that $P\big\{H_{n+j_1} \in I|H_{n+j_2} = h_2, ..., H_{n+j_p} = h_p\big\}$ is

not too different from L(I). It is an assumption easy to check, considering that one has the asymptotic independence of $\big(H_{n+j_1}, H_{n+j_2}, ..., H_{n+j_p}\big)$ (according to the properties of the XORLT : e.g. cf proposition 7.2.2). Then, one has already almost the wished equation. In order to be sure that this equation holds, it is enough to use besides the transformation T_q.

Moreover the curve of probabilities of G'_n is smooth. It is even more the case for H_n. Then, the same method as that one of property 6.3.5 can be applied to the random sequence $X_n = T_q(H_n)$: therefore, one can assume that

$$P\big\{X_1 \in I \mid X_{n+j_2} = x_2, ..., X_{n+j_p} = x_p\big\} = \frac{N(I)}{m}\Big[1 + \frac{O(1)6K'_0}{N(I)}\Big] ,$$

where most of the time, $K'_0 \leq 1$ (and is even much smaller) and where $O(1) \approx 1$. As one wants to avoid any risk of error, one will admit $K'_0 \leq 100$ (of course this increase depends on data). It is easy to understand by using theorem 7 that if S increases K'_0 decreases very quickly [8]. Therefore, generally, this increase is certainly much too strong. Therefore,

$$P\big\{X_1 \in I_k \mid X_{n+j_2} = x_2, ..., X_{n+j_p} = x_p\big\} = L(I_k)\Big[1 + \frac{Ob(1)}{m}\Big]\Big[1 + \frac{O(1)6K'_0}{N(I_k)}\Big]$$

$$= \frac{1}{2^q}\Big[1 + \frac{O(1)6K'_0 2^q}{m}\Big] .$$

Let $Bo = \cup_{k\in\Theta} I_k$. Then,

$$P\big\{X_1 \in Bo \mid X_{n+j_2} = x_2, ..., X_{n+j_p} = x_p\big\} = \sum_{k\in\Theta} \frac{1}{2^q}\Big[1 + \frac{O(1)6K'_0 2^q}{m}\Big]$$

$$= L(Bo)\Big[1 + \frac{O(1)6K'_0 2^q}{m}\Big] .$$

Then, by proposition 4.3.1 of [18]

$$P\big\{B_1 = b | B_2 = b_2 ....., B_p\big\} = \frac{1}{2}\Big[1 + Ob(1)\epsilon\Big] ,$$

where $\epsilon = \frac{12O(1)K'_0 2^q}{m}$ with $O(1) \approx 1$.

Donc, by proposition 4.2.2 of [18],

$$P\Big\{\{B_{n+j_1} = b_1\} \cap ..... \cap \{B_{n+j_p} = b_p\}\Big\} = \big[1/2 + Ob(1)\epsilon/2\big] ..... \big[1/2 + Ob(1)\epsilon/2\big]$$

[8] In theorem 7 independence is assumed. But, clearly, the reasoning remains valid : cf also section 5.5 of [18]

$$\approx \frac{1}{2^p} + \frac{Ob(1)p\epsilon}{2^p} \ .$$

Therefore, if one wants to use analog results to those of theorem 8, one will have to replace $L(Bo)\epsilon_{Bo}^p$ by $p\epsilon/2^p$, therefore to replace ϵ_{Bo}^p by $p\epsilon$, considering that in the case of random bits B_n, $L(Bo) = 1/2^p$. Moreover, by using the same type of proof as in theorem 8, one understands that one will have to replace $\gamma'_{1,p} = \frac{n_0\epsilon_{Bo}^p}{2A(p)}\Big[2^{3/2} + 2L(Bo)\Big]$ by

$$\gamma"_{1,p} = \frac{n_0(p\epsilon)}{2A(p)}\Big[4 + 2L(Bo)\Big] \ .$$

Therefore, so that X_n cannot be differentiated from an IID sequence, it will be necessary to impose

$$\frac{n_0}{2A(p)}\frac{12pK_0'2^q}{m}[4+1] = \frac{30n_0p}{A(p)}\frac{K_0'2^q}{m} << 1 \ .$$

