Empirical Testing of Pseudorandom Number Generators

Diplomarbeit
zur Erlangung des Magistergrades
an der Naturwissenschaftlichen Fakultät
der Universität Salzburg

eingereicht von
Stefan Wegenkittl

March 12, 1996

Contents

Deshalb dürfte man meines Erachtens überflüssiger Weise noch nach den Ursachen des freien Falls nach der Mitte forschen, nachdem ein für allemal auf dem dargelegten Wege aus den Erscheinungen selbst die Tatsache klargestellt ist, daß die Erde den Raum in der Mitte des Weltalls einnimmt, und daß alle schweren Körper auf sie fallen.

– Ptolemäus, Almagest

An dieser Stelle möchte ich mich bei all den Menschen bedanken, mit denen ich die letzten Jahre zusammensein und arbeiten durfte, und durch die ich jene Neugierde gelernt habe, welche die Mathematik erst zum Leben erweckt: Peter, Charly, Hannes, Otmar, Hrn. Thaler, Hrn. Österreicher, Hrn. Czermak, Doris und Roland.

Chapter 1

Introduction

This master's thesis is about the empirical testing of pseudorandom number generators (PRNGs). We will examine the role of empirical testing of pseudorandom numbers and give an example of a more flexible test statistic which can be used to stress the importance of inversive generation methods like ICGs and EICGs in the field of (parallel) stochastic simulation.

In order to address the user of PRNGs as well as the mathematically interested reader, we have tried to separate long mathematical proofs from the elaboration of the basic ideas.

Mathematical theory offers an arsenal of new tools (e.g. *inversive generators*) and tests (e.g. the *overlapping M–tuple test*) in order to increase confidence in stochastic simulations. However, the results are not incorporated sufficiently in everyday stochastic simulation. In our opinion, this *gap between application and theory* results from a permanent misunderstanding between the user and the mathematician.

This master's thesis is a contribution to reduce this gap. It contains information about results offered by the mathematician to the user, but also brings up questions that naturally arise in most applications of PRNGs. We will try to explain the terms that are employed in most of the common literature in the field. The final answer to the 'simple' problem of choosing a 'good' PRNG for a certain stochastic simulation cannot be given but it can be clarified what has to be done in order to get a proper answer:

> Anyone who applies PRNGs in computer simulation should carefully test the used generators. He should use *strongly different generation methods*, in particular inversive generators, to verify the simulation results!

1.1 Outline and summary

The central ideas and results of this thesis are summarized in Section 1.7, which examines the definition of random numbers, in Section 2.2, which concerns statistical testing, and in Section 3.3, where we discuss the practice of empirical testing. Section 4.4 summarizes the empirical results.

When performing a stochastic simulation one faces the following problem: given a stochastic model $T(X_0, X_1, \ldots, X_{N-1}) = T(\mathbf{X})$ depending on a random vector $\mathbf{X}$ for whose components a certain common distribution is assumed, compute as efficiently as possible some numerical properties ξ of the resulting distribution function F_T of T. For example, consider the expectation $E(T)$ or variance $V(T)$. Section 1.3 in Chapter 1 will examine the notion of probability which lies at the heart of this kind of stochastic modelling.

We measure properties by unbiased estimators $\hat{\xi}$ which are random variables defined on a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_k$ of $\mathbf{X}$. The expectation $E(\hat{\xi})$ equals ξ by definition.

If direct mathematical analysis of $\hat{\xi}$ fails to yield estimates for its expectation, one is left to stochastic simulation. Replacing the random vectors $\mathbf{X}_n$ by realizations $\mathbf{x}_n \in \mathbf{R}^N$ we get a numerical value

$$\hat{\xi} = \hat{\xi}(\mathbf{x}_1, \ldots, \mathbf{x}_k),$$

the result of the simulation run, that is used as approximation to ξ.

The realization of a random variable is called *random number*. If we want to stress the fact that the random number has been produced in order to yield desired results within an actual application, i.e. the approximation of the expectation of $\hat{\xi}$, we will use the term *pseudo*random number. Section 1.4 in Chapter 1 examines the notion of a random number and compares it to that of a random variable.

The theory of generating sequences of PRNs (x_i) that can be viewed as realization of iud[1] distributed random variables (X_i) includes as two main parts:

1. The *generation of random numbers according to some algorithm.* A vast amount of so–called (Pseudo–) Random Number Generators (PRNGs) has been proposed for this task. A PRNG is supposed to be a general purpose device[2] for stochastic simulation. We will take a closer look at some popular generation methods and their structural properties in Sections 1.5 and 1.6 of Chapter 1. However, from the mathematical point of view it is clear that no algorithm can lay claim on producing small error *for all possible models* T. Given an arbitrary generator, it is always possible to construct a model that will lead to unsatisfactory results. Thus we will need a second step in order to gain the desired *pseudo*–random numbers:

[1] i.e. independent uniformly distributed

[2] If the domain of application is known one might wish to use more specialized generation methods like (t,m,s)–nets. For the latter notion, see Niederreiter [38].

2. The *testing of the numbers that have been provided by a generator.* The aim of testing is the assertion of little error in the simulation of some special test functions $\hat{\xi}_{T_1}, \hat{\xi}_{T_2}, \ldots, \hat{\xi}_{T_l}$. The statistical concept behind such tests is examined in Chapter 2. We will introduce theoretical and empirical tests and compare their relevance. An overall framework for all such testing will be developed. From this framework we will see that we need very flexible test statistics in order to increase the confidence in a stochastic simulation. We will also see that the mathematical theory is by no means complete within this field.

Chapter 3 contains a general approach to tests for pseudorandom numbers that are used as realizations of uniformly distributed random variables. We will introduce a test that has been proposed by Marsaglia in [31], the so–called *overlapping M–tuple test*, and comment on the design and practice of such tests. The important question of how to interpret results of any statistical test will be treated in detail.

Chapter 4 contains empirical data that have been calculated within the pLab group at the University of Salzburg. We will show that *inversive generation methods are an important and even indispensable contribution to stochastic simulation.*

Chapter 5 contains a proof for the distribution of the test statistic introduced in Chapter 3. Some concepts of probability theory are extended to the multivariate case. In particular, we will consider convergence in distribution, the mapping theorem and central limit theorem in dimensions $s > 1$. Some algebraic tools will be needed in order to transform the multidimensional test statistic into a onedimensional suited for statistical inference. The appendix contains related material.

1.2 Notations and concepts

We will denote random variables (RVs) by capital letters and realizations of RVs by lowercase letters, e.g. x is a realization of X. Vectors will be denoted in boldface in order to distinguish them from scalars. $\mathbf{X}^{(2)}$ denotes the second component of the random vector $\mathbf{X}$. $\emptyset$ is the null vector in $\mathbf{R}^s$. All vectors are column vectors, e.g.

$$\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(s)})^T.$$

We assume that the reader is familiar with the construction of a probability space $(\Omega, \mathcal{A}, P)$ which supports an infinite sequence of independent uniformly distributed random variables $(X_i)_{i \in \mathbf{N}}$. We will further use the following abbreviations:

- $X \sim U([0,1[)$ denotes a random variable distributed uniformly on the half open interval $[0,1[$. The distribution function F_X is

$$F_X(t) := \begin{cases} 0 & : \quad t < 0 \\ t & : \quad 0 \le t \le 1 \\ 1 & : \quad 1 < t \end{cases}$$

- $\mathbf{X} \sim U([0,1[^s)$ indicates an s–dimensional uniform distribution, thus

$$F_{\mathbf{X}}(\mathbf{t}) := \prod_{i=1}^{s} F_X(\mathbf{t}^{(i)})$$

- $Y \sim \mathcal{N}(\mu, \sigma^2)$ is a normal variate with expectation $E(Y) = \mu$ and variance $V(Y) = \sigma^2$. We will examine this distribution in detail in Chapter 5.

- $Z \sim \chi^2_n$ denotes a random variable distributed χ^2 with n degrees of freedom. Its probability density is given by

$$f_Z(t) := \frac{1}{2^{n/2}\Gamma(n/2)} t^{(n/2)-1} e^{-t/2} \quad t > 0$$

  It is well known that the sum of the squares of n independent and identically distributed standard normal variables $Y_1, \ldots, Y_n$, $Y_i \sim \mathcal{N}(0,1)$, is distributed χ^2_n.

- Put $Z = \sum_{i=1}^{n} (Y_i - v_i)^2$ for n constants $v_i \in \mathbf{R}$, then Z is distributed noncentral χ^2 with n degrees of freedom and noncentrality parameter $\delta = \sum v_i^2$. We will denote this distribution by $\chi^2_{(n,\delta)}$. See Chapter 5 for details.

We will denote convergence in distribution by $X_n \Longrightarrow X$. The notation $\hat{F}$ is used for either an estimator for a parameter of a distribution or for the characteristic function. Thus $\hat{F}_X$ denotes the characteristic function of the random variable X and $\hat{\xi}$ denotes an estimator for the parameter ξ. The correct interpretation will always be clear from the context.

We define the function $\delta(\cdot)$ by

$$\delta(\text{relation}) := \begin{cases} 0 & : \text{ relation is not fulfilled} \\ 1 & : \text{ relation is fulfilled} \end{cases}$$

We will call $\delta(\cdot)$ an indicator although it differs from the usual definition of the indicator function of a set. The usual delta function $\delta(x, y)$, for example, can be expressed by $\delta(x = y)$.

We will use $diag(V)$ to denote a matrix formed of the diagonal elements of the matrix V and zero else. $I(s)$ denotes the s–dimensional unity matrix, i.e. $V \cdot V^{-1} = I(dim(V))$, if V is invertible. $dim(V)$ is the dimension of V.

1.3 Random Variables

We will start our study by introducing the concept of random variables (RVs). The mathematical definition of a random variable X is the following:

Definition 1.1 *(Random Variable) A random variable X is a $(\mathcal{A}$ – $\mathcal{B}_{\mathbf{R}})$–measurable function from a probability space $(\Omega, \mathcal{A}, P)$ to* $\mathbf{R}$. $\mathcal{B}_{\mathbf{R}}$ *denotes the standard Borel sigma field on* $\mathbf{R}$.

Thus a random variable is a function assigning a real number to each element ω of the sample space Ω. In that, a random variables serves for measuring desired properties of the elements in Ω on a very general scale. A simple example of the usage of this concept would be putting Ω equal to the set of all possible courses in a betting game and denoting by $X(\omega)$ the winnings (or losses) caused by such a game.

In some cases, we might wish to replace $\mathbf{R}$ by a more abstract set of symbols that serves as sample space for another random variable. Consider for example the field of information theory where typical sources for streams of symbols can be modelled in such a way. We will meet this generalization of random variables in the construction of a test for PRNGs in Chapter 3.

1.3.1 Probability

Up to now the notion of a RV does not involve randomness. We will have to develop an intuitive understanding of the term "probability" before we can go on increasing our knowledge on RVs.

Let us examine the use of the word "probability" in everyday language: we say, that a certain event will occur with a certain (usual only high or low) probability if we do not manage to predict its occurrence in a deterministic way. This inability can be due to a lack of information about either the state of the conditions necessary for, or the mechanisms involved in the occurance of the event.

Even when we know that a system has a totally deterministic behavior and we can measure ingoing parameters very accurately, 'random' behavior is possible due to the inevitable error of measurement. Such systems are commonly called *chaotic* and have the property that even small changes of the input parameters can lead to completely different behavior of the system after some simulation steps. In a notion defined by the theory of ergodicity the system 'forgets information' very quickly[3]. If the deterministic mechanism of such a system cannot be found, a description in terms of probability, e.g. the stable distribution, will provide a quite good formalization of the knowledge on the system.

The metaphysical question whether *chance exists or not* will not be treated in this thesis[4]. That is, we will not make any further assumptions on the cause of randomness. We thus cannot rely on a model for the emergence of probability in order to measure and interpret probability in real world situations. This leaves us with the following interpretation of a probability assigned to an event:

[3] By information we mean the ability to distinguish distributions.

[4] This question would involve physics (e.g. quantum mechanics) as well as philosophic discussions.

Definition 1.2 *The probability assigned to an event expresses the expected*[5] *average rate of occurrences of the event in an unlinked*[6] *sequence of experiments.*

This definition is the relation between the mathematical model for probability, that is real numbers in $[0, 1]$, and the real world. On the one hand we consider the agreement to this definition as being essential for the understanding of some of the material in this thesis. On the other hand the definition is by no means perfect. There exist many other ways to define probability in a way that expresses a relation between a mathematical model and real world phenomena. One might even have to change the mathematical formalization in order to get useful interpretations of a probability. For a good introduction see Fine [13], who discusses –among others– axiomatic comparative probability, axiomatic quantitative probability, computational complexity, logical probability and subjective probability. However, the definition given above has some advantages: it expresses a very common interpretation of probability that arises from dealing with everyday situations like games and errors of measurement. It involves the argument of the long run and is compatible with the experience that the relative frequency seems to stabilize when the number of trials is increased.

The most important advantage may be the target of the relation: the Kolmogorov axioms for probability and the whole theory build upon this formalization of probability. With this theory we are able to express the *laws of chance*. We can calculate new probabilities out of given ones, we can study the increase (Baysian approach) or decrease (ergodic theory and chaotic systems) of information, we can construct models explaining the observed behavior of averaged sums of random values (weak and strong laws of large numbers and central limit theorems).

The theory itself does *not include a possibility to refute or accept assumptions* about a real world experiment. Whenever we want to introduce new probabilities in a mathematical model or interpret derived probabilities we have to rely on the aforementioned definition of the relation between model and reality.

As a consequence new probabilities are introduced into a model by either making strong assumptions like assigning the same probability to all basic events, or by using estimators[7] for the parameters in the models. The setup provides no possibility for verifying such probabilities. This is one main drawback of the axiomatic theory of probability.

Consider for example traditional mechanics. Nobody has ever seen the real world equivalent of the F in Newton's formula $F = m \cdot a$. By measuring the mass and the acceleration of a particle we *are able to* calculate the magnitude of F however. If we predict the behavior of the system by using F, we can compare the results to, say, the actual position

[5] in the sense of 'agreement by some people' or 'made assumption'

[6] We call a sequence of experiments *unlinked*, if the single experiments are supposed to not influence each other. See [13, p.86] for a thorough discussion. In Section 1.3.3 we will discuss two formalizations in terms of random variables.

[7] Estimators again express the frequentists' interpretation of the meaning of probability since they are usually built on relative frequency.

of the particle at a fixed time. If the prediction is wrong, we either have to refine the model – that is the relation $F = m \cdot a$ – or we have to measure F in a more accurate way.

This is not possible for the 'measurement'[8] of a probability p. Given the Kolmogorov setup we only can calculate a confidence interval for p. There is no way to refute or accept p by predictions of the model since such predictions will only occur with a certain probability.

The two levels, mathematical theory and 'real world', should never be mixed up. Mathematical analysis can only transform given probabilities. Within reality we *can only assume* that probabilities with the according laws of change exist. No deterministic method is available for a precise measurement. The language used at the mathematical level often suggests that such tools are available, the typical statements being 'almost sure', or 'for almost every ω'. The interpretation given by Relation 1.2 is that we expect such an event to happen. That the event will happen or not cannot be expressed by probabilities. Even more, typical 'almost sure' arguments will involve the measurement of an infinite sequence of trials. The lack of relevance of such models has also been admitted by Kolmogorov himself[9]:

> "The frequency concept based on the notion of limiting frequency as the number of trials increased to infinity, does not contribute anything to substantiate the applicability of the results of probability theory to real practical problems where we have always to deal with a finite number of trials. "

The notion of randomness is modeled by the combination of Definition 1.2 and the mathematical theory of probability. We do not associate randomness with events or sequences of events but solely associate it with an unknown selection mechanism that selects one ω out of Ω, the set of all possible events (respectively sequences of events), whenever we perform a single experiment (respectively a series of experiments). In this sense, a sequence consisting of nothing but zeros

$$0, 0, 0, 0, 0, \ldots$$

is as random as any other that can be selected by the mechanism. Of course there are other, different approaches to formalize randomness. For example we can interpret randomness as *unpredictability*. This approach is based on complexity theory and defines randomness in the world of events[10]. It will thus exclude the above sequence from the set of random sequences since it is predictable with very little effort. We will see that any interpretation of randomness is important for building models of real world phenomena. However, within the Kolmogorov axioms, randomness is not associated with single sequences and will not be a suitable criterion for selecting some sequences in favor of others: such a selection can only be made with respect to an actual application.

[8] e.g. estimation

[9] A. Kolmogorov, On Tables of Random Numbers, *Sankhya Ser. A*, p.369, 1963, quoted from [13, p.94]

[10] We refer the reader to [32] for an introduction to complexity theory and to [5] for the notion of randomness with respect to finite strings.

1.3.2 The distribution of a random variable

Using Definition 1.2 we are now able to clarify the meaning of random variables. Let us start with the interpretation of the probability space on which a random variable lives. The sigma field $\mathcal{A}$ is a class of sets denoting the single and combined events that we want to distinguish. The sample space Ω sums up all this events[11]. The probability measure P expresses the information about the expected average occurrence of the events in the sigma field.

We introduce the notion of randomness with the help of this probability measure. The probability of each event in the sigma field now can be interpreted using Definition 1.2. As has already been pointed out we can only define such probabilities by either making strong assumptions or by carrying out a sequence of experiments in order to estimate a probability in the form of a relative frequency.

Once the probability space has been defined, a random variable can be used to express the flow of information induced by the transformation: the random variable assigns a certain real number to each basic event, i.e. each element of the sample space. By this and by the measurability of the RV it induces a probability for each 'natural', that is $\mathcal{B}_{\mathbf{R}}$ measurable, set of real values. The information contained in the measure P is transformed to information on the expected average occurrence of numbers in such a set of real numbers. Therefore this set represents a property of the events which is measured on the scale of real numbers. If we take another look at the example given in the last section we can now calculate the probability for the set of real, positive numbers, that is, the probability for 'win'. We will get a number in $[0, 1]$ denoting the expected average occurrence of 'win' in a sequence of unlinked games which should be close to $1/2$ if the game is fair. We will *not* get *any* information if the next, say, 1000 games will lead to even a single 'win'. This is the dilemma with probability theory: the theory can describe laws of chance and the fluctuation of information. It cannot give information about concrete realizations. We will return to this dilemma later in the thesis.

> Thus a random variable is a description of an experiment including assumptions on the probabilities of all events that we are interested in. We can use this device whenever we want to describe information at the level of the *whole* setup of an experiment. It is no use applying random variables if we want to judge on the occurrence of single outcomes. In particular, we cannot 'calculate' the distribution of a random variable from a given sequence $\mathbf{x} = x_0, \ldots, x_{N-1}$ of realizations.

Different solutions to this problem have been proposed. One is to calculate the empirical distribution function of the events in the sequence $\mathbf{x}$ and to use this as distribution function for the random variable describing the experiment. However, if we repeat this procedure for another sequence of realizations we will almost surely get another empirical distribution function. The theorem of Glivenko–Cantelli (Theorem 20.6. in

[11] The theory can also be build without explicitly referring to a sample space. The important object is always the sigma field because it represents the events of interest. In this thesis we will always assume that a sample space is given and that the random variable is defined for all elements in the sample space.

[3]) tells us that if we describe the single experiments by a random variable X and assume certain regularity conditions, especially the independence of the realizations, the empirical distribution will tend to become the distribution of the random variable X in almost all[12] cases if the number of realizations is made arbitrarily large. The setup is consistent in the limit at least. We will examine the notion of empirical distribution function more closely in Section 1.4.2.

Yet, three questions remain unsolved:

1. Who tells us that the actual experiment results from experiments that can be described accurately by *independent* copies of X?
2. Can we dedect the distance between the empirical and the theoretical distribution function after *some* N *realizations* have been observed, and what measure should be used to do so?
3. What criterion assures us that we do not live on a set of measure zero, that is, we actually get a sequence which falls in the region 'almost surely not' of the Glivenko–Cantelli theorem?

Any other solution to the problem of algorithmically finding a model describing the observed data of an experiment will lead to similar problems which are summed up in the following paragraph.

> The realizations of a random variable and the random variable itself are held together only by the range of the observations and the image of the random variable. The Kolmogorov setup provides no method of deterministically relating a specific random variable to a specific sequence of realizations. This is due to the fact that the same sequence of realizations is within the range of many random variables. We can calculate probabilities for the occurrence of a sequence given a model X and maximizing this probability within a family of models $X_i : i \in I$. This amounts to seeking for a 'good explanation' for the sequence, that is, a mathematical model which makes us expect the sequence in the sense of our Definition 1.2. Such probabilities themselves represent *assumptions* on the models X_i.

Both, the number of elements in the sample space and the number of models, seem to be too large to lead to logical[13] statements. The sample space of a vector of unlinked uniformly distributed random variables contains the simple sequence

$$0, 1, 0, 1, 0, 1, 0, 1, \ldots$$

as well as sequences that look much more random. However, is it really necessary to use probability theory to model such a simple behavior? We will discuss this idea in Chapter 2.

[12] with respect to the according probability space

[13] i.e. statements within the axiomatic setup that do not depend on the subjective relation to reality that we have defined in Definition 1.2.

All in all, the use of probability theory is in the calculation of 'laws of change' that transform probabilities, i.e. assumptions made about experiments. Randomness is assigned to a selection mechanism that chooses events with an expected average occurrence, the probability of the event.

1.3.3 Uniformity and independence

With the aforementioned concept of a random variable we are capable of modeling many real world phenomena. Nethertheless it would be practical if we were able to limit our attention to something like a 'most general random variable', a source of randomness from which every other random variable can be deduced. This is possible in a certain sense since a RV with any desired distribution on $\mathbf{R}$ can be achieved by transforming a random variable distributed uniformly on $[0,1[$.

We show the construction only for random variables with distribution functions that are continuous on $\mathbf{R}$, strictly monotone in some interval [a,b] and that assume the values 0 for arguments below a and 1 for arguments above b. Let $X \sim U([0,1[)$ be a random variable distributed uniformly on $[0,1[$. The distribution function of X thus is

$$F_X(x) := \begin{cases} 0 & : \quad x < 0 \\ x & : \quad 0 \leq x \leq 1 \\ 1 & : \quad x > 1 \end{cases}$$

Now, given any distribution function F compatible with the above assumptions, we can construct the corresponding random variable Y by setting

$$Y(\omega) := F^{-1}(X(\omega))$$

The distribution function of Y for $x \in [a,b]$ calculates to

$$\begin{aligned} F_Y(x) &= P(Y \leq x) \\ &= P(F^{-1}(X) \leq x) \\ &= P(X \leq F(x)) \\ &= F_X(F(x)) \\ &= F(x) \end{aligned}$$

For other x, we trivially get $F_Y(x) = 0$ for $x < a$ and $F_Y(x) = 1$ for $x > b$.

However, most stochastic models contain not only one random variable but a sequence of unlinked random variables. We now have to define the meaning of the term unlinkedness precisely. The idea behind this concept is that we repeat an experiment in such a way that the single trials do not influence each other. As long as we want to stay within the Kolmogorov calculus, we have to use the common distribution function of the set of random variables describing the experiments in order to express this property.

Every concept of unlinkedness has to meet two natural requirements. It should provide laws of large numbers explaining the observed behavior of long runs. And it should be

easy to assert that a given series of experiments is unlinked in the given sense. Let us examine two popular formalizations of the notion of unlinkedness.

The first one is called 'exchangeability'. We call a sequence of identically distributed random variables 'exchangeable' if the common distribution function does not depend on the order of the random variables, that is, if the common distribution function is symmetric in its arguments. This formalization goes back to de Finetti [7, p.95–158]. It is strong enough to provide laws of large numbers, but the limit turns out to be a random variable itself, which will not necessarily degenerate at a single value. For details see Fine [13, p.95].

The second formalization is 'independence': we call a sequence of identical experiments independent if the common distribution function can be written as the product of the distribution functions of the single random variables. This is the usual starting point for laws of large numbers[14].

Independence implies exchangeability. It seems more difficult to assess independence for a given experiment, however. In our opinion unlinkedness is a property of experiments which has different formulations in terms of random variables. These cannot be assessed without referring to a certain application. If we stress the flow of information, independence seems to be a proper criterion. Consider the case of two independent events A and B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \overset{(A,B \text{ indep.})}{=} \frac{P(A)P(B)}{P(B)} = P(A),$$

thus the information about - i.e. the probability of - event A does not change if we know that B has happened. The theory of martingales extends this viewpoint to a larger class of sequences of random variables emphasizing the notion of *fairness*.

On the other hand exchangeability is a weaker criterion and therefore can be defended easier. It expresses stability of a large class of functions of random variables, namely the class of weighted sums, under transformations that change the order of the experiments. This property seems to be very natural for experiments that do not influence each other.

However, in the whole field of stochastic simulation we make use of the first characterization of unlinkedness as a matter of fact. We therefore arrive at the following 'most general case' of random variables, which is used to model unknown parameters that are supposed to be unlinked and for which no further information is available:

Definition 1.3 *A sequence of independent random variables distributed uniformly on* $[0,1[$ *is an infinite vector* $(X_0, X_1, X_2, \ldots)$ *of random variables*

$$X_i \sim U([0,1[),$$

such that for every finite set of indices i_s*, the common distribution of the random vector* $(X_{i_1}, X_{i_2}, \ldots, X_{i_s})$ *equals the product of the distribution functions of the components,*

[14] However, the class of sequences of random variables leading to laws of large numbers is much larger. Consider for example ergodic transformations that can lead to dependent random variables, but still suffice to prove laws of large numbers.

that is

$$(X_{i_1}, X_{i_2}, \ldots, X_{i_s}) \sim U\left([0,1[^s\right).$$

We will call such a sequence iud (independent uniformly distributed).

The aspect of 'no further information' is modeled by using uniform distributions whereas the 'unlinkedness' is modeled by independence which is equivalent to claim factorization of the common distribution function into the distribution functions of each component. The existence of such an infinite vector is guaranteed by the consistency theorem of Kolmogorov. A sequence of independent s–dimensional random vectors can be formed by using nonoverlapping s–tuples of the form

$$\mathbf{X}_n := (X_{n\cdot s}, X_{n\cdot s+1}, \ldots, X_{n\cdot s+s-1}).$$

Such a vector is the starting point in almost all stochastic simulations. This definition completes our excursion into the field of random variables.

1.4 Random numbers

We introduce the notion of random numbers in order to denote realizations of a random variable, that is, possible outcomes of an experiment described by a random variable.

Definition 1.4 *(Random numbers) If X is a random variable*

$$X : \Omega \to \mathbf{R},$$

we call realizations

$$x_i \in X(\Omega),$$

where $i \in \{0, 1, \ldots\}$, random numbers.

Example 1.1 *Let X be uniformly distributed on $[0, 1[$, then*

$$0.1,\ 0.76,\ \frac{\pi}{4},\ 0,\ 0.1234567891011$$

is a finite sequence of random numbers for X.

A random number emerges whenever we perform an experiment that yields a real number. In some way random numbers are much more evident and natural than random variables. According to this we should have started with random numbers in order to define random variables: a random variable describes the possible outcomes of an experiment yielding real numbers. The probability it assigns to a random number expresses an information about the expected average number of occurrences of the random number in an unlinked sequence of performances of an experiment.

We thus can make statements like: 'Under the assumption of model X, the random number x has probability p; model X expresses that we expect the relative frequency of x being p'. We cannot make statements like: 'Random number x is a better realization of X than random number y'. *There are no 'good' or 'bad' random numbers.* As long as a random number is within the range of a random variable X, X is a valid model for x and x is a valid realization of X. This statement cannot be made more precise. Statistics seems to master this problem but we will see that this is simply a wrong impression.

Even in the case that x has probability zero and y has probability one with respect to model X we cannot say that y is preferable to x. *All random numbers are equal!* The same is true for sequences of random numbers, since such sequences are simply realizations of a random vector describing an experiment yielding a vector of real numbers.

As pointed out in the last section, the relation between the information expressed by random variables and random numbers can be used to motivate choosing the random variable describing a sequence of unlinked experiments to have the empirical distribution of that sequence. *But this is nevertheless an* assumption *about the experiment.* Any other distribution whose image contains the single results of the experiments is another valid description of the experiment. Putting it the other way round, every sequence of random numbers is a valid realization of a random variable X as long as the range of values is within the image of X. Any further judgement cannot be defended without referring to an application. We will use the term "numerical property" for such characterizations in order to distinguish them from statistical properties that are associated with distributions. The development of useful numerical properties is one of the main tasks in the field of PRN generation!

We would like to mark random numbers with a big red plate[15]:

ATTENTION: Random number.
Avoid contact with distributions, probabilities and random variables!

In comparison to random variables we use random numbers in any situation where we would like to get single outcomes of a stochastic experiment instead of a description of the expected average behavior. For example, imagine a roulette game. Nobody would go to the casino if a roulette wheel would be placed behind bullet-proof glass and some spectacle wearing mathematician would ask you for the number of bets you want to play, make some little calculations and charge the expected loss.

Physical devices like the roulette wheel have been the first generators for random numbers. They have one big advantage: somewhat hiding the intrinsic secret of the mechanism responsible for randomness[16], they have won wide acceptance as being the 'ultima ratio' for generating 'real random numbers'. This is surely related to the ignorance

[15] A short discussion on the importance of such red plates is given by Ian Steward in [44]
[16] By 'randomness' we denote the difficulty of predicting the outcomes of such devices.

of the reason that makes it 'impossible' to determine the next outcome of such a device. Modern mathematics attacks this lack of knowledge proposing new philosophical backgrounds for observed randomness clothed in words like 'ergodicity' and 'chaotic behavior'.

Unfortunately, today's analysts are by no means interested in playing games. Their big number crunchers simulate difficult real world phenomena containing many parameters for which they can only assess an expected asymptotic behavior in modelling them as random variables. The word 'many' stands for numbers of the form 'two to the power of l', where l ranges somewhat between 10 and 60. In carrying out such a stochastic simulation the parameters need to be replaced by random numbers. The appropriate physical device producing such amounts of random numbers is not called roulette wheel but radioactive decay or electronic noise. Many tests have been made to control the quality of random numbers produced by those devices. We refer the interested reader to Bratley [4] and L'Ecuyer [28] for further details. However, they do not seem to fulfil certain important criteria that would make them a reliable source of randomness:

- The aforementioned ignorance of the underlying processes prevents us from guaranteeing a–priory (see also Chapter 2, Section 2.2) any numerical qualities needed in the simulations.
- These physical devices are very sensitive to environmental influences, which makes it necessary to readjust the systems frequently.
- We will never be able to generate the same sequence of random numbers once again (of course: would we call them random otherwise?). This makes it difficult for the scientific community to assess results obtained by stochastic simulation. The only solution to this problem would be something like a CD-ROM or an Internet server storing and providing the random numbers which have been generated once by using such a physical device. See also next item.
- Inefficiency caused by slow (in comparison to the speed of computers) generation or read back from any storage device.

1.4.1 Pseudorandom numbers

To overcome these problems, the concept of *pseudo*random numbers has been proposed: simply substitute the physical device by a computer algorithm that generates numbers in a deterministic, fast and reproducible way. Since the mechanism generating pseudorandomness is now open to mathematical analysis, numerical qualities *can* be guaranteed in principle. These so–called theoretical tests are complemented by empirical tests that should assess qualities for which a mathematical analysis of the algorithm cannot be done (yet).

We now are able to give a first definition of what is understood by PRNs.

Definition 1.5 *(Pseudorandom Numbers) Pseudorandom numbers are deterministically generated rational random numbers having certain numerical properties that are relevant for the actual application where they are used in the place of realizations of random variables.*

The reader may compare this definition with that given by Ripley [43, p.15]. We have adopted the term 'numerical properties' instead of 'statistical properties' since we have already seen that any definition of pseudorandomness using the term 'distribution' or 'statistical' will cause conceptual difficulties in showing the desired properties for a specified sequence of PRNs. We can only calculate the probability assigned to the PRN by a certain model X. But this reflects a property of *both*, the model and the PRN, and cannot be considered a good definition for the PRN itself. Moreover, computer algorithms can only produce periodic sequences due to their finite state space. These sequences can theoretically be shown to be nonrandom in a certain sense[17].

We have to start in the limited world of algorithms in order to characterize the pseudorandom numbers in a way that is both mathematically precise and practically relevant. Forget the idea of real numbers, forget distributions, forget the notion of limits. And forget the rest of this thesis because it will illustrate what has been done since the first of these PRNs have been produced by an algorithm, deliberate of the conscience of the definition given above or not: showing that the PRNs have 'statistical' properties that give reason to treat this numbers as realizations of random variables with a given distribution.

We will try to introduce the lie on which both theoretical and empirical testing (when understood in the sense of giving recommendation for the use of certain PRN sequences) is built. Notice the usage of the terms random number and pseudorandom number in this thesis. One difference between PRNs and RNs is that in the former case we know the deterministic character of the numbers, whereas this is left open to philosophical discussion in the latter case. Since we have formulated the relation between probability and reality in terms excluding the relevance of such discussions, we will not have to distinguish between PRNs and RNs due to this criterion. What remains is the following: Firstly the resolution of the numbers and secondly the notion of 'relevance for an application'.

As to the first point, RNs are real numbers defined as limits of sequences of rational numbers. PRNs are rational numbers. In most cases the resolution of the PRNs will be fine enough to treat them as RNs. Attention has to be paid to the effects of discretizations. The according considerations *have to be made* by anyone who uses computer simulation. Since the discovery that chaos may emerge from the discretization of non–chaotic continuous processes[18] this warning should not be overlooked.

The second point is even more important: whenever we use the term PRN we stress

[17] See [32] for an introduction to complexity theory. When defining randomness as absence of information, periodic behavior clearly cannot be considered random!

[18] Consider the ergodic transformation $4x(1-x)$ which can be viewn as discretization of the continuous equation which describes the growth of a population of bacteria under some limiting conditions.

the relevance for a specific application. *Random numbers* are *innocent* whereas PRNs always can be judged with respect to an application. Roughly speaking *PRNs are random numbers yielding the desired results in an application.*

In order to understand the mentioned lie we have to examine the notion of the empirical distribution function.

1.4.2 Estimation techniques and the empirical distribution function

The empirical distribution function (EDF) of a finite sequence of realizations of a random variable lies at the heart of statistical inference. We will first introduce the concept of estimation and then define the EDF.

Consider a random variable X with distribution function F_X. We call $X_1, \ldots, X_k$ a random sample from F_X if, for all i, $F_{X_i} = F_X$ and the X_i are independent. Let ξ be a property of F_X, e.g. a parameter in the model describing X such as the expectation, the variance or a general statement like the probability that X will assume a value less than t, $P(X \leq t)$.

Definition 1.6 *We call the random variable*

$$\hat{\xi}(X_1, \ldots, X_k) \to \mathbf{R}$$

an unbiased estimator for ξ if

$$E\left(\hat{\xi}\right) = \xi.$$

Estimators are used to measure properties of distribution functions. In the field of statistical inference they are used to construct models from empirical data. We have already commented on the principal difficulties arising from such a setup: the estimations cannot be falsified. However, they are a very important tool for both practical and theoretical statistics. If an estimator is calculated from some k experiments yielding the random numbers $x_1, \ldots, x_k$, the resulting number is a random number itself, i.e. a realization of the random variable $\hat{\xi}$.

The above form of estimators is called 'point estimate' since it will yield a single real value. There also exist so-called 'interval estimates' that yield intervals containing the parameter ξ with a given probability. We will make use of an estimator for the whole distribution function F_X, however. Such an estimator needs to be a function itself. We therefore define the empirical distribution function.

Definition 1.7 *The empirical distribution function for a random sample $X_1, \ldots, X_k$ from F_X is the function*

$$\hat{F}_{X_1,\ldots,X_k}(t) := \frac{1}{k} \# \{i, 1 \leq i \leq k : X_i \leq t\}.$$

The strong law of large numbers (S.L.L.N.) immediately gives pointwise convergence if the random sample is drawn such that the X_i are independent:

$$\forall t \in \mathbf{R} : \lim_{k \to \infty} \hat{F}_{X_1,\ldots,X_k}(t) = F_X(t) \text{ a.s.}$$

The theorem of Glivenko–Cantelli (Theorem 20.6. in [3]) even proves a uniform convergence in t:

$$P\left(\lim_{k \to \infty} \left(\sup_{t \in \mathbf{R}} |\hat{F}_{X_1,\ldots,X_k}(t) - F_X(t)|\right) = 0\right) = 1$$

The empirical distribution thus serves as an estimator for the distribution function of a random variable. It can easily be extended to the multivariate case by defining

$$\hat{F}_{\mathbf{X}_1,\ldots,\mathbf{X}_k}(\mathbf{t}) := \frac{1}{k} \# \left\{i, 1 \leq i \leq k : \mathbf{X}_i \leq \mathbf{t}\right\},$$

where

$$\mathbf{X}_i \leq \mathbf{t} \Leftrightarrow \forall j, 1 \leq j \leq s : \mathbf{X}_i^{(j)} \leq \mathbf{t}^{(j)},$$

and s is the dimension of the vector $\mathbf{X}$.

