W. Freiberger
U. Grenander

Applied
Mathematical
Sciences
6

A Short Course in Computational Probability and Statistics

W. Freiberger
U. Grenander

A Course in Computational Probability and Statistics

With 35 Illustrations

Springer-Verlag New York · Heidelberg · Berlin
1971

Walter Freiberger
Ulf Grenander
Division of Applied Mathematics
Brown University
Providence, Rhode Island

Library of Congress Catalog Card Number 76-176272.
Printed in the United States of America.

ISBN 0-387-90029-2 Springer-Verlag New York • Heidelberg • Berlin
ISBN 3-540-90029-2 Springer-Verlag Berlin • Heidelberg • New York

PREFACE

This book arose out of a number of different contexts, and numerous persons have contributed to its conception and development.

It had its origin in a project initiated jointly with the IBM Cambridge Scientific Center, particularly with Dr. Rhett Tsao, then of that Center. We are grateful to Mr. Norman Rasmussen, Manager of the IBM Scientific Center Complex, for his initial support.

The work is being carried on at Brown University with generous support from the Office of Computing Activities of the National Science Foundation (grants GJ-174 and GJ-710); we are grateful to Dr. John Lehmann of this Office for his interest and encouragement. Professors Donald McClure and Richard Vitale of the Division of Applied Mathematics at Brown University contributed greatly to the project and taught courses in its spirit. We are indebted to them and to Dr. Tore Dalenius of the University of Stockholm for helpful criticisms of the manuscript.

The final stimulus to the book's completion came from an invitation to teach a course at the IBM European Systems Research Institute at Geneva. We are grateful to Dr. J.F. Blackburn, Director of the Institute, for his invitation, and to him and his wife Beverley for their hospitality.

We are greatly indebted to Mrs. Katrina Avery for her splendid secretarial and editorial work on the manuscript.

INTRODUCTION

The purpose of this book is to present an attitude. It has been designed with the aim of making students and perhaps also faculty aware of some of the consequences of modern computing technology for probability theory and mathematical statistics. Not only the increased speed and memory of modern computers are relevant here; of at least equal importance to our subject are the versatile input-output devices and the existence of interactive time-sharing systems and of powerful programming languages. Of the last-mentioned, we have found APL most useful for our purposes.

The work described in these notes was initiated because we felt the time was ripe for a systematic exploitation of modern computing techniques in mathematical statistics and applied probability. Model building, for instance in applied probability, is very different today from what it was in pre-computer days, although this change has not yet fully penetrated to the textbook level. This course is being presented to remedy this situation to some degree; through it, we hope, students will become aware of how computers have increased their freedom of choice of mathematical models and liberated them from the restraints imposed by traditional mathematical techniques.

The project which led to this set of lecture notes is a continuation, although in different form, of an activity organized several years ago at the University of Stockholm. The activity was intended as a complement to the already existing program for training graduate students in mathematical statistics and operations research. It was felt that the students received a well-balanced education in mathematical theory but that something else was lacking: they were not trained for solving real-life problems in raw and unpolished form (in which they usually appear), far removed from that of pure and idealized textbook problems. In particular, little attention was given to the first, difficult stages of problem solving, namely the building of models, the collection of data, and the crucial relation between the mathematical, formal apparatus and the real-life phenomena under study.

To give students an opportunity for acquiring a more realistic orientation and critical attitude towards data, they were exposed to real problems chosen from many fields in industry, government and research. With the help of advice from a teacher

or from older and more experienced students, they were asked to study a problem, formulate their own model, scrutinize data and present an analysis to the person or agency from whom the problem had originated. The results were later discussed in laboratory sessions, often before the analysis was completed, in order to get suggestions and ideas.