Now, by lemma 8.1.2, $\sigma_B^2 \geq A(p)L(Bo)$. Then, in theorem 8,

$$\beta_{1,p} = \frac{\sqrt{n_0}[L^N(Bo) - L(Bo)]}{\sigma_B} \approx \frac{\sqrt{n_0}Ob(1)p\epsilon/2^p}{\sqrt{A(p)/2^p}} \leq \frac{\sqrt{n_0}p\epsilon}{\sqrt{A(p)2^p}} << 1 \ .$$

For example, with $K_0' \leq 100$, let us choose $p \leq p_m = \lfloor Log(n_0)/Log(2) \rfloor$ and $n_0 = 10^6$. Because $log(n_0)/log(2) \approx 19.9$, $\frac{p}{A(p)} \leq \frac{3}{1/8} = 24$. Therefore,

$$\frac{30n_0p}{A(p)}\frac{K_0'2^q}{m} < \frac{24 * 30n_0}{1}\frac{100.2^q}{m} \ .$$

Therefore, if $m \geq 10^{34}$, q=60 ($2^{60} \approx 1.153*10^{18}$), $n_0 = 10^6$, $log(n_0) = 13.815$

$$\frac{24 * 30n_0}{1}\frac{100.2^q}{m} \leq \frac{720 * 10^6}{1}\frac{100 * 1.153 * 10^{18}}{10^{34}} \leq \frac{0.83}{10^5} \ .$$

Remark 11.3.1 *By using the theorem 8 one has used conditions too much strong if it is supposed that one has the asymptotic independence: in fact, one will obtain approximations much better concretely than those described by the previous results.*

Chapter 12

Building of IID sequences : III

12.1 Third method

In this section one uses the convergence of the XORTL (cf theorem 7). One does not apply it to a sequence of numbers as f(i,n), n=1,...,N, but to random numbers of size N, i.e. very large, for example with a sequence of bits of size 100.000.000, these numbers have values in $\{0, 1, ..., 2^{100.000.000} - 1\}$.

12.1.1 Method of construction of the sequence

1) We use again a sequence of data a(j) as in section 10.1. One transforms again it into a sequence of random bits $b^3(n')$ by the same way as in section 10.1.

These $b^3(n')$ are grouped in S lines which we rewrite $bt_i(n')$, i=1,2,..,S, n'=1,2,...,J, each one belonging to files independent of the others.

2) One modifies the lines $bt_i(n')$, $n' = 1, 2, ..., J$, thanks to transformations having a behavior close to that of the permutations. In this aim, one uses other sequences of data $c_i^1(n') \in \{1, 2, ..., J\}$, $n' = 1, 2, ..., J$, where $i = 1, 2, ..., 3S$. Because we use transformations similar to permutations, we set $c_i^1(n') = Perm_i(n')$ in order that the notations are clearer.

2-a) One groups them togheter by sets of three successive sequences $Perm_t^i(n')$ for t=1,2,3, i=1,2,...,S, $n' = 1, 2, ..., J$.

2-b) For each line i, for $n' = 1, 2, ..., J$, one sets, $r_0^i(n') = bt_i(n')$ and, for each t=1,2,3, $r_t^i(n') = bt_i(Perm_t^i(n'))$ for $n' = 1, 2, ..., J$.

2-c) For each line i, we set $r_i(n') = \overline{r_0^i(n') + r_1^i(n') + r_2^i(n') + r_3^i(n')}$ modulo 2 for $n' = 1, 2, ..., J$.

3) One definite g_i as the number with J bits whose writing base 2 is $g_i = \overline{\overline{r_i(1)r_i(2)....r_i(J)}}$.

4) We set $k = \overline{\sum_{i=1}^{S} g_i}$, mod $M_2 + 1$ where $M_2 = 2^J - 1$ (calculations algorithms

are in [18]).

5) Let $k = \overline{\overline{b_1, b_2, ...}}$, the writing of k base 2. Then, the sequence $b_1, b_2, ..., b_J$ is a sequence of random bits.

