A.V. Skorokhod

Basic Principles and Applications of Probability Theory

Springer

Basic Principles and Applications of Probability Theory

A. V. Skorokhod

Basic Principles and Applications of Probability Theory

Edited by Yu. V. Prokhorov

Translated by B. D. Seckler

Springer

A.V. Skorokhod
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824, USA

Yu.V. Prokhorov *(Editor)*
Russian Academy of Science
Steklov Mathematical Institute
ul. Gubkina 8
117966 Moscow, Russia

B. D. Seckler *(Translator)*
19 Ramsey Road
Great Neck, NY 11023-1611, USA
e-mail: bersec@aol.com

Original Russian edition published by Viniti, Moscow 1989

Title of the Russian edition: Teoriya Veroyatnostej 1

Published in the series: Itogi Nauki i Tekhniki. Sovremennye Problemy Matematiki.
Fundamental'nye Napravleniya, Tom 43

Library of Congress Control Number: 2004110444

Mathematics Subject Classification (2000):
60Axx, 60Dxx, 60Fxx, 60Gxx, 60Jxx, 62Cxx, 94Axx

ISBN 3-540-54686-3 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springeronline.com

Printed in Germany

Typeset by Steingraeber Satztechnik GmbH, Heidelberg
using a Springer TEX macro package

Cover design: Erich Kirchner, Heidelberg
Printed on acid-free paper 46/3142LK - 5 4 3 2 1 0

Contents

I. Probability. Basic Notions. Structure. Methods 1

II. Markov Processes
and Probability Applications in Analysis 143

III. Applied Probability 191

Author Index 275

Subject Index 277

Probability. Basic Notions. Structure. Methods

Contents

1 Introduction **5**

1.1 The Nature of Randomness 5

1.1.1 Determinism and Chaos 6

1.1.2 Unpredictability and Randomness 6

1.1.3 Sources of Randomness. 7

1.1.4 The Role of Chance 8

1.2 Formalization of Randomness 9

1.2.1 Selection from Among Several Possibilities. Experiments. Events 9

1.2.2 Relative Frequencies. Probability as an Ideal Relative Frequency 12

1.2.3 The Definition of Probability 13

1.3 Problems of Probability Theory 14

1.3.1 Probability and Measure Theory 15

1.3.2 Independence 15

1.3.3 Asymptotic Behavior of Stochastic Systems 16

1.3.4 Stochastic Analysis 17

2 Probability Space **19**

2.1 Finite Probability Space 19

2.1.1 Combinatorial Analysis 19

2.1.2 Conditional Probability 21

2.1.3 Bernoulli's Scheme. Limit Theorems 24

2.2 Definition of Probability Space 27

2.2.1 σ-algebras. Probability 27

2.2.2 Random Variables. Expectation 29

2.2.3 Conditional Expectation 31

2.2.4 Regular Conditional Distributions 34

2.2.5 Spaces of Random Variables. Convergence 35

2.3 Random Mappings 38

2.3.1 Random Elements 38

2.3.2 Random Functions ... 42
2.3.3 Random Elements in Linear Spaces ... 44
2.4 Construction of Probability Spaces ... 46
2.4.1 Finite-dimensional Space ... 46
2.4.2 Function Spaces ... 47
2.4.3 Linear Topological Spaces. Weak Distributions ... 50
2.4.4 The Minlos-Sazonov Theorem ... 51

3 Independence ... 53
3.1 Independence of σ-Algebras ... 53
3.1.1 Independent Algebras ... 53
3.1.2 Conditions for the Independence of σ-Algebras ... 55
3.1.3 Infinite Sequences of Independent σ-Algebras ... 56
3.1.4 Independent Random Variables ... 57
3.2 Sequences of Independent Random Variables ... 59
3.2.1 Sums of Independent Random Variables ... 59
3.2.2 Kolmogorov's Inequality ... 61
3.2.3 Convergence of Series of Independent Random Variables 63
3.2.4 The Strong Law of Large Numbers ... 65
3.3 Random Walks ... 67
3.3.1 The Renewal Scheme ... 67
3.3.2 Recurrency ... 71
3.3.3 Ladder Functionals ... 74
3.4 Processes with Independent Increments ... 78
3.4.1 Definition ... 78
3.4.2 Stochastically Continuous Processes ... 80
3.4.3 Lévy's Formula ... 83
3.5 Product Measures ... 86
3.5.1 Definition ... 86
3.5.2 Absolute Continuity and Singularity of Measures ... 87
3.5.3 Kakutani's Theorem ... 88
3.5.4 Absolute Continuity of Gaussian Product Measures ... 91

4 General Theory of Stochastic Processes and Random Functions ... 93
4.1 Regular Modifications ... 93
4.1.1 Separable Random Functions ... 94
4.1.2 Continuous Stochastic Processes ... 96
4.1.3 Processes With at Most Jump Discontinuities ... 97
4.1.4 Markov Processes ... 98
4.2 Measurability ... 100
4.2.1 Existence of a Measurable Modification ... 100
4.2.2 Mean-Square Integration ... 101
4.2.3 Expansion of a Random Function in an Orthogonal Series ... 103

4.3 Adapted Processes . . . 104
4.3.1 Stopping Times . . . 105
4.3.2 Progressive Measurability . . . 106
4.3.3 Completely Measurable and Predictable σ-Algebras . . . 107
4.3.4 Completely Measurable and Predictable Processes . . . 108
4.4 Martingales . . . 110
4.4.1 Definition and Simplest Properties . . . 110
4.4.2 Inequalities. Existence of the Limit . . . 111
4.4.3 Continuous Parameter . . . 114
4.5 Stochastic Integrals and Integral Representations of Random Functions . . . 115
4.5.1 Random Measures . . . 115
4.5.2 Karhunen's Theorem . . . 116
4.5.3 Spectral Representation of Some Random Functions . . 117

5 Limit Theorems . . . 119
5.1 Weak Convergence of Distributions . . . 119
5.1.1 Weak Convergence of Measures in Metric Spaces . . . 119
5.1.2 Weak Compactness . . . 122
5.1.3 Weak Convergence of Measures in R^d . . . 123
5.2 Ergodic Theorems . . . 124
5.2.1 Measure-Preserving Transformations . . . 124
5.2.2 Birkhoff's Theorem . . . 126
5.2.3 Metric Transitivity . . . 130
5.3 Central Limit Theorem and Invariance Principle . . . 132
5.3.1 Identically Distributed Terms . . . 132
5.3.2 Lindeberg's Theorem . . . 133
5.3.3 Donsker-Prokhorov Theorem . . . 135

Historic and Bibliographic Comments . . . 139

References . . . 141

1

Introduction

Probability theory arose originally in connection with games of chance and then for a long time it was used primarily to investigate the credibility of testimony of witnesses in the "ethical" sciences. Nevertheless, probability has become a very powerful mathematical tool in understanding those aspects of the world that cannot be described by deterministic laws. Probability has succeeded in finding strict determinate relationships where chance seemed to reign and so terming them "laws of chance" combining such contrasting notions in the nomenclature appears to be quite justified. This introductory chapter discusses such notions as determinism, chaos and randomness, predictibility and unpredictibility, some initial approaches to formalizing randomness and it surveys certain problems that can be solved by probability theory. This will perhaps give one an idea to what extent the theory can answer questions arising in specific random occurrences and the character of the answers provided by the theory.