Note that the *independence of the random variables X_i is essential* to get the desired limiting behavior of the EDF. It can easily be seen that the validity of the Glivenko–Cantelli theorem remains valid if we substitute any measurable function $g(X)$ for X and consider the distance of the EDF $\hat{F}_{g(X_1),\ldots,g(X_k)}$ and the distribution function $F_{g(X)}$. This property shows again the great difference between random variables and random numbers: if we fix a sequence $x_1, x_2, \ldots$ of random numbers, we could get the right limiting behavior, e.g. convergence, for one function g_1 and a wrong behavior for another function g_2! This is due to the fact that our sequence may be contained in the set of measure zero for which the Glivenko–Cantelli theorem does not claim convergence for g_2. This set can differ from the according set for the function g_1.

The term EDF denotes a random function, i.e. an infinite–dimensional random vector, as well as a concrete realization of this vector, which results from substituting random numbers for the random sample. From now on we will always refer to the second meaning of EDF, thus:

> The EDF is a property of a set of realizations of a random variable. It differs from the notion of a distribution function of a random variable since it is not a description of the expected average behavior but a measurement of an observed, finite sequence of realizations. In our opinion it is better to think of the EDF as an "high–dimensional" random number that occurs when we perform an experiment k times.

1.4.3 A sequence of independent uniform (pseudo–) RNs

In Section 1.4.1 we have stated that a sequence of random numbers does not have anything like a distribution. However, everybody in the field of stochastic simulation

speaks of sequences of *P*RNs having this or that distribution and we have to take a closer look at the meaning of such a statement.

A sequence of independent uniform (P)RNs should be a sequence of numbers that behaves somewhat typical for a sequence of independent uniformly distributed random variables. In [26, p.3], Knuth has stated that PRNs should *appear* to be random. In this sense, iud PRNs should appear to be independent and uniformly distributed.

Now given a (finite) sequence of pseudorandom numbers $x_0, x_1, \ldots, x_{N-1}$, where $x_i \in [0,1[$ we have to show that

1. the x_i behave like realizations of a random variable distributed uniformly in $[0,1[$,
2. the x_i behave like independent realizations of such a random variable.

Since an iud sequence of random variables can be defined by using the multidimensional uniform distribution we may restate the two requirements in a single definition:

Definition 1.8 *A sequence of pseudorandom numbers $x_0, x_1, \ldots, x_{N-1}$ is said to be independent uniformly distributed if for every $s \in \{0, 1, \ldots, N-1\}$ and every possible s–tuple $(x_{i_1}, x_{i_2}, \ldots, x_{i_s})$ this tuple behaves like a realization of an s–dimensional uniformly distributed random vector.*

However, this definition remains incomplete until we have clarified what is understood by 'behaves like a realization': such *numerical properties* of the PRNs will have to be developed *with respect to the actual application.*

If we do not mark out such properties any further, the ridiculous, but mathematical correct conclusion is: every possible s–tuple in $[0,1[^s$ behaves like a realization of an s–dimensional uniformly distributed random vector $\mathbf{X}$ since it falls into the range of $\mathbf{X}$. Thus every possible sequence of numbers in $[0,1[$ has the same right to be called a sequence of random numbers. Note that this argument is built on the definition of random numbers and does not account for the fact that the special model 'uniform distribution' assigns the same probability to every sequence of PRNs. The argument is thus valid for any desired distribution. It reflects the difficulty to relate random variables to random numbers.

We have derived this trivial conclusion in a complicated way in order to allow a slight modification, that introduces the numerical properties relevant for an application: since we have a tool for estimating the distribution of a function of random variables, namely the empirical distribution function, *why don't we substitute 'has an empirical distribution function near to the s–dimensional uniform distribution' for 'behaves like a realization' in Definition 1.8?*

This is the key to every sort of statistical inference on PRNs. The usual arguments run the following way: fix a function $g = g(X)$ of a random variable X distributed uniformly

on $[0,1[$. *Define* that a sequence $x_0, \ldots, x_{N-1}$ is 'independent uniformly distributed' if the empirical distribution function $\hat{F}_{g(x_0),\ldots,g(x_{N-1})}(t)$ is near to the distribution function $F_{g(X)}(t)$ of g. As we will see in Chapter 2, this amounts to carrying out a statistical test on the sequence of PRNs.

What is wrong with this definition? It can be criticized in two ways:

- The same sequence of PRNs may be called iud for one function g_1 but may fail the test for another function g_2. This again stresses the fact that PRNs should be chosen with respect to their application.
- We 'abuse the language' by contributing a distribution to random numbers. This has been done for reasons of compatibility to literature only. Random numbers as well as PRNs do not have a distribution in the sense of probability theory. By using the statement mentioned above we express a numerical property, the distance between empirical and theoretical distribution function, that has to be distinguished from the mathematical object of the notion of a distribution of a random *variable*.

Thus statistical inference on PRNs will lead to numerical properties which cannot be used to judge on the PRNs in terms of 'distribution' or 'randomness'. A statement like 'My generator has passed the RUN–test for randomness' does not tell us that the PRNs from the generator are more random than any other sequence of real numbers. We will further examine and develop these kind of numerical properties in Chapter 2 where we will also give examples.

1.5 Pseudorandom number generators

In the last section we have rejected physical devices in favor of deterministic methods as source of (pseudo–) random numbers. Now, let us take a closer look at some typical algorithms and introduce the notations used in the thesis.

We will treat a generator as an algorithm that produces a number in $[0,1[$ on each call. Typically, the algorithms themselves will produce integers in some range $[0, per - 1]$[19] which can be transformed to $[0,1[$ by dividing each number by per. However, in most applications the user will need a sequence of random vectors in the s–dimensional unit cube $[0,1[^s$. Such vectors can either be obtained by special pseudorandom vector generators[20], or by grouping s onedimensional random numbers to form a single vector. In this thesis we will only discuss the second method.

In principal there exist as many algorithms for generating a sequence of N PRNs as there

[19] *per* stands for 'maximal possible period length' and expresses the fact that all known methods produce periodic sequences of integers.

[20] We refer the reader to L'Ecuyer [28, Section 3.6] and to Niederreiter [38].

exist sequences since any enumeration of N PRNs actually is an algorithmic description of this sequence. However, in practice we will have to offer a short and efficient method for calculating the whole sequence when given only initial values. This is what is really meant by the term pseudorandom number generator.

We have to be careful since more complicated algorithms will not necessarily lead to 'better' pseudorandom numbers. This has been illustrated by a nice example in Knuth [26, p.4] with the 'super–random' number generator "algorithm K". Knuth shows that this generator can have a very short period. The device actually gives only *one* PRN when the initial value is set to 6065038420, which is a fixed point of the algorithm.

Since any computer–implemented device will show a periodical behavior due to its finite state space, the maximal possible period length should be regarded as an important quality criterion. However, the simple generator

$$x_n := \frac{n \bmod per}{per} \quad n = 0, 1, \ldots$$

gives a PRN sequence with period per that can be chosen to be any integer that fits into the computer's memory. Without doubt, the sequence will, in almost all cases, not be what we would like to call an pseudo*random* sequence of numbers in the sense of the definition given in the last section.

We thus have to look for methods that combine both, a long period as well as 'random'[21] behavior. Excellent surveys on the most important methods for which a detailed theoretical analysis has been done are L'Ecuyer [28] and Niederreiter [38]. We will concentrate on LCGs, the traditional 'workhorse' in the field of stochastic simulation, ICGs and EICGs. LCGs are *linear* generators, i.e. the algorithm uses linear transformations, whereas ICGs and EICGs are *inversive*. The later have been introduced by Eichenauer-Herrman and Lehn in a series of papers. We refer the interested reader to [8] and [10].

Tables of parameters that yield maximal period length for ICGs can be found in [22]. Implementations in ANSI C can be found on the WEB server [2] of the pLab group in Salzburg, Austria. The next section includes some of the results of the theoretical analysis and will further explain our restriction to these methods.

We have used the following parametrizations for the empirical tests in Chapter 4 because they seem to give a good spot check on the used methods:

- $EICG(2^{31} - 1, 1, 1, 0)$, short "EICG1"
- $EICG(2^{31} - 1, 7, 0, 0)$, short "EICG7"
- $ICG(2^{31} - 1, 1, 1, 0)$, short "ICG"
- $LCG(2^{31} - 1, 950706376, 0, 1)$, recommended by Fishman and Moore in [14]

[21] Again, randomness is understood in the sense of numerical properties relevant for the application.

- $LCG(2^{31}, 1103515245, 12345, 12345)$, the ANSI C system generator
- $LCG(2^{31} - 1, 16807, 0, 1)$, the "minimal standard", see [39]
- $LCG(2^{31}, 65539, 0, 1)$ known as "RANDU", see [39]

With the exception of RANDU, all these generators have about the same period which is of order 2^{31}. RANDU only produces some 2^{29} different numbers. Fast implementations are even possible on 32-bit computers. Without any further knowledge about their application these generators should be treated as being equally good sources of 'randomness'.

A very interesting approach in order to get longer sequences, the so–called compound method which uses some basic generators of the type LCG, EICG or ICG and adds their output modulo one, will not be discussed in this paper. Theoretical analysis has shown that the compound method somewhat preserves the properties of the single generators. A combination of inversive methods will also show typical 'inversive behavior'. We refer the interested reader to [12].

1.6 Structural properties of PRNGs

In order to describe properties of PRNs that are relevant for an actual application we have to develop numerical measures of sequences of PRNs that allow us to distinguish between the different generation methods. Such measures can be divided into two classes, depending on the kind of arguments involved.

So–called statistical tests characterize a sequence of PRNs in comparison to all possible sequences of RNs by rating the region into which the results fall. They will be discussed in the next chapter. In this section, we shall concentrate on structural properties of PRNs.

For certain algorithms we are able to describe in a detailed way the subset of $\mathbf{R}$ or $\mathbf{R}^s$, in which the PRNs fall.

Let us look at a trivial example for such a characterization: if a PRNG first produces integers in the range $\{0, 1, \ldots, per - 1\}$ and divides each number by per, the PRNs will lie in the finite subset

$$\Delta := \left\{ k\frac{1}{per} : k \in \{0, 1, \ldots, per - 1\} \right\}$$

of $[0, 1[$. The maximal possible resolution of these numbers is a known property of the generator and can be used to distinguish it from other generators. It is clear that a finer resolution of the generator will be regarded an advantage in most simulation problems. This is one explanation for the importance of theoretical results that guarantee maximal period length.

All the generators in our test battery have a period that is long enough to put the resolution of the random numbers in the region of the resolution of the typical 'float' types on most computers; apart from using 'double' types or special computing packages a longer period would not contribute any more to the resolution of the values that can be used in common programming languages. However, a long period is still a reasonable criterion for choosing a generator even when the resulting numbers do not gain in resolution. As we will see in Chapters 3 and 4, this is especially the case when many calls to the generators are made[22].

A generalization of the concept of resolution to more dimensions is the term *"lattice"*. A lattice is a set $\Delta \subset \mathbf{R}^s$ of the form

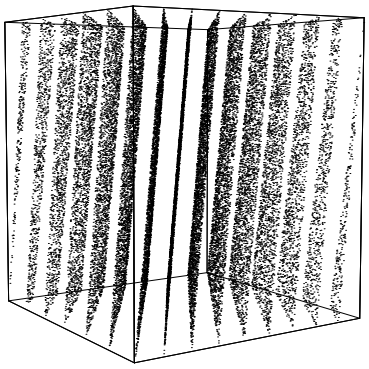
$$\Delta := \left\{ \sum_{i=1}^{s} k_i \mathbf{a}_i : k_i \in \mathbf{Z} \right\},$$

where the $\mathbf{a}_i$ denote s linearly independent vectors in $\mathbf{R}^s$. This definition can be found in [16, Section 1.3, Definition 1].

If we now take nonoverlapping s–tuples of PRNs and form s–dimensional vectors, the same argument as above holds: the pseudorandom numbers of a generator with period length *per* will lie on the lattice

$$\Delta_0 := \left\{ \sum_{i=1}^{s} \frac{k_i}{per} \mathbf{a}_i : k_i \in \{0, 1, \ldots, per - 1\} \right\},$$

where $\mathbf{a}_i$ is the i'th unit vector. Actually, a widely used type of generator, the LCG, produces numbers that lie on a *much coarser* lattice! The following plot of overlapping 3–tuples generated by the famous RANDU needs no accompanying words:

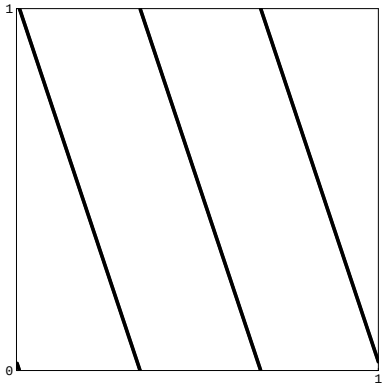


Randu, overlapping 3–tuples.

Pay attention to the fact that RANDU has a period length of 2^{29}! In order to illustrate

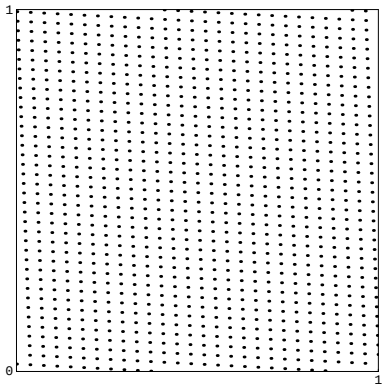
[22] Consider the difference between the distribution of balls drawn from an urn with and without replacement. See also L'Ecuyer [28, p.5]

the sensibility of the LCG with respect to the choice of parameters, we consider two "toy" LCGs[23].



LCG(1024, 1021, 21), overlapping 2–tuples.

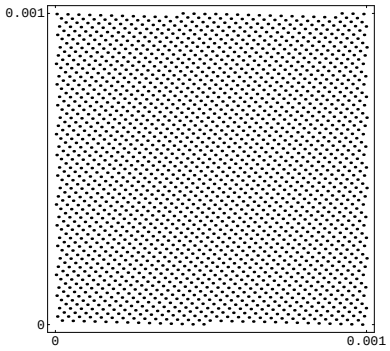
We will not give the proof for this coarse lattice structure of the LCG and refer the reader to Niederreiter [36] and Ripley [42]. However, differently parametrized LCGs produce totally different lattice structures as the following plot of a LCG with the same period shows.



LCG(1024, 997, 21), overlapping 2–tuples.

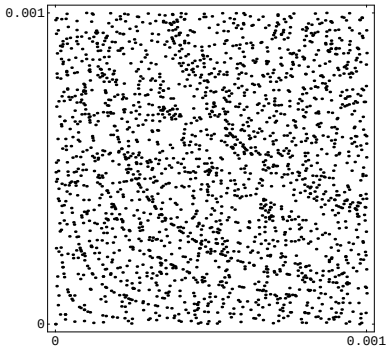
Using a computer–simulated magnifying glass, we even can reveal the lattice structure of a very 'good' LCG of our test battery. Look at the following plot, which contains all overlapping tuples that fall into the square $[0, 0.001[^2$:

[23] We would like to thank Hannes Leeb of the PLAB–group for the permission to reproduce the following five plots.



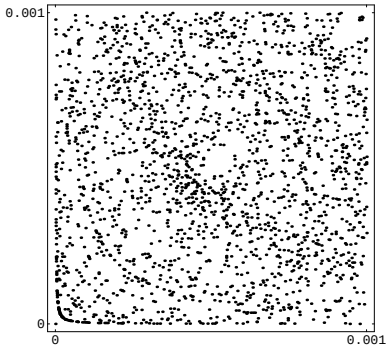
LCG($2^{31}-1, 950706376, 0, 1$), overlapping 2–tuples, whole period.

This is what is produced by an ICG: no lattice seems to be recognizable.



ICG($2^{31}-1, 1, 1, 1$), overlapping 2–tuples, whole period.

The same is true for an EICG:



EICG($2^{31}-1, 1, 0, 0$), overlapping 2–tuples, whole period.

Summing up, we have the following *very important* difference between linear congruential generators and inversive methods:

> PRNs produced by LCGs lie on lattices. The 'coarseness' of the lattice depends on the parameters of the LCG in a very sensitive way. Inverse generators like ICG and EICG produce no lattices.

For a mathematical proof of the statement, we again refer the reader to Niederreiter [36]. A measurement for the 'coarseness' of a lattice is the so called *spectral test* which can be found in Knuth [26, p.89]. Recently, a new approach has been introduced by Hellekalek [19]. The *weighted spectral test* allows a similar theoretical analysis as is possible for the discrepancy. Further, empirical computations for samples are possible.

Well, all this lattice theory is certainly of interest for mathematicians. But what are the consequences for practical stochastic simulation? In our opinion, the following arguments can be built upon the theoretical results:

1. A coarse lattice contains large regions which will never be hit by a point of the generator. There exist simulation problems which will only give reasonable results when these regions are hit by the expected number of points. If somebody expects their simulation problem to be sensitive in this respect, the results should at least be controlled by a second simulation using an inversive generator or a LCG with a fine lattice structure.

2. A lattice expresses strong regularities. We expect certain 'natural' simulation problems to be sensitive to such regularities. Unexpected results can for example result from the superposition of a regular partition of the sample space by the simulation on the lattice which results in aliasing effects or interference. We will meet this argument again in Chapter 4.

3. In the following chapters we will introduce empirical statistical tests that show considerable differences between linear and inversive generators. Many of the results can be explained with the lattice structure of the LCGs.

We have included these remarks in the thesis to make the reader aware of the differences present behind the generation methods for PRNs. *These differences are a big advantage!* Different types of generators *can and should* be used to verify simulation results. From the mathematical viewpoint it is clear that the inversive generators will cause certain simulations to yield bad results, too. Whether these simulations have a structure that appears among 'natural' problems within the field of stochastic simulation cannot be told without further analysis. For the moment all we can do is to increase our confidence in a simulation by evaluating it with strongly different random numbers and comparing the results.

1.7 Summary: All random numbers are equal

- All RNs are equal. The Kolmogorov setup in connection with Definition 1.2 provides *no way to rate* RNs.
- PRNs are RNs designed to have special numerical properties that are relevant for an actual application. This qualities often are called statistical properties[24] but make sense only with respect to certain applications or models.
- There are –among others– *linear and inversive* methods to produce PRNs
- Theoretical analysis has shown that these methods differ strongly in the structure of the PRNs they produce.
- The different methods are an advantage as they can be used to increase confidence in the result of a simulation.

[24]Statistical properties are properties of distributions. RNs have empirical distributions. Empirical distributions are numerical properties of a set of real numbers whereas distributions are assumptions on a model.

Chapter 2

Tests for randomness

Jeder 7. Gesunde war krank.
– Frankfurter Abendpost

In this chapter we will develop further criteria for choosing sequences of PRNs out of the set of possible sequences of RNs. These criteria will always reflect relevance with respect to a simulation problem. We will thus have to take a closer look at such simulations.

Suppose we want to simulate a crossroads. We use N random numbers in the following way: If the i'th random number is greater than $1/2$, a car comes along the street and queues at the crossing. The considerations in the last chapter would now enable us to choose for the first and last time the sequence $0, 0, 0, \ldots, 0$ as realization of N independent uniformly distributed random variables. However, the simulation would by no means reflect the usual rush–hour phenomena; Unfortunately cars happen to appear. What went wrong?

It is clear that any answer has to deal with what we have called 'numerical properties' in the definition of pseudorandom numbers. We will have to take *Pseudorandom* Numbers, not simply RNs, in order to get the desired results!

Put T_c be the number of cars that have arrived in the discretizised[1] time interval

$$[0, N-1] \cap \mathbf{N}.$$

What has to be done in order to carry out the stochastic simulation?

[1] Discretization of a domain is a risky step in forming a model for a real world phenomena since it can lead to unexpected dynamics. However, this is just an introductory example.

2.1 Stochastic simulation

We first have to model the situation expressing the random arrival of cars. This can be done in several ways. A very simple model would be the following: for every 'time' $i \in [0, N-1] \cap \mathbf{N}$ we make an experiment by flipping a fair coin. If head occurs, a car arrives at the crossroads. If tail occurs, no car arrives.

By Definition 1.2, we can model the coin flipping in terms of random variables. The commonly used interpretation would be something like: let $\mathbf{X} : (\Omega, \mathcal{A}, P) \to \mathbf{R}^N \quad \mathbf{X} \sim U\left([0,1[^N\right)$ be a random vector distributed uniformly on $[0,1[^N$. If the i'th component of this vector is greater than $1/2$, a car arrives at time i. Thus, the desired number of cars T_c becomes a function of the random vector $\mathbf{X}$:

$$T_c : (\Omega, \mathcal{A}, P) \to \mathbf{R}, \quad T_c(\omega) := \sum_{i \in [0,N-1] \cap \mathbf{N}} \delta(\mathbf{X^{(i)}}(\omega) > 1/2)$$

This is the mathematical model of the situation. The model has been chosen due to its simplicity and is of course not the 'best' formalization of a crossroads. The principal problems connected with any stochastic simulation can nevertheless been shown by using T_c.

The first question is whether the model using a random variable is clever. The situation is that we do not know the mechanism responsible for the arrival of the cars. Our model treats this situation in a very generous way: every possible sequence of car arrivals can occur and is assigned the same expected average number of occurrences in repeated trials. By this, the trivial sequence

$$0, 0, 0, 0, 0, \ldots, 0$$

is contained in the model as well as the very regular sequence

$$1, 0, 1, 0, 1, 0, \ldots, 1, 0$$

If we would expect the cars to come in such a regular way, we could avoid using random variables and build a deterministic model instead. In our example the value of T_c would calculate to zero for the first sequence and to $\lceil \frac{N}{2} \rceil$ for the second sequence.

The regularities in the sequences above seem to make them non–random. This is true in a certain sense, since we can give a short description of the whole sequence. Such a short description could be turned into a good strategy for avoiding long queuing times. However, since no such description seems in view for the arrival times of cars, we have to include all other possible sequences in our model. The random character of the arrival times is due to the unknown selection mechanism that selects one such sequence out of the sample space. This is the reason for choosing random variables in order to model the situation.

On the other hand, regularities could be a way to reduce the amount of sequences we have to scan in order to evaluate properties of the distribution of T_c if we manage to

describe regularities in a way that allows us to choose such sequences that yield desired results. In our example the regular sequences lead to results spread within the whole range of T_c, however. The first example sequence yields a very untypical number of cars, whereas the second sequence gives values close to the expectation of T_c. In this sense only the second sequence leads to good simulation results.

Every pseudorandom number generator will generate sequences with regularities[2]. One main problem in stochastic simulation is to find out *in advance* if the *special sequence of PRNs we use* contains regularities that lead to (un)desired results with a *specific type of simulation problems.*

Once introduced into the model, the concept of distribution will persecute us throughout the whole simulation. The mathematical model T_c itself has become a random variable. It is nothing but a description of the transformation of information given in the form of distributions. The result of the transformation is the distribution of the random variable T_c, that is, a distribution of the number of cars that arrive at the crossroads.

> As pointed out before, this distribution cannot be calculated from single simulation runs where the parameters are replaced by random numbers and the output of the model thus becomes a random number itself.

This is a hard reality for everyone who considers stochastic simulation in order to gather information on a problem. Conceptually, the formulation using random variables has the advantage that it permits us to introduce parameters into a model for which we cannot give a deterministic description. It has the big disadvantage, though, that single realizations of the model have little to do with the information represented by the whole model, i.e. by the distribution of T_c.

One way to escape from this deadlock would be to calculate the distribution F_{T_c} by either direct mathematical analysis or by the evaluation of all possible ω in the probability space $(\Omega, \mathcal{A}, P)$ on which T_c lives.

Unfortunately, this cannot be done for most problems arising in the field of stochastic simulation due to mathematical difficulty on the one hand and limited computer power on the other hand: Even if the model 'only' were evaluated for all possible sets of N PRNs, which is a finite set, the computations could not be expected to terminate in a reasonable amount of time for usual sample sizes of $N \simeq 2^{10}, \ldots, 2^{60}$.

The other way out is to make only some k or even only one simulation run using PRNs. What can we learn from such results? Without further considerations concerning the problem class and the appropriate sequences of PRNs we cannot learn anything about T_c out of single simulation runs. This is a consequence of the weak relation between random variables and random numbers.

Once again: without further assumptions the result of a simulation run is simply a

[2] Consider lattices, long range correlations, or periodicity for example.

random number out of the set of valid realizations of the random variable T_c. By modeling some parameters using random variables we have decided to be interested in properties of the *distribution* F_{T_c} of the model. The result contains no information from which these properties could be calculated in a deterministic way.

We will now develop an understanding of what is meant by 'numerical' properties of PRNs. The idea starts with simulation problems for which a mathematical analysis can be done: in our example we manage calculating the distribution of T_c which is a binomial distribution with parameters $1/2$ and N:

$$T_c \sim B_{1/2,N}$$

Thus our model expresses that if we agree, that cars come according to fair coin tosses and information about fair coin tosses can be expressed by the given indicator functions of a uniformly distributed random variable, then each number l of cars will occur with an expected average rate of

$$\binom{N}{l}\left(\frac{1}{2}\right)^N.$$

This is exactly the information contained in our model. The distribution assigns more probability to results near the expectation which is $N/2$.

What properties can be shown for a sequence of PRNs using this model? The magic procedure yielding such properties is called 'statistical test'.

2.2 Statistical tests

The idea of statistical tests is quite simple. Every model T for which we *can* directly calculate the distribution will serve as statistical test in the following way: we make an agreement on what we will call the 'typical' results of the model in marking a set of 'bad' values in the range of T as *critical region* $\mathcal{C}$. Values in this critical region are valid realizations of the random variable, but simply not realizations we are interested in. The probability of the critical region, the accumulated probability of all elements $\omega \in \Omega$, for which the value $T(\omega)$ lies in the critical region, is called the *significance level* α of the test. Thus

$$\alpha := P\left(\{\mathbf{x} \in \Omega : T(\mathbf{x}) \in \mathcal{C}\}\right).$$

Usually α is kept at very low rate, say $0.1, 0.5$ or even 0.01, somewhat blurring the 'ad hoc' character of the method.

The variable $\mathbf{x}$ denotes single basic events in the probability space. We may in general assume that T can be written as function of a random vector $\mathbf{X}$, which is distributed uniformly,

$$T = T(\mathbf{X}).$$

We denote the single sequences of PRNs that can be used as realization of this random vector by $\mathbf{x}$.

Using again our simulation problem T_c we might choose the following critical region

$$\mathcal{C} = \{0, 1, \ldots, N\} \setminus \left[\frac{N}{2} - e, \frac{N}{2} + e\right],$$

where e denotes the utmost error with respect to the expected value we will accept. This maximal error determines the significance level α. If we set $N = 100$ and $e = 9$ for example, we get $\alpha \simeq 0.05$.

We then evaluate the simulation using a PRNG. The generator is said to pass the statistical test if the result of the simulation does not fall in the critical region. The simulation problem now is called *test statistic*. Any random variable with known distribution thus defines a test statistic.

> This is understood by a test for randomness. It may seem dexterous, but in principle is only a nice excuse for excluding 'mathematically innocent' random numbers from the set of valid random numbers for a random variable. It is a very bad use of the language to call this procedure 'test for randomness' since it exactly leads to the opposite. We exclude some of the possible sequences, thereby *diminishing the amount of randomness* that is possible.

For the moment we note that a statistical test assigns a numerical property to a sequence of PRNs that is called 'randomness' and means that a certain test statistic yields desired results.

What have we gained? The given test will exclude the sequence

$$0, 0, 0, \ldots, 0$$

from the set of PRNs suitable for this simulation. Moreover, only sequences containing about the same number of PRNs over and under $1/2$ will be accepted. A huge number of valid realizations of the vector $\mathbf{X}$ will be excluded, but the remaining PRN sequences will all lead to desired results for the simulation T_c.

Thus the interpretation of the test is clear: if the generator fails the test, it should not be used to run the concrete simulation. Unfortunately, this interpretation is of no use since we already know the expectation of T_c. What we really want is an information if the generator will also fail in simulating *other models*. Such statements cannot be expected without further information on the models. In his master's thesis [30], Leeb has formulated such conditions on the relation between known simulation problems and the actual model the user is interested in. If one agrees to be interested in the expected value, this conditions provide a formal precision of confidence in PRNs that have passed certain 'statistical tests' and thus have succeeded in simulating expected values in related simulation problems.

Leeb has also shown that these criteria cannot lead to a general class of PRNs that yield good results for any simulation problem. Roughly speaking, the number of problems yielding 'good' and 'bad' results is the *same* for every finite sequence of PRNs. We will examine the relationship between practical applications and tests later in this chapter.

If a sequence of PRNs does not pass a certain statistical test, another often heard interpretation is that the PRNs from the generator do not have 'good' statistical properties. We do not agree with this conclusion because statistical properties is something associated with random variables. A distribution can have statistical properties like contributing a certain probability to a critical region. The laws of large numbers and the central limit theorem are statistical properties of certain random vectors. A random number resulting from a stochastic simulation *does not have any statistical qualities at all.* Statistical properties assign probabilities and cannot lead to judgements like 'good' and 'bad' without referring to a certain problem! Would we say that the S.L.L.Ns. is telling that almost all sequences are good, whereas only a set of measure zero is bad? Such a statement can be made about the sequences if the asymptotic behavior is needed for an application. It is not a fair judgement on the sequences unless we define what is meant by 'good' and 'bad'.

We might wonder why this wrong interpretation has found such a big audience. If the test statistic is an injective function we can exclude any desired set of sequences of RNs only by modifying the critical region. Formally, *a statistical test is equivalent to choosing a set of admissible sequences of PRNs*: the choice of the critical region within the range of the test statistic suggests that the selection of admissible sequences is very natural. This impression is reinforced by keeping the significance level very low. But after all, *this procedure has no right to exclude any sequence for the lack of randomness!*

In our opinion the misunderstanding of the results of statistical tests can be understood by looking at the historical development. Statistical tests had been used in many real world phenomena till the day they had been applied to PRNs. Their great success in the field of decision making on the base of noisy information could have been due to the class of problems they have been applied to, namely problems for which no regularities had been found. Then, suddenly, they rejected a sequence developed precisely to pass such tests: a sequence of PRNs. It's somewhat clear that the sequence was called to account for the failure. The judges thought that events to which the test statistic assigned such a small probability simply *had* not to occur. But this is *lying with statistics* and does not reflect the facts.

Another reason is that a critical region within the range of a test statistic that contains its expectation seems to be natural. It expresses the wish to get simulation results near this expectation. From the view of probability theory there is *no critical region that is more natural than any other*, however.

Statistical tests are *a possibility to express the qualities of PRNs* we are interested in. The tests have to be chosen carefully with respect to the application. Every PRNG will pass and fail the same 'number'[3] of tests. We will call the result of a test a statistical property in order to be compatible to a huge field of literature. The word 'statistical' suggests that we deal with random variables. In fact we are dealing with real numbers. Randomness has been eliminated before any test is applied to PRNs.

[3] See Leeb [30, p.16] for a proof of this statement.

Throughout the literature two classes of (statistical–) tests are distinguished, which are called *theoretical* and *empirical* tests. The following sections discuss the differences between them.

2.2.1 Theoretical tests

By theoretical tests we denote statistical tests that can be evaluated for a certain type of generator using mathematical transformations of the generation algorithm. Consider the following sample generator, the VBPRNG (Very Bad PRNG). It produces the sequence

$$0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, \ldots$$

The generator has a very short period, but its simple behavior allows us to evaluate the sample test T_c given in the last section. The generator will trivially pass the test, if the critical region is set to

$$\mathcal{C} := \{0, 1, \ldots, N\} \setminus \left([\frac{N}{2} - 1, \frac{N}{2} + 1] \cap \mathbf{N} \right).$$

For $N \to \infty$ the significance of the test can even be made arbitrarily high[4]. Thus, the VBPRNG passes a theoretical test at any desired level of significance, provided N is chosen large enough.

Note the a–priori character of this type of tests: without producing a single PRN, we can assess qualities for a certain type of generator.

Of course, mathematical theory can offer us much more sophisticated theoretical tests. In all these tests the generator is treated as an open mechanism. By studying its underlying algorithm the numerical value of certain functions, e.g. test statistics, of the PRNs can be estimated or even calculated exactly. If the theoretical distribution of the test statistic is also known, this amounts to carrying out a statistical test. The generator can *a–priori* be assessed to fail or to pass the test depending on whether the estimates of the test statistic fall in the critical region or not.

We refer the reader to L'Ecuyer [28] for a survey on the available tests. The most important results within this area of research are the spectral test for LCGs and discrepancy estimates for LCGs, EICGs and ICGs depending on the input parameters of the according algorithms. We refer the reader to [36], [9], [14] and [37].

2.2.2 Empirical tests

In contrast to theoretical tests, empirical tests have to be evaluated by a computer. The word 'empirical' has to be seen from the viewpoint of the computer that 'experiments'

[4] That is, α can be made arbitrarily near to zero.

with a certain PRNG. The generator itself is treated as black box, only the sequence of PRNs it generates is taken for evaluating the test statistic. In order to apply an empirical test one has to implement the PRNG and the test statistic. Limitations to empirical testing are mainly imposed by the amount of time and memory needed due to the complexity of the computations.

Note that the distribution of the test statistic has to be evaluated directly by mathematical methods. But apart from this restriction an empirical test can be evaluated for any generator in view. Thus the sample test T can empirically be evaluated for any LCG, ICG or EICG.

As in the field of theoretical tests, literature has proposed numerous tests. We refer the reader to the next chapter, where we will treat a very general class of such tests in detail. Other useful tests can be found in Knuth [26] and Marsaglia [31]. An exhaustive empirical analysis for LCGs has been done in [14].

2.2.3 Comparison

One main drawback of theoretical tests is that they are usually limited to a specific class of generators for which the value of the test statistic can be calculated using the information on the generation method. By this it is impossible to compare the values to those of other generation methods, especially to new ones.

Another problem with theoretical tests is that they usually require a sample size N to be extremely large, often equal to the whole period of the generator[5]. This is definitely not done in most applications; there even exist numerous recommendations that one should only use N less than the square root of the period in applications. Although we do not believe in this rule of thumb, the question remains if 'statistical' properties[6] of the whole sequence also remain valid for smaller fractions of the period.

It has to be said that we have encountered a *strong correlation* between theoretical tests on the whole period and empirical tests on smaller fractions of it. This can be explained for some theoretical measures like the coarseness of the lattice on which the PRNs lie. A coarse lattice will be detected by a large class of empirical tests for example.

The good performance with the theoretical tests gives an a–priory confidence in the generators. No general mathematical technique is currently available that forecasts the performance of the same generators for other sample sizes or different test statistics although we strongly believe that there exist such dependencies. As Hellekalek mentions in [21, p.9], the "generators that have excelled in theoretical tests were the top performers in empirical tests".

[5] Recent results have shown new techniques to overcome this problem. They do not make assertions on one generator, but calculate the arithmetic mean of the test statistic when a large set of generators is applied. We refer the reader to [35], [12] and [11].

[6] in the sense of passing a certain test

One advantage of theoretical tests is that they can be evaluated for interesting test statistics even in higher dimensions. The star–discrepancy, a very important theoretical measure for uniform distribution, cannot be evaluated empirically, since the complexity of the algorithm is exponential in the dimension and leads to unreasonable requirements of both memory and time even for dimension 2. We can calculate estimators for this value for LCGs and inversive methods, however[7].

New theoretical measures are currently under development which can be used to estimate the discrepancy on one hand, but which can also be simulated with computers. We refer the reader to [18, 20, 19].

Empirical tests complement the theoretical methods. They can be adapted to a broader class of simulation problems. One aim of this thesis is to encourage users of PRNs to build their own tests. The tests should reflect some of the methods used in the simulation and should be run with sample sizes that are similar to that used in the simulation. The next chapter will show the design of such tests. Different generators – at least linear and inversive methods – should be tested. The user will get results that will certainly lead to a better simulation by either excluding generators that are sensitive to a certain problem class or by including them in order to check if the tests and the simulation are correlated.