12.1.2 Properties

One uses the properties of the XORLT : One sets $X_i = G_i$ in order to use theorem 7. One supposes that $\sigma^2_{V_r} \leq 1$ (cf remark 7.3.3) where $V^i_{x_n}$ is the sequence of random variables defined on probability space $(\Omega_7, \mathcal{A}_7, Proba_7)$ in hypothesis 7.3.1 .

Then, by theorem 7, for all $y \in \{0, 1, ..., 2^J - 1\}$, with a probability greater than $1 - \Gamma(b)$ approximately,

$$P\{K = y\} \approx (1/2^J)\Big[1 + \frac{bOb(1)}{\sqrt{2^{J(S-1)}}}\Big] .$$

Now, one can write $\{K = y\} = \{B_1 = b_1\} \cap \{B_2 = b_2\} \cap ... \cap \{B_{2^J} = b_{2^J}\}$. Then, if b^0 is large, with a probability infinitely close to 1,

$$P\Big\{\{B_1 = b_1\} \cap \{B_2 = b_2\} \cap ... \cap \{B_{2^J} = b_{2^J}\}\Big\} = (1/2^J)\Big[1 + \frac{b^0 Ob(1)}{\sqrt{2^{J(S-1)}}}\Big] .$$

12.1.3 Permutations and associated transformations

One uses transformations having a behavior close to that of the permutations. Of course, one thinks that one could use Matlab permutations for example. But, it poses a problem: a priori they are not permutations taken randomly. As a matter of fact, one is in the case envisaged by Knuth ([1] : cf also definition 2.1.5) and which it is necessary to avoid. One needs permutations taken randomly.

For that, one want to use nondeterministic sequences of data to define the permutations.

For example let us suppose that one wants to permute a sequence x(j) of size N and that one has data $d(j) \in \{0, 1, ..., N\}$. One would like to be able to define a permutation P by P(j)=d(j). But, there is no reason that P is injective.

One can try to remove the j, j' such that d(j)=d(j'), $j \neq j'$. But if N is large, that can be long. Then, it is easier to use these data differently.

Indeed, it is easy to understand that the technique defined in step 2 allows a mixture of the lines which is as random as it would be the choice of a permutation taken randomly. That is thus adapted perfectly so that we can suppose that the probabilities of each line are chosen randomly.

12.1.4 Example

By using the technique defined in section 12.1.1 with S=5, J=25402545, we have created a real sequence ξ_n. This sequence can be asked to rene.blacher@imag.fr. Soon one will be able to obtain it in a website [1].

Curently, this sequence consists of the secund part of the sequence ξ_n which we have obtained by the three methods studied in this report : $1000000 < n \leq 1.408.040$. Its size is N = 508.040.

One obtains the sequence of bits $b^2(n)$, n=1,2,....,20.402.000 by writing in base 2 these ξ_n in the form $\xi_n = \overline{\overline{b_1^n.....b_{50}^n}}$: $\{b^2(n')\} = \{b_s^n\}$.

Properties

By using theorem 7 (because $J(S-1)/2 \geq 50.000.000$), one understands that in the set of probabilities provided with the distribution such that $\sigma^2_{V_r} \leq 1$, for all p, with a probability infinitely close to 1,

$$P\Big\{\{B_n = b_1\} \cap \{B_{n+j_2} = b_2\} \cap ... \cap \{B_{n+j_p} = b_p\}\Big\} = (1/2^p)\Big[1 + \frac{Ob(1)}{2^{50.000.000}}\Big] .$$

It is a very good approximation! It allows to obtain very fine results about empirical probabilities, e.g. $\beta_{1,p} \leq \frac{1}{2^{49.999.983}}$ and $\gamma_{1,p} \leq \frac{1}{2^{49.999.925}}$ and, for $x \geq \frac{1}{2^{1000}}$,

$$P\Big\{\sqrt{N_1}\big|P_e^B - (1/2)^p\big| \geq \sigma_B x\Big\} \leq \Gamma\Big[\Big(1 - \frac{1}{2^{49.998.982}}\Big)x\Big] .$$

It is quite clear that with such an approximation, nothing could differentiate such a sequence from an IID sequence if one has sample with size 25.402.545.