1.1 The Nature of Randomness

The phrase "by chance" has no single meaning in ordinary language. For instance, it may mean unpremeditated, nonobligatory, unexpected, and so on. Its opposite sense is simpler: "not by chance" signifies obliged to or bound to (happen). In philosophy, necessity counteracts randomness. Necessity signifies conforming to law – it can be expressed by an exact law. The basic laws of mechanics, physics and astronomy can be formulated in terms of precise quantitative relations which must hold with ironclad necessity. True, this state of affairs existed in the classical period when science did not delve into the microworld. But even before, chance had been encountered in everyday life at practicaily every step. Birth and death and even the entire life of a person is a chain of chance occurrences that cannot be computed or foreseen with the aid of determinate laws. What then can be studied and how studied and what sort of answers may be obtained in a world of chance? Science can merely treat the intrinsic in occurrences and so it is important to extract the essential features of a chance occurrence that we shall take into account in what follows.

1.1.1 Determinism and Chaos

In a deterministic world, randomness must be absent – it is absolutely subject to laws that specify its state uniquely at each moment of time. This idea of the world (setting aside philosophical and theological considerations) existed among mathematicians and physicists in the 18th and 19th centuries (Newton, Laplace, etc.). However, such a world was all the same unpredictable because of its complex arrangement. In order to determine a future state, it is necessary to know its present state absolutely precisely and that is impossible. It is more promising to apply determinism to individual phenomena or aggregates of them. There is a determinate relationship between occurrences if one entails the other necessarily. The heating of water to 100°C under standard atmospheric pressure, let us say, implies that the water will boil. Thus, in a determinate situation, there is complete order in a system of phenomena or the objects to which these phenomena pertain. People have observed that kind of order in the motion of the planets (and also the Moon and Sun) and this order has made it possible to predict celestial occurrences like lunar and solar eclipses. Such order can be observed in the disposition of molecules in a crystal (it is easy to give other examples of complete order). The most precise idea of complete order is expressed by a collection of absolutely indistinguishable objects.

In contrast to a deterministic world would be a chaotic world in which no relationships are present. The ancient Greeks had some notion of such a chaotic world. According to their conception, the existing world arose out of a primary chaos. Again, if we confine ourselves just to some group of objects, then we may regard this system to be completely chaotic if the things are entirely distinct. We are excluding the possibility of comparing the objects and ascertaining relationships among them (including even causal relationships). Both of these cases are similar: the selection of one (or several objects) from the collection yields no information. In the first case, we know right away that all of the objects are identical and in the second, the heterogeneity of the objects makes it impossible to draw any conclusions about the remaining ones. Observe that this is not the only way in which these two contrasting situations resemble one another. As might be expected, according to Hegel's laws of logic, these totally contrasting situations describe the exact same situation. If the objects in a chaotic system are impossible to compare, then one cannot distinguish between them so that instead of complete disorder, we have complete order.

1.1.2 Unpredictability and Randomness

A large number of phenomena exist that are neither completely determinate nor completely chaotic. To describe them, one may use a system of nonidentical but mutually comparable objects and then classify them into several groups. Of interest to us might be to what group a given object belongs. We shall illustrate how the existence of differences relates to the absence of complete determinism. Suppose that we are interested in the sex of newborn children. It is known that roughly half of births are boys and half are girls. In other words, the "things" being considered split into two groups. If a strictly valid law existed for the birth of a boy or girl, then it would still be impossible to produce the mechanism which would continually equalize the sexes of babies being born in the requisite proportion (without assuming the effect of the results of prior births on succeeding births, such a premise is meaningless). One may give numerous examples of valid statements like "such a thing happens in such and such fraction of the cases", for instance, "1% of males are color-blind." As in the case of the sex of babies, the phenomenon cannot be explained on the basis of determinate laws. It is advantageous to view a set-up of things as a sequence of events proceeding in time.

The absence of determinism means that future events are unpredictable. Since events can be classified in some sort of way, one may ask to what class will a future event belong? But once again (determinism not being present), one cannot furnish an answer in advance. The question is ill posed in the given situation. The examples cited suggest a proper way to state the question: how often will a phenomenon of a given class occur in the sequence? We shall speak about chance in precisely such situations and it will be natural to raise such questions and to find answers for them.

1.1.3 Sources of Randomness.

We shall now point out a few of the most important existing physical sources of randomness in the real world. In so doing, we view the world to be sufficiently organized (unchaotic) and randomness will be understood as in Sect. 1.1.2.

(a) *Quantum-mechanical laws.* The laws of quantum mechanics are statements about the wave functions of micro-objects. According to these laws, we can specify, for instance, just the wave function of an electron in a field of force. Based on the wave function, only the probability of detecting the electron in some particular region of space may be found – to predict its position is impossible. In exactly the same way, one cannot ascertain the energy of an electron and it is only possible to determine a discrete number of possible energy levels and the probability that the energy of the electron has a specified value. We perceive that the fundamental laws of the microworld make use of the language of probability and thus phenomena in the microworld are random. An important example of a random phenomenon in the microworld is the emission of a quantum of light by an excited atom. Another important example are nuclear reactions.

(b) *Thermal motion of molecules.* The molecules of any substance are in constant thermal motion. If the substance is a solid, then the molecules range close to positions of equilibrium in a crystal lattice. But in fluids and gases, the molecules perform rather complex movements changing their directions of motion frequently as they interact with one another. The presence of such a motion may be ascertained by watching the movement of microscopic particles suspended in a fluid or gas (this is so-called Brownian motion). This motion is of a random nature and the energies of the individual molecules are also random, that is, the energies of the molecules can assume different values and so one talks about the fraction of molecules having an energy within narrow specified bounds. This is the familiar Maxwell distribution in physics. A simple experiment will convince one that the energies of the molecules are different. Take the phenomenon of boiling water: if all of the molecules had the same energy, then the water would become steam all at once, that is, with an explosion, and this does not happen.

(c) *Discreteness of matter.* The discreteness of matter leads to the occurrence of randomness in another way. Items (a) and (b) also considered material particles. The following fact should now be noted: the laws of classical physics have been formulated for macrobodies just as if matter filled up space continuously. The discreteness of matter leads to the occurrence of deviations of the actual values of physical quantities from those predicted by the laws. These deviations or “fluctuations” are of a random nature and they affect the course of a process substantially. Thus, the discreteness of the carriers of electricity in metallic conductors – the electrons – is the source of fluctuation currents which are the reason for internal noise in radios. The discreteness of matter results in the mutual permeation of substances. Furthermore the absence of pure substances, that is, the existence of impurities, also results in random deviations from the calculated flow of phenomena.

(d) *Cosmic radiation.* Experimentation shows that it is irregular (aperiodic and unpredictable) but it conforms to laws that can be studied by probability theory.