It became obvious in the course of this experiment that a major defect in the conventional training of graduate students was the lack of attention paid to the role of computers in the applications of mathematical statistics. Students have often to pass through a more or less painful stage in which they reorient themselves in order to learn what is computationally feasible as distinct from analytically possible. It is desirable that this transition be made easier, quicker and more complete. Although most students will have had some exposure to the computer, they may be inexperienced in its use in, say, applied probability. This will affect their ability to formulate realistic models, forcing them to choose analytically tractable models rather than those which best fit the problem.

The purpose of the present project is to equip students in probability and statistics better for work on real problems. The emphasis in the latter effort is on model building as influenced by computer science.

A growing number of practicing statisticians are today aware of the need to exploit more fully the capability of computers in statistics. This is particularly true in applied statistics, for instance in data analysis, where some research workers have emphasized this for several years. Less attention has been paid to the computational side of probability theory, although the need for a computational reorientation also exists in this subject. We therefore chose to concentrate our efforts on probability theory and applied probability as well as on statistics.

We divided our work into several chapters. Each chapter represents some concept or technique or relation for which a sufficiently rich mathematical structure has been developed and in which, at the same time, the impact of computer science can be expected to be substantial. The chapters together will cover only a part of mathematical statistics, although, we hope, an important one. We are particularly interested in areas in which the interaction between the analytical and computational approach

is strong. This will usually only be the case where the analytical technique has been extended so far that further extension seems possible or worthwhile only through computer use, and makes it necessary that students possess a certain degree of mathematical sophistication. A course designed in such a way should be distinguished from one aiming only at teaching the computer implementation of standard statistical techniques and the writing of statistical programs. A course of the latter type is certainly useful and necessary, but the present project is more ambitious in scope and perhaps also more difficult to accomplish in the classroom. Little, in fact, seems to have been done in the direction described here. We had originally hoped to be able to use some already developed material, but there is disappointingly little available.

The prerequisites for the course are familiarity with the elements of probability and statistics, calculus and linear algebra. It will be assumed that students have some programming experience; most computing in the course will be based on APL which is, from the point of view of our computing tasks, by far the most suitable currently available interactive programming language.

The APL programs in the book should not be interpreted as forming a comprehensive, debugged program library (see section 7.1 in this context). They are presented only to illustrate our approach to computational probability and statistics and undoubtedly contain several mistakes.

Since the degree of mathematical sophistication is expected to vary a good deal among students, more advanced sections are starred, and it is suggested that they be read only by those who feel they have sufficient mathematical background. These sections should be discussed in class, without detailed proofs being given; instead, their interpretation and practical consequences should be discussed by the lecturer.

We strongly recommend that students be encouraged to complete the assignments to help them in the development of a real understanding of the material. The extent and size of the assignments will depend in part upon the computational facilities available during the course.

For further reading, and to find more advanced developments of some of the subjects covered here, we recommend the series of reports published under the NSF-

sponsored "Computational Probability and Statistics" project at Brown University, the titles of which are listed in the References.

The curves in the book were produced by a TSP plotting system on-line with a DATEL terminal, operating under APL/360.

TABLE OF CONTENTS

CHAPTER 1:

RANDOMNESS

1.1 Fundamentals.

The concept of randomness is fundamental in probability theory and statistics, but also most controversial. Among the many interpretations of terms like probability, likelihood, etc., we shall consider two in this course: the usual frequency approach (in this chapter) and the Bayesian one (in chapter 6: "Decision problems").

One should actually not speak of a single frequency approach, since there are several variations of it. That most commonly adopted in the textbook literature is to start from the idea of a random experiment and carry out the mathematical formalization as follows.

Starting from a sample space X that may be completely unstructured, one views the outcome of the random experiment E as a realization of a stochastic variable x described by a probability measure P given on X. The pure mathematician is wont to phrase this as follows: the value of the probability P(S) should be defined for any subset $S \subset X$ belonging to a well-defined σ-algebra of subsets of the sample space. (We shall not, however, go into the measure-theoretical aspects in this course.) This is the mathematical model: the transition to phenomena of the real world is effected through a heuristic principle: the frequency interpretation of probabilities. If the experiment E is repeated n times independently and under equal conditions, then the relative frequency f/n should be close to P(S), where f is the absolute frequency (number of times we get a value $x \in S$), if the sample size, n, is large enough.