Now, the assumptions of theorem 7 are realistic : indeed $r_t^i(n') = bt_i(Perm_t^i(n'))$ has the characteristic of permutations chosen randomly. Therefore, probabilities $p^s_{g_s}$ assocated to each g_s have to be regarded as chosen randomly.

Of course, there will be always models which will not check these assumptions. But there will be of it a negligible number with probability $\Gamma(b)$ introduced into the theorem 7 : i.e. one can suppose that one has an IID sample which also can not check correct assumptions, but with a probability infintely negligible.

One must thus admit that the previous properties are well checked, i.e. one has an extremely fine approximation.

Tests

We have verifed the previous conclusions by making tests : cf [18] section 12.1.10. We have used the classical Diehard tests cf [2], [1] and the higher order corre-

[1] In order to know if this website is created, type the words "Rene Blacher random numbers" in Google for example

lation coefficients cf [10]. Results are in accordance with what we waited: the hypothesis "randomness" is accepted by all these tests.

12.1.5 Conclusion

We thus have one third method to build IID sequences. The advantage is this one has extremely strong mathematical properties and behaves exactly like an IID sequence : it is always possible that the sequence $b^2(n')$ is not good, but only with a very negligible probability.

12.1.6 Comparison of methods II and III

The method defined in this section 12.1.1 has theoretical results much better than those defined in the chapter 11.

But, such a quality of the approximation seems useless since one reasons on samples. In our method defined in chapter 11, we obtained an approximation theoretically less fine and yet, we saw that one can regard it as sufficient.

The improvement made in this section to the method defined in the chapter 11 seems not to mean much. For example, there exists always a probability close to 0.045 such as $\frac{|P_e-(1/2^p)|}{\sigma\sqrt{N}} \geq 2$.

The approximation provided by the method defined in this section 12.1.1 can thus be only one additional guarantee which one can take when one builds a sequence of random bits b_n. It could however to be useful if one wanted to build functions of the $b^2(n')$ with certain mathematical properties

Appendix A

Continous case in dimension 2

We want to prove property 6.3.5 in dimension 2. We keep the notations of section 6.3.3.

Property A.0.1 *Let $m >> 1$. Let h_N be the probability density function of $(Y_1, Y_2) \in F(m)^2$ with respect to $\mu_m \otimes \mu_m$: $\int_{u=0}^{1} \int_{v=0}^{1} h_N(u,v)\mu_m(dv)\mu_m(dv) = 1$. Let h'_N be the probability density function such that $\int_{u=0}^{1} \int_{v=0}^{1} h'_N(u,v).dudv = 1$ and $h'_N = (1/c_0)h_N$.*

Let $K_0 \in \mathbb{R}_+$ such that $|h_N(r_1,r_2) - h_N(r'_1,r'_2)| \leq K_0 Max_s\{|r'_s - r_s|\}$ and $|h'_N(r_1,r_2) - h'_N(r'_1,r'_2)| \leq K_0 Max_s\{|r'_s - r_s|\}$ when $r_1, r_2, r'_1, r'_2 \in [0,1]$.

Assume again that T is a Fibonacci congruence. Then, the following equality holds :

$$P\big\{\big(\overline{T}(mY_1)/m, \overline{T}(mY_2)/m\big) \in I_1 \otimes I_2\big\} = L(I_1)L(I_2)\Big[1 + \frac{O(1)K_0}{Inf_s[N(I_s)]}\Big] \ ,$$

where $inf_s[N(I_s)] \leq m/2$.

Proof We need the following lemmas.