In our opinion testing is a two step procedure. In the first step we search within every available type of generator for parametrizations that pass as many theoretical test as possible. This step leads to parameter tables for LCGs, EICGs, ICGs and other types of generators. In a second step the selected generators have to undergo empirical tests built with respect to the actual simulation problem. The performance within these tests should be used to comment on the results of a stochastic simulation.

We thus consider empirical testing of random numbers to be as important as theoretical testing. Both methods exhibit 'numerical' qualities of the sequences of PRNs produced by PRNGs and every such information should be taken into account when looking for an appropriate generator for a given simulation problem.

2.3 Posing the right questions

It has already been pointed out that any rating of PRNs has to be done with respect to an actual application. This section will examine the relationship between tests and applications more closely. As a result, we will be able to pose appropriate questions. Up to now, mathematic theory cannot provide answers to all of these questions. However, some widely known results can be clarified within this framework.

In order to measure the performance of a sequence of PRNs in an application, we will have to refer to a quantity measuring the error with respect to the desired results. Since

[7] See [34, 35] for the estimation of the star–discrepancy.

any stochastic model $T(\mathbf{X})$, where $\mathbf{X}$ again denotes a random vector distributed independent uniformly on $[0,1[^N$, is a random variable itself and thus represents information in the form of a distribution, every property ξ of this distribution $F_{T(\mathbf{X})}$ can play the role of the desired result. Among such properties are the expectation $\xi_E := E(T(\mathbf{X}))$ and the higher central moments like the variance $\xi_V := V(T(\mathbf{X}))$. Sometimes the object of interest is the distribution *function* $F_{T(\mathbf{X})}$ itself.

If one manages to calculate the desired property of $F_{T(\mathbf{X})}$ by mathematical analysis, it should be done. Any kind of stochastic simulation will not provide any better information. If no direct mathematical treatment seems to be possible, we will have to use a stochastic simulation. We measure the property by the help of an estimator. An estimator $\hat{\xi}$ for property ξ is a random variable defined on some k independent copies of T,

$$\hat{\xi} = \hat{\xi}(T_1, T_2, \ldots, T_k), \tag{2.1}$$

where each T_i obeys the same distribution law as T. We will only consider unbiased estimators $\hat{\xi}$. In this case the expectation $E(\hat{\xi})$ equals ξ by definition. By this, we reduce the set of properties we are interested in to the expectation. Consider for example the property ξ_V, the variance. If we find a random variable $\hat{\xi}_V$ for which we have

$$E(\hat{\xi}_V) = \xi_V,$$

we have reduced the calculation of the variance to the simulation of $\hat{\xi}_V$ in a way that keeps the error with respect to its expectation low.

The variance of the estimator itself should be small since the probability to get small error in a simulation run is higher when the variance is small due to Chebyshev's inequality

$$P(|\hat{\xi} - E(\hat{\xi})| \geq \epsilon) \leq \frac{V(\hat{\xi})}{\epsilon^2}$$

In the evaluation of a simulation we will have to substitute random numbers $\mathbf{x}$ for the random vector $\mathbf{X}$ on which T, and therefore $\hat{\xi}$ is defined. By this, we can simulate the estimator by calculating $\hat{\xi}(\mathbf{x})$. This procedure will be explained in detail in Chapter 3.

The number N of PRNs needed in such an evaluation is called *sample size*. It calculates to $N = k \cdot s$, where s is the dimension of $\mathbf{X}$. The *error* we make is defined by

$$e = ||\hat{\xi}(x_0, x_1, \ldots, x_{N-1}) - \xi||,$$

where $||\cdot||$ denotes any norm in $\mathbf{R}^s$. This error will be used to rate the PRNs $\mathbf{x}$.

Let us consider some examples. A trivial estimator for the expectation of T is $\xi_E = T$ itself. In this case, $k = 1$. However, increasing the number k immediately yields a better estimator. The simple function

$$\xi_E := \frac{1}{k} \sum_{i=0}^{k-1} T_i$$

gives an estimator for $E(T(\mathbf{X}))$ with lower variance than $V(T(\mathbf{X}))$. We will have to simulate T k times using a k times longer sequence of PRNs

$$\mathbf{x}^0, \ldots, \mathbf{x}^{k-1}$$

in order to get an estimate, but the probability for making a certain error will decrease. Note, that the expectation $E(T(\mathbf{X}))$ can be a vector in $\mathbf{R}^r$, $r \in \mathbf{N}$!

The case where we want to estimate the distribution function itself is a little bit more abstract, but the principal idea is the same. An estimator for the distribution function is the empirical distribution function, which is defined on the same product space as in the example above. In Chapter 1, the empirical distribution function has been defined by

$$\hat{F}_{T(\mathbf{X}^1),\ldots,T(\mathbf{X}^k)}(\mathbf{t}) := \frac{1}{k}\#\{i, 1 \leq i \leq k : T(\mathbf{X}^i) \leq \mathbf{t}\},$$

where $T(\mathbf{X}^i) \leq \mathbf{t}$ means $\left(T(\mathbf{X}^i)\right)^{(j)} \leq \mathbf{t}^{(j)}$ for every component j of the random vector $T(\mathbf{X}^i)$. For every $\mathbf{t}$ in $\mathbf{R}^s$, where s is the dimension of T, this function has an expectation equal to the distribution function of T evaluated at $\mathbf{t}$,

$$F_{T(\mathbf{X})}(\mathbf{t}).$$

The difference to the example above is, that now $\hat{F}_{T(\mathbf{X}^1),\ldots,T(\mathbf{X}^k)}(\mathbf{t})$, $\mathbf{t} \in \mathbf{R}^s$, is a function, i.e. a vector with 'infinite' dimension. Nevertheless we can measure the error with respect to the 'expectation' $F_T(\mathbf{t})$ by the supremum norm. The error becomes

$$e = \sup_{\mathbf{t}\in\mathbf{R}^s} |\hat{F}_{T(\mathbf{X}^1),\ldots,T(\mathbf{X}^k)}(\mathbf{t}) - F_T(\mathbf{t})|.$$

Again, the variance of the error becomes smaller when k is made larger.

Up to now we have spoken of the *probability* to make errors. This is of no use, however, in connection with stochastic simulation since no device will tell us the amount of error actually present after we have made a simulation run. We are therefore after deterministic error bounds keeping the error below some critical level e_0 given by the user. Such bounds cannot be expected without restricting both, the number of sequences of PRNs admissible for the simulation, and the class of simulation problems itself.

To see this, we will take a look at two extreme positions. In order to use consistent notation we will mark the values of $\hat{\xi}$ for which the error e is greater than the given maximal error e_0 by a critical region labeled $\mathcal{C}_{\hat{\xi}}$. By fixing $\mathcal{C}_{\hat{\xi}}$, we actually fix a set of sequences of PRNs as unwanted within the application. The two points of view now are:

1. In developing a statistical test, we *fix* the simulation problem, i.e. the test statistic, and calculate the probability that an arbitrary sequence of PRNs passes the test, i.e. it produces results outside the fixed critical region $\mathcal{C}$. This probability is the significance level α of the test. Without limiting the set of possible sequences we cannot give a nontrivial deterministic error bound for the distance of the value of the test statistic to the desired result. In particular, we cannot tell if the result will lie in $\mathcal{C}$ or not.

2. A different approach would be to *fix* the sequence of PRNs and to *vary* the simulation problem. We could for example calculate the probability that an estimator for the expectation of any arbitrary probability distribution yields a result within the complement of a critical region defined by the maximum admissible error e_0.

   The number of distributions for which we get desired and undesired results is about the same[8], however. Without further limiting the class of problems, we cannot tell in advance if the estimation using the fixed sequence of PRNs will lead to an error less or equal to e_0.

The proper way out of the dilemma is to *vary both*, the sequences of PRNs and the simulation problems, but to limit the amount of variation to carefully chosen subsets. The subsets have to be described in a mathematical way that allows deterministic error estimates. It would be convenient and is in fact no restriction if the set of sequences of PRNs is described by using a statistical test T_s.

We thus have the following setup for the relationship between application and test: if the sequences that pass a certain test T_s lead to results outside of $\mathcal{C}_{\hat{\xi}_{T_i}}$ for the estimator $\hat{\xi}$ of a property of the distribution of a certain class $\mathcal{T}$ of applications T_i, the test can be used in order to select an appropriate PRNG. We will call this property a *strong correlation* between test and applications. If we denote the critical region of the test T_s by $\mathcal{C}_{T_s}$, we have

$$T_s, \hat{\xi}\{\mathcal{T}\} \text{ strongly correlated } \Leftrightarrow \forall T \in \mathcal{T} : \hat{\xi}_T(T_s^{-1}(\mathbf{R} \setminus \mathcal{C}_{T_s})) \subset \mathbf{R} \setminus \mathcal{C}_{\hat{\xi}_T}$$

Strong correlation thus leads to deterministic error bounds, which are essential for any numerical method. By studying strong correlations we are able to calculate properties of unknown distributions within a class of applications up to a fixed error.

A similar type of correlation has also been studied by Leeb [30]. He has developed a criterion that is more flexible than the strong correlation but does not lead to deterministic error bounds.

Before we will pose the 'suitable questions' we give a famous example for such a strong correlation, the inequality of Koksma. This inequality is the backbone for number–theoretic numerical integration and Monte Carlo methods.

Let $T = T(X)$ be a random variable where $X \sim U([0,1[)$. If $E(T(X))$ exists, we can write the expectation as an integral,

$$E(T(X)) = \int_\Omega T(X(\omega))dP$$

For the special distribution of X we get $\Omega = [0,1[$, $P = \lambda$, the Lebesgue measure restricted to $[0,1[$, and finally $X(\omega) = \omega \in [0,1[$. Thus

$$E(T(X)) = \int_{[0,1[} T(\omega)d\lambda$$

[8] This has been proved by Leeb in [30, p.28ff]

A commonly used estimator for this expectation –and thus for the integral– has already been introduced:

$$\hat{\xi}(X_0, \ldots, X_{k-1}) = \frac{1}{k} \sum_{i=0}^{k-1} T(X_i),$$

where the X_i are independent uniformly distributed random variables.

We now substitute a sequence of PRNs x_i for the random variables X_i and obtain a value $\hat{\xi}(x_0, \ldots, x_{k-1})$. The inequality of Koksma[9] bounds the error of the approximation of $E(T(X))$:

$$e := |\hat{\xi}(x_0, \ldots, x_{k-1}) - E(T(X))| \leq \hat{V}(T) D_k^*(x_0, \ldots, x_{k-1}),$$

where $\hat{V}(T)$ is the variation of the function $T : [0,1[\rightarrow \mathbf{R}$ and $D_k^*(x_0, \ldots, x_{k-1})$ is the so–called star–discrepancy of the numbers $x_0, \ldots, x_{k-1}$. In order to provide useful estimates, both functions on the right side should have low values. This amounts to restrictions on the problem class as well as the class of admissible sequences. The class of functions with low variation contains smooth functions, and it is somewhat clear that we can approximate the integral of such functions more accurately than the integral of functions that vary considerably.

The second quantity is the more interesting one since it selects a set of PRNs that leads to better error estimates. The star–discrepancy is in fact a numerical property of sequences of PRNs which can be explained well by a statistical point of view since it reflects the distance of the empirical distribution function of the values $x_o, \ldots, x_{k-1}$ from the distribution function of a uniformly distributed random variable $X \sim U([0,1[)$. The asymptotic distribution of the star–discrepancy is known for dimension one and the related test statistic is the Kolmogorov–Smirnov test statistic. We refer the reader to Niederreiter [36] for details and further references.

Since only low values of $D_k^*(x_0, \ldots, x_{k-1})$ lead to a small error bound, the critical region for the K–S test has to include high values of the test statistic. Note that the expectation of the star–discrepancy is not zero but positive. The *appropriate* critical region for this application is not set with respect to the expectation but with respect to the *desired* minimal value.

Let us back up the whole idea again. The approximation of the expectation of a certain set of random variables by the use of a certain estimator is strongly correlated to a statistical test that selects sequences of PRNs that approximate the uniform distribution well.

The example shows that the description of a set of sequences of PRNs by statistical testing can provide deterministic error bounds for a large class of applications that is itself described by a numerical property. The calculation of the significance level of this test is not required for estimating the error, which is bounded by the value of the 'test statistic' itself. Put the other way round, the distribution of a test statistic seems to be

[9] The inequality can be generalized to higher dimensions and to other sets of functions. For an introduction we refer the reader to [36, Chapter 2].

less important for deterministic error bounds in strongly correlated applications. What really counts is the inference of the critical region of the test and the critical regions within the applications that are defined by the maximal admissible error.

The "right questions" thus are:

- What classes of applications are common in the field of stochastic simulation? By application we mean an estimator of a property of any stochastic model for a real world phenomenon.
- How can these classes be described in a way that allows a user to check whether their problem falls into such a class, or not?
- Which test statistic can be used in order to describe the set of PRNs for which a deterministic error bound can be proved?
- Which design of the test leads to best possible error estimates?
- Which generators will pass this test?

The answer to the last question can be given by any user who has access to a computational tool for performing statistical tests and generating random numbers in various ways[10]. The answer to the first question requires "teamwork" between the user and mathematician.

The second question is the central one. In our opinion, much work has to be done in order to close this gap between applications and testing. It is clear, that the answer will strongly depend on the answer to the first question, since we always have to study the relation

$$\hat{\xi}_T(T_s^{-1}(\mathbf{R} \setminus \mathcal{C}_{T_s})) \subset \mathbf{R} \setminus \mathcal{C}_{\hat{\xi}_T}$$

where the set of admissible T's has to be defined. A possible way would be to start with some known tests and to develop strongly correlated classes of applications for the sequences that pass the test. The question is, whether the result of such an attempt will include practical problems.

Another, more promising idea starts with quantifying regularities within sequences of PRNs. A typical example for such regularities is the lattice structure of PRNs produced by LCGs. On the one hand it can be detected by tests, on the other hand we know applications which lead to bad results when a generator with a coarse lattice structure is applied. The author is not aware of any general theory covering the effects of lattices on stochastic simulation, however.

[10] See also Section 3.4 in Chapter 3

2.4 Summary: Some PRNs are more equal than others

- Statistical tests are used to distinguish subsets of sequences of PRNs by means of numerical properties.
- 'Tests for randomness' actually reduce the amount of randomness by selecting only a subset of all valid sequences of RNs.
- If a similar approach is made for selecting subsets of applications, deterministic error bounds can be given in principle.
- Stochastic simulation is used to calculate certain properties of the distribution of given models as exactly as possible. Such properties are measured by estimators.
- A correlation between a test and a class of applications, i.e. a class of estimators, has to be found in order to get deterministic error bounds.
- A famous example is given by the inequality of Koksma in the field of numerical integration of functions of bounded variation with node sets that have small star–discrepancy.
- The whole setup of stochastic simulation should therefore include tests, a class of applications *and a correlation* between tests and applications. By 'correlation' we understand a mathematical theory providing error bounds for the estimators within the class of applications.
- Within such a setup, a test is used to judge PRNs with respect to the class of applications. *As a consequence there are in fact PRNs that are more equal*[11] *than others.*

[11] e.g. better suited to certain applications,

Chapter 3

Empirical tests for uniform PRNs

Alle Daten bestätigen es: Der Mensch ist kein Meerschweinchen!
– Ärzte Zeitung.

In the first two chapters we have examined the mathematical background for empirical testing of PRNs. In this chapter, we will pass to numerical practice. We will develop a class of empirical tests for PRNs. We will also give hints for the design, implementation and interpretation of these tests. As the title of the chapter suggests, we will focus our attention on tests that interpret and complete our definition of iud. PRNs, Definition 1.8 in Chapter 2. Let us recall this definition:

Definition 3.1 *A sequence of pseudorandom numbers $x_0, x_1, \ldots, x_{N-1}$ is said to be independent uniformly distributed if for every $s \in \{0, 1, \ldots, N-1\}$ and every possible s–tuple $x_{i_1}, x_{i_2}, \ldots, x_{i_s}$ this tuple behaves like a realization of an s–dimensional uniformly distributed random vector.*

As it has been explained in the last chapter every test amounts to calculating a numerical property of sequences of PRNs which is assessed by marking a set of unwanted values as the critical region. Sequences that lead to results within that critical region are rejected by the test. The numerical property we are going to discuss in this chapter is related to the distance of the empirical distribution function of a sample of N s–tuples of PRNs to the uniform distribution on $[0, 1[^s$. We interpret the above 'behaves like a realization' by 'has an empirical distribution function *close* to the distribution function of a random variable distributed uniformly on $[0, 1[^s$'.

> This is a very important decision. We just have marked a lot of valid sequences of RNs as unwanted with respect to a stochastic simulation in which we want to use 'iud' PRNs! This decision is made in order to get *desired results* and is of use only if the simulation and the test are correlated in a way that allows such conclusions.

There exist simulation problems which will yield desired results even if the empirical distribution function is far from the theoretical one, as well as there exist problems which will yield bad results although the test has been passed by the numbers.

Within this setup the PRNs are always thought to be used in the place of s–dimensional, uniformly distributed random variables. We will use the notation

$$X_0, X_1, \ldots$$

for independent uniformly distributed random variables

$$X_i \sim U([0,1[)$$

and

$$\mathbf{X}_0, \mathbf{X}_1, \ldots$$

for uniformly distributed random vectors

$$\mathbf{X}_n \sim U([0,1[^s)$$

in the mathematical model from which we derive the distribution of the test statistic. The X_i are simply replaced by the actual PRNs, denoted by x_i, in the evaluation of the test for a generator. The test statistic then becomes a PRN itself.

In order to get a practically computable quantity, we will have to make some simplifications and thereby introduce arbitrariness in the procedure. Every such step should be made with the application of the PRNs in view. If the whole test is similar in structure to the simulation problem for which the PRNs are used, confidence in the above mentioned correlation between results of the test and the simulation will be enlarged.

The mathematical treatment of the test statistic will be given in Chapter 5. An informal description will be given right in the next section.

3.1 A class of statistical tests for uniform PRNs

In order to build a test for s–dimensional random numbers we have to construct such vectors from the sequence of onedimensional random numbers. As any other step in the construction of the test statistic, this transformation will be expressed by a function which can be applied to the random numbers x_i, but also to the random variables X_i. The result of the former are vectors of random numbers which will be tested, the result of the later are random vectors from which we deduce the distribution of the test statistic.

The first simplification that we have to make concerns the transformation of the sequence of random variables and PRNs, respectively, into a sequence of s–dimensional vectors, where $s \in \mathbf{N}$. We will only consider the following two, somewhat natural ways to proceed:

1. *Nonoverlapping s–tuples* are vectors

$$\begin{aligned} \mathbf{X}_n &:= (X_{ns}, X_{ns+1}, \ldots, X_{ns+s-1}) \\ \mathbf{x}_n &:= (x_{ns}, x_{ns+1}, \ldots, x_{ns+s-1}) \end{aligned}$$

2. *Overlapping s–tuples* are vectors

$$\begin{aligned} \mathbf{X}_n &:= (X_n, X_{n+1}, \ldots, X_{n+s-1}) \\ \mathbf{x}_n &:= (x_n, x_{n+1}, \ldots, x_{n+s-1}) \end{aligned}$$

As we are thinking of an empirical test that should be evaluated by a computer in a finite amount of time, we have to limit the number of PRNs that are to be tested. We will only use a finite sample of length[1] $\tilde{N}$ consisting of the vectors $\mathbf{X}_0, \ldots, \mathbf{X}_{\tilde{N}-1}$. Let us denote this sample by $\mathbf{X}$ and a realization of it by $\mathbf{x}$.

The difference between the two methods to form s–dimensional vectors is the different common distribution of the resulting random vectors. Random vectors $\mathbf{X}_n$ formed from nonoverlapping tuples are *independent*[2], whereas random vectors $\mathbf{X}_n$ formed from overlapping tuples are *dependent*. The common distribution function for overlapping tuples will *not* factor to the product of the distributions of the single random variables.

In many applications it will be desirable to have independent vectors and one will choose the nonoverlapping form of tuples. However, in building statistical tests we can use both forms of tuples since we will pay attention to the common distribution while calculating the distribution of the test statistic.

> The restriction to only two special types of s–tuples again expresses qualities we would like the PRNs to have! It is natural in the sense that many applications will form s–tuples of one of these types.

Note that choosing overlapping tuples is more general in that the resulting sequence will also include all vectors resulting from nonoverlapping tuples. However, the information represented by the two possible sequences of random vectors is exactly the same, since every single random variable X_i can be reconstructed from both forms. For the moment let us concentrate on the nonoverlapping case in order to develop the ideas.

The empirical distribution function for such a sample $\mathbf{x}$ is given by

$$\hat{F}_{\mathbf{x}}(\mathbf{t}) := \frac{1}{\tilde{N}} \# \left\{ n, 0 \le n \le \tilde{N} - 1 : \mathbf{x}_n \le \mathbf{t} \right\}$$

for every $\mathbf{t} \in [0,1[^s$, where

$$\mathbf{x}_n \le \mathbf{t} :\Leftrightarrow \forall i, 1 \le i \le s : \mathbf{x}_n^{(i)} \le \mathbf{t}^{(i)}.$$

[1] In this chapter we use $\tilde{N}$ for the number of vectors that are tested. In the other chapters, N denotes the number of (onedimensional) random numbers. These are related by the equation $N = \tilde{N} \cdot s$ in the case of nonoverlapping tuples and by $N = \tilde{N}$ for overlapping tuples.

[2] If two sets of random variables have the property that every random variable in the first set is independent from every RV in the second set, so will be two functions defined on only the first and second set respectively.

When we substitute the sequence $\mathbf{X}$ for the sample $\mathbf{x}$, the Glivenko–Cantelli theorem proves that the empirical distribution function of these random vectors will almost surely converge to the distribution function of $\mathbf{X}_n$, for $N \to \infty$, which is given by

$$F_{\mathbf{X}_n}(\mathbf{t}) := \prod_{i=1}^{s} \mathbf{t}^{(i)}, \quad \mathbf{t} \in [0,1[^s,$$

since the components of $\mathbf{X}_n$ are independent uniformly distributed random variables X_{ns+i}.

Convergence in the above sense is expressed by the distance of the two functions with respect to supremum norm: put d the distance

$$d = d(\mathbf{X}, \tilde{N}) := \sup_{\mathbf{t} \in [0,1[^s} |\hat{F}_{\mathbf{X}}(\mathbf{t}) - F_{\mathbf{X}_n}(\mathbf{t})|,$$

then d will converge to zero almost surely when we increase the sample size[3] $\tilde{N} \to \infty$.

Almost sure is defined with respect to the probability space $(\Omega, \mathcal{A}, P)$ which contains all infinite sequences of random numbers: the described convergence will happen for almost all[4] infinite sequences of random numbers. In other words, d degenerates to a random variable having value zero on a set of measure one if $\tilde{N} \to \infty$. This is a consequence of the strong law of large numbers and the compact range of probabilities.

For the 'esoterical' case $\tilde{N} \to \infty$ we thus have found a test statistic d expressing the numerical properties we consider important in our definition of iud sequences of PRNs. The set of sequences for which d does not converge to zero thus contains valid sequences of random numbers which do not lead to the desired results with our 'simulation problem' d.

But since we have only finite sequences of some length $\tilde{N}$ we have to calculate the distribution of the quantity d for this given sample size. Note that $d(\mathbf{X}, \tilde{N})$ is a random variable which is not only defined on $(\Omega, \mathcal{A}, P)$ but also on the probability space $(\Omega_{\tilde{N}s}, \mathcal{A}_{\tilde{N}s}, P_{\tilde{N}s})$ which contains all sequences of random numbers of length $\tilde{N}s$. Due to the construction of $(\Omega, \mathcal{A}, P)$, the distribution of $d(\mathbf{X}, \tilde{N})$ is the same with respect to these two probability spaces.

If it would be possible to calculate the distribution of d for a given $\tilde{N}$, we could mark large values of d within its range as the critical region and build a statistical test. The test would exclude sequences of vectors of random numbers of length $\tilde{N}$ for which the empirical distribution function differs – in the sense of supremum norm – too much from the theoretical distribution function of iud random vectors.

The weak law of large numbers tells us that the proportion of sequences of length $\tilde{N}$ that approximate the theoretical distribution function up to an error less or equal $\epsilon > 0$ becomes one as $\tilde{N}$ increases towards infinity.

[3] Here we always assume that the sequence $\mathbf{X}$ is a finite leading segment of an infinite sequence of independent uniformly distributed RVs.

[4] with respect to the measure P

Two problems remain. The first one is a mathematical one. The distribution of d is unknown for dimensions $s > 1$. Even in the case $s = 1$ the distribution of d is only asymptotically independent of the shape of $F_{\mathbf{X}_n}$. The second problem is a computational one. The computation of d is *practically impossible* for dimensions $s > 1$ and interesting sample sizes ($\tilde{N} \approx 2^{10}, \ldots, 2^{30}$) since the complexity of the algorithm is about $\tilde{N}^s$, which is exponential in s.

These problems are solved by a two–step procedure. We first partition the range of the random vectors into a finite number of classes thereby simplifying the calculation of the empirical distribution function, which can be expressed by counting the number of hits for every such class.

In the second step, we calculate a certain distance measure χ which expresses the distance of the observed counters to the expected value of each such counter. The distribution of χ can be calculated approximately and provides a test statistic for iud PRNs.

We will define the partition of $[0,1[^s$ by splitting each coordinate into a finite number[5] α of classes

$$\{\mathcal{M}_0, \ldots, \mathcal{M}_{\alpha-1}\}$$

The components of our random vectors $\mathbf{X}_n$ are the random variables X_i and we will use the following notations in order to denote the partition of these: put

$$\mathrm{A} := \{a_0, \ldots, a_{\alpha-1}\}$$

an alphabet of cardinality α containing a symbol for each class of the partition and let

$$a : [0,1[\to \mathrm{A}, \quad a(x) := a_i, \text{ if } x \in \mathcal{M}_i$$

be a function that maps the range of X_i onto the alphabet. Finally put

$$R_i := a(X_i).$$

(R_i) thus becomes a sequence of abstract random variables from $(\Omega, \mathcal{A}, P)$ to A. We will use the notation

$$r_i := a(x_i)$$

in order to denote the partitioned random numbers.

We now again form vectors using nonoverlapping s–tuples of these symbols setting

$$\mathbf{R}_n := (R_{ns}, \ldots, R_{ns+s-1})$$

and

$$\mathbf{r}_n := (r_{ns}, \ldots, r_{ns+s-1})$$

Note that each vector is an element of A^s. The unit cube $[0,1[^s$ has been partitioned into α^s classes which we denote by

$$\mathcal{M}_{(a_0,\ldots,a_0)}, \mathcal{M}_{(a_0,\ldots,a_1)}, \ldots, \mathcal{M}_{(a_{\alpha-1},\ldots,a_{\alpha-1})}$$

[5] Do not confuse this α with the significance level of a statistical test!

A sample of $\tilde{N}$ such vectors will be denoted by $\mathbf{R}$ and $\mathbf{r}$ respectively.

The partition transforms the multidimensional continuous random variables respectively random numbers into discrete random variables. For realizations $\mathbf{r}_n$ of such random variables the information represented by the empirical distribution function, that is the function

$$\hat{F}_{\mathbf{r}}(\mathbf{t}) := \frac{1}{\tilde{N}} \# \left\{ n, 0 \leq n < \tilde{N} : \mathbf{r}_n \leq \mathbf{t} \right\},$$

where $\mathbf{t} \in \mathrm{A}^s$ and $\mathbf{r}_n \leq \mathbf{t}$ means that

$$\mathbf{r}_n^{(i)} \in \left\{ a_0, a_1, \ldots, t^{(i)} \right\}, \forall i, 1 \leq i \leq s,$$

is equivalent to knowing the 'empirical probability density', that is the set of values

$$\hat{f}_{\mathbf{r}}(\mathbf{t}) := \frac{1}{\tilde{N}} \# \left\{ n, 0 \leq n \leq \tilde{N} - 1 : \mathbf{r}_n \in \mathcal{M}_{(\mathbf{t})} \right\}, \mathbf{t} \in \mathrm{A}^s,$$

since either quantity can be constructed from the other one. We thus can simplify the evaluation of the empirical distribution function to counting hits in classes. The empirical probability density is an estimator for the probability density of the random counters of hits in the classes. This function will be written

$$f_{\mathbf{R}}(\mathbf{t}) := E\left(\frac{1}{\tilde{N}} \# \left\{ n, 0 \leq n \leq \tilde{N} - 1 : \mathbf{R}_n \in \mathcal{M}_{(\mathbf{t})} \right\} \right), \quad \mathbf{t} \in \mathrm{A}^s.$$

We again have to introduce arbitrariness into our test, namely the way in which we split the unit cube. An easy and evident way to do so will be to divide $[0,1[$ into l equals segments of length $1/l$, thereby generating l^s half–open mini–cubes in $[0,1[^s$. The number l has to be chosen small enough in order to allow the evaluation and storage of the empirical probability density but also large enough in order to represent actual applications which will usually work with very fine partitions, if they split the vectors at all.

Since we are working with computers, a quite general and easy to implement way to split the unit cube, including the above one for l of the form $2^{\hat{l}}$, are bit manipulations: we will consider the single random numbers x_i as, say, 32–bit integers, that is, we transform them by multiplying[6] with 2^{32}. Thus, our sequence becomes a sequence of 32 bit numbers.

We now cut out some k different bits of every number. These k bits denote a unique symbol in the set $\mathrm{A} := \{a_0, \ldots, a_{\alpha-1}\}$, where $\alpha := \#\mathrm{A}$ denotes the number of combinations possible with k bits, $\alpha = 2^k$.

This can be described by a function

$$a : [0,1[\to \mathrm{A}, \quad a(x) := a_{k(x)},$$

[6] Of course, this will be done without carrying out the multiplication, but by type conversion from the type float to the type long integer in C, or appropriate types in other languages. The computer has not got anything like real numbers, and the mentioned 32 bits are actually present.

where $k(x)$ is the binary number represented by the k bits selected from the random number x.

The same procedure applied to the sequence (X_i) results in a sequence of independent random variables distributed *uniformly* on the alphabet A since the probability for each symbol is the same:

$$\forall a \in \mathrm{A} : P(a(X_i) = a) = \frac{1}{\alpha}.$$

The described bit manipulations always lead to uniformly distributed symbols, since the whole measure on $[0, 1[$ is divided into two parts of equal measure by every bit in a binary representation.

In the tests presented in Chapter 4, we used one of the following bit manipulations:

- Digit($Start$,$Length$): cut out a segment of bits starting at bit number $Start$ and ending at bit number $Start + Length - 1$. This amounts to split the unit cube into $2^{(Start+Length)\cdot s}$ half open mini cubes with edge length $2^{-(Start+Length)}$. Each class of the partition is formed by 2^{Start} such mini cubes which are spread within $[0, 1[^s$.
- BitStream($Number$,$Length$): cut out $Number$ times a bitblock of length $Length$ of every PRN starting from the first bit. This method is used in order to test more bits of every random number without getting too many classes. On the other hand, this amounts to reducing the number of PRNs involved in a test with a fixed sample size $\tilde{N}$ since only $(\tilde{N} \cdot s)/Number$ PRNs are tested in comparison to the $\tilde{N} \cdot s$ PRNs tested with the Digit method.

If you are familiar with statistical testing, you will already have recognized that we are constructing an usual χ^2 goodness of fit test. Such tests are used to decide[7] whether a set of data arises from a specified probabilistic model, or not. The test is based on two properties of the multinomial distribution:

1. Every random variable Y, discrete or continuous, onedimensional or multidimensional, real valued or abstract, can be reduced to a multinomial random variable by splitting its range into a finite set $\mathcal{M}_t, t \in \mathcal{I}$, of classes and calculating the probabilities $p_t = P(Y \in \mathcal{M}_t)$ that the random variable falls into one of these classes. $\mathcal{I}$ denotes a set of indices. If $\tilde{N}$ measurements of Y are taken then the vector counting the number of realizations that fell into each class is distributed multinomial with the parameters $\tilde{N}$ and p_t. Thus the counters $\hat{f}_{\mathbf{R}}(\mathbf{t})$ for our independent uniformly distributed random vectors $\mathbf{R}_n$ should be distributed multinomial with parameters $\tilde{N}$ and

$$p_{(\mathbf{t})} := \frac{1}{\alpha^s},$$

for every $\mathbf{t} \in \mathrm{A}^s$.

[7] in the statistical sense

2. The distribution of a sort of distance between the empirical probability density of a realization of a multinomial random variable and its theoretical probability density can be calculated at least approximately. The approximation is due to the asymptotic convergence in distribution of the multinomial random variable to a multidimensional normal variate. The main tool within this field is the famous theorem of Pearson, see [27, p.439] for example, which establishes the asymptotic distribution of the distance measure.

Let us introduce the distance measure χ. For every class $\mathcal{M}_{(\mathbf{t})}$ in our partition we compute the distance between the actual number of hits $\tilde{N} \cdot \hat{f}_{\mathbf{r}}(\mathbf{t})$ and the expected number of hits

$$\tilde{N} \cdot f_{\mathbf{R}}(\mathbf{t}) = \tilde{N} \frac{1}{\alpha^s}.$$

A weighted sum of the squares of these distances becomes the overall test statistic. The weights are chosen in order to stabilize the mean and variance of the distribution of the test statistic as a function of $\tilde{N}$. This is needed to get a convergence in distribution which in turn is required since we can only calculate the asymptotic distribution. We therefore define

$$\chi(S) := \sum_{\mathbf{t} \in \mathrm{A}^s} \frac{\left(\tilde{N} \cdot \hat{f}_{\mathbf{r}}(\mathbf{t}) - \tilde{N} \cdot f_{\mathbf{R}}(\mathbf{t})\right)^2}{\tilde{N} \cdot f_{\mathbf{R}}(\mathbf{t})}.$$

This really is something like a distance between the empirical distribution function for $\mathbf{r}$ and the theoretical distribution function of $\mathbf{R}$, since the function becomes zero if and only if the mentioned distance is zero. To see this, observe that

$$\begin{aligned} \chi = 0 \quad &\Leftrightarrow \quad \hat{f}_{\mathbf{r}}(\mathbf{t}) = f_{\mathbf{R}}(\mathbf{t}) \quad \forall \mathbf{t} \in \mathrm{A}^s \\ &\Leftrightarrow \quad \hat{F}_{\mathbf{r}}(\mathbf{t}) = F_{\mathbf{R}}(\mathbf{t}) \quad \forall \mathbf{t} \in \mathrm{A}^s \\ &\Leftrightarrow \quad \sup_{\mathbf{t} \in \mathrm{A}^s} |\hat{F}_{\mathbf{r}}(\mathbf{t}) - F_{\mathbf{R}}(\mathbf{t})| = 0. \end{aligned}$$

The original random numbers $\mathbf{x}_n$ have got more freedom to vary since the partition provides only a reduced amount of information. By this, the empirical distribution function of the $\mathbf{x}$ needs only to be near to the theoretical $F_{\mathbf{X}_n}$ in order to get $\chi = 0$.

The theorem of Pearson states, that the test statistic χ will asymptotically be distributed χ^2 with $\alpha^s - 1$ degrees of freedom if we substitute R for r. We can thus perform a statistical test on any set of random numbers by marking a set of values within the range of such a χ^2 variable as the critical region. The quality of the approximation depends on the expectation of the number of hits in the classes and on the total number of such classes. The higher the smallest of these expectations or the higher the number of classes, the better the approximation. We will return to this question in the next section.

As we have constructed the test in order to exclude random numbers whose empirical distribution function is far from the theoretical one, the critical region should be placed at the utmost end of the values of χ. However, traditional statistics often uses a critical

region including values near zero as well as very high values. A test with such a critical region would exclude random numbers whose empirical distribution function is either too close to or too far from the theoretical one. The construction of the critical region $\mathcal{C}$ as well as the selection of the other parameters $\tilde{N}$, s, m and the set of bits is called test design. It will be covered in the next subsection.

Yet another method is the combination of a χ^2 test statistic with a Kolmogorov– Smirnov test. As has already been mentioned, such a test is used to compare the empirical distribution function of realizations of an onedimensional random variable Y to the distribution function F_Y itself by means of the supremum norm

$$d := \sup_{t \in \mathbf{R}} |F_Y(t) - \hat{F}_{y_1, y_1, \ldots, y_k}(t)|.$$

Provided that the distribution function F_Y is continuous, the asymptotic distribution of $\sqrt{k} \cdot d$ for $k \to \infty$ is the Kolmogorov–Smirnov distribution, which is independent of the shape of F_Y. A good approximation is available for sample sizes $k > 4$ using an approximate distribution function which is given in [40]. See also [17, p.184] for further information. We will denote this test statistic by $KS(\chi)$.