1.1.4 The Role of Chance

It is hard to overestimate the role played in our lives by those phenomena that are of a chance nature. The nuclear reactions occurring in the depths of the Sun are the source of the energy sustaining all life on Earth. We are surrounded by the medium of light and the electromagnetic field which are composed of the quanta emitted by the individual atoms of the Sun’s corona. Fluctuations in this emission – the solar flares – affect meteorological processes in a substantial way. Random mechanisms also lead to explosions of supernova stars and to sources of cosmic radiation. Brownian motion results in diffusion and in the mutual permeation of substances and due to it, there are reactions possible and hence even life. Chance mechanisms are responsible for the transmission of hereditary characteristics from parents to children. Cosmic radiation, which is also of a random nature, is one of the sources of mutation of genes due to which we have biological evolution. Many phenomena conform strictly to laws only due to chance and this proves to be the case whenever a phenomenon is dependent upon a large number of independent random microphenomena (for instance, in gases, where there are a huge number of molecules moving randomly and one has the exact Clapeyron law).

1.2 Formalization of Randomness

In order to make chance a subject of mathematical research, it is necessary to construct a formal system which can be interpreted by real phenomena in which chance is observed. This section is devoted to a first discussion.

1.2.1 Selection from Among Several Possibilities. Random Experiments. Events

A most simple scheme in which unpredictable phenomena occur is in the selection of one element from a finite collection. To describe this situation, probability theory makes use of urn models. Let there be an urn containing balls that differ from one another. A ball is drawn from the urn at random. The phrase "at random" means that each ball in the urn can be withdrawn. Later, we shall make at random still more precise. This single selection can be described strictly speaking as being the enumeration of possibilities and furnishes little for discussion. The matter changes substantially when there are a large number of selections. After drawing a ball from the urn and observing what it was, we return it and we again remove one ball from the urn (at random). Observing what the second ball was, we return it to the urn and we repeat the operation again and so on. Let the balls be numbered $1, 2, \ldots, s$ and repeat the selection n times. The results of our operations (termed an experiment in what follows) can be described by the sequence of numbers of the balls drawn: $\alpha_1, \alpha_2, \ldots, \alpha_n$ with $\alpha_n \in \{1, 2, \ldots, s\}$. Questions of interest in probability include this one. How often is the exact same number encountered in such a sequence? At first glance, the question is meaningless: it can still be anything. Nevertheless, although there are certain restrictions, they are based on the following fact. If n_i is the number of times that ball numbered i is drawn, then $n_1 + n_2 + \ldots + n_s = n$. This is of course a trivial remark but, as explained later on, it will serve as a starting point for building a satisfactorily developed mathematical theory. However, there is another nontrivial fact demonstrated by the simplest case $s = 2$. We write out all of the possible results of the n extractions of which there are 2^n. These are all of the possible sequences of digits 1 and 2 of length $n_1 + n_2 = n$, where n_1 is the number of ones in the sequence and n_2 the number of twos. Let N_ε be the amount of those sequences for which $|n_1/n - 1/2| > \varepsilon$. Then $\lim_{n\to\infty} 2^{-n} N_\varepsilon = 0$ for all positive ε. This is an important assertion and it indicates that for large n the fraction of ones in an overwhelming majority of the sequences is close to

1/2. If the same computation is done for s balls, then it can be shown that the fraction of ones is $1/s$ in an overwhelming majority of the sequences. This holds for any $i \leq s$. That the "encounterability" of different numbers in the sequences must be the same can be discerned directly without computation by way of the following symmetry property. If the places of two numbers are interchanged, there are again the same 2^n sequences. Probability theory treats this property as the "*equal likelihood*" of occurrence of each of the numbers in the sequence. Assertions about the relative number of sequences for which n_i/n deviates from $1/s$ by less than ε are examples of the "*law of large numbers*", the class of probability theorems most generally used in applications.

We now consider the notion of "*random experiment*", which is a generalization of the selection scheme discussed above. Suppose that a certain complex of conditions is realized resulting in one of several possible *events*, where generally a different event can occur on iterating the conditions. We then say that we have a random experiment. It is determined by the set of conditions and the set of possible outcomes (observed events). The conditions of the experiment may or may not depend on the will of an experimenter (created artificially) and the presence or absence of an experimenter also plays no role. It is also inessential whether it is possible in principle to observe the outcome of the experiment. Any sufficiently complicated event can generally be placed under the concept of random experiment if one chooses as conditions those that do not determine its course completely. The pattern of its course is then a result of the experiment. The main thing for us in a random experiment is the possibility of repeating it indefinitely. Only for large series of iterated experiments is it possible to obtain meaningful assertions. Examples of physical phenomena have already been given above in which randomness enters. If we consider radioactive decay, for example, then each individual atom of a radioactive element undergoes radioactive conversion in a random fashion. Although we cannot follow each atom, a conceptual experiment can be performed which can help establish which of the atoms have already undergone a nuclear reaction and which still have not. In the same way, by considering a volume of gas, we can conceive an experiment which can ascertain the energies of all of the molecules in the gas. If the possible outcomes of an experiment are known, then we can imagine the experiment as choosing from among several possibilities. Again considering an urn containing balls, we can assume that each ball has one of the possible outcomes of the pertinent experiment written on it and any possibility has been written on one of the balls. On drawing one of the balls, we ascertain which one of the possibilities has been realized. Such a description of an experiment is advantageous because of its uniformness. We point out two difficulties arising in associating an urn model with an experiment. First, it is easy to imagine an experiment which in principle has infinitely many different outcomes. This will always be the case whenever an experiment is measuring a continuously varying quantity (position, energy, etc.). However, in practical situations a continuously varying quantity is measured with a certain accuracy. Second, there is a definite symmetry among the possibilities in the urn model, which was discussed above. It would be unnatural to expect every experiment to have this property. However, the symmetry can be broken by increasing the number of balls and viewing some of them as identical. The indistinguishable balls correspond to one and the same outcome of the experiment but the number of such balls varies from outcome to outcome. Say that an experiment has two outcomes and one ball corresponds to outcome 1 and two balls to outcome 2. Then in a long run of trials, outcome 2 should be encountered twice as often as outcome 1.

In discussing the outcomes of an experiment above, we meant all possible mutually exclusive outcomes. They are usually called "*elementary events*" or "*sample points*". They can be used to construct an "algebra of events" that are observable in an experiment. Events that are observable in an experiment will be denoted by $A, B, C, \ldots$. We now define operations on events. The sum or union of two events A and B is the event that occurs if and only if at least one of A or B occurs and it is denoted by $A \cup B$ or $A + B$. The product or intersection of two events A and B is the event that both A and B occur (simultaneously) and it is denoted by $A \cap B$ or AB. An event is said to be impossible if it can never occur in an experiment (we denote it by $\emptyset$) and to be sure if it always occurs (we denote it by U). The event $\bar{A}$ is the complement of A and corresponds to A not happening. The event $A \cap \bar{B}$ is the difference of A and B and is denoted by $A \setminus B$.