Lemma A.0.2 *The following equality holds :*

$$c_0 = 1 + \frac{O(1)K_0}{m} \ .$$

Proof The following equalities hold :

$$1 = \sum_{t,t'} \int_{u=t/m}^{(t+1)/m} \int_{v=t'/m}^{(t'+1)/m} h'_N(u,v).dudv$$

$$= \sum_{t,t'} \int_{u=t/m}^{(t+1)/m} \int_{v=t'/m}^{(t'+1)/m} \Big[h'_N(t/m, t'/m) + Ob(1)K_0/m\Big] dudv$$

$$= \frac{1}{m^2}\sum_{t,t'} h'_N(t/m, t'/m) + \frac{Ob(1)K_0}{m}$$

$$= \int_{u=0}^{1}\int_{v=0}^{1} h'_N(u,v)\mu_m(dv)\mu_m(dv) + \frac{Ob(1)K_0}{m} \ .$$

Then, $\int_{u=0}^{1}\int_{v=0}^{1} h'_N(u,v).\mu_m(dv)\mu_m(dv) = 1 + \frac{Ob(1)K_0}{m}$. Therefore,

$$1 = \int_{u=0}^{1}\int_{v=0}^{1} h_N(u,v)\mu_m(dv)\mu_m(dv) = c_0\int_{u=0}^{1}\int_{v=0}^{1} h'_N(u,v).\mu_m(dv)\mu_m(dv)$$

$$= c_0\Big[1 + \frac{Ob(1)K_0}{m}\Big] \ . \blacksquare$$

Lemma A.0.3 *The following equality holds :*

$$\frac{1}{N(I_1)N(I_2)}\sum_{r,r'} h_N\big(r/N(I_1), r'/N(I_2)\big) = 1 + \frac{2Ob(1)K_0}{\inf_s[N(I_s)]} \ .$$

Proof The following equalities hold :

$$1 = \sum_{r,r'}\int_{u=r/N(I_1)}^{(r+1)/N(I_1)}\int_{v=r'/N(I_2)}^{(r'+1)/N(I_2)} h'_N(u,v)dudv$$

$$= \sum_{r,r'}\int_{u=r/N(I_1)}^{(r+1)/N(I_1)}\int_{v=r'/N(I_2)}^{(r'+1)/N(I_2)} \Big[h'_N\big(r/N(I_1), r'/N(I_2)\big) + Ob(1)K_0 Max_s\Big(\frac{1}{N(I_s)}\Big)\Big]dudv$$

$$= \sum_{r,r'}\int_{u=r/N(I_1)}^{(r+1)/N(I_1)}\int_{v=r'/N(I_2)}^{(r'+1)/N(I_2)} \Big[h'_N\big(r/N(I_1), r'/N(I_2)\big) + \frac{Ob(1)K_0}{\inf_s[N(I_s)]}\Big]dudv$$

$$= \frac{1}{N(I_1)N(I_2)}\sum_{r,r'} h'_N\big(r/N(I_1), r'/N(I_2)\big) + \frac{Ob(1)K_0}{\inf_s[N(I_s)]} \ .$$

Therefore

$$c_0 = \frac{1}{N(I_1)N(I_2)}\sum_{r,r'} h_N\big(r/N(I_1), r'/N(I_2)\big) + \frac{Ob(1)c_0K_0}{\inf_s[N(I_s)]} \ .$$

Therefore, by lemma A.0.2,

$$c_0 = 1 + \frac{O(1)K_0}{m} = \frac{1}{N(I_1)N(I_2)}\sum_{r,r'} h_N\big(r/N(I_1), r'/N(I_2)\big) + \frac{Ob(1)[1 + \frac{O(1)K_0}{m}]K_0}{\inf_s[N(I_s)]} \ .$$

Because $m >> 1$ and $Inf_s[N(I_s]) \leq m/2$, we deduce the lemma. ■

Then, the following property holds.

Property A.0.4 *Let* $I_1 = [c_1/m, c_1'/m[$ *and* $I_2 = [c_2/m, c_2'/m[$. *Let* $g_N(k,k') = h_N\big(\overline{T}^{-1}(k)/m, \overline{T}^{-1}(k')/m\big)$. *Then, the following approximation holds*

$$\frac{1}{N(I_1)N(I_2))}\sum_{k=c_1}^{c_1'-1}\sum_{k'=c_2}^{c_2'-1} g_N(k,k') = 1 + \frac{6Ob(1)K_0}{\inf_s[N(I_s)]} \ .$$