The combined test is performed by first evaluating k times the χ test statistic. On these approximately χ^2 distributed values a final K-S test is calculated. The critical region of the K-S test is set to include high values since we want to exclude sequences of PRNs that do not approximate the χ^2 distribution within a fixed amount of error d. Usually the critical region is set to the interval $[1.58, \infty[$ which has a probability of approximately 0.01 if $k = 32$.

The sample tests in Chapter 4 have all been performed using such combined tests. In our opinion, they are most likely to yield sequences of PRNs that are useful in the class of applications where the user makes not only one but k simulation runs in order to get a good approximation to the distribution of their model.

3.2 Test design

The procedure of selecting the crucial values $(\tilde{N}, s, m, \mathcal{C}, \ldots)$ again amounts to defining numerical properties of the random numbers that pass such a test. There exists no mathematical reason to defend either selection of such parameters in favor to another one as long as we do not have a certain application in mind, see the arguments developed in the first two chapters. Thus, the design should always be made with the application in view.

Put another way, we should not speak of a statistical test unless the test statistic itself, the parameters and the critical region have all been selected. A 'statistical test' then strictly defines certain numerical properties of the numbers that pass the test. By this,

our above test statistic should be written

$$\chi_{(\tilde{N},s,\mathcal{C},\{\mathcal{M}_{(\mathbf{t})}:\mathbf{t}\in \mathrm{A}^s\})} := \sum_{\mathbf{t}\in \mathrm{A}^s} \frac{\left(\tilde{N}\cdot \hat{f}_{\mathbf{r}}(\mathbf{t}) - \tilde{N}\cdot f_{\mathbf{R}}(\mathbf{t})\right)^2}{\tilde{N}\cdot f_{\mathbf{R}}(\mathbf{t})}.$$

The design has to meet some preliminary requirements resulting from the approximative character of the distributions of our test statistics. In order to let the distribution of χ be close to a χ^2, we have to select $\tilde{N}$ and the number of classes α^s in a way that gives an expected number of hits

$$E\left(\tilde{N}\cdot \hat{f}_{\mathbf{R}}(\mathbf{t})\right) = \tilde{N}\cdot f_{\mathbf{R}}(\mathbf{t}) \geq 5 \quad \forall \mathbf{t}\in \mathrm{A}^s.$$

This rule of thumb is very inaccurate but secure. We will be able to observe the approximation quantitatively in Chapter 4. By this, the rule of thumb will not be necessary for us.

If a combined χ – K-S test is used, we also have to use enough samples k in order to get the desired approximation to the K-S distribution. For a detailed treatment of the numerical stability we refer the reader to [40, p.623ff] and [17, p.184].

We have constructed a set of tests build upon the notion of the distance of the empirical distribution function to the theoretical distribution function. But what theory tells us, how do the parameters affect the numerical properties and how are such properties correlated to an actual application? In our opinion, there exist only very few mathematical tools that help the user answer these questions. The mathematicians have to be blamed for ignoring the demands that arise in context with this kind of statistical reasoning.

For the χ^2 goodness–of–fit test there exists a trivial set of applications for which we know the answer to the above question: for a linear transformation of the test statistic χ,

$$G(\mathbf{r}) := c\chi(\mathbf{r}) + d,$$

we know in advance how the test results will affect the results of the application since we can directly calculate the distance between the expectation of G and an evaluation with a sequence $\mathbf{r}$.

Without doubt this result is useless. We just have computed a new test statistic G for which we can calculate the distribution. In such a case, stochastic simulation is no longer necessary since we already have every information on the problem G.

Most simulation problems that arise in real world applications of stochastic simulation seem to avoid the type of G, however, and we are left to stochastic simulation for finding out information on them. In the last chapter we have introduced the strong correlation between integrating sets of well–behaved functions and a test based on the star–discrepancy. This result is a much better generalization than the trivial one mentioned for the χ^2 test since many problems in stochastic simulation amount to calculate expected values of smooth functions. However, the corresponding tests are not within computational reach due to their complexity.

This is the dilemma. On the one hand, there exists a mathematical concept which guarantees good results for certain applications if we provide some information on properties of it and choose PRNs according to a certain test. Unfortunately, the test is not computable for interesting sizes of s and N.

On the other hand, there exist tests that are computable. But only very few results[8] are known that provide error estimates for a class of applications that occurs in reality.

For the moment, we can only try to make the test as general as possible in order to include more applications, or approximations to applications at least. An important step towards this direction is to admit sequences $\mathbf{r}$ resulting from partitioning *overlapping* s–tuples of random numbers x_i. Since the random variables $\mathbf{R}_n$ are not independent in this case we will have to recalculate the distribution of the test statistic. However, only a little modification of the test statistic will lead to the so–called *overlapping M–tuple test*[9], which again is distributed χ^2. This test has been proposed by Marsaglia in [31]. We will give a formal proof of the distribution of this test statistic in Chapter 5.

The modification takes into account that the vector of counters for s–tuples resulting from overlapping tuples lives only on an affine subspace of $\mathbf{N}^{\alpha \cdot s}$. Some components linearly depend on other components. The vector will thus not be as close to its expectation as in the nonoverlapping case. By this the distance measure χ would yield values that are too large. We have to subtract a payoff for the overlapping tuples in order to get a χ^2 distributed value. The number of degrees of freedom also is reduced as a consequence of the dependence structure. For the 'o'verlapping M–tuple test we have

$$\begin{aligned} \chi_o(\mathbf{r}) \quad := \quad & \sum_{\mathbf{t} \in \mathrm{A}^M} \frac{(\tilde{N} \cdot \hat{f}_{\mathbf{r}}(\mathbf{t}) - \tilde{N} \frac{1}{\alpha^M})^2}{\tilde{N} \frac{1}{\alpha^M}} - \\ - \quad & \sum_{\mathbf{t}' \in \mathrm{A}^{M-1}} \frac{(\tilde{N} \cdot \hat{f}'_{\mathbf{r}'}(\mathbf{t}') - \tilde{N} \frac{1}{\alpha^{M-1}})^2}{\tilde{N} \frac{1}{\alpha^{M-1}}}, \end{aligned}$$

where the second sum is the χ value for the same sequence of random numbers r_i, this time forming $(M-1)$–tuples $\mathbf{r}'_n$ and counting their number of occurrences.

χ_o is asymptotically distributed with $\alpha^M - \alpha^{M-1}$ degrees of freedom. The single classes are now counted from the sequences

$$\begin{aligned} \mathbf{r} \quad &:= \quad \mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_{\tilde{N}-1} \\ \mathbf{r}' \quad &:= \quad \mathbf{r}'_0, \mathbf{r}'_1, \ldots, \mathbf{r}'_{\tilde{N}-1}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{r}_n \quad &:= \quad (r_n, \ldots, r_{n \oplus M \ominus 1}) \\ \mathbf{r}'_n \quad &:= \quad (r_n, \ldots, r_{n \oplus M \ominus 2}), \end{aligned}$$

[8] Consider the weighted spectral test, Hellekalek [19], for example.

[9] M stands for the dimension of the tuples and thus has the same meaning as s in the χ^2 goodness–of–fit test.

and

$$x \oplus y := (x + y) \bmod \tilde{N}, \quad x \ominus y := (x - y) \bmod \tilde{N}.$$

Be aware of the fact that we only use the first [10] $\tilde{N}$ elements of the sequence (r_i). We will examine the test statistic in detail in Chapter 5.

The generalization to overlapping tuples makes the χ^2 test a very flexible tool since many applications have a similar structure. Further generalizations would affect the way the unit cube in $\mathbf{R}^M$ is partitioned. Such tests then could be fitted to even a larger set of application problems. We are not yet aware of a strong correlation to an appropriate class of applications, however. We consider the development of such a class as being very important for applying test results obtained with the overlapping M–tuple test.

Such a correlation would also give methods for an appropriate design of the test. Since we cannot rely on such hints, we give the following, somewhat natural suggestions.

- Let s or M be a dimension, and $\tilde{N}$ be a sample size that are equal to the dimension and amount of vectors of random numbers within the application. Attention: for the nonoverlapping χ^2 test, the number of PRNs scanned by the test will be of order $s \cdot \tilde{N}$, whereas only $\tilde{N}$ numbers are tested by the χ_o test.

- Select the classes in a way that reflects the problem structure of the application. If the number of significant bits in the application seems to be high and the partition thus becomes too fine e.g. the number of components in the counter vector becomes too large, think of using the *BitStream* instead of the *Digit* method.

- Be aware that the approximation of the distribution of χ by the χ^2 distribution becomes better if the expected value in each class becomes larger. Thus you will have to accept higher values of $\tilde{N}$ for more classes α. A commonly used rule of thumb is to choose $\tilde{N}$ such that the expected value in each class is at least 5. If you do not want to trust such rules of thumb, observe the whole behavior of the test statistic as a function of $\tilde{N}$ for $\tilde{N}$ going from zero to the desired size. For examples refer to Chapter 4.

- Compare the resulting value of the test statistic to its range. If you need PRNs that approximate each class as good as possible, the value should be close to zero and $\mathcal{C}$ will include higher values. If you want to get average behavior, $\mathcal{C}$ should be set to values far from the expectation.

- If you want to approximate the distribution of your application model, a test should be made by combining the χ test with a K-S test. The number of samples k of this test should equal the number of runs of your stochastic simulation. Put $k > 4$ in order to get good approximation with the formula given in [40, p.624].

[10] In this case $\tilde{N}$ equals N.

3.3 Interpretation of the results

One main problem within the field of empirical and theoretical testing of PRNs is the proper interpretation of the test results. We will discuss the results of the tests we have constructed and applied to our battery of generators in Chapter 4. However, this section contains some "beware of's" and sums up the whole chapter in short.

- The result of applying a test to a certain sequence of PRNs is the information whether the result of the test statistic lies in the critical region or not. Since we have seen the arbitrariness involved in the selection of a critical region, every test report should include the absolute value of the test statistic too[11]. Using this value, subsequent users can predict the performance of the sequence in their own tests by simply comparing the result to the critical region they use.

- Results within the critical region can be due to the approximative character of the distribution that has been given for the test statistics. In order to check if the approximation is valid, we will use "bath-tub" kind of graphics performing the same test for different sample sizes $\tilde{N}$. See also the corresponding section in Chapter 4.

- If a generator fails the test and the result is not due to bad approximation we can conclude that the generator will also fail in simulating *correlated* problems. The result of a test can never be an assessment on PRNs *without* referring to a class of applications.

- If you use combined tests like a $KS(\chi)$, also check the single outcomes of the sub–test statistic χ. A possible graphical rendering of such values is given in Chapter 4 using gray–scale plots.

- By calculating the so–called upper tail probability $U = P(X > t)$ for a realization t of X, you can make results of different tests comparable. The results become easier to plot since the range of U is exactly $[0, 1]$ and also easier to interpret since the distribution of the upper tail statistic is always $U \sim U[0, 1]$. We will use the upper tail probability in order to plot gray scale plots for the single χ values within a K-S test.

- In Chapter 4 we will see that different generation methods lead to different advantages and disadvantages even within the same test setup. By this, the importance of carefully choosing the *design* of the test is stressed again.

- In the whole field of PRN generation one speaks of 'good' and 'bad' algorithms. This is true in the empirical context since generators that pass the given tests have outperformed other generators in many interesting stochastic simulations. This correlation has *not* been proven mathematically, however. Without such a correlation, the first 'natural' simulation problem that yields unwanted results

[11] This and also further guidelines for reporting results in connection with computer based statistical testing can be found in [23].

with the 'best' known generation methods could be found tomorrow. By this, skepticism should always accompany stochastic simulation:

- test different generators,
- use different generators, and
- compare the results.

Always keep in mind that you are working with random variables. You have to find out information on their distribution! 'Typical behavior' in the sense defined by the expectation of a RV is *only one property* of such a distribution.

3.4 pLab

In this section we want to introduce a software system that has been developed by the pLab group at the university of Salzburg in Austria. The main design and programming tasks have been performed by Leeb who gives an overview of the system in [29]. The system keeps on growing since new theoretical result are incorporated continuously. At the current state, we are able to perform a huge number of tests on very different generation processes for PRNs including LCG, ICG, EICG, and compound generators. The software is currently running in C++, Mathematica and Smalltalk versions.

All the tests in Chapter 4 have been evaluated using this software. We strongly recommend every user of PRNs to test their PRNs with such a tool. The pLab software is especially useful since it will be available via internet. Refer to our WEB server [2] for further information.

This is a small example of a test with pLab. We use the C++ version. The test calculates the value of the overlapping M–tuple test for a BitStream method cutting out 3 times a block of 3 bits of every PRN. The dimension varies from 2 to 7, the generator varies within our test battery. The sample size $\tilde{N} = N$ is set such that the expected hits in every class are 6. We feed 32 samples of the overlapping M–tuple statistic into the K-S test. The resulting value is printed to a file with extension '.ks'.

```
[...]

gens[0] = (PrimaryProcess*) new
     EICG ((unsigned int)(pow(2,31)-1), 1, 1, 0);
gens[1] = (PrimaryProcess*) new
     EICG ((unsigned int)(pow(2,31)-1), 7, 0, 0);
gens[2] = (PrimaryProcess*) new
     ICG ((unsigned int)(pow(2,31)-1), 1, 1, 0);
gens[3] = (PrimaryProcess*) new
     LCG ((unsigned int)pow(2,31)-1, 950706376, 0, 1);
```

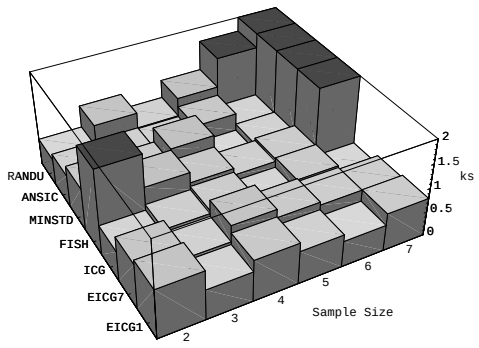
```
gens[5] = (PrimaryProcess*) new
     SystemGenerator;
gens[4] = (PrimaryProcess*) new
     LCG ((unsigned int)pow(2,31)-1, 16807, 0, 1);
gens[6] = (PrimaryProcess*) new
     LCG ((unsigned int)pow(2,31), 65539, 0, 1);

for (dim=2;dim<8;dim++)
     for (i=0;i<GENERATORS;i++)
            {

//                  ( master, number of blocks, length )
            BitStream digit(*gens[i],3,3);
            points = (unsigned int)((double) 6.0 * (pow((double)2.0,
                     (double)3.0 * (double)ms)));
            EMTupleStatistic tstat(digit,dim,points);
            KSStatistic ks(tstat,32);
            Printer prn_ks(ks,ks_output);
            prn_ks.produceValues(1);

            };
```

This is a Mathematica plot of the values in the '.ks' file. We will interpret such results in the next chapter.



Overlapping M–tuple sample test.

Chapter 4

Results

There are three kinds of lies: lies, damned lies and statistics.
– Benjamin Disraeli

This chapter contains results of the overlapping M–tuple test which has been applied to PRNs resulting from the following generators that have been described in Chapter 1.

- $EICG(2^{31}-1,1,1,0)$, short "EICG1"
- $EICG(2^{31}-1,7,0,0)$, short "EICG7"
- $ICG(2^{31}-1,1,1,0)$, short "ICG"
- $LCG(2^{31}-1,950706376,0,1)$, recommended by Fishman and Moore in [14]
- $LCG(2^{31},1103515245,12345,12345)$, the ANSI C system generator
- $LCG(2^{31}-1,16807,0,1)$, the "minimal standard", see [39]
- $LCG(2^{31},65539,0,1)$ known as "RANDU", see [39]

The tests are grouped into three sections. The first section demonstrates the phenomenon of the 'bath–tub'.

The second section contains tests that directly show the effect of the lattice structure of LCGs by partitioning the unit cube finer than the resolution of the lattice on which the PRNs of LCGs lie. The partition is achieved by using bit blocks at later positions in the single random numbers. The tests show, that inversive methods and LCGs with a known good lattice structure pass such tests, whereas linear methods with coarse lattice structure fail. The tests are somewhat unnatural in that usual applications will rarely use such combinations of bits.

The third section contains tests that also show different behavior of linear and inversive methods. However, this time the tests are constructed in a very natural way dealing only with the first bits of every random number. We consider these results to be the most exciting and important ones.

In all the tests we use the following notations: a K-S test applied to k samples of an overlapping M–tuple test with sample size N, dimension M and one of the two partition methods[1]

$$BitStream(Number, Length)$$

or

$$Digit(Start, Length)$$

is denoted by

$$KS(\chi_o(Dim, Method, N), k)$$

The parameters in the method determine the number of classes for which we count hits. The number of classes calculates to

$$2^{Dimension \cdot Length}$$

for both methods.

In all our tests we have set k to 32. In the plots of the K-S test statistic values within the critical region are colored red in Postscript. The critical region has been set to include high values $KS > 1.58$ such that the significance level of the test is approximately $\alpha = 0.01$.

Pay attention to the fact that values of $KS > 2$ are cut within the graphic. The worst values in the test statistics have been around 5! But since the K-S distribution assigns only a very small total weight to such values, we have felt free to cut them off.

The sample size N will either be included in the description of the test or in the graphic itself. In the later case we use scientific notation, e.g. '26' means $N = 2^{26}$. Keep also in mind that the actual number of PRNs that is required from the generator for every χ_o value is $N + Dimension$ for $Digit$ and $N/Number$ for $BitStream$.

For most tests we have also provided a plot of the 32 χ_o–values from which a single KS value has been calculated. The χ_o values have been normalized by calculating the upper tail probability $P(Y > \chi_o)$ for a random variable $Y \sim \chi^2$. In the graphic for these upper tail probabilities, which we denote by

$$U(\chi_o(Dim, Method, N), k),$$

we have always grouped the 32 U–values within a small rectangle that has the same position within the whole graphic as has the bar for the KS values in the plot of these.

[1] These have been described in Chapter 3

4.1 Bath–tub

The following pictures demonstrate the principal empirical behavior of generators in a

$$KS(\chi_o(Dim, Method, N), k = 32)$$

test. Every graphic sums up the values of the test statistic KS for every generator in the test battery and for different sample sizes N ranging from 2^3 to 2^{26}. The basic observations are:

- For low values of N we know that the theoretical distribution of χ_o, F_{χ_o} is far from the asymptotic χ^2. The KS statistic thus yields high values even if the PRNs have an empirical distribution function near to F_{χ_o}. The supremum norm used in the K–S test statistic also permits the situation of a generators empirical distribution function being close to the asymptotic χ^2 and yielding good[2] KS values although the distance to the theoretical F_{χ_o} is far and should lead to bad KS values.

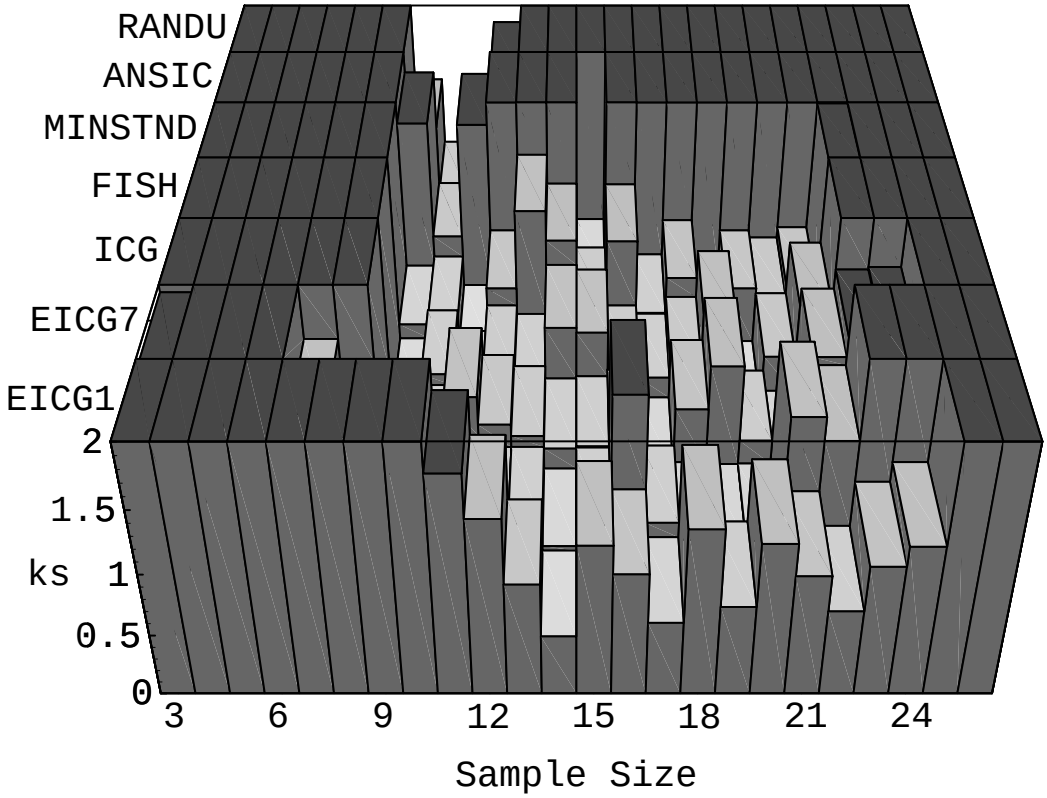
- The region of bad approximation is marked by constant values of the U statistic near to $1/2$ indicated by gray rectangles in the plots.

- Once having entered the range of sample sizes for which the approximation by the χ^2 distribution is good, a generator gives reasonable values of the KS statistic until some upper bound which differs from generator to generator.

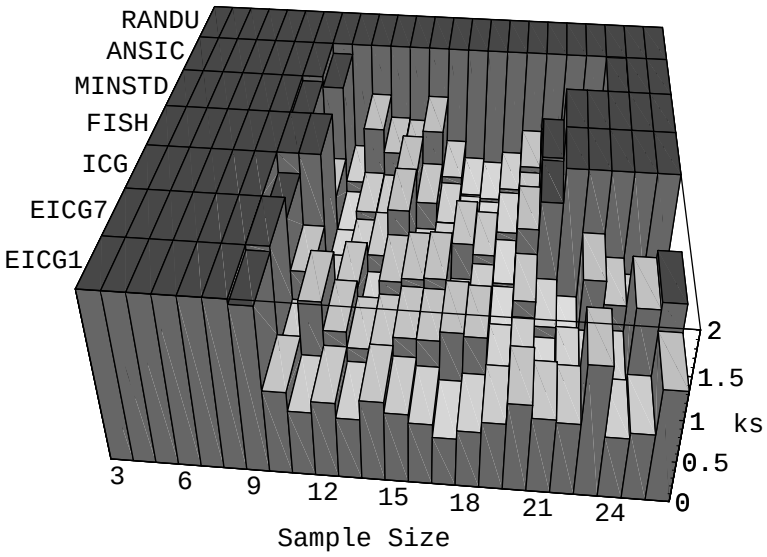
- We thus can speak of a 'bath–tub' whose bounding values determine the range of sample sizes N for which we can recommend using the generator if an application is correlated to the $KS(\chi_o(Dim, Method, N), 32)$ test statistic.

The first example shows the performance of the generators in the battery within an $KS(\chi_o(5, BitStream(8, 4), N), 32)$ test. The second example uses $Digit$ instead of $BitStream$ testing only the first 4 bits of every PRN. For the second example we provide the plot of the corresponding U statistic, too.

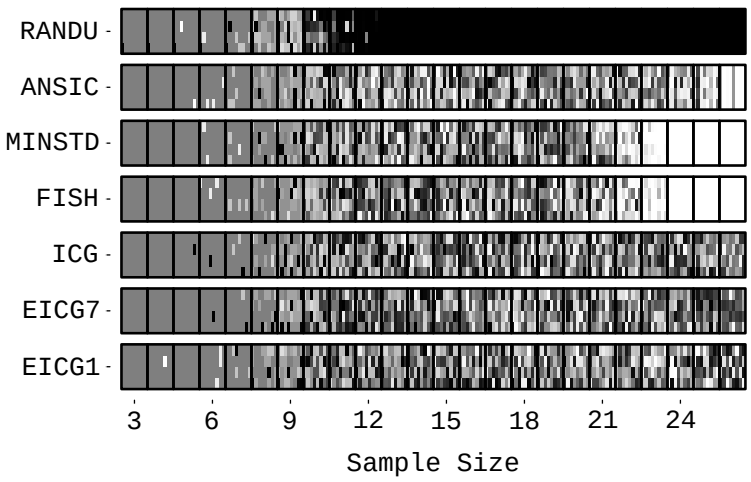
[2] that is, small



Example 1: $KS(\chi_o(5, BitStream(8,4), N), 32)$



Example 2a: $KS(\chi_o(5, Digit(1,4), N), 32)$

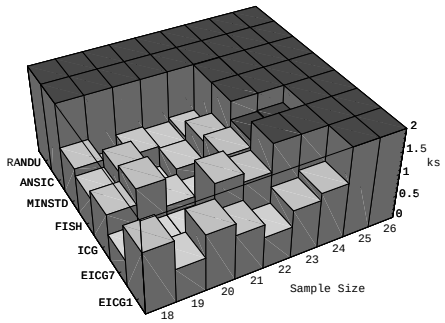


Example 2b: $U(\chi_o(5, Digit(1,4), N))$

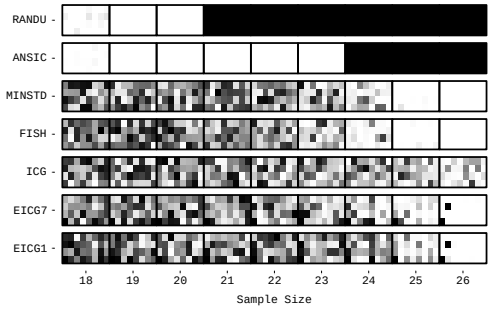
The following test uses 32 bits of every PRN. This is definitively more than most applications will do[3]. The test can be used to illustrate the reason for the bad performance of the generators for high sample sizes $N \simeq 2^{26}$. The MINSTND for example has a good empirical distribution of the $U(\chi_o)$ values for sample sizes $N < 2^{23}$. The gray–scale plot shows a good approximation to the uniform distribution for such sample sizes. Then something strange happens: the U values get closer and closer to one; For $N = 2^{26}$ we have an almost white rectangle in the graphic. This means that the probability of getting χ_o values higher than the those resulting from a sequence of PRNs produced by the MINSTND gets close to one. In other words, we have produced very small χ_o values. This means that the empirical distribution function of overlapping 5–tuples is very close to the theoretical one.

The classes have all been hit by about the expected number of hits. The generator has lost the ability to 'randomize' this counters. This could be due to the periodic structure of the generators since the same behavior is observed for linear and inversive methods.

RANDU even fails the test for sample sizes $N = 2^{18}$. This can theoretically been explained with the very short period of the last bits of PRNs from this generator[4]. However, as N gets larger, the U values suddenly jumps to zero indicating that the χ_o statistic has yielded very large values. The hits in the classes are far from the expected number. In our opinion, this results from the interference of the regular lattice of RANDU and the regular partition formed by *BitStream*. Some classes contain to many points of the lattice. As N gets larger, the counter for such classes will always be higher than their expectation.



Exp. 3: $KS(\chi_o(5, BitStream(8,4), N), 32)$



$U(\chi_o(5, BitStream(8,4), N))$

4.2 Uncovering lattice structure

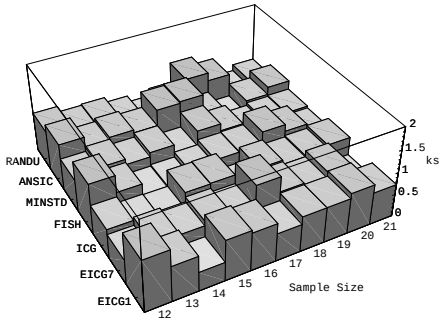
The next tests show the influence of the number of bits in each PRN that are tested. In order to keep the number of classes within reasonable bounds, we use the *BitStream*

[3] In an application that treats the PRNs as real numbers the user has to decide whether later bits will influence the results of the stochastic model. We have already mentioned the importance of such considerations.

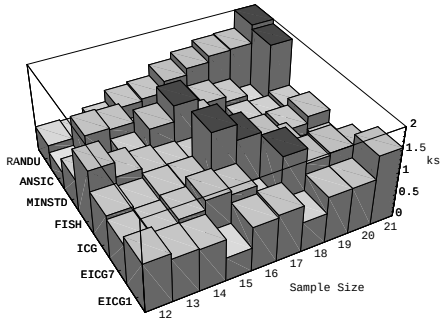
[4] See Altman [1] for a detailed discussion.

method. The effect of testing more bits can be observed by comparing the graphics for different values of the *number* parameter in the *BitStream* method. The higher this value is set, the more bits of every PRN are tested. The edge length of the finest partition imposed on the unit cube calculates to $2^{-Number \cdot Length}$.

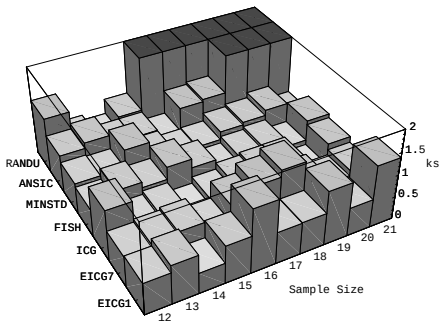
The test actually reveals the bad lattice structure of the RANDU and ANSI–C generators in dimension 3 by partitioning the unit cube into classes that are finer than the resolution of the lattice of these generators. The number of hits in some of the classes will thus always be zero, yielding the 'bad' KS–values.



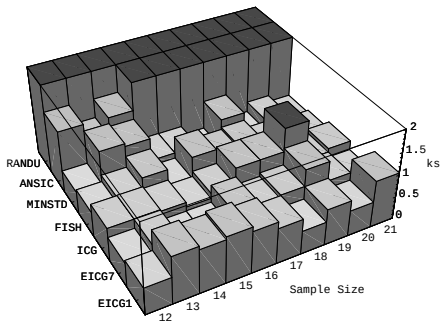
$KS(\chi_o(3, BitStream(4,4), N), 32)$



$KS(\chi_o(3, BitStream(5,4), N), 32)$



$KS(\chi_o(3, BitStream(6,4), N), 32)$



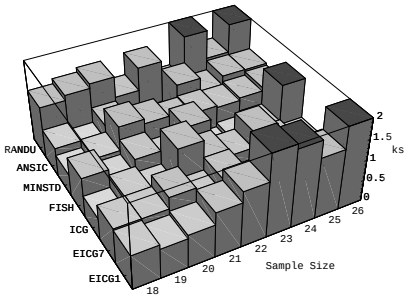
$KS(\chi_o(3, BitStream(7,4), N), 32)$

4.3 Inverse methods vs. linear methods

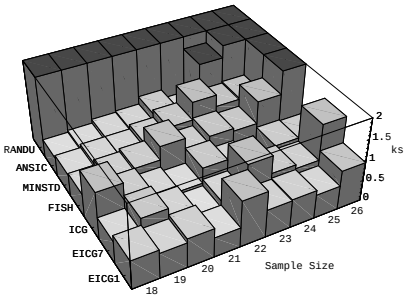
In this final section we want to introduce results that we consider being of utmost relevance for any application of PRNGs. The tests use the first 4 bits of every PRNs and have been performed in dimensions $2, 3, 4$ and 5. The unit cube is partitioned using $Digit(1, 4)$.

The tests show that *we have found a model that distinguishes between linear and inversive methods*: even LCGs with a very good lattice structure fail the tests in dimension 4 and 5 and sample sizes $N \geq 2^{23}$. This result is due to the overall regularities inherent in the sequences of PRNs that emerge from LCGs. In our opinion it can be explained by the interference between the lattice and the way *Digit* partitions the unit cube. This interference leads to the defects we measure with the KS–value. Take a look

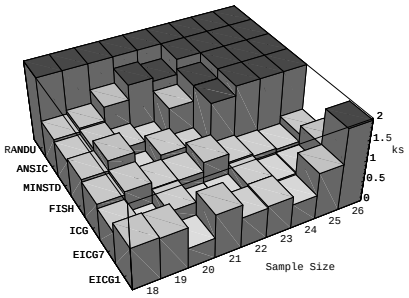
at the following figures. You can observe the usual bad performance of RANDU in 3 dimensions. The important result however is the general behavior of LCGs and inversive methods in the last two plots.



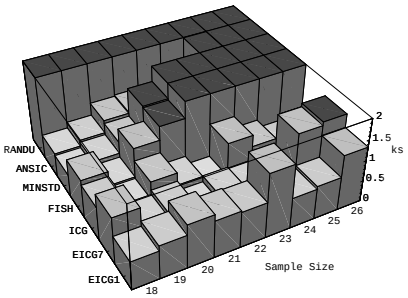
$KS(\chi_o(2, Digit(1,4), N), 32)$



$KS(\chi_o(3, Digit(1,4), N), 32)$

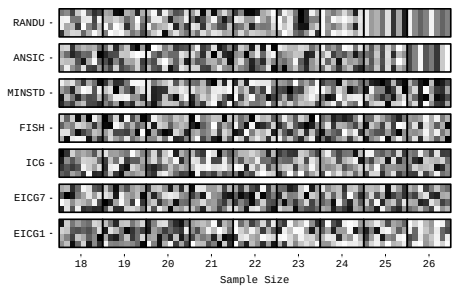


$KS(\chi_o(4, Digit(1,4), N), 32)$

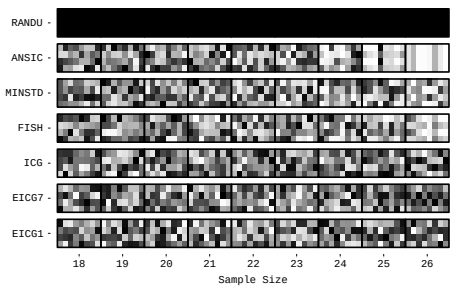


$KS(\chi_o(5, Digit(1,4), N), 32)$

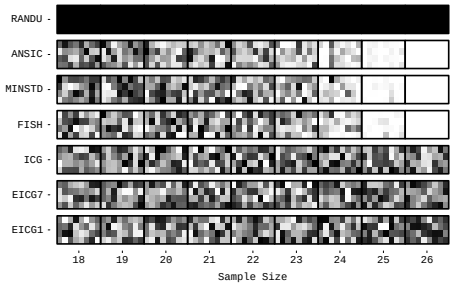
We also provide the values of the corresponding $U(\chi_o)$ statistics:



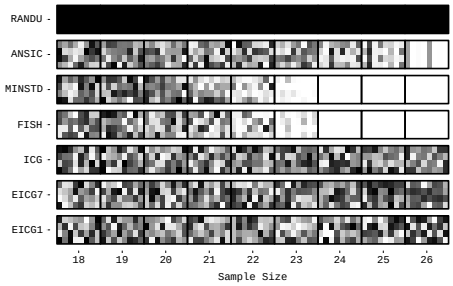
$U(\chi_o(2, Digit(1,4), N))$



$U(\chi_o(3, Digit(1,4), N))$



$U(\chi_o(4, Digit(1,4), N))$



$U(\chi_o(5, Digit(1,4), N))$

> The graphics again show the problems we get when using LCGs and large sample sizes. Regularities are present in every sequence of PRNs. The question is always if the actual application detects these regularities and yields undesired results. We did not encounter a natural test statistic that detects such regularities within the class of inversive generation methods, however. The given overlapping M–tuple test is natural with respect to the way we partition the random numbers and choose the dimension, the sample size and the number of samples. The notion of distance between empirical and theoretical distribution function has been chosen with respect to a wide class of applications. This makes the results remarkable.

We can construct models that detect the regularities in inversive PRNs too. Consider for example a construction that uses inversive operations in order to select a subsequence of PRNs. Such a sequence will be rejected by a large class of statistical tests. We do not expect real world simulations to have such a structure, however.

> LCGs produce sequences of valid RNs. However, the structure of many application will detect the sort of regularities inherent in such sequences. In order to enhance the confidence in results of any application, linear *and* inversive methods should be applied.

We finish this chapter with a short summary.