A collection $\mathcal{A}$ of events observable in an experiment is called an *algebra of events* if together with each A it contains $\bar{A}$ and together with each pair A and B it contains $A \cup B$ (the collection $\mathcal{A}$ is nonempty). Since $A \cup \bar{A} = U$, $U \in \mathcal{A}$ and $\emptyset = \bar{U} \in \mathcal{A}$. If A and $B \in \mathcal{A}$, then $A \cap B = \overline{(\bar{A} \cup \bar{B})} \in \mathcal{A}$ and $A \cap \bar{B} \in \mathcal{A}$. Thus the operations on events introduced above do not lead out of the algebra. Let $A_1, A_2, \ldots, A_m$ be a set of events. A smallest algebra of events exists containing these events. We introduce the natural assumption that the events that are observable in an experiment form an algebra. If $A_1, A_2, \ldots, A_m$ are all elementary events of a given experiment, then the algebra of events observable in the experiment comprises events of the form

$$A = \bigcup_{k \in \Lambda} A_k, \quad \Lambda \subset \{1, 2, \ldots, m\}\,, \tag{1.2.1}$$

where Λ is any subset of the segment of integers $\overline{1, m}$; if $\Lambda = \emptyset$, then A is considered to be the impossible event. Let Ω denote the set of elementary events or *sample space*. Every event may be viewed as a subset of Ω. More precisely, one can associate with each event A the set of elementary events A_k occurring in the union on the right of (1.2.1).

As a result there is a one-to-one correspondence between the events in an experiment and the subsets of Ω in which a sum of events corresponds to a union of sets, a product of events to an intersection of sets and the opposite event to the complement of a set in Ω. The relation $A \subset B$ for subsets of Ω has the probabilistic meaning that the event A implies event B because B occurs whenever A occurs. The interpretation of events as subsets of a set enables us to make set theory the basis of our probability-theoretic development and to avoid in what follows such indefinite terminology as "event", "occurs in an experiment" and so on.

1.2.2 Relative Frequencies. Probability as an Ideal Relative Frequency

Consider some experiment and let Ω be the set of elementary events that can occur in the experiment. Let $\mathcal{A}$ be an algebra of observable events in the experiment. $\mathcal{A}$ is a collection of subsets of which together with A contains $\Omega \setminus A$ and together with each pair of sets A and B contains $A \cup B$. The elements of Ω will be denoted by $\omega, \omega_1, \omega'$, etc. Suppose that the experiment is repeated n times. Let ω_k denote the outcome in the k-th experiment; the n-fold repetition of the experiment determines a sequence $(\omega_1, \ldots, \omega_n)$, or in other words, a point of the space Ω^n (the n-th Cartesian power of Ω). An event A occurred in the k-th experiment if $\omega_k \in A$. Let $n(A)$ denote the number of occurrences of A in these n experiments. The quantity

$$\nu_n(A) = \frac{n(A)}{n} \tag{1.2.2}$$

is the *relative frequency* of A (in the stated series of experiments). The relative frequency of A characterizes a connection between A and the conditions of the experiment. Thus, if the conditions of the experiment always imply the occurrence of A, that is, the connection between the conditions of the experiment and A is determinate, then $\nu_n(A) = 1$. If A is impossible under the conditions of the experiment, then $\nu_n(A) = 0$. The closer $\nu_n(A)$ is to 1 or 0, the more "strictly" is the occurrence (nonoccurrence) of A tied to the conditions of the experiment.

We now indicate the basic properties of a relative frequency.

1. $0 \le \nu_n(A) \le 1$ with $\nu_n(\emptyset) = 0$ and $\nu_n(U) = 1$. Two events A and B are said to be disjoint or mutually exclusive if $A \cap B = \emptyset$, that is, they cannot occur simultaneously.
2. If A and B are mutually exclusive events, then $\nu_n(A \cup B) = \nu_n(A) + \nu_n(B)$. Thus the relative frequency is a non-negative additive set-function defined on $\mathcal{A}$ and it is normalized: $\nu_n(\Omega) = \nu_n(U) = 1$.

Relative frequency is a function of the sequence of outcomes of an experiment:

$$\nu_n(A) = n^{-1} \sum_{k=1}^{n} I_A(\omega_k), \tag{1.2.3}$$

where I_A is the indicator function of A. If another sequence of outcomes is considered, the relative frequency can change. In the discussion of the urn model, it was said that for a large number n of observations, the fraction of sequences $(\omega_1, \ldots, \omega_n)$ for which a relative frequency differs little from a certain number approaches 1. Therefore the variability of relative frequency does not preclude some "ideal" value around which it fluctuates and which it approaches in some sense. This ideal value of the relative frequency of an event is then its *probability.* Our discussion has a very vague meaning and it may be viewed as a heuristic argument. Just as actual cats are imperfect "copies" of an ideal cat (the idea of a cat) according to Plato, relative frequencies are likewise realizations of an absolute (ideal) relative frequency – the probability. The sole pithy conclusion that can be drawn from the above heuristic discussion is that probability must preserve the essential properties of relative frequency, that is, it should be a non-negative additive function of events and the probability of the sure event should be 1.

1.2.3 The Definition of Probability

The preceding considerations can be used in different ways to define probability. The initial naive view of the matter was that probabilities of events exist objectively and therefore probability needs no defining. The question was how to calculate a probability.

(a) *The classical definition of probability.* Games of chance and the analysis of testimony of witnesses were originally the basic areas of application of probability theory. Games of chance involving cards, dice and flipping coins naturally permitted the creation of appropriate random experiments (this terminology first appeared in the twentieth century) so that their outcomes had symmetry in relation to the conditions of the experiment. These outcomes were treated as "*equally likely*" and they were assigned the same probabilities. Thus, if there are s outcomes in the experiment, each elementary event was assigned a probability of $1/s$ (it is easy to see that an elementary event has that probability using the additivity of probability and the fact that the sure event has probability one). If an event is expressed as the union of r elementary events ($r \leq s$), then the probability of A is r/s by virtue of the additivity. Thus we arrive at the definition of probability that has been in use for about two centuries.

The probability of an event A is the quotient of the number of outcomes favorable to A and the number of all possible outcomes. The outcomes favorable to A are understood to be those that imply A.

This is the classical definition of probability. With this definition as a starting point, it is possible to establish that probability has the properties indicated in Sect. 1.2.2. The definition is convenient, consistent and allows results obtained by the theory to have a simple interpretation. A deficiency is the impossiblity of extending it to experiments with infinitely many outcomes or to any case in which the outcomes are asymmetric in relation to the conditions of the experiment. In particular, the classical set-up has no events with irrational probabilities.

(b) *The axioms of von Mises.* The German mathematician R. von Mises proposed as the definition of probability the second of the properties mentioned for urn models – the convergence of a relative frequency to some limiting value in the sense indicated there. Von Mises gave a system of probability axioms whose first one postulates the existence of the limit of a relative frequency and this limit is called the probability of an event. Such a system of axioms results in considerable mathematical difficulties. On the one hand, there is the possibility of varying the sequence of experiments and on the other hand, the definition is too empirical and so it hardly accommodates mathematical study. The ideas of von Mises can be used in some interpretations of the results of probability but they are untenable for constructing a mathematical theory.

