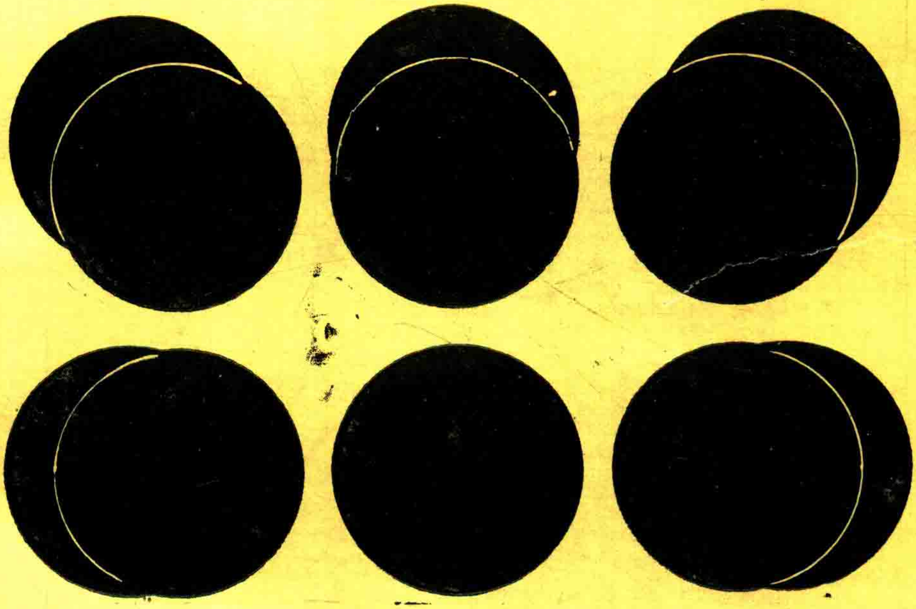
Hoel
Port
Stone

Introduction to Probability Theory



Introduction to Probability Theory

Paul G. Hoel
Sidney C. Port
Charles J. Stone
University of California, Los Angeles

HOUGHTON MIFFLIN COMPANY BOSTON
Atlanta Dallas Geneva, Illinois Hopewell, New Jersey
Palo Alto London

PRINTED IN THE U.S.A.

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 74-136173

ISBN: 0-395-04636-X

The Houghton Mifflin Series in Statistics
under the Editorship of Herman Chernoff

LEO BREIMAN
Probability and Stochastic Processes: With a View Toward Applications
Statistics: With a View Toward Applications

PAUL G. HOEL, SIDNEY C. PORT, AND CHARLES J. STONE
Introduction to Probability Theory
Introduction to Statistical Theory
Introduction to Stochastic Processes

PAUL F. LAZARSFELD AND NEIL W. HENRY
Latent Structure Analysis

GOTTFRIED E. NOETHER
Introduction to Statistics—A Fresh Approach

Y. S. CHOW, HERBERT ROBBINS, AND DAVID SIEGMUND
Great Expectations: The Theory of Optimal Stopping

I. RICHARD SAVAGE
Statistics: Uncertainty and Behavior

General Preface

This three-volume series grew out of a three-quarter course in probability, statistics, and stochastic processes taught for a number of years at UCLA. We felt a need for a series of books that would treat these subjects in a way that is well coordinated, but which would also give adequate emphasis to each subject as being interesting and useful on its own merits.

The first volume, *Introduction to Probability Theory*, presents the fundamental ideas of probability theory and also prepares the student both for courses in statistics and for further study in probability theory, including stochastic processes.

The second volume, *Introduction to Statistical Theory*, develops the basic theory of mathematical statistics in a systematic, unified manner. Together, the first two volumes contain the material that is often covered in a two-semester course in mathematical statistics.

The third volume, *Introduction to Stochastic Processes*, treats Markov chains, Poisson processes, birth and death processes, Gaussian processes, Brownian motion, and processes defined in terms of Brownian motion by means of elementary stochastic differential equations.

Preface

This volume is intended to serve as a text for a one-quarter or one-semester course in probability theory at the junior-senior level. The material has been designed to give the reader adequate preparation for either a course in statistics or further study in probability theory and stochastic processes. The prerequisite for this volume is a course in elementary calculus that includes multiple integration.