While the idea behind this ancient principle is quite appealing, the above formulation is not quite clear on three points:

a) what is meant by "independently"?

b) how should one interpret the phrase "under equal conditions"?

c) how large is "large enough"?

It has often been argued that this sort of vagueness must always be expected when any mathematical model is interpreted in terms of the physical world. For example, when we use Euclidean plane geometry to describe and analyze measurements

of length, angles and areas, we meet the same sort of difficulty when trying to relate notions like points, lines and areas to physical data. We continue to use the model only as long as no logical inconsistency is found or no serious discrepancy between model and data has been established empirically. This pragmatic attitude has been widely accepted, but doubts have been voiced by critics who claim that a more profound analysis is possible. To understand better how this can be done we shall take a look at the manner in which we actually use the model.

Simplifying drastically, we could say that from the above point of view probability and mathematical statistics are the study of bounded measures. While such a statement would undoubtedly describe much research activity in these fields quite accurately, it is a superficial point of view and of little help when we want to discuss the relation between theory and its application.

Randomness enters on three different levels. The first can be exemplified by a set of measurements of some physical constant like the speed of light. Here we would think of the variation encountered among the data as caused by imperfections in the experimental arrangement, imperfections which could, at least in principle, be eliminated or reduced by building better equipment or using a more efficient design for the experiment. We describe this variation in probabilistic terms, but probability plays here only a marginal role. On the second level randomness plays a more fundamental role. Assume that we measure the toxicity of a drug and use guinea pigs in the experiment. We would have to consider the apparent randomness caused by the biological variation in the population of animals used. We would always expect such variation, although its form and extent might vary between different populations. In a well-designed experiment we would like to be able to make precise statements about this variation; to eliminate it entirely does not seem possible. Most of experimental statistics falls into this category. To illustrate the third level, let us think of a Monte Carlo experiment in which we try to find the properties of a statistical technique by applying it to artificial data and studying the result, or in which a probabilistic limit theorem is examined by simulating it for large samples. Here we need randomness, and try to generate data in some way so that they have properties we would expect from random sequences. Another example bringing out this feature perhaps even more clearly occurs in the design of experiments when we inject randomness

intentionally into the experimental arrangement. Here randomness is desirable and necessary.

In recent years simulation of probabilistic phenomena has become a useful and often-applied tool, especially in the area of operations research. Almost all of this work is being done on computers and it is obvious that this fact has greatly influenced our interest in computer generation of random sequences. There is a close relation between certain results in analytical probability theory, some of which have been known for many years, and the more algorithmically-oriented computer studies carried out recently. This relation will be examined in the following sections.

1.2 Random Number Generation.

Let us now turn to the question of how one could generate randomness on a computer.

We shall try to construct an algorithm that generates numbers $x_1, x_2, x_3, \ldots$ in the interval (0,1) such that the sequence $\{x_i\}$ behaves as one would expect from a sample from the rectangular (uniform) distribution over the same interval (0,1).

What properties should we ask for? Let us mention a few. We know from the law of large numbers that the average

$$\bar{x} = \frac{1}{n} \sum_{1}^{n} x_i \tag{1.2.1}$$

should tend in probability to 1/2 and we would require $\bar{x}$ to be close to 1/2 in some sense that has to be left somewhat vague at present. Similarly, the empirical variance

$$s_n^2 = \frac{1}{n} \sum_{1}^{n} (x_i - \bar{x})^2 \tag{1.2.2}$$

should tend to 1/12 in probability. We can think of other similar properties; let us mention some others. We certainly want the sample $(x_1, x_2, \ldots, x_n)$ to have an approximately uniform distribution over (0,1), e.g. in the following sense. We divide the interval into the m intervals $I_\nu = (\frac{\nu-1}{m}, \frac{\nu}{m})$, $\nu = 1,2,\ldots,m$, where $m << n$, and call N_ν the number of x_i that fall in I_ν; then we would expect N_ν/n to be close to $1/m$ for $\nu = 1,2,\ldots,m$.