Proof Let k^n, $n = 1, 2, .., c_1'-c_1$, and h^n, $n = 1, 2, .., c_2'-c_2$, be two permutations of $I_1 \cap F(m) = \{c_1/m, (c_1+1)/m, ..., (c_1'-1)/m\}$ and $I_2 \cap F(m) = \{c_2/m, (c_2+1)/m, ..., (c_2'-1)/m\}$, respectively such that $\overline{T}^{-1}(k^1) < \overline{T}^{-1}(k^2) < \overline{T}^{-1}(k^3) < ..... < \overline{T}^{-1}(k^{c_1'-c_1})$ and $\overline{T}^{-1}(h^1) < \overline{T}^{-1}(h^2) < \overline{T}^{-1}(h^3) < ..... < \overline{T}^{-1}(h^{c_2'-c_2})$. Then, for all numerical simulations which we executed, one has always obtained

$$|\overline{T}^{-1}(k^r)/m - r/N(I_1)| \leq 4/N(I_1)$$

and therefore

$$|\overline{T}^{-1}(h^r)/m - r/N(I_2)| \leq 4/N(I_2) \ .$$

We deduce that

$$|g_N(k^r, h^{r'}) - h_N\big(r/N(I_1), r'_N(I_2)\big)| \leq 4K_0.Max_s\Big(\frac{1}{N(I_s)}\Big) = \frac{4K_0}{Inf_s[N(I_s)]} \ .$$

Therefore, by lemma A.0.3,

$$\frac{1}{N(I_1)N(I_2)}\sum_{k=c_1}^{c_1'-1}\sum_{k'=c_2}^{c_2'-1} g_N(k,k') = \frac{1}{N(I_1)N(I_2)}\sum_{r,r'} g_N(k^r, h^{r'})$$

$$= \frac{1}{N(I_1)N(I_2)}\sum_{r,r'} h_N\big(r/N(I_1), r'/N(I_2)\big) + \frac{1}{N(I_1)N(I_2)}\sum_{r,r'}\Big[g_N(k^r, h^{r'}) - h_N\big(r/N(I_1), r'/N(I_2)\big)\Big]$$

$$= \frac{1}{N(I_1)N(I_2)}\sum_{r,r'} h_N\big(r/N(I_1), r'/N(I_2)\big) + \frac{4Ob(1)K_0}{\inf_s[N(I_s)]}$$

$$= 1 + \frac{2Ob(1)K_0}{\inf_s[N(I_s)]} + \frac{4Ob(1)K_0}{\inf_s[N(I_s)]} \ . \ ■$$

Proof of property A.0.1 By the previous equalities,

$$P\big\{\big(\overline{T}(mY_1)/m, \overline{T}(mY_2)/m\big) \in I_1 \otimes I_2\big\} = \frac{1}{m^2}\sum_{k,k'} g_N(k,k')$$

$$= \frac{N(I_1)N(I_2)}{m^2}\Big[1+\frac{6Ob(1)K_0}{\inf_s[N(I_s)]}\Big] = L(I_1)L(I_2)\Big[1+\frac{Ob(1)}{m}\Big]\Big[1+\frac{Ob(1)}{m}\Big]\Big[1+\frac{6Ob(1)K_0}{\inf_s[N(I_s)]}\Big]$$

$$= L(I_1)L(I_2)\Big[1+\frac{O(1)K_0}{Inf_s[N(I_s)]}\Big] \ . \ \blacksquare$$

Bibliography

[1] KNUTH D.E. (1998) the Art of Computer Programming; Vol 2. Third Edition Addison-Wesley, Reading, Massachusetts.

[2] GENTLE J. (1984) Random Number Generation and Monte Carlo Method, Springer 13, 61-81.

[3] MENEZES A., VAN OORSCHOT P. , VANSTONE S. (1996) Handbook of Applied Cryptography, CRC Press, 1996.

[4] VON NEUMANN J . (1951) Various techniques used in connection with random digits. Monte Carlo method, Applied Mathematics series N12, US National Bureau of Standards, Washington DC, 36-38.

[5] SCHNEIER B (1996) Applied Cryptography 2nd Edition, John Wiley and sons, Inc

[6] ROSENBLATT M. (1972) Uniform ergodicity and strong mixing. Z. Wahrsch. Werw. Gebiete. 24, 79-84.