We have endeavored to present only the more important concepts of probability theory. We have attempted to explain these concepts and indicate their usefulness through discussion, examples, and exercises. Sufficient detail has been included in the examples so that the student can be expected to read these on his own, thereby leaving the instructor more time to cover the essential ideas and work a number of exercises in class.

At the conclusion of each chapter there are a large number of exercises, arranged according to the order in which the relevant material was introduced in the text. Some of these exercises are of a routine nature, while others develop ideas introduced in the text a little further or in a slightly different direction. The more difficult problems are supplied with hints. Answers, when not indicated in the problems themselves, are given at the end of the book.

Although most of the subject matter in this volume is essential for further study in probability and statistics, some optional material has been included to provide for greater flexibility. These optional sections are indicated by an asterisk. The material in Section 6.2.2 is needed only for Section 6.6; neither section is required for this volume, but both are needed in *Introduction to Statistical Theory*. The material of Section 6.7 is used only in proving Theorem 1 of Chapter 9 in this volume and Theorem 1 of Chapter 5 in *Introduction to Statistical Theory*. The contents of Chapters 8 and 9 are optional; Chapter 9 does not depend on Chapter 8.

We wish to thank our several colleagues who read over the original manuscript and made suggestions that resulted in significant improvements. We also would like to thank Neil Weiss and Luis Gorostiza for obtaining answers to all the exercises and Mrs. Ruth Goldstein for her excellent job of typing.

Table of Contents

1 Probability Spaces

Probability theory is the branch of mathematics that is concerned with random (or chance) phenomena. It has attracted people to its study both because of its intrinsic interest and its successful applications to many areas within the physical, biological, and social sciences, in engineering, and in the business world.

Many phenomena have the property that their repeated observation under a specified set of conditions invariably leads to the same outcome. For example, if a ball initially at rest is dropped from a height of d feet through an evacuated cylinder, it will invariably fall to the ground in $t = \sqrt{2d/g}$ seconds, where $g = 32$ ft/sec^2 is the constant acceleration due to gravity. There are other phenomena whose repeated observation under a specified set of conditions does not always lead to the same outcome. A familiar example of this type is the tossing of a coin. If a coin is tossed 1000 times the occurrences of heads and tails alternate in a seemingly erratic and unpredictable manner. It is such phenomena that we think of as being random and which are the object of our investigation.

At first glance it might seem impossible to make any worthwhile statements about such random phenomena, but this is not so. Experience has shown that many nondeterministic phenomena exhibit a *statistical regularity* that makes them subject to study. This may be illustrated by considering coin tossing again. For any given toss of the coin we can make no nontrivial prediction, but observations show that for a large number of tosses the proportion of heads seems to fluctuate around some fixed number p between 0 and 1 (p being very near 1/2 unless the coin is severely unbalanced). It appears as if the proportion of heads in n tosses would converge to p if we let n approach infinity. We think of this limiting proportion p as the "probability" that the coin will land heads up in a single toss.

More generally the statement that a certain experimental outcome has probability p can be interpreted as meaning that if the experiment is repeated a large number of times, that outcome would be observed "about" $100p$ percent of the time. This interpretation of probabilities is called the relative frequency interpretation. It is very natural in many applications of probability theory to real world problems, especially to those involving the physical sciences, but it often seems quite artificial. How, for example, could we give a relative frequency interpretation to the probability that a given newborn baby will live at least 70 years? Various attempts have been made, none of them totally acceptable, to give alternative interpretations to such probability statements.

For the mathematical theory of probability the interpretation of probabilities is irrelevant, just as in geometry the interpretation of points, lines, and planes is irrelevant. We will use the relative frequency interpretation of probabilities only as an intuitive motivation for the definitions and theorems we will be developing throughout the book.

1.1. Examples of random phenomena

In this section we will discuss two simple examples of random phenomena in order to motivate the formal structure of the theory.

Example 1. A box has s balls, labeled $1, 2, \ldots, s$ but otherwise identical. Consider the following experiment. The balls are mixed up well in the box and a person reaches into the box and draws a ball. The number of the ball is noted and the ball is returned to the box. The outcome of the experiment is the number on the ball selected. About this experiment we can make no nontrivial prediction.