If we want the algorithm to have a simple form and still generate long (or even

infinite) sequences, it is almost necessary that it be recursive. The simplest would be to set

$$x_{i+1} = f(x_i) , \quad i = 1,2,\ldots \tag{1.2.3}$$

where f is a predetermined function taking values in (0,1) (one could also let f depend on more than one of the x's preceding x_{i+1}). It must also be simple to compute, since (1.2.3) will be repeated many times. Starting from a value $x_i \varepsilon (0,1)$ and applying (1.2.3) recursively, we get a sequence of unlimited length.

To start with, let us choose f as

$$f(x) = \frac{1}{2} + 4(x - \frac{1}{2})^3 ; \tag{1.2.4}$$

it will soon become obvious why we have chosen a function of this form (which is graphed in figure 1).

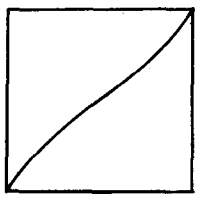

Figure 1.

Assignment: Write a program in APL to generate a series of length n. Calculate $\bar{x}$, s_n^2 and the values of N_v/n for some suitable value of n. Work with small values of n, say 20.

Let us run the program and note what happens.

To avoid the clearly non-random clustering observed in the above example, let us use

$$f(x) = \{x + a\} , \tag{1.2.5}$$

starting with $x_1 = 0$, where a is some fixed number in (0,1). Here {x} stands for the fractional part of x or, in APL, x-⌊x.

Let us now write the same program as before and execute it at the terminal. What behavior do we observe? What about special values of a?

Let us now do the same with

(1.2.6) $$f(x) = \{2x\}$$

Let us also try larger values of n, say 100, and see what happens then. We may also be interested in replacing (1.2.6) by f(x) = {Mx}, M = any natural number, or by more "chaotic" functions, such as

(1.2.7) $$f(x) = \begin{cases} 2x & 0 \leq x \leq \frac{1}{3} \\ 2 - 3x & \frac{1}{3} \leq x \leq \frac{2}{3} \\ 3x - 2 & \frac{2}{3} \leq x \leq 1 \end{cases}$$

There may be other suitable choices of f with which to run your program. What can we say about the behavior of the sequence generated?

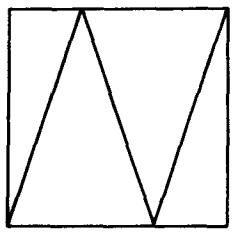

Figure 2

The reason why (1.2.4) failed is obvious. Indeed, by looking at the graph of the function plotted in figure 2, it is clear that if $1/2 < x_i < 1$ then $1/2 < x_{i+1} < x_i$, and that the sequence converges to 1/2; the same is true if $0 < x_i < 1/2$. The three values x = 0, 1/2, 1 play the role of fixed points: the middle one is stable and the other two are unstable. We would, of course, like to avoid fixed points. The function is not sufficiently irregular to generate the chaotic behavior we want. Actually, for any f that is continuous and maps [0,1] onto itself there is at least one fixed point. This leads us to introduce discontinuous functions.

Consider instead the function (1.2.5); it has a discontinuity at x = 1-a (plot the function!). The following beautiful result due to Herman Weyl illustrates the situation well.

Theorem. For any irrational $a \in (0,1)$ the sequence $x_1, x_2, x_3, \ldots$ with $x_i = \{ia\}$ is equidistributed over $(0,1)$ (this relates to additive congruence generators; see below).