4.4 Summary: Different PRNGs and flexible tests

- There exist two kinds of simulation problems that *do not* lead to the *desired* results when applying LCGs:
  - Simulation problems that are *sensitive* with respect to the *lattice structure* of LCGs in the sense, that they detect the regions which will never be hit by a point of the generator.
    Since *inversive methods do not have a lattice* structure, generators like EICG and ICG can lead to appropriate results.
  - Simulation problems that detect the lattice structure of LCGs when using a *huge amount of PRNs* due to effects causing from the superposition of a

regular partitioning of $[0,1[^s$ and the lattice of the PRNs. *Inversive methods* seem to be much *more stable*.

- *Inversive methods do not lead to significant higher cost of a simulation*, because the additional effort to produce the PRNs is small in comparison to the evaluation of the simulation problem itself. The evaluation of the highest sample size in the test $KS(\chi_o(5, Digit(1,4), N), 32)$ for the slowest generator, EICG7, took 44335 seconds. The fastest generator, ANSIC, took 18736 seconds[5]. Note, that the calculation of the test statistic itself is very simple and consumes only a negligible amount of time. The scrambling of the PRNs in typical stochastic simulations will likely contribute a larger amount of time to the overall computations. The different time complexity of the generators will thus be of less importance.

- The M–tuple test provides a *very broad concept* for scrambling numbers in a way *many simulations* will do.

- *In all the tests* that have been performed yet, *inversive methods have never led to results worse than linear methods*. It is clear that there exist problems for which linear methods will lead to better results, but we expect the problems to have a structure which is different from that of usual stochastic simulations.

- The bath–tub is used instead of rules of thumb. A generator should be *judged by the length and placement* of the sample sizes, for which the desired results are obtained *instead of using rules of thumb* like recommending setting the expected hits to 5 or using at most about the square root of the period PRNs from the generator. This second recommendation has been shown to be empirically false in many cases for both types, inversive generators as well as some LCGs.

Based on these arguments we recommend every user of PRNs to

- *test* PRNs with sample size and problem structure *similar* to that of the *simulation* problem!

- use more *sophisticated statistical* tests like the M–tuple test, since they can be adapted to a broader class of simulation problems than usual χ^2–tests!

- use *different generation methods*, including inversive generators, to verify the simulation results!

[5] The time statistic was measured on a DEC 3000 alpha workstation.

Chapter 5

The distribution of χ_o

Die Mathematiker –
sie tun das der inneren Schönheit der Dinge wegen.
– Persi Diaconis

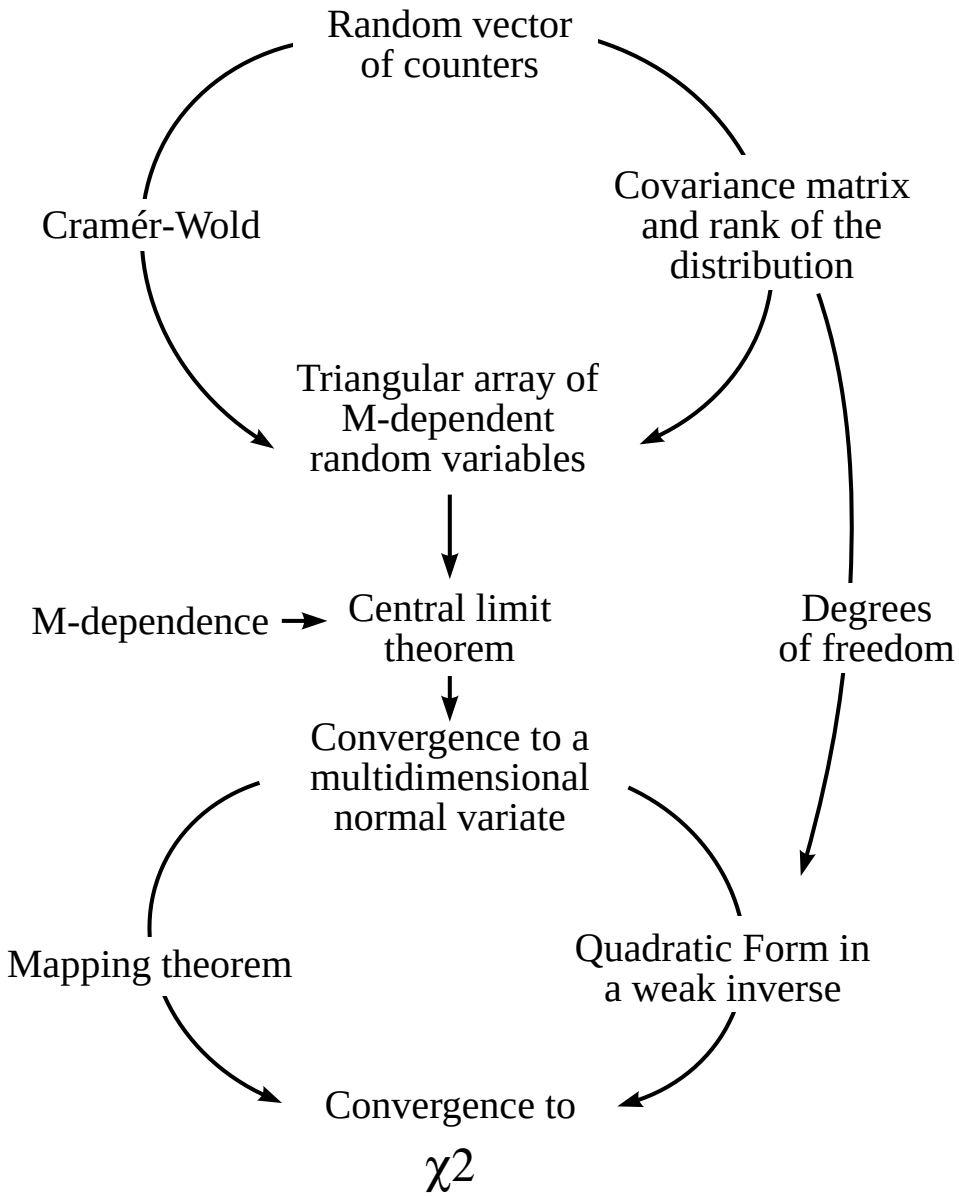
In this chapter we will provide a detailed analysis of the distribution of the random variable χ_o which has been introduced in Chapter 3.

We will first examine the role of the covariance matrix and give a geometrical interpretation of it. We then define the multidimensional normal variate and establish some basic properties. The M–tuple test statistic is computed from special counter variables, for which we calculate the covariance matrix and the asymptotic distribution, which will be multivariate normal. We show that the covariance of the counters for *overlapping* M–tuples causes this asymptotic distribution to *degenerate* to a lower dimension.

In order to construct a useful test statistic for the degenerate multivariate normal distribution we compute a onedimensional fingerprint which is related to the distance between theoretical and empirical distribution function of the counter vector. The transformation is based on the concept of weak inverses. We will study the distribution of quadratic forms of normally distributed random vectors in such weak inverses. Putting all together, we can show that the asymptotic distribution of the M–tuple statistic is χ^2.

The technique used to derive the asymptotic distribution of the test statistic can be extended to a large class of statistical problems. It applies whenever a test statistic is made in the form of counting events that arise from successive, almost independent trials.

The following flow chart can be used to follow up the dependencies of the single results within this chapter.



5.1 The covariance matrix

Throughout this section $\mathbf{X}$ denotes a random vector $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)})^T$ defined on the probability space $(\Omega, \mathcal{A}, P)$.

We assume that the reader is familiar with the concept of moments in the onedimensional

case. The k'th moment of the random variable $\mathbf{X}^{(i)}$ is defined by

$$M^k\left(\mathbf{X}^{(i)}\right) := \int_\Omega (\mathbf{X}^{(i)})^k(\omega)dP,$$

where we denote the first moment, e.g. the expectation, by $E(\mathbf{X}^{(i)})$. The notion of expectation simply extends to the multidimensional case by

$$E(\mathbf{X}) := \left(E(\mathbf{X}^{(1)}), \ldots, E(\mathbf{X}^{(n)})\right)^T$$

The amount in which the values of $\mathbf{X}^{(i)}$ differ from their expectation is measured with the second *central* moment or 'variance'

$$V(\mathbf{X}^{(i)}) := \int_\Omega \left(\mathbf{X}^{(i)} - E(\mathbf{X}^{(i)})\right)^2 dP = E\left(\left(\mathbf{X}^{(i)} - E(\mathbf{X}^{(i)})\right)^2\right)$$

The variance is a very important characterization of a random variable. The Chebysev inequality for example shows the relation between the variance and the probability that the random variable will assume a value differing from its expectation by more than a given distance ϵ:

$$P\left(\left|\mathbf{X}^{(i)} - E(\mathbf{X}^{(i)})\right| \geq \epsilon\right) \leq \frac{V(\mathbf{X}^{(i)})}{\epsilon^2}$$

In order to generalize this concept to higher dimensions we have to calculate not only the variance of the single components of the vector, but the whole $n \times n$-matrix

$$V(\mathbf{X}) := \begin{pmatrix} v_{11} & v_{12} & \cdots \\ v_{21} & v_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

where

$$v_{ij} := E\left((\mathbf{X}^{(i)} - E(\mathbf{X}^{(i)}))(\mathbf{X}^{(j)} - E(\mathbf{X}^{(j)}))\right) = Cov(\mathbf{X}^{(i)}, \mathbf{X}^{(j)})$$

This matrix contains important information on the random vector $\mathbf{X}$ and has some remarkable properties.

Lemma 5.1 *The covariance matrix is symmetric.*

This is a trivial consequence from the definition and the fact that the product is commutative.

Lemma 5.2 *The quadratic form in the covariance matrix is positive semidefinite.*

Proof: Let $\mathbf{t} \in \mathbf{R}^n, \mathbf{t} = (\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(n)})^T$. If we define a new onedimensional random variable

$$Y := \sum_{i=1}^{n} \mathbf{t}^{(i)}\mathbf{X}^{(i)}$$

the expectation of Y is according to the linearity of E

$$E(Y) = \sum_{i=1}^{n} \mathbf{t}^{(i)} E(\mathbf{X}^{(i)})$$

The variance of Y resolves to

$$\begin{aligned} V(Y) &= E(Y - E(Y))^2 \\ &= E\left(\sum_{i=1}^{n} \mathbf{t}^{(i)}\mathbf{X}^{(i)} - (\sum_{i=1}^{n} \mathbf{t}^{(i)} E(\mathbf{X}^{(i)}))\right)^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{t}^{(i)}\mathbf{t}^{(j)} E\left((\mathbf{X}^{(i)} - E(\mathbf{X}^{(i)}))(\mathbf{X}^{(j)} - E(\mathbf{X}^{(j)}))\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}\mathbf{t}^{(i)}\mathbf{t}^{(j)} \\ &= \mathbf{t}^T V \mathbf{t}, \end{aligned}$$

where $V = (v)_{ij} := V(\mathbf{X})$. The quadratic form in the covariance matrix is thus equivalent to the variance of the random variable Y. Therefore, it is nonnegative. □

Consider now the following interpretation of an onedimensional random variable Y with zero variance:

$$V(Y) = 0 \Leftrightarrow P(Y = E(Y)) = 1 \tag{5.1}$$

In other words: Y is a random variable from $\Omega \to \mathbf{R}$ with $E(Y) = \mu$ and $V(Y) = 0$, if and only if the probability $P(Y = \mu) = 1$. The relation can be shown by observing at first

$$\int_\Omega Y^2 dP = \int_{\{Y=E(Y)\}} Y^2 dP + \int_{\{Y\neq E(Y)\}} Y^2 dP$$

Now, if $P(Y = E(Y)) = 1$, the second integral is defined on a set of P–measure zero and its integrand can be set to an arbitrary finite value, say $E^2(Y)$, giving

$$\int_\Omega Y^2 dP = \int_{\{Y=E(Y)\}} E^2(Y) dP + \int_{\{Y\neq E(Y)\}} E^2(Y) dP$$

We thus have shown that

$$\int_\Omega Y^2 dP = \left(\int_\Omega Y dP\right)^2$$

and therefore we have $V(Y) = \int_\Omega Y^2 dP - (\int_\Omega Y dP)^2 = 0$. If, on the other hand, $V(Y) = 0$, then, assuming a set A that has P–measure $P(A) > 0$ and $Y(a) \neq E(Y)$ for every $a \in A$, this implies

$$E(Y - E(Y))^2 = \int_\Omega (Y - E(Y))^2 dP \geq \int_A \underbrace{(Y - E(Y))^2}_{>0} > 0$$

contradicting $V(Y) = 0$.

Now, looking for a multidimensional analogue of this interpretation of zero variance, let $\mathbf{t}_1 \in \mathbf{R}^n, \mathbf{t}_1 = (\mathbf{t}_1^{(1)}, \ldots, \mathbf{t}_1^{(n)})^T$ and define the onedimensional random variable Y_1 as an inner product:

$$Y_1 := \sum_{i=1}^{n} \mathbf{t}_1^{(i)} \mathbf{X}^{(i)} = \mathbf{t}_1^T \cdot \mathbf{X}$$

Now suppose that Y_1 has a vanishing variance, i.e. $V(Y_1) = 0$. Then we have by (5.1)

$$\begin{aligned} P\left(Y_1 = E(Y_1)\right) &= P\left(\sum_{i=1}^{n} \mathbf{t}_1^{(i)} \mathbf{X}^{(i)} = \sum_{i=1}^{n} \mathbf{t}_1^{(i)} E(\mathbf{X}^{(i)})\right) \\ &= P\left(\underbrace{(\mathbf{X} - E(\mathbf{X}))^T \cdot \mathbf{t}_1 = 0}_{(*)}\right) = 1 \end{aligned} \tag{5.2}$$

From a geometrical point of view zero variance of Y_1 means that, with probability one, the random vector $\mathbf{X}$ lies on the hyperplane ϵ_1 defined by

$$\epsilon_1 := \{\xi \in \mathbf{R}^n : (\xi - E(\mathbf{X}))^T \cdot \mathbf{t}_1 = 0\}$$

If on the other hand $P(\mathbf{X} \in \epsilon_1) = 1$, then the variance of the random variable $\mathbf{X}^T \cdot \mathbf{t}_1$ has to vanish, yielding $V(\mathbf{X}^T \cdot \mathbf{t}_1) = \mathbf{t}_1^T V \mathbf{t}_1 = 0$, where $V = V(\mathbf{X})$ again. We thus have the equivalence

$$P(\mathbf{X} \in \epsilon_1) = 1 \Leftrightarrow \mathbf{t}_1^T V \mathbf{t}_1 = 0$$

Some little calculations will show that the right hand side of this equivalence can be reduced even to a linear form. Let us now denote $E(\mathbf{X}^{(j)})$ by $E^{(j)}$ for convenience. If $(*)$ in (5.2) is fulfilled and $\mathbf{t}_1 \neq \emptyset$, we can multiply both sides by $\mathbf{X}^{(j)} - E^{(j)}$, where $j \in \{1, \ldots, n\}$. We then get

$$P\left((\mathbf{X}^{(j)} - E^{(j)})(\mathbf{X} - E(\mathbf{X}))^T \cdot \mathbf{t}_1 = (\mathbf{X}^{(j)} - E^{(j)}) \cdot 0\right) = 1 \tag{5.3}$$

If for any integrable random variable Z, $P(Z = 0) = 1$, then the expectation of Z has to equal zero. Taking the expectation in (5.3) thus yields

$$\begin{aligned} E((\mathbf{X}^{(j)} - E^{(j)})(\mathbf{X} - E(\mathbf{X}))^T \cdot \mathbf{t}_1) &= E(\sum_{i=1}^{n} \mathbf{t}_1^{(i)} (\mathbf{X}^{(i)} - E^{(i)})(\mathbf{X}^{(j)} - E^{(j)})) \\ &= \sum_{i=1}^{n} \mathbf{t}_1^{(i)} E((\mathbf{X}^{(i)} - E^{(i)})(\mathbf{X}^{(j)} - E^{(j)})) \\ &= \sum_{i=1}^{n} \mathbf{t}_1^{(i)} Cov(\mathbf{X}^{(i)}, \mathbf{X}^{(j)}) \\ &= E\left((\mathbf{X}^{(j)} - E^{(j)}) \cdot 0\right) = 0 \end{aligned}$$

We therefore have the inclusion:

$$\forall j \in \{1, \ldots, n\} : \quad P(\mathbf{X} \in \epsilon_1) = 1 \Rightarrow \sum_{i=1}^{n} \mathbf{t}_1^{(i)} v_{ij} = 0$$

If on the other hand $\forall j \in \{1,\ldots,n\} : \sum_{i=1}^{n} \mathbf{t}_1^{(i)} v_{ij} = 0$, we can multiply the left hand side with $\mathbf{t}_1^{(j)}$ and take the sum over all $j \in \{1,\ldots,n\}$ and arrive at

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{t}_1^{(i)} v_{ij} \mathbf{t}_1^{(j)} = V(Y_1) = 0$$

implying that $P(\mathbf{X} \in \epsilon_1) = 1$. By this we have shown that

$$P(\mathbf{X} \in \epsilon_1) = 1 \Leftrightarrow \forall j \in \{1,\ldots,n\} : \underbrace{\sum_{i=1}^{n} \mathbf{t}_1^{(i)} v_{ij} = 0}_{(**)}$$

Note, that the quadratic form has been reduced to a linear one on the right hand side of the equivalence. Given a random vector $\mathbf{X}$ with covariance matrix V we may ask, how much linearly independent vectors $\mathbf{t}_k \in \mathbf{R}^n \setminus \emptyset$ can we find that fulfill $(**)$? Clearly all such vectors are solutions of the linear system

$$V\mathbf{t}_k = \emptyset$$

and linear algebra tells us that there are $n - R(V)$ nontrivial linearly independent solutions, where $R(V)$ denotes the rank of the covariance matrix V. Every such $\mathbf{t}_k$ defines an $(n-1)$–dimensional hyperplane in $\mathbf{R}^n$, which we denote by

$$\epsilon_k := \{\xi \in \mathbf{R}^n : (\xi - E(\mathbf{X}))^T \cdot \mathbf{t}_k = 0\}.$$

Setting $l := n - R(V)$, we conclude that if $l \geq 0$ then

$$P(\mathbf{X} \in \epsilon_k) = 1 \quad \text{for every } k \in \{1,\ldots,l\}$$

and thus

$$P(\mathbf{X} \in \bigcap_{k=1}^{l} \epsilon_k) = 1$$

implying that the distribution of $\mathbf{X}$ is concentrated on an $R(V)$–dimensional subspace of the n–dimensional Euclidean space. We say that the distribution of $\mathbf{X}$ is $R(V)$–dimensional. The result is summed up in a lemma due to Frisch, which can be found in [45, p.202]:

Lemma 5.3 *The joint probability distribution of random variables*

$$\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(n)}$$

is l–dimensional if and only if the covariance matrix $V = V(\mathbf{X})$ is of rank l, $\mathbf{X} = (\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(n)})^T$.

Remark 5.1 *If* $\mathbf{X}$ *is a random vector having a distribution with lower dimension than the number of components it has, it cannot have a probability density. Consider for example the onedimensional random vector*

$$X(\omega) := \begin{cases} 1 & \omega = \omega_0 \\ 0 & \omega \neq \omega_0 \end{cases}$$

which has a distribution with dimension zero concentrated at ω_0*. The distribution function* F_X *of* X *is*

$$F_X(x) := \begin{cases} 0 & x < X(\omega_0) \\ 1 & x \geq X(\omega_0) \end{cases}$$

The measure induced by X *on* $\mathbf{R}$ *is singular with respect to the Lebesgue measure, since the whole measure is concentrated at one point. We call such random vectors* degenerate.

We conclude our observations with a lemma that shows the effect of linear mappings on the covariance matrix.

Lemma 5.4 *If* $\mathbf{X}$ *is an* n*–dimensional random vector with covariance matrix* V *and expectation* $\mathbf{E}$*,* M *is an* $m \times n$ *matrix, and* $\xi \in \mathbf{R}^m$*, then* $\mathbf{Y} = M\mathbf{X} + \xi$ *is a random vector with covariance matrix* MVM^T *and expectation* $M\mathbf{E} + \xi$

Proof: The i'th component of $\mathbf{Y}$ is $(\sum_{k=1}^{n} M_{ik}\mathbf{X}^{(k)}) + \xi^{(i)}$, which has expectation

$$E(\mathbf{Y}^{(i)}) = E\left(\sum_{k=1}^{n} M_{ik}\mathbf{X}^{(k)} + \xi^{(i)}\right) = \left(\sum_{k=1}^{n} M_{ik}E(\mathbf{X}^{(k)})\right) + \xi^{(i)}.$$

Thus $E(\mathbf{Y}) = M\mathbf{E} + \xi$ by the linearity of the expectation. Let us calculate

$$\begin{aligned}
Cov(\mathbf{Y}^{(i)}, \mathbf{Y}^{(j)}) &= E\left(\left(\mathbf{Y}^{(i)} - E(\mathbf{Y}^{(i)})\right)\left(\mathbf{Y}^{(j)} - E(\mathbf{Y}^{(j)})\right)\right) \\
&= E\left(\left(\sum_{k=1}^{n} M_{ik}\mathbf{X}^{(k)} - \sum_{k=1}^{n} M_{ik}E(\mathbf{X}^{(k)})\right) \cdot \right. \\
&\quad \left. \cdot \left((\sum_{l=1}^{n} M_{jl}\mathbf{X}^{(l)} - \sum_{l=1}^{n} M_{jl}E(\mathbf{X}^{(l)})\right)\right) \\
&= E\left(\sum_{k=1}^{n} M_{ik}(\mathbf{X}^{(k)} - E(\mathbf{X}^{(k)})) \cdot \right. \\
&\quad \left. \cdot \sum_{l=1}^{n} M_{jl}(\mathbf{X}^{(l)} - E(\mathbf{X}^{(l)})\right) \\
&= \sum_{k=1}^{n}\sum_{l=1}^{n} M_{ik}v_{kl}M_{jl} \\
&= (MVM^T)_{ij}
\end{aligned}$$

□

The last equality can be proved straightforward by using the definition of matrix multiplication. ξ cancels out in the first line. The lemma has some application in the next section, where we want to calculate the distribution of a multidimensional random vector which is the image of a vector, whose components are independent standard normal variates, under a linear mapping.

We have seen that the covariance matrix of a random vector tells us about the dimension of the joint distribution of its components and that a linear subspace of $\mathbf{R}^n$ on which the random vector "lives" is defined by the intersection of hyperplanes which are specified by linearly independent solutions of the equation

$$\mathbf{t}^T V \overset{(V \text{ is symmetric})}{=} V\mathbf{t} = \emptyset$$

and the expectation of the random vector. If we transform the random vector using a linear mapping, the expectation and the covariance matrix can be calculated easily by matrix multiplication.

5.2 The covariance matrix of $\mathbf{W}_N$

We now calculate the covariance of the counter variables used in the overlapping M–tuple statistic. Let us recall the notations introduced in Chapter 3.

Let $r = (r_i)_{i \geq 0}$ be an infinite sequence of symbols r_i taken from an alphabet $\mathrm{A} := \{a_0, \ldots, a_{\alpha-1}\}$. For given positive integers N and M, define

$$x \oplus y := (x + y) \bmod N, \quad x \ominus y := (x - y) \bmod N$$

and put, for every $n \in \{0, \ldots, N-1\}$,

$$\begin{aligned} \mathbf{r}_n &:= (r_n, \ldots, r_{n \oplus M \ominus 1}) \\ \mathbf{r}'_n &:= (r_n, \ldots, r_{n \oplus M \ominus 2}) \end{aligned}$$

We now count for every possible M–tuple $\mathbf{a} \in \mathrm{A}^M$ the number of occurrences of the tuple in the finite sequence of overlapping M–tuples $(\mathbf{r}_n)$.

$$\mathbf{W}_N^{\mathbf{a}}(r) := \#\{n, 0 \leq n < N : \mathbf{r}_n = \mathbf{a}\}$$

Finally put

$$\mathbf{W}_N(r) = \begin{pmatrix} \mathbf{W}_N^{(a_0,\ldots,a_0,a_0)}(r) \\ \mathbf{W}_N^{(a_0,\ldots,a_0,a_1)}(r) \\ \mathbf{W}_N^{(a_0,\ldots,a_0,a_2)}(r) \\ \vdots \\ \mathbf{W}_N^{(a_{\alpha-1},\ldots,a_{\alpha-1},a_{\alpha-1})}(r) \end{pmatrix}.$$

In calculating $\mathbf{W}_N$, we use only the first N elements of the sequence (r_i). The tuples $\mathbf{r}_n$ can be regarded as if they arise from a cyclic sequence $(\hat{r}_n)$ defined by $\hat{r}_{Nj+i} = r_i, \quad j \in$

$\mathbf{N}_0, i \in \{0, 1, \ldots, N-1\}$. The function $\mathbf{W}_N$ thus is defined for infinite sequences (r_i) as well as for finite sequences of length N, say $p \in \mathrm{A}^N$. The value of $\mathbf{W}_N(p)$ does not change if we rotate the N symbols in p treating the latter as a cyclic sequence. We will frequently make use of this fact in later calculations. Note, that the definition of $\mathbf{r}_n$ depends on the value of N thus $\mathbf{r}_n = \mathbf{r}_n(r, N, n)$ and $\mathbf{W}_N = \mathbf{W}_N(r, N, M, \mathrm{A}) \in \mathbf{R}^{\alpha^s}$. The vector $\mathbf{W}_N$ equals a vector whose components are the counters $\hat{f}_{\mathbf{r}}(\mathbf{a})$ which have been introduced in Chapter 3. The order of the components has been chosen the alphabetical one.

The calculation of the covariance matrix of the vector $\mathbf{W}_N$ is based on the assumption that the projectors $r_i \quad i \in \{0, 1, 2, \ldots N-1\}$ of the sequence r are independent random variables distributed uniformly on A:

$$r_i \sim U(\mathrm{A}), \text{ iid.}$$

The existence of a probability space $(\Omega, \mathcal{A}, P)$ with $\Omega = \mathrm{A}^\infty$ and the above properties for every $N \in \mathbf{N}$ is assured by the consistency theorem of Kolmogorov for which we refer the reader to [3, p.510]. In the following proofs, we will make use of two properties of the probability model.

- The range of $\mathbf{W}_N^{\mathbf{a}}(r)$ equals the set $\{0, 1, \ldots, N\}$.
- In calculating probabilities for $\mathbf{W}_N$, we only need to consider finite sequences of length N, that is, we work with a copy of $\mathbf{W}_N$ defined on the probability space $(\Omega_N, \mathcal{A}_N, P_N)$ which contains only sequences of length N, an appropriate sigma–field, and the equidistribution on the single sequences. Every sequence in this model has probability $1/\alpha^N$, since this is the measure of the cylinder generated by all infinite sequences in A^∞ that share the same initial sequence of length N. The two copies of $\mathbf{W}_N$ thus have the same distribution and cannot be distinguished by means of the probability calculus.

At first, we show two little lemmas.

Lemma 5.5 $\sum_{x=0}^{N} x \cdot \delta(\mathbf{W}_N^{\mathbf{a}}(r) = x) = \mathbf{W}_N^{\mathbf{a}}(r)$

This can easily be verified by observing that the indicator is nonzero if and only if $\mathbf{W}_N^{\mathbf{a}}(r) = x$.

Lemma 5.6 $\sum_{p \in \mathrm{A}^N} \mathbf{W}_N^{\mathbf{a}}(p) = N\alpha^{N-M}$

This can be proved by counting the number of occurrences of the M–tuple $\mathbf{a}$ in two different ways. The first is given by the left hand side in the above equation, counting at first all occurrences of $\mathbf{a}$ in the finite sequence p and then summing up over all such sequences of length N. The second way is counting at first all sequences p, where the

M–tuple $\mathbf{a}$ occurs at a specific position i such that $\mathbf{p}_i = \mathbf{a}$. This is true for α^{N-M} sequences, since all the other symbols may be chosen arbitrary. Now sum up over all possible positions $i = \{0, \ldots, N-1\}$. This gives the right hand side in the lemma.

We can now calculate the expectation and covariance matrix of $\mathbf{W}_N$ by discrete "integration" over the finite probability space with equidistribution. The expectation of $\mathbf{W}_N^{\mathbf{a}}$, where $\mathbf{a}$ denotes an arbitrary M–tuple, is

$$\begin{aligned} E(\mathbf{W}_N^{\mathbf{a}}) &= \frac{1}{\alpha^N} \sum_{p \in \mathrm{A}^N} \mathbf{W}_N^{\mathbf{a}}(p) \\ &= \frac{1}{\alpha^N} N \alpha^{N-M} \\ &= \frac{N}{\alpha^M} \end{aligned}$$

The expectation of the product of two counters equals

$$E(\mathbf{W}_N^{\mathbf{a}} \mathbf{W}_N^{\mathbf{b}}) = \frac{1}{\alpha^N} \sum_{p \in \mathrm{A}^N} \mathbf{W}_N^{\mathbf{a}}(p) \mathbf{W}_N^{\mathbf{b}}(p) \tag{5.4}$$

Thus

$$\begin{aligned} Cov(\mathbf{W}_N^{\mathbf{a}}, \mathbf{W}_N^{\mathbf{b}}) &= E(\mathbf{W}_N^{\mathbf{a}} \mathbf{W}_N^{\mathbf{b}}) - E(\mathbf{W}_N^{\mathbf{a}}) E(\mathbf{W}_N^{\mathbf{b}}) \\ &= \frac{1}{\alpha^N} \sum_{p \in \mathrm{A}^N} \mathbf{W}_N^{\mathbf{a}}(p) \mathbf{W}_N^{\mathbf{b}}(p) - (\frac{N}{\alpha^M})^2 \end{aligned} \tag{5.5}$$

In this form, the covariance matrix will be used within the next sections. However, if a little more effort is invested in (5.4) the underlying structure reveals itself better. We thus continue, setting $\mathbf{W}_N^{\mathbf{a}}(p) = \sum_{i=0}^{N-1} \delta(\mathbf{p}_i = \mathbf{a})$ in (5.4). This yields

$$\begin{aligned} \mathbf{W}_N^{\mathbf{a}}(p) \mathbf{W}_N^{\mathbf{b}}(p) &= \left(\sum_{i=0}^{N-1} \delta(\mathbf{p}_i = \mathbf{a}) \right) \left(\sum_{j=0}^{N-1} \delta(\mathbf{p}_j = \mathbf{b}) \right) \\ &= \sum_{i=0}^{N-1} \sum_{d=-\lfloor \frac{N-1}{2} \rfloor}^{\lceil \frac{N-1}{2} \rceil - 1} \delta(\mathbf{p}_i = \mathbf{a}) \cdot \delta(\mathbf{p}_{(i+d)} = \mathbf{b}) \end{aligned}$$

The counter d can be interpreted as the distance between the two positions at which the sequence p is scanned by the indicators. It runs through a complete system of residues modulo N. The expectation of the product of $\mathbf{W}_N^{\mathbf{a}}$ and $\mathbf{W}_N^{\mathbf{b}}$ becomes

$$E(\mathbf{W}_N^{\mathbf{a}} \mathbf{W}_N^{\mathbf{b}}) = \frac{1}{\alpha^N} \sum_{i=0}^{N-1} \sum_{d=-\lfloor \frac{N-1}{2} \rfloor}^{\lceil \frac{N-1}{2} \rceil - 1} \underbrace{\sum_{p \in \mathrm{A}^N} \delta(\mathbf{p}_i = \mathbf{a}) \cdot \delta(\mathbf{p}_{(i+d)} = \mathbf{b})}_{=: f(i, \mathbf{a}, \mathbf{b}, d)}$$

where $f(i, \mathbf{a}, \mathbf{b}, d)$ is the number of sequences that have M–tuple $\mathbf{a}$ at position i and M–tuple $\mathbf{b}$ d positions ahead. Clearly, for far distances (in the sense of the topology

of the *cyclic* treatment of p), f should be constant, because a fixed amount $(2M)$ of symbols is defined by $\mathbf{a}$ and $\mathbf{b}$, and the remaining $(N-2M)$–symbols can be chosen arbitrarily. On the other hand, if d is small, the conditions imposed on the symbols in p by $\mathbf{a}$ and $\mathbf{b}$ may be impossible to fulfill simultaneously. Consider for example the alphabet $\mathrm{A} := \{0,1\}$ and the two 3–tuples $\mathbf{a} := (0,1,0)$ and $\mathbf{b} := (1,0,0)$. These tuples coincide for $d = 1$ or $d = -2$ but not for $d \in \{-1,0,2\}$.

We write

$$\delta(\mathbf{a},\mathbf{b},d) := \begin{cases} 1 & : \ \mathbf{a} \text{ and } \mathbf{b} \text{ coincide at distance } d \\ 0 & : \ \text{else} \end{cases}$$

and therefore get

$$f(i,\mathbf{a},\mathbf{b},d) = \begin{cases} \alpha^{N-2M} & : \ d \le -M \vee d \ge M \\ \alpha^{N-(M+|d|)}\delta(\mathbf{a},\mathbf{b},d) & : \ \text{else} \end{cases}$$

Note that f does not depend on i. We can now rewrite the expectation

$$\begin{aligned} E(\mathbf{W}_N^{\mathbf{a}}\mathbf{W}_N^{\mathbf{b}}) &= \frac{1}{\alpha^N} N \Big[(N-2M+1)\alpha^{N-2M} + \\ &\quad + \sum_{d=-M+1}^{M-1} \delta(\mathbf{a},\mathbf{b},d)\alpha^{N-(M+|d|)}\Big] \\ &= N \cdot \left[(N-2M+1)\alpha^{-2M} + \sum_{d=-M+1}^{M-1} \delta(\mathbf{a},\mathbf{b},d)\alpha^{-(M+|d|)}\right] \end{aligned}$$

We get the following representation for the covariance

$$\begin{aligned} Cov(\mathbf{W}_N^{\mathbf{a}},\mathbf{W}_N^{\mathbf{b}}) &= N \cdot \left(\frac{1-2M}{\alpha^{2M}} + \sum_{d=-M+1}^{M-1} \delta(\mathbf{a},\mathbf{b},d)\alpha^{-(M+|d|)}\right) \\ &=: N \cdot v_{\mathbf{ab}} \end{aligned} \tag{5.6}$$

The covariance matrix of the standardized vector $\frac{1}{\sqrt{N}}\mathbf{W}_N$ thus is constant according to Lemma 5.4:

$$V\left(\frac{1}{\sqrt{N}}\mathbf{W}_N\right) = v_{\mathbf{ab}}$$

We will use this result in the calculation of the asymptotic distribution of $\mathbf{W}_N$.

5.2.1 The linear dependence of the $\mathbf{W}_N^{\mathbf{a}}$

The covariance matrix will help us to calculate the dimension of the joint distribution of the $\mathbf{W}_N$ later on. Consider however the strong dependencies of two successive $\mathbf{r}_n$. We should expect a relation of the following nature: $\mathbf{W}_N^{\mathbf{a}}$ should be determined by values of 'neighbors' $\mathbf{W}_N^{\mathbf{i}_1},\ldots,\mathbf{W}_N^{\mathbf{i}_k}$, where neighborhood is measured by the similarities in the tuples $\mathbf{a}$ and $\mathbf{i}_1,\ldots,\mathbf{i}_k$. Actually, the $\mathbf{W}_N^{\mathbf{a}}$ depend in such a strong way, that at least

α^{M-1} of them can be expressed linearly by the remaining ones. This means, that the vector $\mathbf{W}_N$ lies in an affine subspace of dimension lower or equal to $\alpha^M - \alpha^{M-1}$. The joint distribution of the $\mathbf{W}_N^{\mathbf{a}}$ cannot have a higher dimension, as we have seen in Section 5.1. The following computations only serve to clarify the way, in which the $\mathbf{W}_N^{\mathbf{a}}$ depend on each other and will not be used in the later sections.