Suppose we repeat the above experiment n times. Let $N_n(k)$ denote the number of times the ball labeled k was drawn during these n trials of the experiment. Assume that we had, say, $s = 3$ balls and $n = 20$ trials. The outcomes of these 20 trials could be described by listing the numbers which appeared in the order they were observed. A typical result might be

$$1, 1, 3, 2, 1, 2, 2, 3, 2, 3, 3, 2, 1, 2, 3, 3, 1, 3, 2, 2,$$

in which case

$$N_{20}(1) = 5, \qquad N_{20}(2) = 8, \qquad \text{and} \qquad N_{20}(3) = 7.$$

The relative frequencies (i.e., proportion of times) of the outcomes 1, 2, and 3 are then

$$\frac{N_{20}(1)}{20} = .25, \qquad \frac{N_{20}(2)}{20} = .40, \qquad \text{and} \qquad \frac{N_{20}(3)}{20} = .35.$$

As the number of trials gets large we would expect the relative frequencies $N_n(1)/n, \ldots, N_n(s)/n$ to settle down to some fixed numbers $p_1, p_2, \ldots, p_s$ (which according to our intuition in this case should all be $1/s$).

By the relative frequency interpretation, the number p_i would be called the probability that the ith ball will be drawn when the experiment is performed once ($i = 1, 2, \ldots, s$).

We will now make a mathematical model of the experiment of drawing a ball from the box. To do this, we first take a set Ω having s points that we place into one-to-one correspondence with the possible outcomes of the experiment. In this correspondence exactly one point of Ω will be associated with the outcome that the ball labeled k is selected. Call that point ω_k. To the point ω_k we associate the number $p_k = 1/s$ and call it the probability of ω_k. We observe at once that $0 \leq p_k \leq 1$ and that $p_1 + \cdots + p_s = 1$.

Suppose now that in addition to being numbered from 1 to s the first r balls are colored red and the remaining $s - r$ are colored black. We perform the experiment as before, but now we are only interested in the color of the ball drawn and not its number. A moment's thought shows that the relative frequency of red balls drawn among n repetitions of the experiment is just the sum of the relative frequencies $N_n(k)/n$ over those values of k that correspond to a red ball. We would expect, and experience shows, that for large n this relative frequency should settle down to some fixed number. Since for large n the relative frequencies $N_n(k)/n$ are expected to be close to $p_k = 1/s$, we would anticipate that the relative frequency of red balls would be close to r/s. Again experience verifies this. According to the relative frequency interpretation, we would then call r/s the probability of obtaining a red ball.

Let us see how we can reflect this fact in our model. Let A be the subset of Ω consisting of those points ω_k such that ball k is red. Then A has exactly r points. We call A an event. More generally, in this situation we will call any subset B of Ω an event. To say the event B occurs means that the outcome of the experiment is represented by some point in B.

Let A and B be two events. Recall that the union of A and B, $A \cup B$, is the set of all points $\omega \in \Omega$ such that $\omega \in A$ or $\omega \in B$. Now the points in Ω are in correspondence with the outcomes of our experiment. The event A occurs if the experiment yields an outcome that is represented by some point in A, and similarly the event B occurs if the outcome of the experiment is represented by some point in B. The set $A \cup B$ then represents the fact that the event A occurs or the event B occurs. Similarly the intersection $A \cap B$ of A and B consists of all points that are in both A and B. Thus if $\omega \in A \cap B$ then $\omega \in A$ and $\omega \in B$ so $A \cap B$ represents the fact that both the events A and B occur. The complement A^c (or A') of A is the set of points in Ω that are not in A. The event A does not occur if the experiment yields an outcome represented by a point in A^c.

Diagrammatically, if A and B are represented by the indicated regions in Figure 1a, then $A \cup B$, $A \cap B$, and A^c are represented by the shaded regions in Figures 1b, 1c, and 1d, respectively.