(c) *The axioms of Kolmogorov.* The set of axioms of A.N. Kolmogorov has been universally recognized as the starting point for the development of probability theory. He proposed them in his book "Fundamental Concepts of Probability Theory." These axioms employ only the most general properties which are inherent to probability about which we spoke above. First of all, Kolmogorov considered the set-theoretic treatment already discussed above and also the notion of random experiment. He postulated the existence of the probability of each event occurring in a random experiment. Probability was assumed to be a nonnegative additive function on the algebra of events with the probability of the sure event equal to 1. Thus a random experiment is formally specified by a triple of things: 1. a sample space Ω of elementary events; 2. an algebra $\mathcal{A}$ of its subsets, the members of $\mathcal{A}$ being the random events; 3. a nonnegative additive function $\mathbf{P}(A)$ defined on $\mathcal{A}$ for which $\mathbf{P}(\Omega) = 1$; $\mathbf{P}(A)$ is termed the probability of A. If random experiments with infinitely many outcomes are considered, then it is natural to require that $\mathcal{A}$ be a *σ-algebra* (or *σ-field*). In other words, together with each sequence of events A_n, $\mathcal{A}$ also contains the countable union $\bigcup_n A_n$ and $\mathbf{P}(A)$ must be a countably-additive function on $\mathcal{A}$: if $A_n \cap A_m = \emptyset$ for $n \neq m$, then $\mathbf{P}(\bigcup_n A_n) = \sum_n \mathbf{P}(A_n)$. This means that $\mathbf{P}$ is a measure on $\mathcal{A}$ and since $\mathbf{P}(\Omega) = 1$, the measure is normalized.

1.3 Problems of Probability Theory

Initially, probability theory was the study of ways of computing probabilities of events knowing the probabilities of other given events. The techniques developed for computing the probabilities of certain classes of events now form a constituent unit of probability but only partly and far from the main part. However, as before, probability theory only deals with the probabilities of events independently of what meaningful sense can be invested in the words "the probability of event A is p". This means that probability theory itself does interpret its results meaningfully but in so doing it does not exclude the term "probability". There is no statement like "A always occurs" but rather the statement "A occurs with probability one".

1.3.1 Probability and Measure Theory

Kolmogorov's axioms make probability theory a special part of measure theory namely finite measure theory (being finite and being normalized are clearly essentially equivalent since any finite measure may be converted into a normalized measure by multiplication by a constant). If this is so, is probability theory unnecessary? The answer to this question has already been given by the development of probability theory following the introduction of Kolmogorov's axioms. Probability theory does employ measure theory in an essential way but classical measure theory really involves the construction of a measure by extension and the development of the integral and its properties including the Radon-Nikodym theorem. Probability theory has inspired new problems in measure theory: the convergence of measures and construction of a measure fibre (”conditional” measure); these now belong traditionally to probability theory. A completely new area of measure theory is the analysis of absolute continuity and singularity of measures. The Radon-Nikodym theorem of measure theory serves merely as a starting point for the development of the very important theory of absolute continuity and singularity of probability measures (also of consequence in applications). Its meaningfulness lies in the broad class of special probability measures that it examines. Finally, the specific classes of measures in probability theory, say, product measures or fibre bundles of measures, establish the nature of its position in relation to general measure theory. This manifests itself in the concepts utilized such as independence, weak dependence and conditional dependence, which are more associated with certain physical ideas at the basis of our probabilistic intuition. These same concepts lead to problems whose reformulations in the language of measure theory prove to be cumbersome, unclear and perplexing making one wonder where these problems arose. (For individuals familiar with probability theory, as an example, it is suggested that one formulate the degeneracy problem for the simplest branching process in terms of measure theory.) Nonetheless, there are a number of sections of probability that can relate immediately to measure theory, for instance, measure theory in infinite-dimensional linear spaces. Having originated in probability problems, they remain traditionally within the framework of probability theory.

1.3.2 Independence

Independence is one of the basic concepts of probability theory. According to Kolmogorov, it is exactly this that distinguishes probability theory from measure theory. Independence will be discussed more precisely later on. For the moment, we merely point out that stochastic independence and physical independence of events (one event having no effect on another) are identical in content. Stochastic independence is a precisely-defined mathematical concept to be given below. At this point, we note that independence was already used in latent form in the definition of random experiment. One of the requirements imposed on an experiment is the possibility of iterating it indefinitely. To iterate it assumes that the conditions of the experiment can be reconstructed after which the one just performed and all of the prior ones have no affect on the outcome of the next experiment. This means that the events occurring in different experiments must be independent.

Probability theory also studies laws of large numbers for independent experiments. One such law has already been stated on an intuitive level. An example is Bernoulli's form of the law of large numbers: "Given a series of independent trials in each of which an event A can occur with probability p and $\nu_n(A)$ the relative frequency of A in the first n trials. Then the probability that $|\nu_n(A) - p| > \varepsilon$ tends to zero as $n \to \infty$ for any positive ε." Observe that the value of $\nu_n(A)$ is random and so the fulfillment of the inequality in this theorem is a random event. The theorem is a precise statement of the fact that the relative frequency of an event approaches its probability. As will be seen below, the proof of this assertion is strictly mathematical. It may seem paradoxical that it is possible to use mathematics to obtain precise knowledge about randomly-occurring events (that it is possible to do so in a determinate world, say, to calculate the dates of lunar eclipses, is quite natural). In fact, the choice of p is supposedly arbitrary and only the fulfillment of Kolmogorov's axioms is required. However, something interesting can be extracted from Bernoulli's theorem only if events of small probability actually rarely occur in practice. It is precisely these kinds of events (or events whose probability is close to 1) that interest us primarily in probability. If one comes to the point of view that events of probability 0 practically never occur and events of probability 1 practically always occur, then the kind of conclusions that may be drawn from random premises will be of interest.

1.3.3 Asymptotic Behavior of Stochastic Systems

Many physical, engineering and biological objects may be viewed as randomly evolving systems. Such a system is in one of its possible states (frequently viewable as finitely many) and with the passage of time the system changes its state at random. One of the major problems of probability is to study the asymptotic behavior of these systems over unbounded time intervals. We give one of the possible results in order to demonstrate the problems arising here. Let $T_t(E)$ be the total time that a system spends in the state E on the time interval $[0, t]$. Then the nonrandom

$$\lim_{t\to\infty} \frac{1}{t} T_t(E) = \pi(E)$$

exists with probability 1; $\pi(E)$ is the probability that the system will be found in the state E after a sufficiently long time. More precisely, the probability that the system is in the state E at time t tends to $\pi(E)$ as $t \to \infty$. This assertion holds of course under certain assumptions on the system in question. We cannot state them at this point since the needed concepts still have not been introduced. Assertions of this kind are lumped together under the generic name of *ergodic theorems*. Just as for the laws of large numbers, they provide reliable conclusions from random premises. One may be interested in a more exact behavior of the sojourn time in a given state, for instance, in studying the behavior of the difference $[t^{-1}T_t(E) - \pi(E)]$ multiplied by a suitable increasing function of t (the difference itself tends to zero). Under very broad assumptions, this difference multiplied by $\sqrt{t}$ behaves primarily the same way for all systems. We have now the second most important probability law (after the law of large numbers), which may be called the *law of normal fluctuations*. It holds also for relative frequencies and says that the deviation of a relative frequency from a probability after multiplication by a suitable constant behaves the same way in all cases (this is expressed precisely by the phrase "has a normal distribution"; what this means will be explained later on). Among the practically important problems involving stochastic systems is "predicting" their behavior from observations of their past behavior.