[7] PINSKER M.S. (1964) Information and information stability of random variables and processes. Holden Day, San Francisco.

[8] ELIAS P. (1972) The efficient construction of an unbiased random sequence. Annals Math Stat, vol 43, n3, 865-870.

[9] LANCASTER H. O. (1960) Orthogonal models for contingency tables. Developments in statistics. Academic Press, New York.

[10] BLACHER R. (1993) Higher Order Correlation Coefficients. Statistics 25, 1-15.

[11] BLACHER R. (1995) Central limit theorem by polynomial dependence coefficients. Journal of computational and applied mathematics 57, 45-56.

[12] BLACHER R. (1990) Theoreme de la limite centrale par les moments. Compte rendus de l'Academie des Sciences de Paris. t-311 serie I, p 465-468.

[13] BLACHER R. (1983) Quelques propriétés des congruences linéaires considérées comme générateur de nombres pseudo-aléatoires. Rapport de recherche n 345 IMAG, Université Joseph Fourier de Grenoble.

[14] BLACHER R. (2007) Central Limit Theorem by moments. Statistics and Probability Letters, 2007; 77 (17) 1647-1651

[15] BLACHER R. (2007) Une nouvelle condition d'independance pour le theoreme de la limite centrale http://hal.archives-ouvertes.fr/hal-00144878/en/ HAL: hal-00144878, version 1

[16] BLACHER R. (2002) Transformation d'une suite aléatoire q-dépendante. Rapport de Recherche LMC-IMAG RR 1054-M, Universite de Grenoble.

[17] BLACHER R. (1988) A new form for the chi-squared test of independence. Statistics 19,4, 519-536

[18] BLACHER R. (2009) A Perfect Random Number Generator. Rapport de Recherche LJK Universite de Grenoble. http://hal.archives-ouvertes.fr/hal-00426555/fr/

[19] GNEDENKO B.V., KOLMOGOROV A.N. (1968) Limit distributions for sums of independent random variables, Addison Wesley Publishing Co, Reading, Massachussets, London.

[20] MARSAGLIA G (1995) CD ROM. Florida State University, site internet http://stat.fsu.edu/pub/diehard/

[21] IBRAGIMOV I.A. LINNIK Yu. V.(1971) Independent and stationary sequences of random variables. Wolters-Noordhoff, Groningen.

[22] BRADLEY R.C. (1984) On a very weak Bernouilli condition. Stochastics, 13, 61-81.

[23] DEHLING H. DENKER M. PHILLIPPS W. (1984) Versik Processes and very weak Bernouilli processes with summable rates are independent. Proc. Amer. Math Soc. 91, 618-624.

[24] WITHERS C. S. (1981) Central limit theorems for dependent random variables. I. Z. Wahrsch. verw. Gebiete, 54, 509-534.

[25] COGBURN R. (1960) Asymptotic properties of stationary sequences. Univ. Calif. Publ. Statist. 3, 99-146.

[26] ROSENBLATT M. (1972) Uniform ergodicity and strong mixing. Z. Wahrsch. Werw. Gebiete. 24, 79-84.

[27] DOUKHAN P, LOUHICHI S. (1999) A new weak dependence condition and application to moments inequalities. Stochastics Processes and their Applications. (84) 313-342.

[28] HALL P. , HEYDE C.C. (1980) Martingale Limit Theory and Its Application, Academic Press, London

[29] SERFLING R.J. (1980) Approximation theorems of mathematical statistics. Wiley, New York

[30] JOHNSON N.L. KOTZ S. (1969) Discrete distributions. Wiley, New York

[31] JOHNSON N.L. KOTZ S. (1970) Continuous univariate distributions. Wiley, New York

[32] SANTHA M. , VAZIRANI U. V. (1984). Generating quasi-random sequences from slightly-random sources. Proceedings of the 25th IEEE Symposium on Foundations of Computer Science: pages 434440, University of California. ISBN 0-8186-0591-X.

[33] SANTHA M. , VAZIRANI U. V. (1986). Generating quasi-random sequences from semi-random sources, Journal of Computer and System Sciences, v.33 n.1, p.75-87.