The idea behind the calculation is that the sum over all counters referring to M–tuples $(b_0, \ldots, b_{M-1})$, that contain a given $(M-1)$–tuple, say $(c_0, \ldots, c_{M-2})$, can be obtained in two ways, by either summing over the first or over the last symbol. The formula holds because of the cyclic treatment of the first N symbols in the sequence r that is scanned by the counter vector:

$$\begin{aligned}
\sum_{b_0 \in \mathrm{A}} \mathbf{W}_N^{(b_0, c_0, \ldots, c_{M-2})}(r) &= \sum_{b_0 \in \mathrm{A}} \sum_{i=0}^{N-1} \delta(\mathbf{r}_i = (b_0, c_0, \ldots, c_{M-2})) = \\
&= \sum_{i=0}^{N-1} \sum_{b_0 \in \mathrm{A}} \delta(\mathbf{r}_i = (b_0, c_0, \ldots, c_{M-2})), \\
\text{and, because of the cyclic treatment} \quad &= \sum_{i=0}^{N-1} \sum_{b_{M-1} \in \mathrm{A}} \delta(\mathbf{r}_i = (c_0, \ldots, c_{M-2}, b_{M-1})) \\
&= \sum_{b_{M-1} \in \mathrm{A}} \mathbf{W}_N^{(c_0, \ldots, c_{M-2}, b_{M-1})}(r) \qquad (5.7)
\end{aligned}$$

In order two show the linear dependence, we will calculate all counters referring to M– tuples of the form $(a_0, x_1, \ldots, x_{M-1})$ from counters of M–tuples of the form $(x_0 \neq a_0, x_1, \ldots, x_{M-1})$, where a_0 is the first symbol in the alphabet. From the above identity we have

$$\begin{aligned}
\mathbf{W}_N^{(a_0, x_1, \ldots, x_{M-1})} &= \sum_{b \in \mathrm{A}} \mathbf{W}_N^{(b, x_1, \ldots, x_{M-1})} - \sum_{c \in \mathrm{A} \setminus \{a_0\}} \mathbf{W}_N^{(c, x_1, \ldots, x_{M-1})} \\
&= \sum_{b \in \mathrm{A}} \mathbf{W}_N^{(x_1, \ldots, x_{M-1}, b)} - \sum_{c \in \mathrm{A} \setminus \{a_0\}} \mathbf{W}_N^{(c, x_1, \ldots, x_{M-1})}
\end{aligned}$$

If $x_1 \neq a_0$, we have successfully calculated the counter from other counters that are not indicated by an M–tuple having a_0 in the first position. If $x_1 = a_0$, the procedure goes on as follows: first observe, that the identity (5.7) can be extended to shorter tuples contained in other ones. By the same arguments as above, it follows that

$$\sum_{b_0 \in \mathrm{A}} \sum_{b_1 \in \mathrm{A}} \mathbf{W}_N^{(b_0, b_1, x_0, \ldots, x_{M-3})}(r) = \sum_{b_{M-2} \in \mathrm{A}} \sum_{b_{M-1} \in \mathrm{A}} \mathbf{W}_N^{(x_0, \ldots, x_{M-3}, b_{M-2}, b_{M-1})}(r)$$

Now let us calculate

$$\begin{aligned}
\mathbf{W}_N^{(a_0, a_0, x_2, \ldots, x_{M-1})} &= \sum_{(b_0, b_1) \in \mathrm{A}^2} \mathbf{W}_N^{(b_0, b_1, x_2, \ldots, x_{M-1})} - \\
&- \sum_{(c_0, c_1) \in \mathrm{A}^2 \setminus \{(a_0, a_0)\}} \mathbf{W}_N^{(c_0, c_1, x_2, \ldots, x_{M-1})} \\
&= \sum_{(b_0, b_1) \in \mathrm{A}^2} \mathbf{W}_N^{(x_2, \ldots, x_{M-1}, b_0, b_1)} -
\end{aligned}$$

$$- \sum_{(c_0,c_1)\in \mathrm{A}^2\setminus\{(a_0,a_0)\}} \mathbf{W}_N^{(c_0,c_1,x_2,\ldots,x_{M-1})}$$

which only depends on counters indicated by M–tuples not having a_0 in the first position as long as $x_2 \neq a_0$, since if $c_0 = a_0$ in the second sum then $c_1 \neq a_0$, which denotes a counter that has already been reconstructed above. Now iterate this procedure for counters $(a_0, a_0, a_0, x_3, \ldots, x_{M-1})$, and so on. The last step will yield

$$\begin{aligned}\mathbf{W}_N^{(a_0,\ldots,a_0,x_{M-1})} &= \sum_{(b_0,\ldots,b_{M-2})\in \mathrm{A}^{M-1}} \mathbf{W}_N^{(x_{M-1},b_0,\ldots,b_{M-2})} - \\ &- \sum_{(c_0,\ldots,c_{M-2})\in \mathrm{A}^{M-1}\setminus\{(a_0,\ldots,a_0)\}} \mathbf{W}_N^{(c_0,\ldots,c_{M-2},x_{M-1})}\end{aligned}$$

In the second sum only such counters appear, that have already been reconstructed in former steps of the procedure. Thus we can reduce all but one counters indicated by tuples of the form $(a_0, x_1, \ldots, x_{M-1})$ from counters indicated by M–tuples of the form $(x_0 \neq a_0, x_1, \ldots, x_{M-1})$. The remaining last counter $(a_0, \ldots, a_0)$ can be computed by observing that the total sum of all counters equals N. This completes the proof.

5.3 The multidimensional normal variate

In this section we take a closer look at the multidimensional equivalent to the normal variate. Recall that a random variable Y is said to be *standard normal* with expectation 0 and variance 1, iff

$$F_Y(t) = \int_{x=-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Its characteristic function is

$$\hat{F}_Y(t) = e^{-\frac{t^2}{2}}$$

Suppose $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ are n independent copies of Y. Again, the existence of such a vector is guaranteed by Kolmogrov's consistency theorem. Since the components of the vector are independent, the joint distribution of them equals the product of the distribution functions of every component, that is

$$F_{\mathbf{X}}(\mathbf{t}) = \prod_{i=1}^{n} F_{\mathbf{X}^{(i)}}(\mathbf{t}^{(i)})$$

Now recall the rule that if the $\mathbf{X}^{(i)}$ are independent, the characteristic function of the vector $\mathbf{X}$ equals the product of the characteristic functions of the $\mathbf{X}^{(i)}$. By this, we have

$$\hat{F}_{\mathbf{X}}(\mathbf{t}) = \prod_{i=1}^{n} \hat{F}_{\mathbf{X}^{(i)}}(\mathbf{t}^{(i)}) = e^{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{t}^{(i)})^2}$$

If we denote the inner product by '$\cdot$' we may restate the last result by

$$\hat{F}_{\mathbf{X}}(\mathbf{t}) = e^{-\frac{1}{2}\mathbf{t}^T\cdot\mathbf{t}} \tag{5.8}$$

This distribution plays the role of the standard normal variate in dimension n. Since the $\mathbf{X}^{(i)}$ are independent, we have $E(\mathbf{X}^{(i)}\mathbf{X}^{(j)}) = \delta(i,j)$ and consequently

$$V(\mathbf{X}) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

If we apply the definition of the characteristic function of a multidimensional random vector to the left side in (5.8), we have for every $\mathbf{t} \in \mathbf{R}^n$

$$\hat{F}_{\mathbf{X}}(\mathbf{t}) := E(e^{i\mathbf{t}^T\mathbf{X}}) \overset{(5.8)}{=} e^{-\frac{1}{2}\mathbf{t}^T\cdot\mathbf{t}} \tag{5.9}$$

under the assumption, that $\mathbf{X}$ is a vector of *independent* standard normal distributed random variables.

In order to extend our definition of a multidimensional normal variate, we consider a linear mapping A represented by an $n \times n$ matrix and define

$$\mathbf{Y} = A\mathbf{X}$$

Thus, for every $\mathbf{t} \in \mathbf{R}^n$,

$$\begin{aligned} \hat{F}_{\mathbf{Y}}(\mathbf{t}) &= E(e^{i\mathbf{t}^T\mathbf{Y}}) \\ &= E(e^{i\mathbf{t}^T(A\mathbf{X})}) \\ &= E(e^{i(\mathbf{t}^TA)\mathbf{X}}) \\ &\qquad e^{-\frac{1}{2}\mathbf{t}^T\underbrace{AA^T}_{:=C}\mathbf{t}} \\ \text{recall (5.9)} &= e \end{aligned}$$

The formal equivalence to the multivariate standard normal is obvious. The matrix C turns out to be the covariance matrix of the random vector $\mathbf{Y}$, because

$$\begin{aligned} Cov(\mathbf{Y}^{(i)}, \mathbf{Y}^{(j)}) &= E(\mathbf{Y}^{(i)} \cdot \mathbf{Y}^{(j)}) \\ &= E\left((\sum_{k=1}^{n} a_{ik}\mathbf{X}^{(k)}) \cdot (\sum_{l=1}^{n} a_{jl}\mathbf{X}^{(l)})\right) \\ &= \sum_{k=1}^{n}\sum_{l=1}^{n} a_{ik}a_{jl}E(\mathbf{X}^{(k)}\mathbf{X}^{(l)}) \end{aligned}$$

Since $E(\mathbf{X}^{(k)}\mathbf{X}^{(l)}) = \delta(k,l)$, the covariance reduces to

$$Cov(\mathbf{Y}^{(i)}, \mathbf{Y}^{(j)}) = \sum_{k=1}^{n} a_{ik}a_{jk} = (AA^T)_{ij}$$

This property gives reason for the following definition:

Definition 5.1 $\mathbf{Y} \sim \mathcal{N}(\mu, C)$ *iff*

$$\hat{F}_{\mathbf{Y}-\mu}(\mathbf{t}) = e^{-\frac{1}{2}\mathbf{t}^T C \mathbf{t}}$$

where C is a symmetric, nonnegative matrix.

Such a $\mathbf{Y}$ can always be represented by $\mathbf{Y} = A\mathbf{X} + \mu$ where $\mathbf{X}$ is a vector of independent standard normal variates, because a real symmetric matrix has a decomposition of the form $C = AA^T$: let C be symmetric and real, then the so–called spectral theorem for hermitian matrices (see [24], Theorem 4.1.5) states that there exists an orthogonal matrix U with $U^T C U =: D$ being a diagonal matrix containing the eigenvalues of C. Define D_0 as a diagonal matrix containing the square roots of the eigenvalues, then A can be defined by $A = UD_0$, so that $AA^T = UD_0D_0^TU^T = UDU^T = C$.

We therefore have for every nonnegative, symmetric matrix[1] C a unique central multidimensional normal variate with this covariance matrix. If C is singular with rank $R(C) =: r < n$, then $\mathbf{Y} = A\mathbf{X}$ has a r–dimensional distribution and no probability density in $\mathbf{R}^n$.

As linear mappings did occur in the definition of the multivariate normal, we should expect this class of distributions to be closed under such linear mappings. This is stated by

Lemma 5.7 *If M is an $j \times k$ matrix and $\mathbf{Y}$ from $\Omega \to \mathbf{R}^k$ is distributed central multivariate normal, $\mathbf{Y} \sim \mathcal{N}(0, C)$, then*

$$M \cdot \mathbf{Y} \sim \mathcal{N}(0, MCM^T)$$

Proof: Since we can represent $\mathbf{Y}$ by $\mathbf{Y} = A\mathbf{X}$, where $\mathbf{X}$ is a vector of independent standard normal variates and the variance C of $\mathbf{Y}$ equals to $A \cdot A^T$, we have by (5.9)

$$\begin{aligned}
\hat{F}_{(M\mathbf{Y})}(\mathbf{t}) &= E(e^{i\mathbf{t}^T(M\cdot A)\mathbf{X}}) \\
&= e^{-\frac{1}{2}\mathbf{t}^T(M\cdot A)(M\cdot A)^T\mathbf{t}} \\
&= e^{-\frac{1}{2}\mathbf{t}^T M(AA^T)M^T\mathbf{t}} \\
&= e^{-\frac{1}{2}\mathbf{t}^T(MCM^T)\mathbf{t}}
\end{aligned}$$

where (MCM^T) is a symmetric, nonnegative matrix. □

5.4 Central limit theorems

Our goal is to show that a standardized vector of counter variables converges in distribution. In the onedimensional case, e.g. for only one component of the vector, we would

[1] That is, for every possible covariance matrix

expect this to be the normal distribution, because subsequent counts do not depend too strongly. More precisely, the model of our counters forgets information after M steps, because then the M–tuple that is actually counted depends on entirely different $\mathbf{r}_n$. Let us now define what is understood by convergence in distribution and state some propositions and a theorem that sum up the most important results of the theory for the onedimensional case.

Convergence in distribution, $Y_N \Longrightarrow Y$ as $N \to \infty$, is defined as pointwise convergence of the distribution function in every continuity point of the asymptotic distribution function, that is $F_{Y_N}(x) \to F_Y(x)$ for all x, where F_Y is continuous. By this the random variables Y_N may be defined on entirely different probability spaces. In order to show a convergence in distribution we only need a well–defined distribution function for every $N \in \mathbf{N}$. The asymptotic distribution also can be defined on an arbitrary probability space capable of supporting random variables with the desired distribution[2]. We can work with the finite and discrete model discussed when calculating the expectation and covariance of the counter variables and switch the probability space for every N.

Another type of convergence is *convergence in probability*: let the random variables Y_N, Y all be defined on the same probability space $(\Omega, \mathcal{A}, P)$. If

$$\forall \epsilon > 0 : \quad \lim_{N \to \infty} P[|Y_N - Y| > \epsilon] = 0$$

then the Y_N are said to converge in probability to Y. We will need the following proposition, that combines the two types of convergence.

Proposition 5.8 *Let Y_N, Y all be defined on the same probability space. If Y_N converges to Y in probability, then $Y_N \Longrightarrow Y$.*

For a proof, see [3], Theorem 25.2. A related result is

Proposition 5.9 *If $Y_N \Longrightarrow Y$, and $X_N - Y_N \Longrightarrow 0$, then $X_N \Longrightarrow Y$.*

For a proof see ibid., Theorem 25.4. The next proposition relates asymptotic zero expectation to convergence in probability.

Proposition 5.10 *If*

$$\lim_{N \to \infty} E\left((Y_N)^2\right) = 0$$

then Y_N converges to zero in probability.

Proof: If $\lim_{N \to \infty} E\left((Y_N)^2\right) = 0$, we have for every $\delta > 0$ an index N_0, such that for every $N \geq N_0$ and for every $\epsilon > 0$

$$\int_\Omega (Y_N)^2 dP = \int_{|Y_N| \leq \epsilon} (Y_N)^2 dP + \int_{|Y_N| > \epsilon} (Y_N)^2 dP < \delta$$

[2] In this sense, the strong law of large numbers is a less trivial mathematical statement than the central limit theorem!

The last integral is always positive and thus can be omitted, giving

$$\int_{|Y_N|>\epsilon} (Y_N)^2 dP < \delta$$

which can be estimated by

$$\int_{|Y_N|>\epsilon} \epsilon^2 dP < \delta,$$

and, thus,

$$\epsilon^2 P(|Y_N| > \epsilon) < \delta.$$

By setting $\delta := \delta_1 \epsilon$ and applying the above results we have for every $\delta_1 > 0$ and for every $\epsilon > 0$ an $N_0 \in \mathbf{N}$ such, that for every $N \geq N_0$,

$$P(|Y_N| > \epsilon) < \delta_1.$$

Hence we have shown that

$$\forall \epsilon > 0 \quad \lim_{N \to \infty} P(|Y_N| > \epsilon) = 0,$$

i.e. convergence in distribution. □

An important role in the theory of convergence in distribution is played by the characteristic function of a random variable Y, which is defined by

$$\hat{F}_Y(t) := E(e^{itY}), \quad t \in \mathbf{R}$$

and uniquely defines the distribution of Y. The famous continuity theorem states that we may change the order of going to the limit and transforming the distribution of a random variable back from its characteristic function:

Theorem 5.11 (Continuity Theorem) *A necessary and sufficient condition for*

$$Y_N \Longrightarrow Y$$

is that

$$\hat{F}_{Y_N}(t) \to \hat{F}_Y(t) \quad \forall t \in \mathbf{R}$$

The central limit theorem tells us about the behavior of the sum of independent random variables. We will extend this theorem in first applying it to multidimensional random vectors and in secondly weakening the condition of independence. In order to do so we need a more general central limit theorem for the onedimensional case that allows the use of triangular arrays of random variables. If, for each $N \in \mathbf{N}$

$$Y_N^0, \ldots Y_N^{r_N}$$

is a set of independent random variables defined on the same probability space that may vary with N, we call this a triangular array. We put

$$S_N := \sum_{i=0}^{r_N} Y_N^i$$

and ask for the asymptotic distribution of a correctly standardized S_N. A theorem due to Lindeberg establishes sufficient conditions on the random variables that guarantee asymptotic normality of S_N.

Theorem 5.12 (Central Limit Theorem for triangular arrays) .

Suppose that the random variables Y_N^i are defined as above, and that

$$E(Y_N^i) = 0,\ \sigma_{N,i}^2 := E((Y_N^i)^2),\ s_N^2 := \sum_{i=0}^{r_N} \sigma_{N,i}^2$$

If the so–called Lindeberg condition

$$\lim_{N\to\infty} \sum_{i=0}^{r_N} \frac{1}{s_N^2} \int_{|Y_N^i| \geq \epsilon s_N} (Y_N^i)^2 dP = 0$$

is fulfilled for all positive ϵ, then

$$\frac{S_N}{s_N} \Longrightarrow \mathcal{N}(0,1)$$

For a proof see [3, Theorem 27.2]. This completes our excursion into the field of onedimensional probability theory and we will now consider convergence in distribution in the multivariate case.

Definition 5.2 *The sequence of random vectors $\mathbf{Y}_N$ is said to converge in distribution to the random vector $\mathbf{Y}$, iff the corresponding distribution functions $F_{\mathbf{Y}_N}$ converge weakly to $F_{\mathbf{Y}}$, that is iff*

$$\lim_{N\to\infty} F_{\mathbf{Y}_N}(\mathbf{x}) = F_{\mathbf{Y}}(\mathbf{x})$$

for every continuity point $\mathbf{x} \in \mathbf{R}^k$ of $F_{\mathbf{Y}}$.

We express this fact by $\mathbf{Y}_N \Longrightarrow \mathbf{Y}$. Note that the distribution function of a multivariate normal has non–continuity points, if the dimension of the distribution is less than the dimension of the vector.

As in the onedimensional case, the main tool for proving central limit theorems are characteristic functions. We will skip this (main) part of theory, because we will use another tool that allows the reduction to the onedimensional case. The following theorem can be found in [3, p.397].

Theorem 5.13 (Cramér - Wold) *For random vectors $\mathbf{Y}_N = (\mathbf{Y}_N^{(1)}, \ldots, \mathbf{Y}_N^{(k)})$ and $\mathbf{Y} = (\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(k)})$, a necessary and sufficient condition for*

$$\mathbf{Y}_N \Longrightarrow \mathbf{Y}$$

is that

$$\mathbf{t}^T \cdot \mathbf{Y}_N \Longrightarrow \mathbf{t}^T \cdot \mathbf{Y}$$

for every $\mathbf{t} = (\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(k)})^T$ *in* $\mathbf{R}^k$.

Proof: We only proof the sufficiency of the condition. If $\mathbf{t}^T \cdot \mathbf{Y}_N \Longrightarrow \mathbf{t}^T \cdot \mathbf{Y}$ for every $\mathbf{t} \in \mathbf{R}^k$ then the characteristic functions of $\mathbf{t}^T \cdot \mathbf{Y}_N$ has to converge to that of $\mathbf{t}^T \cdot \mathbf{Y}$ by the continuity theorem for the onedimensional case. Thus

$$E\left(e^{i \cdot s \cdot \mathbf{t}^T \cdot \mathbf{Y}_N}\right) \to E\left(e^{i \cdot s \cdot \mathbf{t}^T \cdot \mathbf{Y}}\right)$$

for all real s. Setting $s = 1$ we arrive at

$$E\left(e^{i \cdot \mathbf{t}^T \cdot \mathbf{Y}_N}\right) \to E\left(e^{i \cdot \mathbf{t}^T \cdot \mathbf{Y}}\right)$$

where the left side is the characteristic function of $\mathbf{Y}_N$ and the right side is the characteristic function of $\mathbf{Y}$. Since this equation holds for all $\mathbf{t} \in \mathbf{R}^k$, the characteristic function converges pointwise, which is equivalent to the convergence in probability. □

With this device many central limit theorems in $\mathbf{R}^k$ can be proved. We give only the following simple example.

Example 5.1 *Let* $\mathbf{Y}_N \in \mathbf{R}^k$ *be independent identically distributed random vectors. Suppose, that* $E\left((\mathbf{Y}_N^{(i)})^2\right) < \infty$ *and set* $\mu := E(\mathbf{Y}_N)$. *Finally let* $V = V(\mathbf{Y}_N)$ *be the covariance matrix and put* $\mathbf{S}_N = \mathbf{Y}_1 + \cdots + \mathbf{Y}_N$, *then*

$$\frac{\mathbf{S}_N - N\mu}{\sqrt{N}} \Longrightarrow \mathbf{Y} :\sim \mathcal{N}(0, V)$$

Proof: We will only sketch the proof. For a given $\mathbf{t} \in \mathbf{R}^k$ let $Z_N = \mathbf{t}^T \cdot (\mathbf{Y}_N - \mu)$ and $Z = \mathbf{t}^T \cdot \mathbf{Y}$. The variance of Z_N calculates to

$$V(Z_N) = \mathbf{t}^T V \mathbf{t} N,$$

such that

$$V(\frac{1}{\sqrt{N}} Z_N) = \mathbf{t}^T V \mathbf{t}.$$

From Section 5.1 we know that the $\frac{1}{\sqrt{N}} Z_N$ are random variables with expectation and variance that fulfill the conditions of the Lindeberg theorem (see Theorem 5.12). Thus

$$\frac{1}{\sqrt{N}} Z_N \Longrightarrow \mathcal{N}(0, \mathbf{t}^T V \mathbf{t}).$$

But this is exactly the distribution of Z. Since $\mathbf{t}$ has been chosen arbitrarily, Theorem 5.13 proves the desired multidimensional convergence in distribution. □

5.5 The asymptotic distribution of $\mathbf{W}_N$

Let us define

$$\tilde{\mathbf{W}}_N := \mathbf{W}_N - E(\mathbf{W}_N),$$

such that $E(\tilde{\mathbf{W}}_N) = 0$ and $V(\tilde{\mathbf{W}}_N) = V(\mathbf{W}_N) = N \cdot V$ where V has been calculated in equation (5.6). We want to show that $\tilde{\mathbf{W}}_N$ converges – if correctly standardized – in distribution to a multivariate normal having the same covariance matrix V and zero expectation. In order to apply the Cramér - Wold device, we need to consider the onedimensional random variable $\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N$, where $\mathbf{t} = (\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(k)})^T \in \mathbf{R}^k$. Here, k is the number of components in the random vectors $\mathbf{W}_N$ respectively $\tilde{\mathbf{W}}_N$, that is, the number of counters. Since a counter exists for every possible M–tuple of symbols out of the alphabet A, we have $k = \#\mathrm{A}^M = \alpha^M$. The expectation of $\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N$ is

$$E(\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N) = \mathbf{t}^T \cdot \Big(E(\tilde{\mathbf{W}}_N)\Big) = 0$$

The variance of $\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N$ calculates to

$$V(\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N) = \mathbf{t}^T V(\tilde{\mathbf{W}}_N)\mathbf{t} = \mathbf{t}^T(nV)\mathbf{t} = n(\mathbf{t}^T V \mathbf{t})$$

In order to apply a central limit theorem we will have to rewrite the vector $\tilde{\mathbf{W}}_N$ as a sum of random vectors. Remember the construction of $\mathbf{W}_N$. We have defined the $\mathbf{a}$'th component of $\mathbf{W}_N$ as the number of occurrences of the M–tuple $\mathbf{a}$ in the overlapping M–tuples $\mathbf{r}_n$ constructed from a cyclic copy of the first N symbols in the sequence r. Every $\mathbf{r}_n$ increases exactly one component in $\mathbf{W}_N$ by one. Let us define

$$\mathbf{I}_n := \begin{pmatrix} 0 & : & \mathbf{a}_0 \neq \mathbf{r}_n \\ & \vdots & \\ 0 & : & \mathbf{a}_{i-1} \neq \mathbf{r}_n \\ 1 & : & \mathbf{a}_i = \mathbf{r}_n \\ 0 & : & \mathbf{a}_{i+1} \neq \mathbf{r}_n \\ & \vdots & \\ 0 & : & \mathbf{a}_{\alpha^M - 1} \neq \mathbf{r}_n \end{pmatrix}$$

where $\mathbf{a}_i$ indicates the i'th M–tuple in A^M with respect to alphabetical order. Thus

$$\mathbf{W}_N := \sum_{n=0}^{N-1} \mathbf{I}_n$$

and

$$\tilde{\mathbf{W}}_N := \sum_{n=0}^{N-1} \tilde{\mathbf{I}}_n,$$

where $\tilde{\mathbf{I}}_n := \mathbf{I}_n - E(\mathbf{I}_n)$.

Now, rewrite $\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N$ using $\tilde{\mathbf{I}}_n$:

$$\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N = \mathbf{t}^T \cdot \sum_{n=0}^{N-1} \tilde{\mathbf{I}}_n = \sum_{n=0}^{N-1} (\mathbf{t}^T \cdot \tilde{\mathbf{I}}_n)$$

Thus, the onedimensional random variable $\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N$ can be written as sum of N onedimensional random variables and fulfills two important conditions for a central limit theorem, namely zero expectation and finite second central moment.

However, $\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N$ is not a sum of *independent* random variables since different $\mathbf{r}_n$ may depend on some common random variables in r. Furthermore, the definition of $\mathbf{I}_n$ depends on the value of N as does the value of $\mathbf{r}_n$: the fact that the calculation of $\mathbf{W}_N$ was based on a cyclic copy of the first n elements of r was a relief when calculating the covariance matrix but aggravates the treatment of $\tilde{\mathbf{W}}_N$ as sum of random vectors in this section.

For example, $\tilde{\mathbf{I}}_n$ is based on $r_0, \ldots, r_{M-1}$ in the calculation of $\mathbf{W}_N$ if $N = n$. If we now switch to $N = n+1$, $\tilde{\mathbf{I}}_n$ will be calculated from $r_n, r_0, r_1, \ldots, r_{M-2}$. Thus, we need a triangular array and the corresponding central limit theorem, Theorem 5.12. Let us denote $\tilde{\mathbf{I}}_n$ by $\tilde{\mathbf{I}}_n^N$, if we consider the random variable based on the cyclic copy of the first N symbols in r.

It is now easy to show, that $\tilde{\mathbf{I}}_i^N$ and $\tilde{\mathbf{I}}_j^N$ $(i < j)$ are independent, if either $i < N - M + 1$ and $j - i \geq M$ or $i \geq N - M + 1$ and $j \geq i + M - N$: since $\tilde{\mathbf{I}}_i^N$ depends on the random variables $r_i, r_{i\oplus 1}, \ldots, r_{i\oplus(M-1)}$ and $\tilde{\mathbf{I}}_j^N$ respectively on $r_j, r_{j\oplus 1}, \ldots, r_{j\oplus(M-1)}$, the r_l, $l \in \{0, 1, 2, \ldots N-1\}$, being independent random variables, they depend on mutually distinct sets of independent random variables, if the above conditions are fulfilled. Moreover, if $(i_k)_{k\in\{0,1,\ldots,k^*\}} \subset \{0, 1, \ldots, N-1\}$ is a sequence of indices with $i_k + M - 1 < i_{k+1}$ and the last index $i_{k^*} + M - N \leq i_0$, then the associated set of random variables $\tilde{\mathbf{I}}_{i_k}$ obviously is a set of independent random variables. This property is closely related to the definition of the so–called M–dependence, see [6, p.196], but takes into account the cyclic structure imposed on r. We thus define the term cyclic M–dependence as follows.

Definition 5.3 (Cyclic M–dependence) *A sequence*

$$(X_n)_{n\in\{0,\ldots,N-1\}}$$

of random variables is said to be cyclic M–dependent, iff

$$\forall i \in \mathbf{Z} : M \leq |i| < N - M \Rightarrow X_n, X_{n\oplus i} \text{ are independent.}$$

Note that in the reference text [6], M–dependence is what we would call $(M-1)$– dependence. Common literature uses both forms of definitions. Since the random variables $\mathbf{t}^T \cdot (\tilde{\mathbf{I}}_n^N)$ depend only on $\tilde{\mathbf{I}}_n^N$, they are also cyclic–M–dependent.

For our 'central limit tool' we will need yet another condition. The $\mathbf{t}^T \cdot (\tilde{\mathbf{I}}_i^N)$ have to be uniformly bounded: there has to exist a constant $U \in \mathbf{R}$, so that for every $i \in \mathbf{N}$, $|\mathbf{t}^T \cdot (\tilde{\mathbf{I}}_i^N)| \leq U$. This is fulfilled trivially by choosing $U = ||\mathbf{t}^T||_\infty$. We can use a slightly modified version of Theorem 7.3.1 in [6, p.196].

Theorem 5.14 (Central Limit Theorem) *Suppose that $\{X_i^N\}$ is a triangular array with lines containing cyclic–M–dependent, uniformly bounded random variables and put*

$$S_N := \sum_{i=0}^{N-1} X_i^N$$

If

$$\frac{\sqrt{V(\sum_{i=0}^{N-1} X_i^N)}}{N^{\frac{1}{3}}} \to \infty$$

as $N \to \infty$ then

$$\frac{\sum_{i=0}^{N-1} X_i^N - E(\sum_{i=0}^{N-1} X_i^N)}{V(\sum_{i=0}^{N-1} X_i^N)} \Longrightarrow \mathcal{N}(0,1)$$

Proof: Without loss of generality, we may assume that $E(X_i^N) = 0$. For an integer $l \geq 1$ let $i_j := [\frac{jN}{l}]$ $(0 \leq j < l)$ and put

$$\begin{aligned} Y_j^N &:= X_{i_j+1}^N + X_{i_j+2}^N + \cdots + X_{i_{j+1}-M}^N \\ Z_j^N &:= X_{i_{j+1}-M+1}^N + \cdots + X_{i_{j+1}}^N \end{aligned}$$

Thus $S_N = \sum_{j=0}^{l-1} Y_j^N + \sum_{j=0}^{l-1} Z_j^N = S'_N + S''_N$, say. From this definitions it follows that the Y_j^N are independent random variables for every $N > N_0 := (2M+1)l$, because the original sequence is cyclic–M–dependent and the calculation of the last Y_{l-1}^N needs utmost the random variable X_{N-M}^N, which is independent of $X_0^N, X_1^N, \ldots, X_{i_1-M}^N$. The same is true for Z_j^N, since every Z_j^N depends on exactly M X_i^N's, which are separated from the X_i^N used for the next Z_{j+1}^N by at least $\left[\frac{N}{l} - M\right]$ X_N^i's used in the calculation of Y_j^N, which are more than M RVs, if $n > n_0$ is chosen. The random variables S'_N and S''_N do not have to be independent, but we will see that S''_N is comparatively negligible.

Since the X_i^N are uniformly bounded, there exists a constant say $U > 0$ for which $|X_i^N| \leq U$. According to the cyclic–M–dependence of the whole sequence each X_i^N in the calculation of S''_N can depend on no more than M of the X_j^N's, for which we have

$$\left|E(X_i^N X_j^N)\right| \leq \left|\int_\Omega X_i^N X_j^N dP\right| \leq \int_\Omega \left|X_i^N X_j^N\right| dP \leq U^2.$$

For every other independent X_j^N, clearly $\left|E(X_i^N X_j^N)\right| = 0$. The sum S''_N involves lM different X_j^N's, say $X_{j_k}^N$ for $k \in \{1, 2, \ldots, l \cdot M\}$ and thus

$$\begin{aligned} |E(S'_N S''_N)| &= \left|\int_\Omega S'_N \sum_{k=1}^{lM} X_{j_k}^N dP\right| \\ &\leq \sum_{k=1}^{lM} \int_\Omega |S'_N||X_{j_k}^N| dP \\ \text{(cyclic } M\text{–dependence)} &\leq \sum_{k=1}^{lM} MU^2 \\ &= M^2 U^2 l \end{aligned}$$

Due to the independence of the Z_j^N we further get

$$\begin{aligned}\left|E((S_N'')^2)\right| &= \left|\sum_{i=0}^{l-1}\sum_{j=0}^{l-1}\int_\Omega Z_i^N Z_j^N dP\right| \\ &= \left|\sum_{j=0}^{l-1}\int_\Omega (Z_j^N)^2 dP\right| \\ &= \left|\sum_{j=0}^{l-1} E\left((Z_j^N)^2\right)\right| \\ &\leq lM^2U^2\end{aligned}$$

since

$$\begin{aligned}\left|E\left((Z_j^N)^2\right)\right| &= \left|\int_\Omega \left(\sum_{k=1}^{M} X^N_{i_{j+1}-M+k}\right)^2 dP\right| \\ &\leq \sum_{k=1}^{M}\sum_{l=1}^{M}\int_\Omega |X^N_{i_{j+1}-M+k}||X^N_{i_{j+1}-M+l}|dP \\ &\leq \sum_{l=1}^{M} MU^2 \\ &= M^2U^2\end{aligned}$$

With $E((S_N)^2) = E((S_N')^2) + 2E((S_N')^2(S_N'')^2) + E((S_N'')^2)$, we conclude that

$$\left|E((S_N)^2) - E((S_N')^2)\right| \leq 3lM^2U^2 \tag{5.10}$$

Now choose l with regard to N, say $l = l_N := [N^{\frac{2}{3}}]$. This does not conflict with our conditions on N_0 since for such an l, N should be greater or equal to $(2m+1)l$, which is certainly fulfilled for each $N \geq (2m+1)^3$. We denote $E((S_N)^2)$ by s_N and similarly $s_N' := E((S_N')^2)$, and $s_N'' := E((S_N'')^2)$. Now for $N \to \infty$

$$\frac{s_N'}{s_N} = 1 - \frac{s_N - s_N'}{s_N} \to 1 \tag{5.11}$$

since $|s_N - s_N'| \leq cN^{\frac{2}{3}}$ from (5.10) for a constant c and $\frac{N^{2/3}}{s_N} \to 0$ from the hypothesis of our theorem: since $V(\sum_{i=0}^{N-1} X_i^N) = \sqrt{s_N}$, we have $\frac{\sqrt{s_N}}{N^{1/3}} \to \infty$, and hence $\frac{s_N}{N^{2/3}} \to \infty$, or $\frac{N^{2/3}}{s_N} \to 0$. Next, we have

$$\left|E\left(\left(\frac{S_N''}{s_N}\right)^2\right)\right| \leq \frac{N^{2/3}M^2U^2}{(s_N)^2} \to 0$$

and hence applying proposition 5.10: $\frac{S_N''}{s_N} \to 0$ in probability. Thus, since

$$\frac{S_N}{s_N} = \frac{s_N'}{s_N}\frac{S_N'}{s_N'} + \frac{S_N''}{s_N},$$

$\frac{S_N}{s_N}$ will converge to $\mathcal{N}(0,1)$, if $\frac{S'_N}{s'_N}$ does, as stated by propositions 5.8 and 5.9. In fact, S''_N is comparatively negligible for the asymptotic distribution. To prove the statement, we only have to show that the triangular scheme of the Y_j^N fulfills the conditions for the Lindeberg theorem given in (5.12). First observe that the single rows of the triangular scheme contain *independent* random variables defined on the same probability space if $N > N_0$ due to the cyclic–M–dependence. Since each Y_j^N is the sum of no more than $[\frac{N}{l_N}] + 1$ of the X_i^N's,

$$|Y_j^N| \leq \left(\frac{N}{l_N} + 1\right) U = O(N^{\frac{1}{3}}) = o(s_N) = o(s'_N)$$

by first the hypothesis in the theorem and second the relation (5.11). By this we have for every $\epsilon > 0$ an index N_1 such, that for every $N > N_1$

$$\frac{|Y_j^N|}{s'_N} < \epsilon$$

or

$$|Y_j^N| < \epsilon s'_N$$

Thus for each $\epsilon > 0$, we have for sufficiently large $N > N_1$

$$\int_{|x|>\epsilon s'_N} x^2 dF_{Y_j^N}(x) = E\left((Y_j^N)^2 \delta(|Y_j^N| > \epsilon s'_N)\right) = 0$$

Hence Lindeberg's condition is satisfied for the triangular scheme $\frac{Y_j^N}{s'_N}$ and we conclude that $\frac{S'_N}{s'_N}$ converges to $\mathcal{N}(0,1)$ in distribution. This proves the theorem. □

The theorem immediately shows the asymptotic behavior of $\mathbf{W}_N$. Since

$$\frac{V(\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N)}{N^{\frac{1}{3}}} = \frac{N(\mathbf{t}^T V \mathbf{t})}{N^{\frac{1}{3}}} \to \infty \quad \text{if } N \to \infty$$

the conditions of the theorem are fulfilled, and we have

$$\frac{\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N - E(\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N)}{\sqrt{V(\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N)}} = \frac{\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N}{\sqrt{N(\mathbf{t}^T V \mathbf{t})}} \Longrightarrow \mathcal{N}(0,1)$$

or equivalently

Proposition 5.15

$$\frac{\mathbf{t}^T \cdot \tilde{\mathbf{W}}_N}{\sqrt{N}} \Longrightarrow \mathcal{N}(0, (\mathbf{t}^T V \mathbf{t}))$$

If we now set $\tilde{\mathbf{W}} \sim \mathcal{N}(0, V)$, a random vector that we expect to have the asymptotic distribution of $\frac{\tilde{\mathbf{W}}_N}{\sqrt{N}}$, we have from lemma (5.7) for every $s \in \mathbf{R}^1$

$$\hat{F}_{\mathbf{t}^T \cdot \tilde{\mathbf{W}}}(s) = e^{-\frac{1}{2} s (\mathbf{t}^T V \mathbf{t}) s}$$

Thus $\mathbf{t}^T \cdot \tilde{\mathbf{W}} \sim \mathcal{N}(0, (\mathbf{t}^T V \mathbf{t}))$, and we get

$$\mathbf{t}^T \cdot \frac{\tilde{\mathbf{W}}_N}{\sqrt{N}} \Longrightarrow \mathbf{t}^T \cdot \tilde{\mathbf{W}} \quad \forall \mathbf{t} \in \mathbf{R}^k$$

Using the Cramér - Wold device, we arrive at the desired

$$\frac{\tilde{\mathbf{W}}_N}{\sqrt{N}} \Longrightarrow \tilde{\mathbf{W}}$$

or equivalently

$$\tilde{\mathbf{W}}_N \Longrightarrow \mathcal{N}(0, NV)$$

Now going back to the original $\mathbf{W}_N$ by adding the expectation gives

Proposition 5.16 *The asymptotic distribution of* $\mathbf{W}_N$ *is the multidimensional normal variate* $\mathcal{N}(E(\mathbf{W}_N), V(\mathbf{W}_N))$*:*

$$\frac{\mathbf{W}_N - E(\mathbf{W}_N)}{\sqrt{N}} \Longrightarrow \mathcal{N}(0, V)$$

These calculations have shown that we can construct a test statistic which is asymptotically distributed multivariate normal. However, testing a multidimensional distribution by applying something like a Kolmogrov–Smirnov test statistic is computationally out of reach. We need to transform the vector $\mathbf{W}_N$ into an onedimensional random variable and calculate its distribution. We thus have to deal with the following question: if

$$\mathbf{X}_N \Longrightarrow \mathbf{X} \quad \mathbf{X}_N, \mathbf{X} \in \mathbf{R}^k,$$

can we conclude that

$$h(\mathbf{X}_N) \Longrightarrow h(\mathbf{X}), \quad \text{where } h : \mathbf{R}^k \to \mathbf{R}^1$$

for h fulfilling certain regularity conditions? The answer for $k = 1$ is provided by the so–called mapping theorem. We will use a formulation in terms of random variables.