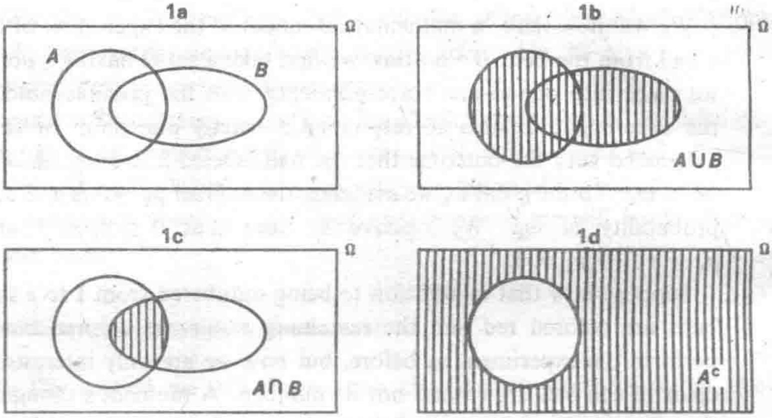


Figure 1

To illustrate these concepts let A be the event "red ball selected" and let B be the event "even-numbered ball selected." Then the union $A \cup B$ is the event that either a red ball or an even-numbered ball was selected. The intersection $A \cap B$ is the event "red even-numbered ball selected." The event A^c occurs if a red ball was not selected.

We now would like to assign probabilities to events. Mathematically, this just means that we associate to each set B a real number. A priori we could do this in an arbitrary way. However, we are restricted if we want these probabilities to reflect the experiment we are trying to model. How should we make this assignment? We have already assigned each point the number s^{-1}. Thus a one-point set $\{\omega\}$ should be assigned the number s^{-1}. Now from our discussion of the relative frequency of the event "drawing a red ball," it seems that we should assign the event A the probability $P(A) = r/s$. More generally, if B is any event we will define $P(B)$ by $P(B) = j/s$ if B has exactly j points. We then observe that

$$P(B) = \sum_{\omega_k \in B} p_k ,$$

where $\sum_{\omega_k \in B} p_k$ means that we sum the numbers p_k over those values of k such that $\omega_k \in B$. From our definition of $P(B)$ it easily follows that the following statements are true. We leave their verification to the reader.

Let $\varnothing$ denote the empty set; then $P(\varnothing) = 0$ and $P(\Omega) = 1$. If A and B are any two disjoint sets, i.e., $A \cap B = \varnothing$, then

$$P(A \cup B) = P(A) + P(B).$$

Example 2. It is known from physical experiments that an isotope of a certain substance is unstable. In the course of time it decays by the emission of neutrons to a stable form. We are interested in the time that it takes an atom of the isotope to decay to its stable form. According to the laws of physics it is impossible to say with certainty when a specified atom of the isotope will decay, but if we observe a large number N of atoms initially, then we can make some accurate predictions about the number of atoms $N(t)$ that have not decayed by time t. In other words we can rather accurately predict the fraction of atoms $N(t)/N$ that have not decayed by time t, but we cannot say which of the atoms will have done so. Since all of the atoms are identical, observing N atoms simultaneously should be equivalent to N repetitions of the same experiment where, in this case, the experiment consists in observing the time that it takes an atom to decay.

Now to a first approximation (which is actually quite accurate) the rate at which the isotope decays at time t is proportional to the number of atoms present at time t, so $N(t)$ is given approximately as the solution of the differential equation

$$\frac{df}{dt} = -\lambda f(t), \qquad f(0) = N,$$

where $\lambda > 0$ is a fixed constant of proportionality. The unique solution of this equation is $f(t) = Ne^{-\lambda t}$, and thus the fraction of atoms that have not decayed by time t is given approximately by $N(t)/N = e^{-\lambda t}$. If $0 \leq t_0 \leq t_1$, the fraction of atoms that decay in the time interval $[t_0, t_1]$ is $(e^{-\lambda t_0} - e^{-\lambda t_1})$. Consequently, in accordance with the relative frequency interpretation of probability we take $(e^{-\lambda t_0} - e^{-\lambda t_1})$ as the probability that an atom decays between times t_0 and t_1.

To make a mathematical model of this experiment we can try to proceed as in the previous example. First we choose a set Ω that can be put into a one-to-one correspondence with the possible outcomes of the experiment. An outcome in this case is the time that an atom takes to decay. This can be any positive real number, so we take Ω to be the interval $[0, \infty)$ on the real line. From our discussion above it seems reasonable to assign to the interval $[t_0, t_1]$ the probability $(e^{-\lambda t_0} - e^{-\lambda t_1})$. In particular if $t_0 = t_1 = t$ then the interval degenerates to the set $\{t\}$ and the probability assigned to this set is 0.