1.3.4 Stochastic Analysis

Moving on from the concept of random event, one could "randomize" any mathematical object. Such randomization is widely employed and studied in probability. The new objects do not result in idle philosophizing. They come about in an essential way and nontrivial important theorems are associated with them that find extensive application in the natural sciences and engineering. The first thing of this kind is the random number (or random variable in the accepted terminology). Such variables appear in experiments in which one or more characteristics of the experimental results are being measured. Following this, it is natural to consider the arithmetic of these variables and then to extend the concepts of mathematical analysis to them: limit, functional dependence and so on. Thus we arrive at the notions of random function, random operator, random mapping, stochastic integral, stochastic differential equation, etc. This is a comparatively new rather intensively developing area of probability theory. Despite their stochastic coloration, the problems that arise here are often analogous to problems of ordinary analysis.

2

Probability Space

The probability space is the basic object of study in probability theory and formalizes the notion of random experiment. A *probability space* is defined by three things: the space Ω of elementary events or sample space, a σ-algebra $\mathcal{A}$ of subsets of Ω called events, and a countably-additive nonnegative normalized set function $\mathbf{P}(A)$ defined on $\mathcal{A}$, which is called probability. A probability space defined by this triple is denoted by $(\Omega, \mathcal{A}, \mathbf{P})$.

2.1 Finite Probability Space

A finite probability space is one whose sample space is a finite set and $\mathcal{A}$ comprises all of the subsets of Ω. The probability is defined by its values on the elementary events.

2.1.1 Combinatorial Analysis

Suppose that the probabilities of all of the elementary events are the same (they are equally likely). To find the probability of an event A, it is necessary to know the overall number of elementary events and the number of those elementary events which imply A. The number of elements in a finite set can be calculated using direct methods that sort out all of the possibilities or combinatorial methods. Only the latter are of mathematical interest. We consider some examples applying them.

(a) *Allocation of particles in cells.* Problems of this kind arise in statistical physics. Given n cells in which N particles are distributed at random. What is the distribution of the particles in the cells? The answer depends on what are considered to be the elementary events.

Maxwell-Boltzmann statistics. We assume that all of the particles are distinct and all allocations of particles are equally likely. An elementary event is given by the sequence $(k_1, k_2, \ldots k_N)$, where k_i is the number of the cell into which the particle numbered i has fallen. Since each k_i assumes n distinct values, the number of such sequences is n^N. The probability of an elementary event is n^{-N}.

Bose-Einstein statistics. The particles are indistinguishable. Again all of the allocations are equally likely. An elementary event is given by the sequence $(\ell_1, \ldots, \ell_n)$, where $(\ell_1 + \ldots + \ell_n = N$ and ℓ_i is the number of particles in the i-th cell, $i \leq n$. The number of such sequences can be calculated as follows. With each $(\ell_1, \ldots, \ell_n)$ associate a sequence of zeroes and ones $(i_1, \ldots, i_{k+n-1})$ with zeroes in the positions numbered $\ell_1+1, \ell_1+\ell_2+2, \ldots, \ell_1+\ell_2+\ldots+\ell_{n-1}+n-1$ (there are $n-1$ of them) and ones in the remaining positions. The number of such sequences is equal to the number of combinations of $N+n-1$ things taken $n-1$ at a time. The probability of an elementary event is $\binom{N+n-1}{n-1}^{-1}$.

Fermi-Dirac statistics. In this case $N < n$ and each cell contains at most one particle. Then the number of elementary events is $\binom{n}{N}^{-1}$.

For each of the three statistics, we find the probability that a given cell (say, number 1) has no particle. Each time the number of favorable elementary events equals the number of allocations of the particles into $n-1$ cells. Therefore if we let p_1, p_2, and p_3 be the probabilities of the specified event for each statistics (in order of discussion), we have

$$\begin{aligned} p_1 &= (n-1)^N/n^N = \left(1 - \frac{1}{n}\right)^N , \\ p_2 &= \binom{N+n-2}{n-2} \Big/ \binom{N+n-1}{n-1} = \frac{n-1}{N+n-1} . \\ p_3 &= \binom{n-1}{N} \Big/ \binom{n}{N} = 1 - \frac{N}{n} . \end{aligned}$$

If $N/n = \alpha$ and $n \to \infty$, then

$$p_1 = e^{-\alpha}, \quad p_2 = \frac{1}{1+\alpha}, \quad p_3 = 1 - \alpha .$$

For small α, these probabilities coincide up to $O(\alpha^2)$. α characterizes the "average density" of the particles. If α is small, then the three probabilities are primarily equal.

(b) *Samples.* A sample may be defined in general as follows. There are m finite sets $A_1, A_2, \ldots, A_m$. From each set, we choose an element $a_i \in A_i$ one by one. The collection $(a_1, \ldots, a_m)$ is then the sample. Samples are distinguished by identification rules (let us say, we are not interested in the order of the elements in a sample). Each sample is regarded as an elementary event and the elementary events are considered to be equally likely.

1. *Sampling with replacement.* In this instance, the A_i coincide: $A_i = A$ and the number of samples is n^m, where n is the number of elements in A.
2. *Sampling without replacement.* A sample is constructed as follows. $A_1 = A$, $A_2 = A\backslash\{a_1\}, \ldots, A_k = A\backslash\{a_1, \ldots, a_{k-1}\}$. In other words, only samples $(a_1, \ldots, a_m)$, $a_i \in A$, are considered in which all of the elements are distinct. If A has n elements, then the number of samples without replacement is $n(n-1)\ldots(n-m+1)/m! = \binom{n}{m}$.
3. *Sampling without replacement from intersecting sets.* In this instance, the A_i have points in common but we are considering samples in which all of the elements are distinct. The number of such samples may be computed as follows. Consider the set $A = \bigcup_{k=1}^{m} A_k$ and the algebra $\mathcal{A}$ of subsets of it generated by $A_1, \ldots, A_m$. This is a finite algebra. Let $B_1, B_2, \ldots, B_N$ be atoms of the algebra, that is, they each have no subsets belonging to the algebra other than the *empty set* and themselves. Let $n(B_{i_1}, \ldots, B_{i_m})$ denote the number of samples without replacement from $B_{i_1}, \ldots, B_{i_m}$, where each B_{i_k} may be any atom. The value of $n(B_{i_1}, \ldots, B_{i_m})$ depends on the distinct sets encountered in the sequence and on the number of times these sets are repeated. Let $n(\ell_1, \ell_2, \ldots, \ell_N)$ be the number of samples from such a sequence, where B_1 occurs ℓ_1 times, B_2 occurs ℓ_2 times and so on, $\ell_i \geq 0, \ell_1 + \ldots + \ell_N = m$. If B_i has n_i elements, then

$$n(\ell_1, \ldots, \ell_N) = \prod_{i=1}^{N} \frac{n_i!}{(n_i - \ell_i)!} \ .$$

The number of samples of interest to us equals

$$\sum_{B_{i_1} \subset A_1, \ldots, B_{i_m} \subset A_m} n(B_{i_1}, \ldots, B_{i_m}) \ .$$