Theorem 5.17 (Mapping theorem) *Suppose that* $h : \mathbf{R}^1 \to \mathbf{R}^1$ *is measurable and that the set* D_h *of its discontinuities is measurable. If* $\mathbf{X}_N \Longrightarrow \mathbf{X}$ *and* $P(\mathbf{X} \in D_h) = 0$*, then*

$$h(\mathbf{X}_N) \Longrightarrow h(\mathbf{X})$$

A proof can be found in [3, p.343]. Now consider the case $k > 1$. We prove the following

Theorem 5.18 (Mapping theorem in $\mathbf{R}^k$) *Suppose that* $h : \mathbf{R}^k \to \mathbf{R}^l$ *is measurable and that the set* D_h *of its discontinuities is measurable. Let* $\mathbf{X}_N$ *and* $\mathbf{X}$ *be random vectors from* $(\Omega, \mathcal{A}, P) \to \mathbf{R}^k$*. If* $\mathbf{X}_N \Longrightarrow \mathbf{X}$ *and* $P(\mathbf{X} \in D_h) = 0$*, then*

$$h(\mathbf{X}_N) \Longrightarrow h(\mathbf{X})$$

Proof: By Skorohod's theorem in $\mathbf{R}^k$ (see [3, p.399], Theorem 29.6), we may work with copies $\mathbf{Y}_N$ of $\mathbf{X}_N$ and $\mathbf{Y}$ of $\mathbf{X}$,

$$\mathbf{Y}_N, \mathbf{Y} : (\Omega_*, \mathcal{A}_*, P_*) \to \mathbf{R}^k,$$

sharing the same distribution and having the property that $\mathbf{Y}_N(\omega) \to \mathbf{Y}(\omega)$ for each $\omega \in \Omega_*$. If $\mathbf{Y}(\omega) \notin D_h$, this implies that $h(\mathbf{Y}_N(\omega)) \to h(\mathbf{Y}(\omega))$. But since $P(\{\omega : \mathbf{Y}(\omega) \in D_h\}) = 0$, we have

$$h(\mathbf{Y}_N(\omega)) \to h(\mathbf{Y}(\omega)) \quad \text{with probability one}$$

implying convergence in probability and hence convergence in distribution:

$$h(\mathbf{Y}_N) \Longrightarrow h(\mathbf{Y})$$

Since our copies share the same distribution they cannot be distinguished by ' $\Longrightarrow$ ', and we arrive at

$$h(\mathbf{X}_N) \Longrightarrow h(\mathbf{X}),$$

which completes the proof. □

We are now in the position to design a test by constructing a onedimensional 'fingerprint' of the multivariate normal distribution using a transformation h that fulfills the above conditions. If we choose

$$h : \mathbf{R}^k \to \mathbf{R}^1 \quad h(\xi) := \xi^T A \xi$$

where A is some $k \times k$ matrix, then D_h is the empty set and h is measurable because it is continuous. If we succeed in calculating the distribution of $h(\mathbf{Y})$, where $\mathbf{Y} \sim \mathcal{N}(E(\mathbf{W}_N), V(\mathbf{W}_N))$, we have by the mapping theorem

$$h(\mathbf{W}_N) \Longrightarrow h(\mathbf{Y}).$$

5.6 A generalized χ^2–test

Let us start with an arbitrary multidimensional normal variate

$$\mathbf{Y} \sim \mathcal{N}(\mu, C) \quad \text{where } C = BB^T$$

and hence

$$\mathbf{Y} = \mu + B \cdot \mathbf{X} \quad \text{with } \mathbf{X}^{(k)} \sim \mathcal{N}(0,1) \quad \text{iid.}$$

Roughly speaking we will somehow rotate $\mathbf{Y}$ in order to get access to the original coordinates of $\mathbf{X}$, whose squares can be summed up and are distributed χ^2 with $R(B) = dim(\mathbf{X})$ degrees of freedom. More precisely, we will study the quadratic form

$$\chi := h(\mathbf{Y} - \mu) = (\mathbf{Y} - \mu)^T A (\mathbf{Y} - \mu)$$

for some appropriate matrix A.

Consider the trivial example $\mathbf{Y} = B \cdot \mathbf{X}$, where $B = I_{dim(\mathbf{X})}$. Then

$$\chi = \mathbf{X}^T\mathbf{X} = \sum_{i=1}^{dim(\mathbf{X})} (\mathbf{X}^{(i)})^2 \sim \chi^2_{(dim(\mathbf{X}))}$$

In the following two sections we will first discuss an algebraic description for the properties of A with respect to B and then give an example for the application of such matrices.

5.6.1 Weak inverses

Let A be a nonsingular square matrix with rank n. If we denote the inverse of A by A^{-1}, we can solve linear equations in A, that is, given a certain $\mathbf{y} \in \mathbf{R}^n$, we know that

$$\mathbf{x} := A^{-1}\mathbf{y}$$

solves the equation

$$A\mathbf{x} = \mathbf{y}$$

Linear Algebra proves that only this $\mathbf{x}$ will solve the equation and that for every $\mathbf{y}$ exactly one $\mathbf{x}$ exists such that $A\mathbf{x} = \mathbf{y}$.

Now suppose that the $n \times m$ matrix A is not of full rank. We denote the range of A by $\mathcal{M}(A)$, that is

$$\mathcal{M}(A) := \{A \cdot \mathbf{x} : \mathbf{x} \in \mathbf{R}^m\}$$

For a given $\mathbf{y} \in \mathbf{R}^n$, there are two possibilities: if $\mathbf{y} \in \mathcal{M}(A)$, there exists at least one $\mathbf{x} \in \mathbf{R}^m$, such that $A\mathbf{x} = \mathbf{y}$. If $\mathbf{y} \notin \mathcal{M}(A)$, the equation has no solution. Nevertheless, it would be convenient to have something like an inverse, a linear mapping that yields one solution $\mathbf{x}$ if there exists one a least. Such an inverse can exist in a suitable sense: let us denote our 'equation–solver' by A^- and require the property that, for every $\mathbf{y} \in \mathcal{M}(A)$, $\mathbf{x} = A^-\mathbf{y}$ is a solution for $A\mathbf{x} = \mathbf{y}$.

We show the following little lemma, which can be found in [41].

Lemma 5.19 *Given a square matrix A, a linear transformation A^- with the desired property exists, iff the relation $AA^-A = A$ holds.*

Proof: Assume first that A^- exists in the way described above. Let $\mathbf{z} \in \mathbf{R}^m$ be arbitrary. Then $\mathbf{y} = A\mathbf{z}$ is in $\mathcal{M}(A)$. Thus $\mathbf{x} = A^-\mathbf{y} = A^-A\mathbf{z}$ is a vector with $A\mathbf{x} = \mathbf{y}$, that is, for every $\mathbf{z} \in \mathbf{R}^m$, we have

$$AA^-A\mathbf{z} = A\mathbf{z} \text{ and thus } AA^-A = A$$

If on the other hand $AA^-A = A$, and $\mathbf{y} \in \mathcal{M}(A)$, then there is some $\mathbf{x}$ giving $A\mathbf{x} = \mathbf{y}$. Since $AA^-A = A$, also $AA^-A\mathbf{x} = A\mathbf{x}$ by which $AA^-\mathbf{y} = \mathbf{y}$. This means, that $A^-\mathbf{y}$ is a solution to $A\mathbf{x} = \mathbf{y}$. □

We now can define what we will call a *weak inverse.*

Definition 5.4 *Let A be an arbitrary real $n \times m$ matrix. If there exists an $m \times n$ matrix A^- with*

$$AA^-A = A$$

this matrix is called weak inverse of A. In some of the literature, the term generalized inverse is also used.

It should be mentioned that the theory of generalized inverses is a broad field in linear algebra. The definition above gives only a special case of weak inverses. Such inverses can be constructed in some different ways serving for other purposes. See for example [41], [33] or [25]. However, we will only need some algebraic properties of A^-, that will serve in calculating the distribution of quadratic forms. These are stated in the following lemmas.

Lemma 5.20 *If A is an $n \times m$, and B is an $m \times n$ matrix, then $tr(AB) = tr(BA)$*

Proof: Since AB is $n \times n$ and BA is $m \times m$, straightforward calculation of the sum of the diagonal elements in each product yields

$$\begin{aligned} \sum_{j=1}^{n}(AB)_{(j,j)} &= \sum_{j=1}^{n}\sum_{i=1}^{m} a_{(j,i)}b_{(i,j)} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} b_{(i,j)}a_{(j,i)} \\ &= \sum_{i=1}^{m}(BA)_{(i,i)}. \end{aligned}$$

□

Corollary 5.21 *If A is idempotent, $AA = A$, then A is square, say $n \times n$, and thus for any $n \times m$ matrix B, we have*

$$tr(B^T(AB)) = tr(ABB^T),$$

since B^T is $m \times n$ and AB is $n \times m$.

Lemma 5.22 *If H is an idempotent matrix then the trace of H equals the rank of H and H is diagonizable, that is, there exists a regular matrix S for which we have $S^{-1}HS = D$, where D is a diagonal matrix containing the eigenvalues of H.*

The proof requires some additional calculations and is therefore postponed to Appendix A.

Lemma 5.23 *A^-A is idempotent and the rank of A thus is equal to the trace of A^-A*

Proof: A^- exists $\Rightarrow AA^-A = A \Rightarrow A^-AA^-A = A^-A$ showing that A^-A is idempotent. Furthermore, $R(A) \geq R(A^-A) \geq R(AA^-A) = R(A)$ and thus $R(A) = R(A^-A) = tr(A^-A)$. □

Lemma 5.24 *If A is a symmetric $n \times n$ matrix, so is A^-*

Proof: We apply the spectral theorem for symmetric matrices [24, Theorem 4.1.5], to A and obtain

$$A = P^T D P,$$

where P is orthogonal and $D := diag(d_1, \ldots, d_N)$ is diagonal. Now we set

$$A^- := P^T D^- P,$$

with

$$D^- := diag(i(d_1), \ldots, i(d_N)),$$

where

$$i(x) := \begin{cases} 0 & : \quad x = 0 \\ \frac{1}{x} & : \quad x \neq 0 \end{cases}$$

This is an appropriate choice for A^-:

$$\begin{aligned} AA^-A &= P^T D P P^T D^- P P^T D P \\ &= P^T D D^- D P \\ &= P^T D P \\ &= A, \end{aligned}$$

since $DD^-D = D$, which can be checked observing that the first product is a diagonal matrix containing only values zero and one, where zero can only arise in the case, when the corresponding d_i is zero. □

5.6.2 Quadratic forms in weak inverses

Employing these concepts we now are able to deduce a generalization of the theorem of Pearson. We will consider an arbitrary multidimensional normal variate $\mathbf{Y} \sim \mathcal{N}(\mu, C)$. This section contains the proof of the following

Theorem 5.25 *Let $\mathbf{Y} \sim \mathcal{N}(\mu, C)$ and set $r := R(C)$ the dimension of the distribution. Then the quadratic form $(\mathbf{Y} - \mu)^T C^- (\mathbf{Y} - \mu)$ has the following distribution,*

$$(\mathbf{Y} - \mu)^T C^- (\mathbf{Y} - \mu) \sim \chi^2_r,$$

providing a test-statistic for the multidimensional normal variate.

The proof for the theorem will be undertaken up in several steps. Let us first introduce some notations and a lemma.

Definition 5.5 (Noncentral χ^2 distribution)
If $X_1, \ldots, X_k$ are independent and identically distributed normal variables with zero mean and unit variance and $x_1, \ldots, x_k \in \mathbf{R}$, then the variable

$$X := (X_1 - x_1)^2 + \cdots + (X_k - x_k)^2$$

is distributed noncentral $\chi^2_{(k,\delta)}$, where k denotes the degrees of freedom and $\delta = \sum_{i=1}^k x_i^2$ is called the noncentrality parameter. If $\delta = 0$, this is the usual χ^2 distribution.

Lemma 5.26 *Let $Y_i \sim \mathcal{N}(\mu_i, 1), \quad i \in \{1, \ldots, n\}$, be independent normal variates, then*

$$\sum_{i=1}^n \lambda_i Y_i^2 + 2\sum_{i=1}^n b_i Y_i + c =: Y \sim \chi^2_{(k,\delta)}$$

where $k = \sum_{i=1}^n \lambda_i$, and $\delta = \sum_{i=1}^n \lambda_i(\mu_i + b_i)^2$, if

$$\lambda_i \in \{0, 1\}$$

$$b_i = \begin{cases} 0 & : \quad \lambda_i = 0 \\ \in \mathbf{R} & : \quad otherwise \end{cases}$$

$$c = \sum_{i=1}^n b_i^2$$

Proof: Without loss of generality, let $\lambda_i = 1$ for every i, and put $c = \sum_{i=1}^n b_i^2$. Then we have

$$Y = \sum_{i=1}^n (Y_i + b_i)^2 = \sum_{i=1}^n Z_i^2$$

where $Z_i \sim \mathcal{N}(\mu_i + b_i, 1)$ and the Z_i are independent normal variates. Thus Y is distributed χ^2 with noncentrality parameter δ and k degrees of freedom. □

The next lemma contains the main part of the proof.

Lemma 5.27 *Let $\mathbf{Y} \sim \mathcal{N}(\mu, I_n)$, where I_n denotes the n–dimensional identity matrix, then*

$$\mathbf{Y}^T A \mathbf{Y} + 2\mathbf{b}^T \mathbf{Y} + c \sim \chi^2_{(k,\delta)}$$

if

$$A = A^T, A^2 = A, \mathbf{b} \in \mathcal{M}(A), \text{ and } c = \mathbf{b}^T \mathbf{b}$$

where $k = R(A)$ and $\delta = (\mathbf{b} + \mu)^T A(\mathbf{b} + \mu)$.

Proof: Since A is symmetric, there exists an orthogonal matrix P such that $P^TAP = D$, where D denotes a diagonal matrix containing the eigenvalues of A. If we now set $\mathbf{Z} = P^T\mathbf{Y}$, where $\mathbf{Z} \sim \mathcal{N}(P^T\mu, P^TP = I)$, we get, substituting $\mathbf{Y} = P\mathbf{Z}$,

$$\mathbf{Y}^TA\mathbf{Y} + 2\mathbf{b}^T\mathbf{Y} + c = \mathbf{Z}^TD\mathbf{Z} + 2\mathbf{b}^TP\mathbf{Z} + c \tag{5.12}$$

This is of the form discussed in lemma 5.26. Thus (5.12) is distributed $\chi^2_{(k,\delta)}$ if

1. D is diagonal containing only values of zero or one
2. $c = \mathbf{b}^TPP^T\mathbf{b}$
3. the i'th coordinate in $\mathbf{b}^TP$ is zero, if the i'th diagonal element in D is zero

The first condition is fulfilled if $A^2 = A$ since the eigenvalues of an idempotent matrix can only be zero or one. The second condition also is fulfilled setting $c = \mathbf{b}^T\mathbf{b}$, since $PP^T = I$, the identity matrix. The third condition is implied by $\mathbf{b} \in \mathcal{M}(A)$: if $\mathbf{b} \in \mathcal{M}(A)$, there exists an $\mathbf{y}$ with $A\mathbf{y} = \mathbf{b}$. Since P is orthogonal, $\mathbf{y}$ can be expressed by $\mathbf{y} = P\mathbf{x}$ and multiplying with P^T from the left yields $P^TAP\mathbf{x} = P^T\mathbf{b}$, or $D\mathbf{x} = P^T\mathbf{b}$. Thus $P^T\mathbf{b} \in \mathcal{M}(D)$ which implies the third condition. The parameters k and δ calculate to

$$\begin{aligned} k = \sum_{i=1}^{n} \lambda_i = tr(D) = tr(P^TAP) = tr((APP^T)^T) \\ = tr(A^T) = tr(A) = R(A), \end{aligned}$$

since A is idempotent and we can apply (5.21), and

$$\delta = (P^T\mathbf{b} + P^T\mu)^TD(P^T\mathbf{b} + P^T\mu) = (\mathbf{b} + \mu)^TA(\mathbf{b} + \mu)$$

This completes the proof. □

Now switch to an arbitrary multidimensional normal variate $\mathbf{Y} \sim \mathcal{N}(\mu, C)$. The next lemma extends Lemma 5.27 to this case.

Lemma 5.28 *Let $\mathbf{Y} \sim \mathcal{N}(\mu, C)$ and denote by $r := R(C)$ the dimension of the distribution, and by n the number of coordinates, that is the dimension of the vector $\mathbf{Y}$. The statistic*

$$\mathbf{Y}^TA\mathbf{Y} + 2\mathbf{b}^T\mathbf{Y} + c$$

is distributed $\chi^2_{(k,\delta)}$, where

$$k = tr(AC), \quad \delta = (\mathbf{b} + A\mu)^TCAC(\mathbf{b} + A\mu),$$

if

1. *$CACAC = CAC$, and $A = A^T$*

2. $C(A\mu + \mathbf{b}) \in \mathcal{M}(CAC)$
3. $(A\mu + \mathbf{b})^T C(A\mu + \mathbf{b}) = \mu^T A\mu + 2\mathbf{b}^T\mu + c$.

Proof: First, note that A is square. We have already seen that $\mathbf{Y}$ can be represented by $\mathbf{Y} = \mu + B\mathbf{Z}$, where $\mathbf{Z} \sim \mathcal{N}(\mu, I_m)$ and $BB^T = C$: since C is symmetric, we can decompose into in the form $C = PDP^T$, where P is orthogonal and D is diagonal, due to the already cited spectral theorem (see [24], Theorem 4.1.5). Now denote by $\bar{D}$ a diagonal matrix containing the square roots of the diagonal elements in D, that is

$$\bar{D}\bar{D} = \bar{D}\bar{D}^T = D$$

then

$$C = P\bar{D}\bar{D}^T P^T = BB^T$$

where $B = P\bar{D}$. Putting r the rank of C, $r = R(C)$, we can assume without loss of generality that the r nonzero elements of D, respectively $\bar{D}$ appear in the first r rows and lines. This is possible, because line– and row–switching is an orthogonal operation. We now rewrite B as $n \times r$ matrix and B^T as $r \times n$ matrix by simply cutting $\bar{D}$ to a $r \times r$ matrix. The rank of B is r, since $r \geq R(B) \geq R(BB^T) = R(C) = r$. Notice, that $B^T B$ is an $r \times r$ matrix with full rank, since $B^T B = \bar{D}^T P^T P\bar{D} = \bar{D}^T\bar{D}$, which is an $r \times r$ diagonal matrix containing nonzero elements in the diagonal. This implies that there exists an inverse $(B^T B)^{-1}$. Now substitute in the above expression

$$\begin{aligned} \mathbf{Y}^T A\mathbf{Y} + 2\mathbf{b}^T\mathbf{Y} + c &= (\mu + B\mathbf{Z})^T A(\mu + B\mathbf{Z}) + 2\mathbf{b}^T(\mu + B\mathbf{Z}) + c \\ &= \mathbf{Z}^T B^T AB\mathbf{Z} + 2(A\mu + \mathbf{b})^T C\mathbf{Z} + \\ &\quad + \mu^T A\mu + 2\mathbf{b}^T\mu + c \end{aligned}$$

From Lemma 5.27 we know that this expression is $\chi^2_{(k,\delta)}$, if

1. $B^T AB$ is idempotent and symmetric
2. $B^T(A\mu + \mathbf{b}) \in \mathcal{M}(B^T AB)$
3. $(A\mu + \mathbf{b})^T BB^T(A\mu + \mathbf{b}) = \mu^T A\mu + 2\mathbf{b}^T\mu + c$

Now if $CACAC = CAC$, then $BB^T ABB^T ABB^T = BB^T ABB^T$. Multiplying with B^T from the left and with B from the right yields

$$B^T BB^T ABB^T ABB^T B = B^T BB^T ABB^T B$$

Since $B^T B$ is nonsingular we can multiply with its inverse from the left and from the right,

$$B^T ABB^T AB = B^T AB,$$

proving the first part of condition (1). The second part is fulfilled trivially since

$$(B^T AB)^T = B^T A^T B = B^T AB$$

Next, if $C(A\mu + \mathbf{b}) \in \mathcal{M}(CAC)$, then

$$C(A\mu + \mathbf{b}) = CAC\xi$$

for some ξ. Substituting BB^T for C results in

$$BB^T(A\mu + \mathbf{b}) = BB^TABB^T\xi$$

Setting $\xi' := B^T\xi$ and multiplying with B^T from the left yields

$$B^TBB^T(A\mu + \mathbf{b}) = B^TBB^TAB\xi'$$

Now, since B^TB is nonsingular, we get

$$B^T(A\mu + \mathbf{b}) = B^TAB\xi'$$

or, equivalently,

$$B^T(A\mu + \mathbf{b}) \in \mathcal{M}(B^TAB)$$

At last, the third condition is fulfilled since $C = BB^T$ in the third hypothesis of our lemma. Applying the formulas for k and δ in Lemma 5.27 and Corollary 5.21, we get

$$k = R(B^TAB) = tr(B^TAB) = tr(ABB^T) = tr(AC)$$

and

$$\delta = (\mathbf{b} + A\mu)^TBB^TABB^T(\mathbf{b} + A\mu) = (\mathbf{b} + A\mu)^TCAC(\mathbf{b} + A\mu).$$

□

We are now able to prove Theorem 5.25. Setting $A := C^-$, the hypotheses in Lemma 5.28 are fulfilled since the inverse of a symmetric matrix has been shown to be symmetric itself and

$$CACAC = CC^-CC^-C = CC^-C = CAC.$$

Since $(\mathbf{Y} - \mu) \sim \mathcal{N}(0, C)$, we have

$$C(C^-\emptyset + \emptyset) = \emptyset \in \mathcal{M}(CC^-C) = \mathcal{M}(C)$$

and, trivially,

$$0 + c = 0 + 0 = 0$$

Because C^-C is idempotent, the parameters calculate to

$$k = tr(C^-C) = R(C)$$

and

$$\delta = 0$$

This completes the proof of the theorem. □

5.7 A weak inverse for V

Let us sum up the results obtained in the previous sections. We have shown that the vector of counters for the overlapping M–tuples in a cyclic sequence of length N converges asymptotically to a multivariate normal distribution. The expectation E and covariance matrix $N \cdot V$ of this distribution have been calculated. It has also been shown that this convergence in distribution stays valid if we apply a transformation to the counter vectors and to the asymptotic distribution as long as the set of discontinuities of the transformation has measure zero. In particular, if

$$\mathbf{W}_N \Longrightarrow \mathbf{Y}_N \quad \mathbf{Y}_N \sim \mathcal{N}(E, NV)$$

then

$$h(\mathbf{W}_N) \Longrightarrow h(\mathbf{Y}_N)$$

for $h(\xi) = (\xi - E)^T (NV)^- (\xi - E)$, where $(NV)^-$ denotes a weak inverse of NV. From the last section, we know the distribution of $h(\mathbf{Y}_N)$: it is χ^2 with $R(NV) = R(V)$ degrees of freedom. The remaining problems are calculating the weak inverse for NV, that is, finding a matrix $(NV)^-$ for which we have

$$(NV)(NV)^-(NV) = (NV),$$

and determining the rank of V. From the linear dependencies calculated in Section 5.2.1, we conclude, that $R(V) \leq \alpha^M - \alpha^{M-1}$. Some tricky operations will show that the inequality is sharp. A weak inverse for NV was given by Good in [15] and by Marsaglia in [31]. The statistic $h(\mathbf{W}_N)$ turns out to be remarkably easy to evaluate. Denote by $\mathbf{a}$ and $\mathbf{b}$ arbitrary M–tuples, by $\mathbf{a}^\star$ the $(M-1)$–tuple generated by the first $(M-1)$ symbols in $\mathbf{a}$, and by $\mathbf{a}^{(M)}$ the last symbol in $\mathbf{a}$.

Proposition 5.29 (A weak inverse for NV) *A weak inverse for NV is given by*

$$(NV)^-_{\mathbf{a},\mathbf{b}} := \begin{cases} 0 & \text{if} \quad \mathbf{a}^\star \neq \mathbf{b}^\star \\ -\frac{\alpha^{M-1}}{N} & \text{if} \quad \mathbf{a}^\star = \mathbf{b}^\star \text{ and } \mathbf{a}^{(M)} \neq \mathbf{b}^{(M)} \\ \frac{\alpha^M}{N} - \frac{\alpha^{M-1}}{N} & \text{if} \quad \mathbf{a} = \mathbf{b} \end{cases}$$

The rank of NV is

$$R(V) = \alpha^M - \alpha^{M-1}$$

This weak inverse yields the formula given for the M–tuple statistic in Chapter 3. If we set

$$\chi_o := h(\mathbf{W}_N) = (\mathbf{W}_N - E)^T (NV)^- (\mathbf{W}_N - E)$$

and carry out the matrix multiplication, we arrive at

$$\begin{aligned} \chi_o \quad := \quad & \sum_{\mathbf{a} \in \mathrm{A}^M} \frac{(\mathbf{W}_N^{\mathbf{a}} - E(\mathbf{W}_N^{\mathbf{a}}))^2}{E(\mathbf{W}_N^{\mathbf{a}})} \\ - \quad & \sum_{\mathbf{a}' \in \mathrm{A}^{M-1}} \frac{(\mathbf{W}_N'^{\mathbf{a}'} - E(\mathbf{W}_N'^{\mathbf{a}'}))^2}{E(\mathbf{W}_N'^{\mathbf{a}'})}, \end{aligned}$$

where $\mathbf{W}_N'^{\mathbf{a}'}$ denotes a counter variable counting the $(M-1)$– tuple $\mathbf{a}'$. Note that the first summand is identical to the test statistic for nonoverlapping tuples, e.g. the usual goodness–of–fit test. Since the linear dependencies force the counter vector to live on an affine subspace with lower dimension than α^M, the vector will, in average, differ from its expectation by a greater value than the original χ^2 test statistic. We thus have to correct the statistic by subtracting the 'distance' of the $(M-1)$–tuples from their expectation. It is clear that we have to pay with a loss of some degrees of freedom for the estimation of the average error.

The statements in the proposition are proved subsequently.

5.7.1 Part 1: $(NV)(NV)^-(NV) = (NV)$

The proof of this relation requires only elementary mathematics. Note, that

$$E(\mathbf{W}_N^{\mathbf{a}}) = \frac{N}{\alpha^M}$$

for every M–tuple $\mathbf{a}$, and that

$$(NV)_{\mathbf{a},\mathbf{b}} = E(\mathbf{W}_N^{\mathbf{a}}\mathbf{W}_N^{\mathbf{b}}) - E(\mathbf{W}_N^{\mathbf{a}})E(\mathbf{W}_N^{\mathbf{b}})$$

For convenience denote NV by C. For arbitrary M–tuples $\mathbf{u}$ and $\mathbf{v}$, we have to calculate

$$\begin{aligned}
(CC^-C)_{\mathbf{u},\mathbf{v}} &= \sum_{\mathbf{i}\in \mathrm{A}^M}\sum_{\mathbf{j}\in \mathrm{A}^M} C_{\mathbf{u},\mathbf{i}}C^-_{\mathbf{i},\mathbf{j}}C_{\mathbf{j},\mathbf{v}} \\
&= \sum_{\mathbf{i}}\sum_{\mathbf{j}} C^-_{\mathbf{i},\mathbf{j}} \left(\frac{1}{\alpha^N}\sum_{p\in \mathrm{A}^N} \mathbf{W}_N^{\mathbf{u}}(p)\mathbf{W}_N^{\mathbf{i}}(p) - \frac{N^2}{\alpha^{2M}} \right) \cdot \\
&\quad \cdot \left(\frac{1}{\alpha^N}\sum_{q\in \mathrm{A}^N} \mathbf{W}_N^{\mathbf{j}}(q)\mathbf{W}_N^{\mathbf{v}}(q) - \frac{N^2}{\alpha^{2M}} \right) \\
&= \sum_{\mathbf{i}}\sum_{\mathbf{j}} C^-_{\mathbf{i},\mathbf{j}} \frac{1}{\alpha^{2N}} \left\{ \Sigma_1 - \Sigma_2 + \Sigma_3 \right\}
\end{aligned}$$

where

$$\begin{aligned}
\Sigma_1 &:= \sum_p\sum_q \mathbf{W}_N^{\mathbf{u}}(p)\mathbf{W}_N^{\mathbf{i}}(p)\mathbf{W}_N^{\mathbf{j}}(q)\mathbf{W}_N^{\mathbf{v}}(q) \\
\Sigma_2 &:= \frac{N^2\alpha^N}{\alpha^{2M}} \left(\sum_p \mathbf{W}_N^{\mathbf{u}}(p)\mathbf{W}_N^{\mathbf{i}}(p) + \sum_q \mathbf{W}_N^{\mathbf{j}}(q)\mathbf{W}_N^{\mathbf{v}}(q) \right) \\
\Sigma_3 &:= \frac{N^4\alpha^{2N}}{\alpha^{4M}}
\end{aligned}$$

According to the definition of the values of C^-, the expression $\sum_i \sum_j C^-_{i,j}\frac{1}{\alpha^{2N}}$ can be written as a sum over the diagonal elements of C^- and over a block-structure. Therefore

$$\begin{aligned}
(CC^-C)_{\mathbf{u},\mathbf{v}} &= \tau_1 \sum_{\mathbf{i}\in \mathrm{A}^M} \left\{ \Sigma_1 - \Sigma_2 + \Sigma_3 \right\} + \\
&+ \tau_2 \sum_{\mathbf{a}'\in \mathrm{A}^{M-1}} \sum_{k\in \mathrm{A}} \sum_{l\in \mathrm{A}} \left\{ \Sigma_1 - \Sigma_2 + \Sigma_3 \right\},
\end{aligned}$$

where we set $\mathbf{j} = \mathbf{i}$ in the first sum, and $\mathbf{i} = \mathbf{a}', k$ and $\mathbf{j} = \mathbf{a}', l$ in the second.The values τ_1 and τ_2 can be computed from $(NV)^-$

$$\begin{aligned} \tau_1 &= \frac{\alpha^{M-2N}}{N} \\ \tau_2 &= -\frac{\alpha^{M-2N-1}}{N} \end{aligned}$$

and we have $\tau_1 = -\tau_2\alpha$

We will now carry out the summation in six parts for each combination of τ and Σ. The symmetries in the problem lead us to expect the cancellation of the main part of the combinations.

Combination 1: τ_1 and Σ_2

$$\begin{aligned} C1 &= \tau_1\left(-\frac{N^2\alpha^N}{\alpha^{2M}}\right)\sum_{\mathbf{i}}\left(\sum_p \mathbf{W}^{\mathbf{u}}_N(p)\mathbf{W}^{\mathbf{i}}_N(p) + \sum_q \mathbf{W}^{\mathbf{i}}_N(q)\mathbf{W}^{\mathbf{v}}_N(q)\right) \\ &= \tau_1\left(-\frac{N^2\alpha^N}{\alpha^{2M}}\right)\left\{\left(\sum_p \mathbf{W}^{\mathbf{u}}_N(p)\underbrace{(\sum_{\mathbf{i}}\mathbf{W}^{\mathbf{i}}_N(p))}_{=N}\right) + \right. \\ &\quad \left. + \left(\sum_q \mathbf{W}^{\mathbf{v}}_N(q)\underbrace{(\sum_{\mathbf{i}}\mathbf{W}^{\mathbf{i}}_N(q))}_{=N}\right)\right\} \\ &= \tau_1\left(-\frac{N^2\alpha^N}{\alpha^{2M}}\right)N\left(\underbrace{\sum_p \mathbf{W}^{\mathbf{u}}_N(p)}_{=N\alpha^{N-M}} + \underbrace{\sum_q \mathbf{W}^{\mathbf{v}}_N(q)}_{=N\alpha^{N-M}}\right) \\ &= -\frac{2N^3}{\alpha^{2M}} \end{aligned}$$

Combination 2: τ_1 and Σ_3

$$\begin{aligned} C2 &= \tau_1\frac{N^4\alpha^{2N}}{\alpha^4 M}\left(\sum_{\mathbf{i}} 1\right) \\ &= \tau_1\frac{N^4\alpha^{2N}}{\alpha^4 M}\alpha^M \\ &= \frac{N^3}{\alpha^{2M}} \end{aligned}$$

Thus

$$C1 + C2 = \frac{-N^3}{\alpha^{2M}}$$

Combination 3: τ_2 and Σ_2

$$\begin{aligned}
C3 &= \tau_2 \left(-\frac{N^2\alpha^N}{\alpha^{2M}}\right) \sum_{\mathbf{a}' \in \mathrm{A}^{M-1}} \sum_{k \in \mathrm{A}} \sum_{l \in \mathrm{A}} \left(\sum_p \mathbf{W}_N^{\mathbf{u}}(p) \mathbf{W}_N^{\mathbf{a}',k}(p) + \right. \\
&\quad \left. + \sum_q \mathbf{W}_N^{\mathbf{a}',l}(q) \mathbf{W}_N^{\mathbf{v}}(q) \right) \\
&= \tau_2 \left(-\frac{N^2\alpha^N}{\alpha^{2M}}\right) \left\{ \underbrace{\left(\sum_l 1\right)}_{=\alpha} \left(\underbrace{\sum_p \mathbf{W}_N^{\mathbf{u}}(p)}_{=N\alpha^{N-M}} \underbrace{\left(\sum_{\mathbf{a}'} \sum_k \mathbf{W}_N^{\mathbf{a}',k}(p) \right)}_{=N, \text{ indep. of } p} \right) + \right. \\
&\quad \left. + \underbrace{\left(\sum_k 1\right)}_{=\alpha} \left(\underbrace{\sum_q \mathbf{W}_N^{\mathbf{v}}(q)}_{=N\alpha^{N-M}} \underbrace{\left(\sum_{\mathbf{a}'} \sum_l \mathbf{W}_N^{\mathbf{a}',l}(q) \right)}_{=N, \text{ indep. of } p} \right) \right\} \\
&= \tau_2 \left(-\frac{N^2\alpha^N}{\alpha^{2M}}\right) \{\alpha N^2 \alpha^{N-M} 2\} \\
&= \frac{2N^3}{\alpha^{2M}}
\end{aligned}$$

Combination 4: τ_2 and Σ_3

$$\begin{aligned}
C4 &= \tau_2 \frac{N^4\alpha^{2N}}{\alpha^{4M}} \left(\sum_{\mathbf{a}'} 1\right) \left(\sum_k 1\right) \left(\sum_l 1\right) \\
&= -\frac{N^3}{\alpha^{2M}}
\end{aligned}$$

Thus

$$C3 + C4 = \frac{N^3}{\alpha^{2M}}$$

As has been expected, the first four combinations cancel to zero and the value of

$$(CC^-C)_{\mathbf{u},\mathbf{v}}$$

will be determined only by the last two combinations.