In our previous example Ω had only finitely many points; however, here Ω has a (noncountable) infinity of points and each point has probability 0. Once again we observe that $P(\Omega) = 1$ and $P(\varnothing) = 0$. Suppose A and B are two disjoint intervals. Then the proportion of atoms that decay in the time interval $A \cup B$ is the sum of the proportion that decay in the time interval A and the proportion that decay in the time interval B. In light of this additivity we demand that in the mathematical model, $A \cup B$ should have probability $P(A) + P(B)$ assigned to it. In other words, in the mathematical model we want

$$P(A \cup B) = P(A) + P(B)$$

whenever A and B are two disjoint intervals.

1.2. Probability spaces

Our purpose in this section is to develop the formal mathematical structure, called a probability space, that forms the foundation for the mathematical treatment of random phenomena.

Envision some real or imaginary experiment that we are trying to model. The first thing we must do is decide on the possible outcomes of the experiment. It is not too serious if we admit more things into our consideration than can really occur, but we want to make sure that we do not exclude things that might occur. Once we decide on the possible outcomes, we choose a set Ω whose points ω are associated with these outcomes. From the strictly mathematical point of view, however, Ω is just an abstract set of points.

We next take a nonempty collection $\mathscr{A}$ of subsets of Ω that is to represent the collection of "events" to which we wish to assign probabilities. By definition, now, an *event* means a set A in $\mathscr{A}$. The statement *the event A occurs* means that the outcome of our experiment is represented by some point $\omega \in A$. Again, from the strictly mathematical point of view, $\mathscr{A}$ is just a specified collection of subsets of the set Ω. Only sets $A \in \mathscr{A}$, i.e., events, will be assigned probabilities. In our model in Example 1, $\mathscr{A}$ consisted of all subsets of Ω. In the general situation when Ω does not have a finite number of points, as in Example 2, it may not be possible to choose $\mathscr{A}$ in this manner.

The next question is, what should the collection $\mathscr{A}$ be? It is quite reasonable to demand that $\mathscr{A}$ be closed under finite unions and finite intersections of sets in $\mathscr{A}$ as well as under complementation. For example, if A and B are two events, then $A \cup B$ occurs if the outcome of our experiment is either represented by a point in A or a point in B. Clearly, then, if it is going to be meaningful to talk about the probabilities that A and B occur, it should also be meaningful to talk about the probability that either A or B occurs, i.e., that the event $A \cup B$ occurs. Since only sets in $\mathscr{A}$ will be assigned probabilities, we should require that $A \cup B \in \mathscr{A}$ whenever A and B are members of $\mathscr{A}$. Now $A \cap B$ occurs if the outcome of our experiment is represented by some point that is in both A and B. A similar line of reasoning to that used for $A \cup B$ convinces us that we should have $A \cap B \in \mathscr{A}$ whenever $A, B \in \mathscr{A}$. Finally, to say that the event A does not occur is to say that the outcome of our experiment is not represented by a point in A, so that it must be represented by some point in A^c. It would be the height of folly to say that we could talk about the probability of A but not of A^c. Thus we shall demand that whenever A is in $\mathscr{A}$ so is A^c.

We have thus arrived at the conclusion that $\mathscr{A}$ should be a nonempty collection of subsets of Ω having the following properties:

(i) If A is in $\mathscr{A}$ so is A^c.

(ii) If A and B are in $\mathscr{A}$ so are $A \cup B$ and $A \cap B$.

An easy induction argument shows that if $A_1, A_2, \ldots, A_n$ are sets in $\mathscr{A}$ then so are $\bigcup_{i=1}^n A_i$ and $\bigcap_{i=1}^n A_i$. Here we use the shorthand notation

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n$$

and

$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n.$$

Also, since $A \cap A^c = \varnothing$ and $A \cup A^c = \Omega$, we see that both the empty set $\varnothing$ and the set Ω must be in $\mathscr{A}$.