2.1.2 Conditional Probability

The *conditional probability* of an event A given event B having positive probability has occurred is the quantity

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \ . \qquad (2.1.1)$$

As a function of A, $\mathbf{P}(A|B)$ possesses all of the properties of a probability. The meaning of conditional probability may be explained as follows. Together with the original experiment, consider a conditional probability experiment which is performed if event B has happened in the original experiment. Thus if the original experiment has been done n times and B has happened n_B times, then this sequence contains n_B conditional experiments. The event A will have occurred in the conditional experiment if A and B occur simultaneously, i.e., if $A \cap B$ occurs. If $n_{A\cap B}$ is the number of experiments in which the event $A \cap B$ is observed (of the n carried out), then the relative frequency of occurrence in the n_B conditional experiments is $n_{A\cap B}/n_B = \nu_n(A\cap B)/\nu_n(B)$. If we replace the relative frequencies by the probabilities, then we have the right-hand side of (1.2.1).

(a) *Formula of total probability. Bayes's theorem.* A finite collection of events $H_1, H_2, \ldots, H_r$ is said to form a complete group of events if they are pairwise disjoint and their union is the sure event: 1. $H_i \cap H_j = \emptyset$ if $i \neq j$; 2. $\bigcup_i H_i = \Omega$. One can consider a supplementary experiment in which the H_i are the elementary events and the original experiment is viewed as a compound experiment: first one clarifies which H_i has occurred and then knowing H_i, one performs a conditional experiment under the assumption that H_i has occurred. An event A occurs in the conditional experiment with probability $P(A|H_i)$, the conditional probability of A given H_i. In many problems, the H_i are called the *causes* or *hypotheses* and the conditional probabilities given the causes are prescribed. The following relation expressing the probability of an event in terms of these conditional probabilities and the probabilities of causes is called the *formula of total probability*:

$$\mathbf{P}(A) = \sum_{i=1}^{r} \mathbf{P}(A|H_i)\mathbf{P}(H_i) \,. \tag{2.1.2}$$

On the basis of (2.1.1) the right-hand side becomes $\sum_{i=1}^{r} \mathbf{P}(A \cap H_i)$ and since the events $A \cap H_i$ are mutually exclusive and $\cup H_i = \Omega$, it follows that

$$\sum_{i=1}^{r} \mathbf{P}(A \cap H_i) = \mathbf{P}\left(\bigcup_{i=1}^{r}(A \cap H_i)\right) = \mathbf{P}\left(A \cap \bigcup_{i=1}^{r} H_i\right) = \mathbf{P}(A) \,.$$

Formula (2.1.2) is really useful when considering a compound experiment.

Example. There are r urns containing black and white balls. The probability of drawing a white ball from the urn numbered i is p_i. One of the urns is chosen at random and then a ball is drawn from it. By formula (2.1.2), we determine the probability of drawing a white ball. In our case, $\mathbf{P}(H_i) = 1/r$, $\mathbf{P}(A|H_i) = p_i$ and hence $\mathbf{P}(A) = r^{-1}\sum_{i=1}^{r} p_i$.

The formula of total probability leads to an important result called *Bayes's theorem.* It enables one to find the conditional probabilities of the causes given that an event A has occurred:

$$\mathbf{P}(H_k|A) = \mathbf{P}(A|H_k)\mathbf{P}(H_k) \Big/ \sum_{i=1}^{r} \mathbf{P}(A|H_i)\mathbf{P}(H_i) \,. \tag{2.1.3}$$

This formula is commonly interpreted as follows. The conditional probabilities of an event given each of the causes $H_1, \dots, H_r$ and the probabilities of the causes are assumed to be known. If the experiment has resulted in the occurrence of event A, then the probabilities of the causes have changed: once we know that A has already occurred, then it is natural to treat the probabilities of the causes as their conditional probabilities given A. The $\mathbf{P}(H_i)$ are called the apriori probabilities of the causes and the $\mathbf{P}(H_i|A)$ are their aposteriori probabilities. Bayes's theorem expresses the aposteriori probabilities of the causes in terms of their apriori probabilities and the conditional probabilities of an event given the various causes.

Example. There are two urns of which the first contains 2 white and 8 black balls and the second 8 white and 2 black balls. An urn is selected at random and a ball is drawn from it. It is white. What is the probability that the first urn was chosen? Here we have $\mathbf{P}(H_1) = \mathbf{P}(H_2) = 1/2$, $\mathbf{P}(A|H_1) = 1/5$ and $\mathbf{P}(A|H_2) = 4/5$. By (2.1.3),

$$\mathbf{P}(H_1|A) = 1/2 \cdot 1/5/(1/2 \cdot 1/5 + 1/2 \cdot 4/5) = 1/5\ .$$

(b) *Independence.* An event A does not depend on an event B if the conditional probability $\mathbf{P}(A|B)$ equals the unconditional probability $\mathbf{P}(A)$. In that case,

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)\ , \tag{2.1.4}$$

which shows that the property of *independence* is symmetric. Formula (2.1.4) could serve as a definition of independence of two events A and B. The first definition is more meaningful: the fact that B has occurred has no affect on the probability of A and it is reasonable to assume that A does not depend on B. It follows from (2.1.4) that the independence of A and B implies the independence of A and $\bar{B}$, $\bar{A}$ and B, and $\bar{A}$ and $\bar{B}$ ($\bar{A}$ is the negation of the event A). Independence is defined for several events as follows. $A_1, A_2, \dots, A_m$ are said to be mutually independent if

$$\mathbf{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_K}) = \mathbf{P}(A_{i_1}) \ldots \mathbf{P}(A_{i_k}) \tag{2.1.5}$$

for any $k \leq m$ and $i_1 < i_2 \ldots < i_k \leq m$. Thus for three events A, B and C their independence means that the following four equalities hold: $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$, $\mathbf{P}(A \cap C) = \mathbf{P}(A)\mathbf{P}(C)$, $\mathbf{P}(B \cap C) = \mathbf{P}(B)\mathbf{P}(C)$ and $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$.

Bernstein's example. The sample space consists of four elements E_1, E_2, E_3, and E_4 with $\mathbf{P}(E_k) = 1/4$, $k = 1, 2, 3, 4$. Let $A_i = E_i \cup E_4, i = 1, 2, 3$. Then $A_1 \cap A_2 = A_1 \cap A_3 = A_2 \cap A_3 = A_1 \cap A_2 \cap A_3 = E_4$. Therefore $\mathbf{P}(A_1 \cap A_2) = \mathbf{P}(A_1)\mathbf{P}(A_2)$, $\mathbf{P}(A_1 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_3)$ and $\mathbf{P}(A_2 \cap A_3) = \mathbf{P}(A_2)\mathbf{P}(A_3)$. But $\mathbf{P}(A_1 \cap A_2 \cap A_3) \neq \mathbf{P}(A_1)\mathbf{P}(A_2)\mathbf{P}(A_3)$. The events are pairwise independent but they are not mutually independent.