Combination 5: τ_1 and Σ_1

Suppose again that p and q are sequences of length N consisting of symbols in A. We will write $\delta(\mathbf{p}_k = \mathbf{q}_l)$ to denote a function that equals one, if the $M-$ tuples

$$(p_k, p_{k\oplus 1}, \ldots, p_{k\oplus(M-1)})$$

and

$$(q_l, q_{l\oplus 1}, \ldots, q_{l\oplus(M-1)})$$

are equal, and zero otherwise.

$$\begin{aligned}
C5 &= \tau_1 \sum_{\mathbf{i}\in \mathrm{A}^M} \sum_{p\in \mathrm{A}^N} \sum_{q\in \mathrm{A}^N} \mathbf{W}^{\mathbf{u}}_N(p)\mathbf{W}^{\mathbf{i}}_N(p)\mathbf{W}^{\mathbf{i}}_N(q)\mathbf{W}^{\mathbf{v}}_N(q) \\
&= \tau_1 \sum_p \sum_q \mathbf{W}^{\mathbf{u}}_N(p)\mathbf{W}^{\mathbf{v}}_N(q) \sum_{\mathbf{i}\in \mathrm{A}^M} \left(\sum_{k=0}^{N-1} \sum_{l=0}^{N-1} \delta(\mathbf{p}_k = \mathbf{i} \wedge \mathbf{q}_l = \mathbf{i}) \right) \\
(*1) &= \tau_1 \sum_p \sum_q \mathbf{W}^{\mathbf{u}}_N(p)\mathbf{W}^{\mathbf{v}}_N(q) \sum_k \sum_l \delta(\mathbf{p}_k = \mathbf{q}_l) \\
&= \tau_1 \sum_p \mathbf{W}^{\mathbf{u}}_N(p) \sum_q \sum_{k,l,m=0}^{N-1} \delta(\mathbf{p}_k = \mathbf{q}_l \wedge \mathbf{q}_m = \mathbf{v})
\end{aligned}$$

The line marked $(*1)$ follows from the fact that the indicator is nonzero for at most one $\mathbf{i} = \mathbf{p}_k$, but only if $\mathbf{p}_k = \mathbf{q}_l$. We will now substitute the summation index m by d, where d expresses the distance to index l. The substitution takes into account the cyclic structure of p and q. Thus

$$\begin{aligned}
C5 &= \tau_1 \sum_p \mathbf{W}^{\mathbf{u}}_N(p) \sum_{k,l=0}^{N-1} \sum_{d=-\lfloor \frac{N}{2} \rfloor}^{\lceil \frac{N}{2} \rceil - 1} \sum_q \delta(\mathbf{p}_k = \mathbf{q}_l \wedge \mathbf{q}_{l\oplus d} = \mathbf{v}) \\
(*2) &= \tau_1 \alpha^{N-2M} \sum_p \mathbf{W}^{\mathbf{u}}_N(p) \sum_{k,l,d} \alpha^{c(\mathbf{p}_k, \mathbf{v}, d)} \delta^*(\mathbf{p}_k, \mathbf{v}, d) \\
&= \tau_1 N \alpha^{N-2M} \sum_p \mathbf{W}^{\mathbf{u}}_N(p) \sum_{k,d} \alpha^{c(\mathbf{p}_k, \mathbf{v}, d)} \delta^*(\mathbf{p}_k, \mathbf{v}, d)
\end{aligned}$$

Now, what has happened in $(*2)$? In the sequence q, $2M - c(\mathbf{p}_k, \mathbf{v}, d)$ symbols are fixed by the M–tuples $\mathbf{p}_k$ and $\mathbf{v}$, where we denote by $c(\mathbf{p}_k, \mathbf{v}, d)$ the number of symbols of the two M–tuples, that overlap at distance d, that is, the number of symbols in q that will be fixed by symbols in the two tuples simultaneously, i.e., in 'c'ommon. However, the indicator in the preceding line only gives a nonzero value, if for every such position the symbols imposed on q by the two M–tuples are equal. This is expressed by the new indicator function $\delta^*(\mathbf{p}_k, \mathbf{v}, d)$, which becomes one iff the symbols in $\mathbf{p}_k$ and $\mathbf{v}$ that overlap at the given distance d, match each other. In this case the indicator in the preceding line will be one for exactly $\alpha^{N-2M+c(\cdot)}$ sequences q, since this is the number of symbols in the sequence that are not occupied by any symbol in $\mathbf{p}_k$ or $\mathbf{v}$.

The formal definition for $\delta^*(\cdot)$ is given by

Definition 5.6

$$\delta^*((b_0,b_1,\ldots,b_{k_1}),(c_0,c_1,ldots,c_{k_2}),d) := \prod_{k=max\{0,d\}}^{min\{k_1,k_2+d\}} \delta(b_k = c_{k-d})$$

Note, that the tuples **b** and **c** do not need to have the same length!

We will continue evaluating this sum later. Using the same ideas, we treat the last remaining term of CC^-C:

Combination 6: τ_2 and Σ_1

We will use the notation $\mathbf{a}^\star$ in order to denote an $M-1$–tuple generated by truncating the last component of **a**. By $\mathbf{a}^\star, j$ we denote the M–tuple which results by changing the last symbol in **a** to $j \in \mathrm{A}$.

$$\begin{aligned} C6 &= \tau_2 \sum_{\mathbf{i}\in \mathrm{A}^M} \sum_{j\in\mathrm{A}} \sum_{p,q\in\mathrm{A}^n} \mathbf{W}^{\mathbf{u}}_N(p)\mathbf{W}^{\mathbf{i}}_N(p)\mathbf{W}^{\mathbf{i}^\star,j}_N(q)\mathbf{W}^{\mathbf{v}}_N(q) \\ &= \tau_2 \sum_p \mathbf{W}^{\mathbf{u}}_N(p) \sum_q \sum_{k,l=0}^{N-1} \sum_{\mathbf{i}\in\mathrm{A}^M} \sum_{j\in\mathrm{A}} \delta(\mathbf{p}_k = \mathbf{i} \wedge \mathbf{q}_l = \mathbf{i}^\star, j) \\ &= \tau_2 \sum_p \mathbf{W}^{\mathbf{u}}_N(p) \sum_{k,l,m=0}^{N-1} \sum_q \delta(\mathbf{p^\star}_k = \mathbf{q^\star}_l \wedge \mathbf{q}_m = \mathbf{v}) \end{aligned}$$

The last line resulted from the observation that the indicator in the preceding line can only be nonzero for one $\mathbf{i} = \mathbf{p}_k$, in which case $\mathbf{p^\star}_k$ has to be equal to $\mathbf{q^\star}_l$. We can always find exactly one j satisfying $\mathbf{q^\star}_l = \mathbf{i}^\star, j$ then. Now continue introducing the distance d and the functions c and $\delta^*()$ as in the previous combination:

$$\begin{aligned} C6 &= \tau_2 \sum_p \mathbf{W}^{\mathbf{u}}_N(p) \sum_{k,l=0}^{N-1} \sum_{d=-\lfloor \frac{N}{2} \rfloor}^{\lceil \frac{N}{2} \rceil - 1} \sum_q \delta(\mathbf{p^\star}_k = \mathbf{q^\star}_l \wedge \mathbf{q}_{l\oplus d} = \mathbf{v}) \\ &= \tau_2 \alpha^{N-2M+1} N \sum_p \mathbf{W}^{\mathbf{u}}_N(p) \sum_k \sum_d \alpha^{c(\mathbf{p^\star}_k, \mathbf{v}^\star, d)} \delta^*(\mathbf{p^\star}_k, \mathbf{v}^\star, d) \\ (*3) &= -\tau_1 \alpha^{N-2M} N \sum_p \mathbf{W}^{\mathbf{u}}_N(p) \sum_k \sum_d \alpha^{c(\mathbf{p^\star}_k, \mathbf{v}^\star, d)} \delta^*(\mathbf{p^\star}_k, \mathbf{v}^\star, d) \end{aligned}$$

In $(*3)$, we made use of $\tau_1 = -\tau_2\alpha$. We now can write the last two combinations in one sum.

$$C5 + C6 = \tau_1 N \alpha^{N-2M} \sum_{p\in\mathrm{A}^N} \mathbf{W}^{\mathbf{u}}_N(p) \sum_{k=0}^{N-1} \sum_{d=-\lfloor \frac{N}{2} \rfloor}^{\lceil \frac{N}{2} \rceil - 1} (\Upsilon_1 - \Upsilon_2) \tag{5.13}$$

where

$$\begin{aligned} \Upsilon_1 &:= \alpha^{c(\mathbf{p}_k, \mathbf{v}, d)} \delta^*(\mathbf{p}_k, \mathbf{v}, d) \\ \Upsilon_2 &:= \alpha^{c(\mathbf{p^\star}_k, \mathbf{v}^\star, d)} \delta^*(\mathbf{p^\star}_k, \mathbf{v}^\star, d) \end{aligned}$$

For every $d < -M+1$ or $d > M-1$ we have

$$c(\mathbf{p}_k, \mathbf{v}, d) = c(\mathbf{p}^\star{}_k, \mathbf{v}^\star, d) = 0$$

because such distances prevent the tuples from overlapping each other, and thus trivially

$$\delta^*(\mathbf{p}_k, \mathbf{v}, d) = \delta^*(\mathbf{p}^\star{}_k, \mathbf{v}^\star, d) = 1$$

Hence $\Upsilon_1 = \Upsilon_2$ for these values of d and the terms cancel from the expression. If $-M+1 \leq d < 0$, we have

$$c(\mathbf{p}_k, \mathbf{v}, d) = c(\mathbf{p}^\star{}_k, \mathbf{v}^\star, d) = M - d,$$

since $M-d$ is the number of symbols that overlap, and

$$\delta^*(\mathbf{p}_k, \mathbf{v}, d) = \delta^*(\mathbf{p}^\star{}_k, \mathbf{v}^\star, d),$$

since the last symbol in $\mathbf{p}_k$ respectively $\mathbf{p}^\star{}_k$ will not overlap any symbol in $\mathbf{v}$ respectively $\mathbf{v}^\star$. Thus we again have $\Upsilon_1 = \Upsilon_2$ and no contribute to the sum of $C5$ and $C6$.

Now consider the two cases $d = 0$ and $d = M-1$:

- $d = 0$: In this case, the tuples in Υ_1 overlap in M symbols giving

  $$\Upsilon_1 = \alpha^M \delta(\mathbf{p}_k = \mathbf{v}).$$

  Summing up over all other variables, Υ_1 contributes for $d = 0$ the following value Ψ_1 to the sum of $C5$ and $C6$:

  $$\begin{aligned} \Psi_1 &= \tau_1 N \alpha^{N-2M} \alpha^M \sum_p \mathbf{W}_N^{\mathbf{u}}(p) \sum_{k=0}^{N-1} \delta(\mathbf{p}_k = \mathbf{v}) \\ &= \frac{1}{\alpha^N} \sum_p \mathbf{W}_N^{\mathbf{u}}(p) \mathbf{W}_N^{\mathbf{v}}(p) \\ &= E(\mathbf{W}_N^{\mathbf{u}} \mathbf{W}_N^{\mathbf{v}}) \end{aligned}$$

- $d = M-1$: Now the tuples in Υ_2 do not overlap since the distance is greater than the length of the $(M-1)$–tuples considered in Υ_2. We therefore get $\Upsilon_2 = \alpha^0 \cdot 1$, contributing Ψ_2 to the sum of $C5$ and $C6$, where

  $$\begin{aligned} \Psi_2 &= -\tau_1 N \alpha^{N-2M} \alpha^0 \sum_p \mathbf{W}_N^{\mathbf{u}}(p) \sum_{k=0}^{N-1} 1 \\ &= -\frac{N^2}{\alpha^{2M}} \\ &= -E(\mathbf{W}_N^{\mathbf{u}}) E(\mathbf{W}_N^{\mathbf{v}}) \end{aligned}$$

Note, that $\Psi_1 + \Psi_2$ already equals $C_{\mathbf{u},\mathbf{v}}$. We will have to show that the remaining sum, say, Ψ_3 of the terms in (5.13) cancels to zero. Writing down the according indices

($d = 1, \ldots, M-1$ for Υ_1, and $d = 0, \ldots, M-2$ for Υ_2) and summing up over all other summands yields

$$\begin{aligned}\Psi_3 &= \tau_1 N \alpha^{N-2M} \sum_p \mathbf{W}_N^{\mathbf{u}}(p) \sum_{k=0}^{N-1} \sum_{d=1}^{M-1} \cdot \\ &\quad \cdot \left(\alpha^{c(\mathbf{p}_k, \mathbf{v}, d)} \delta^*(\mathbf{p}_k, \mathbf{v}, d) - \alpha^{c(\mathbf{p^\star}_k, \mathbf{v^\star}, d-1)} \delta^*(\mathbf{p^\star}_k, \mathbf{v^\star}, d-1) \right)\end{aligned}$$

where we get

$$c(\mathbf{p}_k, \mathbf{v}, d) = M - d = c(\mathbf{p^\star}_k, \mathbf{v^\star}, d-1)$$

by counting the length of the segments where the two tuples overlap. Hence

$$\begin{aligned}\Psi_3 &= \tau_1 N \alpha^{N-2M} \sum_p \mathbf{W}_N^{\mathbf{u}}(p) \sum_{k=0}^{N-1} \sum_{d=1}^{M-1} \cdot \alpha^{M-d} \left(\delta^*(\mathbf{p}_k, \mathbf{v}, d) - \delta^*(\mathbf{p^\star}_k, \mathbf{v^\star}, d-1) \right) \\ &= \tau_1 N \alpha^{N-2M} \sum_p \mathbf{W}_N^{\mathbf{u}}(p) \sum_{d=1}^{M-1} \cdot \alpha^{M-d} \sum_{k=0}^{N-1} \left(\delta^*(\mathbf{p}_k, \mathbf{v}, d) - \delta^*(\mathbf{p^\star}_k, \mathbf{v^\star}, d-1) \right) \\ &= \tau_1 N \alpha^{N-2M} \sum_p \mathbf{W}_N^{\mathbf{u}}(p) \sum_{d=1}^{M-1} \cdot \alpha^{M-d} \sum_{k=0}^{N-1} \left(\delta^*(\mathbf{p}_k, \mathbf{v}, d) - \delta^*(\mathbf{p}_k, \mathbf{v}, d) \right) \\ &= 0\end{aligned}$$

because

$$\begin{aligned}\sum_{k=0}^{N-1} \delta^*(\mathbf{p^\star}_k, \mathbf{v^\star}, d-1) &= \sum_{k=0}^{N-1} \delta(p_{k \oplus d \ominus 1} = \mathbf{v}^{(1)}, \ldots, p_{k \oplus M \ominus 2} = \mathbf{v}^{(M-d)}) \\ \text{(cyclic structure)} &= \sum_{k=0}^{N-1} \delta(p_{k \oplus d} = \mathbf{v}^{(1)}, \ldots, p_{k \oplus M \ominus 1} = \mathbf{v}^{(M-d)}) \\ &= \sum_k \delta^*(\mathbf{p}_k, \mathbf{v}, d)\end{aligned}$$

We have succeeded in showing that $C5 + C6$ equals $C_{\mathbf{u},\mathbf{v}}$. Thus

$$CC^-C = C$$

This completes the proof. □

5.7.2 Part two: $R(V) = \alpha^M - \alpha^{M-1}$

Using the same type of calculations we will now show that the rank of the distribution of the vector of counter variables is $\alpha^M - \alpha^{M-1}$. Consider first the trace

$$tr(C^-C) = \sum_{\mathbf{v} \in \mathrm{A}^M} \sum_{\mathbf{i} \in \mathrm{A}^M} C^-_{\mathbf{v},\mathbf{i}} C_{\mathbf{i},\mathbf{v}}$$

We rewrite this expression using the formulas for the covariance given in Section 5.1 and get

$$\begin{aligned} tr(C^-C) &= \sum_{\mathbf{v}}\sum_{\mathbf{i}} C^-_{\mathbf{v},\mathbf{i}} \frac{1}{\alpha^N} \sum_{p\in \mathrm{A}^N} \left(\mathbf{W}^{\mathbf{i}}_N(p)\mathbf{W}^{\mathbf{v}}_N(p) - \frac{N^2}{\alpha^{2M}} \right) \\ &= \sum_{\mathbf{v}}\sum_{\mathbf{i}}\sum_{p} C^-_{\mathbf{v},\mathbf{i}} \frac{1}{\alpha^N} \left(\mathbf{W}^{\mathbf{i}}_N(p)\mathbf{W}^{\mathbf{v}}_N(p) - \frac{N^2}{\alpha^{2M}} \right) \end{aligned}$$

Let us recall the definition and the values of C^-: $C^-_{\mathbf{i},\mathbf{v}}$ is nonzero only if $\mathbf{i}^\star = \mathbf{v}^\star$. This enables us to substitute $\mathbf{v}^\star, j$ for $\mathbf{i}$ and sum up over all symbols in the alphabet.

$$\begin{aligned} tr(C^-C) &= \sum_{\mathbf{v}}\sum_{j\in \mathrm{A}}\sum_{p} C^-_{\mathbf{v},(\mathbf{v}^\star,j)} \frac{1}{\alpha^N} (\mathbf{W}^{\mathbf{v}^\star,j}_N(p)\mathbf{W}^{\mathbf{v}}_N(p) - \frac{N^2}{\alpha^{2M}}) \\ &= \sum_{\mathbf{v}}\sum_{p} \frac{\alpha^{M-N}}{N} (\mathbf{W}^{\mathbf{v}}_N(p)\mathbf{W}^{\mathbf{v}}_N(p) - \frac{N^2}{\alpha^{2M}}) + \\ &\quad + \sum_{\mathbf{v}}\sum_{j\in \mathrm{A}}\sum_{p} - \frac{\alpha^{M-N-1}}{N} (\mathbf{W}^{\mathbf{v}^\star,j}_N(p)\mathbf{W}^{\mathbf{v}}_N(p) - \frac{N^2}{\alpha^{2M}}) \end{aligned}$$

The last line results from applying the values of C^-. Put $\tau_3 := \frac{\alpha^{M-N}}{N}$, and $\tau_4 := -\frac{\alpha^{M-N-1}}{N}$. The sum splits into two parts, one for each definition case of C^-, where C^- has a nonzero value. The constants $\frac{N^2}{\alpha^{2M}}$ cancel from the total sum since the second sum has α times the number of summands than the first one and that $\tau_3 = -\tau_4\alpha$. The remaining terms are

$$\begin{aligned} tr(C^-C) &= \sum_{\mathbf{v}}\sum_{p}\sum_{k,l=0}^{N-1} \left(\tau_3 \delta(\mathbf{p}_k = \mathbf{v} \wedge \mathbf{p}_l = \mathbf{v}) + \tau_4 \sum_j \delta(\mathbf{p}_k = \mathbf{v}^\star, j \wedge \mathbf{p}_l = \mathbf{v}) \right) \\ &= \sum_{\mathbf{v}}\sum_{k,l=0}^{N-1}\sum_{p} (\tau_3 \delta(\mathbf{p}_k = \mathbf{v} \wedge \mathbf{p}_l = \mathbf{v}) + \tau_4 \delta({\mathbf{p}^\star}_k = \mathbf{v}^\star \wedge \mathbf{p}_l = \mathbf{v})) \\ &= \sum_{\mathbf{v}}\sum_{k} \sum_{d=-\lfloor \frac{N}{2} \rfloor}^{\lceil \frac{N}{2} \rceil - 1} \Big(\tau_3 \alpha^{N-2M+c(\mathbf{v},\mathbf{v},d)} \delta^*(\mathbf{v},\mathbf{v},d) + \\ &\quad \tau_4 \alpha^{N-2M+1+c(\mathbf{v}^\star,\mathbf{v},d)} \delta^*(\mathbf{v}^\star,\mathbf{v},d) \Big) \\ &= \frac{\alpha^{-M}}{N} \sum_{\mathbf{v}}\sum_{k}\sum_{d} (\Upsilon_3 - \Upsilon_4) \\ &= \frac{1}{\alpha^M} \sum_{\mathbf{v}}\sum_{d} (\Upsilon_3 - \Upsilon_4) \end{aligned}$$

where

$$\begin{aligned} \Upsilon_3 &:= \alpha^{c(\mathbf{v},\mathbf{v},d)} \delta^*(\mathbf{v},\mathbf{v},d) \\ \Upsilon_4 &:= \alpha^{c(\mathbf{v}^\star,\mathbf{v},d)} \delta^*(\mathbf{v}^\star,\mathbf{v},d) \end{aligned}$$

We now use exactly the same arguments as in the last section and conclude that for $d < 0$ and for $d > M-1$ we have $\Upsilon_3 = \Upsilon_4$, canceling these from the total sum. In the case $d = 0$, we have

$$\Upsilon_3 = \alpha^M \cdot 1 = \alpha^M$$

and

$$\Upsilon_4 = \alpha^{M-1} \cdot 1 = \alpha^{M-1}$$

For the remaining $d, 0 < d \leq M-1$, we evaluate the rest of the total sum:

$$\begin{aligned}
\frac{1}{\alpha^M} \sum_{\mathbf{v}} \sum_{d} (\Upsilon_3 - \Upsilon_4) &= \frac{1}{\alpha^M N} \sum_{\mathbf{v}} \sum_{d=1}^{M-1} \sum_{k} \underbrace{\alpha^{c(\mathbf{v},\mathbf{v},d)}}_{\alpha^{M-d}} \delta^*(\mathbf{v},\mathbf{v},d) - \underbrace{\alpha^{c(\mathbf{v}^\star,\mathbf{v},d)}}_{\alpha^{M-d-1}} \delta^*(\mathbf{v}^\star,\mathbf{v},d) \\
&= \frac{1}{\alpha^M N} \alpha^{M-d} \sum_{\mathbf{v}} \sum_{d=1}^{M-1} \sum_{k} \left(\delta^*(\mathbf{v},\mathbf{v},d) - \frac{1}{\alpha} \delta^*(\mathbf{v}^\star,\mathbf{v},d) \right) \\
&= \frac{1}{\alpha^M N} \alpha^{M-d} \sum_{d=1}^{M-1} \sum_{k} \sum_{\mathbf{v}' \in \mathrm{A}^{M-1}} \sum_{j \in \mathrm{A}} (\delta^*((\mathbf{v}',j),(\mathbf{v}',j),d) - \\
&\quad - \frac{1}{\alpha} \delta^*(\mathbf{v}',(\mathbf{v}',j),d) \Big)
\end{aligned}$$

The first $\delta^*()$ in the last line is nonzero only for one single j and if $\delta^*(\mathbf{v}',\mathbf{v}',d) = 1$. The second $\delta^*()$ is fulfilled for all j, that is α times, if $\delta^*(\mathbf{v}',\mathbf{v}',d) = 1$. Thus the two $\delta^*()$ cancel and we get

$$tr(C^- C) = \alpha^M - \alpha^{M-1},$$

thereby completing the proof. □

5.8 Putting the puzzle together

Let us sum up the results of this chapter. We have shown that the counter vector of the overlapping M–tuples is asymptotically distributed multivariate normal, see Proposition 5.16. We have calculated the covariance matrix of this normal distribution in Section 5.3 and have described a weak inverse for the covariance matrix. We know from Theorem 5.25 that a quadratic form in a weak inverse of the covariance matrix of a normal vector is distributed χ^2 with the rank of the covariance matrix degrees of freedom. This rank has been calculated in the last section. The quadratic form in the weak inverse turns out to be remarkably easy to evaluate if one considers the usual χ^2 test statistic for the M–tuples and subtracts a χ^2 test statistic for the $(M-1)$–tuples. The Mapping Theorem, Theorem 5.18, has been used to link the convergence in distribution before and after the transformation by the quadratic form.

The theory developed above can be used for similar test-statistics. It 'contains' a proof for the theorem of Pearson as well as for other usual pseudorandom number tests, e.g. the well–known run test proposed for example by Donald E. Knuth in [26]. The difficult part of such proofs is finding a (weak) inverse for the covariance matrix and calculating the rank of it. If the dimension of the distribution is maximal this can be done by common

mathematical software. [31] contains some further results for such test statistics. It also considers the very interesting case of sparse occupancy: if the number of counters in the vector (α^M) is very big, the computer memory needed to keep the array counting the hits will become unmanageably large. In addition, one will have to scan very long sequences in order to obtain enough expected hits per counter that guarantee a good approximation of the distribution of the test statistic by the asymptotic distribution. In the sparse occupancy test only one bit is maintained for every counter. The bit is set if the counter has been hit at least once. The test statistic is calculated from the number of counters that have been hit during the scanning period.

The family of χ^2 tests is suited well to testing pseudorandom numbers because the open structure of the tests permits us to construct rigid tests in order to show some numerical properties of common generators. The easy geometric interpretation of the counter variables makes the tests open to public acceptance. Last but not least, the evaluation of the test can be done in almost $O(N)$ steps, where N is the sample size, because the calculation of the test-statistic for reasonable α^M can be carried out very efficiently.

Appendix A

In this appendix, we provide the proof for the following

Lemma A.1 *If H is an idempotent matrix, $HH = H$, then the trace of H equals the rank of H and H is diagonizable, that is, there exists a regular matrix S for which we have $S^{-1}HS = D$, where D is a diagonal matrix containing the eigenvalues of H.*

Proof: First, H is square since it is idempotent. Assume that H is $n \times n$. If H is the zero matrix, the lemma holds trivially. Now assume that H is not zero. Let $p(\cdot)$ be a polynomial. It is clear what is understood by the i'th power of a square matrix A, and we thus can calculate the value of $p(A)$. We further say that $p(\cdot)$ annihilates A, if $p(A)$ is the zero matrix. The following theorem can be found in [24, Theorem 1.1.6]:

Theorem A.2 *Let A be a square matrix and $p(\cdot)$ a polynomial. If λ is an eigenvalue of A and ξ is a corresponding eigenvector, then $p(\lambda)$ is an eigenvalue of $p(A)$ with eigenvector ξ.*

Now, if $H^2 = H$, the polynomial

$$p(t) = t^2 - t = t(t-1)$$

annihilates H. Thus, if λ is an eigenvalue of H, $\lambda(\lambda - 1)$ is an eigenvalue of the zero matrix, which has only the eigenvalue zero. The only solutions for

$$\lambda(\lambda - 1) = 0$$

are one and zero. Thus $\lambda \in \{0, 1\}$. Note, that for a diagonal matrix containing only values of zero and one, the rank trivially equals the trace. Now suppose that H is diagonalizable, that is, it is similar to a diagonal matrix D containing the eigenvalues of H,

$$D = S^{-1}HS$$

for some appropriate S. Since both, rank and trace of a matrix are invariant under similarity, and D is diagonal containing only values of zero and one, we have

$$R(H) = R(D) = \sum_{i=1}^{n} \lambda_i = tr(D) = tr(H)$$

It remains to prove, that H is diagonalizable. We again cite [24, Corollary 3.3.10]:

Theorem A.3 *If every root of the minimal polynomial of a square matrix A has multiplicity one, then A is diagonalizable.*

The minimal polynomial of A is the polynomial with the lowest degree and leading coefficient one that annihilates A. We allready know that $p(t) = t^2 - t$ annihilates H. Since H is not zero, no polynomial of degree one will annihilate H and we have already found the minimal polynomial. Since every root of $p(t) = t^2 - t$ has multiplicity one, H is diagonalizable. This completes the proof. □

Bibliography

[1] N. S. Altman. Bit–wise behavior of random number generators. *SIAM J. Sci. Stat. Comput.*, **9**(5):941–949, 1988.

[2] T. Auer, K. Entacher, P. Hellekalek, H. Leeb, O. Lendl, and S. Wegenkittl. The PLAB www-server. `http://random.mat.sbg.ac.at`. Also accessible via ftp.

[3] P. Billingsley. *Probability and Measure*. Wiley and Sons, New York, second edition, 1986.

[4] P. Bratley, B. Fox, and L.E. Schrage. *A Guide to Simulation*. Springer-Verlag, New York Berlin Heidelberg Tokyo, 1983.

[5] G.J. Chaitin. Randomness and mathematical proof. *Sci. Amer.*, **232**:47–52, 1975.

[6] K. L. Chung. *A Course in Probability Theory*. Harcourt, Braa and World Inc., New York, 1968.

[7] B. de Finetti. *Studies in Subjective Probability*. Wiley, New York, 1964.

[8] J. Eichenauer and J. Lehn. A non-linear congruential pseudo random number generator. *Statist. Papers*, **27**:315–326, 1986.

[9] J. Eichenauer-Herrmann. Inversive congruential pseudorandom numbers: a tutorial. *Int. Statist. Rev.*, **60**:167–176, 1992.

[10] J. Eichenauer-Herrmann. Statistical independence of a new class of inversive congruential pseudorandom numbers. *Math. Comp.*, **60**:375–384, 1993.

[11] J. Eichenauer-Herrmann and F. Emmerich. Compound inversive congruential numbers: an average-case analysis. *preprint*.

[12] J. Eichenauer-Herrmann and F. Emmerich. A review of compound methods for pseudorandom number generation. In P. Hellekalek, G. Larcher, and P. Zinterhof, editors, *Proceedings of the 1st Salzburg Minisymposium on Pseudorandom Number Generation and Quasi-Monte Carlo Methods, Salzburg, Nov 18, 1994*, Technical Report Series. ACPC – Austrian Center for Parallel Computation, Universität Wien, Austria, 1995.

[13] T. L. Fine. *Theories of Probability*. Academic Press, New York, 1973.

[14] G.S. Fishman and L.R. Moore. An exhaustive analysis of multiplicative congruential random number generators with modulus $2^{31} - 1$. *SIAM J. Sci. Statist. Comput.*, **7**:24–45, 1986. see also the Erratum, ibib. **7**(1986), p. 1058.

[15] I. J. Good. The serial test for sampling numbers and other tests for randomness. *Proc. Cambridge Philosophical Society*, **49**:276–284, 1953.

[16] P. M. Gruber and C. G. Lekkerkerker. *Geometry of Numbers*. Elsevier Science Publishers B. V., Amsterdam, second edition, 1987.

[17] J. Hartung, B. Elpelt, and K. H. Klösener. *Statistik*. R. Oldenburg, Munich, 9th edition, 1993.

[18] P. Hellekalek. General discrepancy estimates IV: the dyadic diaphony. *Preprint, Institute of Mathematics, University of Salzburg, Austria*, 1994.

[19] P. Hellekalek. Correlations between pseudorandom numbers: theory and numerical practice. In P. Hellekalek, G. Larcher, and P. Zinterhof, editors, *Proceedings of the 1st Salzburg Minisymposium on Pseudorandom Number Generation and Quasi-Monte Carlo Methods, Salzburg, Nov 18, 1994*, Technical Report Series, pages 43–73. ACPC – Austrian Center for Parallel Computation, Universität Wien, Austria, 1995.

[20] P. Hellekalek. General discrepancy estimates V: diaphony and the spectral test. *Preprint, Institute of Mathematics, University of Salzburg, Austria*, 1995.

[21] P. Hellekalek and K. Entacher. Revised implementation and testing of the algorithms for IMP-polynomials. Report D5H-3, CEI-PACT Project, WP5.1.2.1.2, Research Institute for Software Technology, University of Salzburg, Austria, 1995.

[22] P. Hellekalek and K. Entacher. Tables of IMP-polynomials. Report D5H-4, CEI-PACT Project, WP5.1.2.1.2, Research Institute for Software Technology, University of Salzburg, Austria, 1995.

[23] D. C. Hoaglin and D. F. Andrews. The reporting of computation-based results in statistics. *The American Statistician*, **29**(3):122–126, 1975.

[24] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1988, reprinted 1990.

[25] A. B. Israel and T. Greville. *Generalized Inverses: Theory and Applications*. Wiley Interscience Publications. Wiley and Sons, New York, 1974.

[26] D. E. Knuth. *The Art of Computer Programming*, volume 2: Seminumerical Algorithms. Addison-Wesley, Reading, MA, second edition, 1981.

[27] S. Kotz et al., editor. *Encyclopedia of statistical sciences*, volume 1, pages 439–457. Wiley and Sons, wiley interscience publication edition, 1982.

[28] P. L'Ecuyer. Uniform random number generation. *Annals of Operations Research*, **53** :77–120, 1994.

[29] H. Leeb. PLAB – a system for testing random numbers. In M. Vajteršic and P. Zinterhof, editors, *Proceedings of the International Workshop on Parallel Numerics '94, Smolenice, Sept. 19–21*, pages 89–99, Slovakia, 1994. Slovak Academy of Sciences, Institute for Informatics. Available on the internet at `http://random.mat.sbg.ac.at`.

[30] H. Leeb. Random numbers for computer simulation. Master's thesis, University of Salzburg, 1995. Available on the internet at `http://random.mat.sbg.ac.at`.

[31] G. Marsaglia. A current view of random number generators. In L. Billard, editor, *Computer Science and Statistics: The Interface*, pages 3–10. Elsevier Science Publishers B.V., 1985.

[32] L. Ming and P. Vitàny. *An Introduction To Kolmogorov Complexity And Its Applications*. Texts and Monographs in Computer Science. Springer Verlag, New York, 1993.

[33] M. Z. Nashed, editor. *Generalized Inverses and Applications*. Academic Press, New York, 1976.

[34] H. Niederreiter. Pseudo-random numbers and optimal coefficients. *Adv. in Math.*, **26**:99–181, 1977.

[35] H. Niederreiter. Quasi-Monte Carlo methods and pseudo-random numbers. *Bull. Amer. Math. Soc.*, **84**:957–1041, 1978.

[36] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia, USA, 1992.

[37] H. Niederreiter. Pseudorandom vector generation by the inversive method. *ACM Trans. Modeling and Computer Simulation*, 4:191–212, 1994.

[38] H. Niederreiter. New developments in uniform pseudorandom number and vector generation. In H. Niederreiter and P.J.-S. Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, volume 106 of *Lecture Notes in Statistics*. Springer-Verlag, Heidelberg New York, 1995.

[39] S. K. Park and K. W. Miller. Random number generators: good ones are hard to find. *Comm. ACM*, **31** :1192–1201, 1988.

[40] W. H. et al Press. *Numerical Recipes in C*. The Art of Scientific Computing. Press Syndicate of the University of Cambridge, Cambridge, 1992.

[41] C. R. Rao and S. K. Mitra. *Generalized Inverse of Matrices and its Applications*. Wiley Series in Probability and Mathematical Statistics. Wiley and Sons, 1971.

[42] B. D. Ripley. The lattice structure of pseudo-random number generators. *Proc. R. Soc. London Ser. A*, **389**:197–204, 1983.

[43] B.D. Ripley. *Stochastic Simulation*. John Wiley, New York, 1987.

[44] I. Steward. *Spielt Gott Roulette?* Birkhäuser, Basel, 1990.

[45] S. Zubrzycki. *Lectures in Probability Theory.* American Elsevier Pub. Company, New York, 1972.

Curriculum vitae

Name: Stefan Wegenkittl

Date of birth: July 8, 1969

Place of birth: Salzburg, Austria

Parents: Renate and Willibald Wegenkittl

Education:

1975-1979: Volksschule in Salzburg
1979-1987: Gymnasium in Salzburg
1988-1995: University studies (M.Sc. in Mathematics)
at the University of Salzburg