A nonempty collection of subsets of a given set Ω that is closed under finite set theoretic operations is called a *field of subsets* of Ω. It therefore seems we should demand that $\mathscr{A}$ be a field of subsets. It turns out, however, that for certain mathematical reasons just taking $\mathscr{A}$ to be a field of subsets of Ω is insufficient. What we will actually demand of the collection $\mathscr{A}$ is more stringent. We will demand that $\mathscr{A}$ be closed not only under finite set theoretic operations but under countably infinite set theoretic operations as well. In other words if $\{A_n\}$, $n \geq 1$, is a sequence of sets in $\mathscr{A}$, we will demand that

$$\bigcup_{n=1}^{\infty} A_n \in \mathscr{A} \quad \text{and} \quad \bigcap_{n=1}^{\infty} A_n \in \mathscr{A}.$$

Here we are using the shorthand notation

$$\bigcup_{n=1}^{\infty} A_n = A_1 \cup A_2 \cup \cdots$$

to denote the union of all the sets of the sequence, and

$$\bigcap_{n=1}^{\infty} A_n = A_1 \cap A_2 \cap \cdots$$

to denote the intersection of all the sets of the sequence. A collection of subsets of a given set Ω that is closed under countable set theory operations is called a σ-field of subsets of Ω. (The σ is put in to distinguish such a collection from a field of subsets.) More formally we have the following:

Definition 1 *A nonempty collection of subsets $\mathscr{A}$ of a set Ω is called a σ-field of subsets of Ω provided that the following two properties hold:*

(i) *If A is in $\mathscr{A}$, then A^c is also in $\mathscr{A}$.*

(ii) *If A_n is in $\mathscr{A}$, $n = 1, 2, \ldots$, then $\bigcup_{n=1}^{\infty} A_n$ and $\bigcap_{n=1}^{\infty} A_n$ are both in $\mathscr{A}$.*

We now come to the assignment of probabilities to events. As was made clear in the examples of the preceding section, the probability of an event is a nonnegative real number. For an event A, let $P(A)$ denote this number. Then $0 \leq P(A) \leq 1$. The set Ω representing every possible outcome should, of course, be assigned the number 1, so $P(\Omega) = 1$. In our discussion of Example 1 we showed that the probability of events satisfies the property that if A and B are any two disjoint events then $P(A \cup B) = P(A) + P(B)$. Similarly, in Example 2 we showed that if A and B are two disjoint intervals, then we should also require that

$$P(A \cup B) = P(A) + P(B).$$

It now seems reasonable in general to demand that if A and B are disjoint events then $P(A \cup B) = P(A) + P(B)$. By induction, it would then follow that if $A_1, A_2, \ldots, A_n$ are any n mutually disjoint sets (that is, if $A_i \cap A_j = \varnothing$ whenever $i \neq j$), then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

Actually, again for mathematical reasons, we will in fact demand that this additivity property hold for countable collections of disjoint events.

Definition 2 *A probability measure P on a σ-field of subsets $\mathscr{A}$ of a set Ω is a real valued function having domain $\mathscr{A}$ satisfying the following properties:*

(i) $P(\Omega) = 1$.
(ii) $P(A) \geq 0$ *for all* $A \in \mathscr{A}$.
(iii) *If* A_n, $n = 1, 2, 3, \ldots$, *are mutually disjoint sets in* $\mathscr{A}$, *then*

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

A probability space, denoted by $(\Omega, \mathscr{A}, P)$, is a set Ω, a σ-field of subsets $\mathscr{A}$, and a probability measure P defined on $\mathscr{A}$.

It is quite easy to find a probability space that corresponds to the experiment of drawing a ball from a box. In essence it was already given in our discussion of that experiment. We simply take Ω to be a finite set having s points, $\mathscr{A}$ to be the collection of all subsets of Ω, and P to be the probability measure that assigns to A the probability $P(A) = j/s$ if A has exactly j points.

Let us now consider the probability space associated with the isotope disintegration experiment (Example 2). Here it is certainly clear that $\Omega = [0, \infty)$, but it is not obvious what $\mathscr{A}$ and P should be. Indeed, as we will indicate below, this is by no means a trivial problem, and one that in all its ramifications depends on some deep properties of set theory that are beyond the scope of this book.