2.1.3 Bernoulli's Scheme. Limit Theorems

Let $A_1, A_2, \ldots, A_r$ be a complete group of events. An event B is independent of this group if it does not depend on any of the events $A_k, k = 1, \ldots, r$. Let $\mathcal{A}$ be the algebra generated by the events $A_1, \ldots, A_r$; it comprises the impossible event and all unions of the form $\bigcup_k A_{i_k}$, $i_k \leq r$. Then B is independent of the algebra $\mathcal{A}$, that is, it does not depend on any event $A \in \mathcal{A}$. Two *algebras* of events $\mathcal{A}_1$ and $\mathcal{A}_2$ are said to be *independent* if A_1 and A_2 are independent for each pair of events $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$. Algebras of events $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m$ are independent if $A_1, A_2, \ldots, A_m$ are mutually independent, where $A_i \in \mathcal{A}_i$, $i \leq m$. To this end, it suffices that

$$\mathbf{P}\left(\bigcap_{i=1}^{m} A_i\right) = \prod_{i=1}^{m} \mathbf{P}(A_i) \tag{2.1.6}$$

for any choice of $A_i \in \mathcal{A}_i$. (This definition simplifies as compared to that of independent events in (2.1.5) because some A_i may be chosen to be Ω.)

Consider several experiments specified by the probability spaces $(\Omega_k, \mathcal{A}_k, \mathbf{P}_k)$, $k = 1, 2, \ldots, n$. We now form a new probability space $(\Omega, \mathcal{A}, \mathbf{P})$. Ω is taken to be the Cartesian product $\Omega_1 \times \Omega_2 \times \ldots \times \Omega_n$. The algebra $\mathcal{A}$ is the *product of algebras* $\mathcal{A}_1 \otimes \mathcal{A}_2 \otimes \ldots \otimes \mathcal{A}_n$ of subsets of Ω generated by sets of the form $A_1 \times A_2 \times \ldots \times A_n$ with $A_k \in \mathcal{A}_k$, $k = 1, 2, \ldots, n$ (an algebra is said to be generated by a collection of sets if it is the smallest algebra containing that collection). Finally, the measure $\mathbf{P}$ is the product of measures $\mathbf{P}_k$: $\mathbf{P} = \prod_{k=1}^{n} \mathbf{P}_k$, that is, $\mathbf{P}(A_1 \times A_2 \ldots \times A_n) = \mathbf{P}(A_1)\mathbf{P}(A_2)\ldots\mathbf{P}(A_n)$. The probability space $(\Omega, \mathcal{A}, \mathbf{P})$ corresponds to a compound experiment in which each of the n experiments specified above is performed independently.

(a) *Bernoulli's scheme* involves a series of independent and identical experiments (trials). This just means that a probability space $(\Omega_1 \times \ldots \times \Omega_n, \mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n, \prod_{i=1}^{n} \mathbf{P}_i)$ is defined for every n in which each probability space $(\Omega_k, \mathcal{A}_k, \mathbf{P}_k)$ coincides with the exact same space $(\Omega, \mathcal{A}, \mathbf{P})$. (As we shall see below, it is possible to consider an infinite product of such probability spaces right away; it will not be finite if the given space is nontrivial, that is, Ω contains more than one element.) Let $A \in \mathcal{A}$. The event $\Omega \times \ldots \times A \times \ldots \times \Omega$, where A is in the k-th position and the remaining factors are Ω, is interpreted as the event "A occurred in the k-th experiment." Let $p_n(m)$ denote the probability that A happens exactly m times in n independent trials. Then

$$p_n(m) = \binom{n}{m} p^m (1-p)^{n-m}, \quad p = \mathbf{P}(A)\,. \tag{2.1.7}$$

Indeed, our event of interest is the union of events of the form $A \times \bar{A} \times \ldots \times A \times \ldots \times \bar{A} \times \ldots \times A$, where A occurs in the product m times and $\bar{A}$ occurs $n - m$ times. There are $\binom{n}{m}$ such distinct products and the probability of one such event is $p^m(1-p)^{n-m}$.

Let $A_1, A_2, \dots A_r$ be a complete group of events in an algebra $\mathcal{A}$. Let $p_n(k_1, \dots, k_r)$ be the probability that in n independent trials A_i occurs k_i times, $i = 1, \dots, r$ and $k_1 +, \dots, +k_r = n$. Similarly to the preceding, one can establish that

$$p_n(k_1, \dots, k_r) = \frac{n!}{k_1! \dots k_r!} p_1^{k_1} \dots p_r^{k_r}, \quad p_i = \mathbf{P}(A_i) \ , \quad i = 1, \dots, r \ . \tag{2.1.8}$$

The probabilities (2.1.7) are called the *binomial probabilities* and (2.1.8) the *multinomial probabilities.*

(b) *The law of large numbers.* This law has been mentioned several times in the introductory chapter. We are now in a position to prove it.

Bernoulli's Theorem. *Let ν_n be the number of occurences of an event A in n independent trials having probability p in each trial, $0 < p < 1$. Then for any positive ε,*

$$\lim_{n \to \infty} \mathbf{P} \left\{ \left| \frac{1}{n} \nu_n - p \right| > \varepsilon \right\} = 0 \ . \tag{2.1.9}$$

Proof. For fixed n, the event $\{\nu_n = k\}$ has probability $p_n(k)$. For different values of k, these events are mutually exclusive. Therefore

$$\mathbf{P} \left\{ \left| \frac{1}{n} \nu_n - p \right| > \varepsilon \right\} = \sum_{k < n(p - \varepsilon)} p_n(k) + \sum_{k > n(p + \varepsilon)} p_n(k) \ .$$

Starting with (2.1.7), we find that

$$\frac{p_n(k+1)}{p_n(k)} = \frac{n-k}{k+1} \frac{p}{1-p} \ .$$

Therefore for $k > n(p + \varepsilon)$,

$$\frac{p_n(k+1)}{p_n(k)} < \frac{n - n(p+\varepsilon)}{np} \frac{p}{1-p} = 1 - \frac{\varepsilon}{1-p} \ .$$

Let $k*$ denote the smallest value of k satisfying $k > n(p + \varepsilon)$ and let k_* be the smallest value of k for which $(n-k)p/[(k+1)(1-p)] < 1$. Then $p_n(k+1) < p_n(k)$ for $k \geq k_*$. Therefore

$$\sum_{k > n(p+\varepsilon)} p_n(k) = \sum_{k \geq k^*} p_n(k) < p_n(k^*) \sum_{m=0}^{\infty} \left(1 - \frac{\varepsilon}{1-p} \right)^m = \frac{(1-p)}{\varepsilon} p_n(k^*) \ .$$

Next,

$$1 \geq \sum_{k=k_*}^{k^*} p_n(x) \geq (k^* - k_*) p_n(k^*)$$