Note: by equidistribution in this context - for sequences of numbers and not for stochastic variables - is meant the following. For any a,b with $0 \leq a < b \leq 1$ we should have

$$\lim_{n\to\infty} [\frac{1}{n} \times \text{number } \{x_i : a \leq x_i \leq b;\ i=1,2,\ldots,n\}] = b - a \tag{1.2.8}$$

*Proof of theorem (asterisks denote the beginning and end of an advanced section). Consider the trigonometric sums

$$T_k = \frac{1}{n} \sum_{j=1}^{n} e^{2\pi ik\{ja\}} = \frac{1}{n} \sum_{j=1}^{n} e^{2\pi ikja} \tag{1.2.9}$$

We have $T_0 = 1$ and for $k \neq 0$ by summing the geometric series

$$|T_k| \leq \frac{1}{n} \frac{1}{\sin \pi ka} \tag{1.2.10}$$

Since a is irrational, the value ka is not an integer, so that $\sin \pi ka \neq 0$. We get from (1.2.10)

$$\lim_{n\to\infty} T_k = 0 \tag{1.2.11}$$

This implies that for any trigonometric polynomial $P(x) = \sum_k a_k e^{2\pi ikx}$ we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} P(x_j) = \int_0^1 P(x)\, dx \tag{1.2.12}$$

Since any continuous function on $(0,1)$ can be uniformly approximated by trigonometric polynomials, (1.2.12) holds when P is an arbitrary continuous function. Introduce, for $\varepsilon > 0$, the two stepwise linear functions f^+ and f^- (see figure 3). They are con-

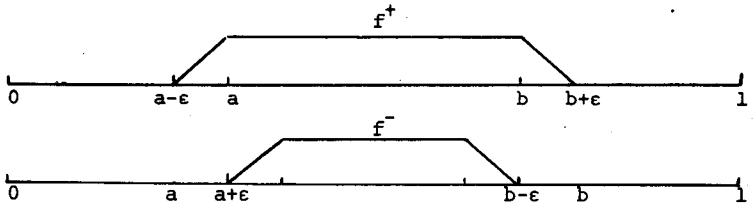

Figure 3

tinuous so that (1.2.12) holds for them. But the quantity in (1.2.8) is included between the two limits we just got; this proves the theorem.*

Note: this theorem does not tell us for which values of x_1 we get equidistribution, only that equidistribution is the typical case.

There is also the following result:

Theorem. For almost all initial values $x_1 \in (0,1)$, the sequence generated by (1.2.6) is equidistributed (this relates to multiplicative congruence generators; see below).

*Proof. The mapping Tx, where

$$T: x \to \{2x\} \tag{1.2.13}$$

of (0,1) onto itself preserves Lebesgue measure. Indeed, let f(x) be an arbitrary periodic (with period 1) bounded and measurable function. Then

$$\int_0^1 f(Tx)dx = \int_0^1 f(2x)dx = \frac{1}{2}\int_0^1 f(y)dy = \int_0^1 f(y)dy \tag{1.2.14}$$

Taking f as the indicator function of an arbitrary subinterval of (0,1) proves that T is measure-preserving in the sense that $m(T^{-1}S) = m(S)$ for any Borel set $S \in [0,1]$. This implies that if x_1 is given a uniform distribution over (0,1) the sequence x_i (now consisting of stochastic variables) forms a stationary stochastic process and we can apply Birkhoff's individual ergodic theorem (see ref. 13, p. 105) and find that for any $g \in L_1(0,1)$

$$\hat{g} = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} g(x_n) \tag{1.2.15}$$

exists for almost all choices of x_1. It remains only to prove that the limit $\hat{g}$ is essentially constant; it must then be equal to

$$\bar{g} = \int_0^1 g(x)dx \qquad \text{(almost certainly)} \tag{1.2.16}$$

which would prove the equidistribution. To prove that g is constant it is enough to prove metric transitivity (ergodicity) of T (see ref. 13, p. 105). Let I be an invariant set with the indicator function I(x):

$$I(Tx) = I(\{2x\}) = I(x), \quad \text{all } x \in (0,1) \tag{1.2.17}$$