Probability Theory:
The Logic of Science

by

E. T. Jaynes

Wayman Crow Professor of Physics

Washington University

St. Louis, MO 63130, U. S. A.

Dedicated to the Memory of Sir Harold Jeffreys,
who saw the truth and preserved it.

EDITORS FORWARD

E. T. Jaynes died April 30, 1998. Before his death he asked me to finish and publish his book on probability theory. I struggled with this for some time, because there is no doubt in my mind that Jaynes wanted this book finished. Unfortunately, most of the later Chapters, Jaynes' intended volume 2 on applications, were either missing or incomplete and some of the early also Chapters had missing pieces. I could have written these latter Chapters and filled the missing pieces, but if I did so, the work would no longer belong to Jaynes; rather, it would be a Jaynes-Bretthorst hybrid with no way to tell which material came from which author. In the end, I decided that the missing Chapters would have to stay missing—the work would remain Jaynes'.

There were a number of missing pieces of varying length that Jaynes had marked by inserting the phrase "MUCH MORE COMING." I could have left these comments in the text, but they were ugly and they made the book looks very incomplete. Jaynes intended this book to serve as both a reference and a text book. Consequently, there are question boxes scattered throughout most Chapters. In the end, I decided to replace the "MUCH MORE COMING" comments by introducing an "editors" question box. If you answer these questions, you will have filled in the missing material. You will be able to identify these questions because I used a shaded box for the editors questions, while Jaynes' question boxes are not shaded.

Jaynes' wanted to include a series of computer programs that implemented some of the calculations in this book. I had originally intended to include these programs. But as time went on, it became increasingly obvious that many of the programs were not available and the ones that were, were written in a particularly obscure form of BASIC (it was the programs that were obscure, not the BASIC). Consequently, I removed references to these programs and, where necessary, inserted a few sentences to direct people to the necessary software tools to implement the calculations.

Finally, while I am the most obvious person who has worked on getting this book into publication, I am not the only person to do so. Some of Jaynes' closest friends have assisted me in completing this work. These include Tom Grandy, Ray Smith, Tom Loredo, Myron Tribus and John Skilling, and I would like to thank them for their assistance. I would also like to thank Joe Ackerman for allowing me to take the time necessary to get this work published.

G. Larry Bretthorst, Editor
May 2002

PROBABILITY THEORY – THE LOGIC OF SCIENCE

VOLUME I – PRINCIPLES AND ELEMENTARY APPLICATIONS

VOLUME II – ADVANCED APPLICATIONS

PREFACE

The following material is addressed to readers who are already familiar with applied mathematics at the advanced undergraduate level or preferably higher; and with some field, such as physics, chemistry, biology, geology, medicine, economics, sociology, engineering, operations research, etc., where inference is needed.† A previous acquaintance with probability and statistics is not necessary; indeed, a certain amount of innocence in this area may be desirable, because there will be less to unlearn.

We are concerned with probability theory and all of its conventional mathematics, but now viewed in a wider context than that of the standard textbooks. Every Chapter after the first has "new" (*i.e.* not previously published) results that we think will be found interesting and useful. Many of our applications lie outside the scope of conventional probability theory as currently taught. But we think that the results will speak for themselves, and that something like the theory expounded here will become the conventional probability theory of the future.

History: The present form of this work is the result of an evolutionary growth over many years. My interest in probability theory was stimulated first by reading the work of Harold Jeffreys (1939) and realizing that his viewpoint makes all the problems of theoretical physics appear in a very different light. But then in quick succession discovery of the work of R. T. Cox (1946), C. E. Shannon (1948) and G. Pólya (1954) opened up new worlds of thought, whose exploration has occupied my mind for some forty years. In this much larger and permanent world of rational thinking in general, the current problems of theoretical physics appeared as only details of temporary interest.

The actual writing started as notes for a series of lectures given at Stanford University in 1956, expounding the then new and exciting work of George Pólya on "Mathematics and Plausible Reasoning." He dissected our intuitive "common sense" into a set of elementary qualitative desiderata and showed that mathematicians had been using them all along to guide the early stages of discovery, which necessarily precede the finding of a rigorous proof. The results were much like those of James Bernoulli's "Art of Conjecture" (1713), developed analytically by Laplace in the late 18'th century; but Pólya thought the resemblance to be only qualitative.

However, Pólya demonstrated this qualitative agreement in such complete, exhaustive detail as to suggest that there must be more to it. Fortunately, the consistency theorems of R. T. Cox were enough to clinch matters; when one added Pólya's qualitative conditions to them the result was a proof that, if degrees of plausibility are represented by real numbers, then there is a uniquely determined set of quantitative rules for conducting inference. That is, any other rules whose results conflict with them will necessarily violate an elementary—and nearly inescapable—desideratum of rationality or consistency.

But the final result was just the standard rules of probability theory, given already by Bernoulli and Laplace; so why all the fuss? The important new feature was that these rules were now seen as uniquely valid principles of logic in general, making no reference to "chance" or "random variables"; so their range of application is vastly greater than had been supposed in the conventional probability theory that was developed in the early twentieth century. As a result, the imaginary distinction between "probability theory" and "statistical inference" disappears, and the field achieves not only logical unity and simplicity, but far greater technical power and flexibility in applications.

† By "inference" we mean simply: deductive reasoning whenever enough information is at hand to permit it; inductive or plausible reasoning when—as is almost invariably the case in real problems—the necessary information is not available. But if a problem can be solved by deductive reasoning, probability theory is not needed for it; thus our topic is the optimal processing of incomplete information.

In the writer's lectures, the emphasis was therefore on the quantitative formulation of Pólya's viewpoint, so it could be used for general problems of scientific inference, almost all of which arise out of incomplete information rather than "randomness." Some personal reminiscences about George Pólya and this start of the work are in Chapter 5.

But once the development of applications started, the work of Harold Jeffreys, who had seen so much of it intuitively and seemed to anticipate every problem I would encounter, became again the central focus of attention. My debt to him is only partially indicated by the dedication of this book to his memory. Further comments about his work and its influence on mine are scattered about in several Chapters.

In the years 1957-1970 the lectures were repeated, with steadily increasing content, at many other Universities and research laboratories.[‡] In this growth it became clear gradually that the outstanding difficulties of conventional "statistical inference" are easily understood and overcome. But the rules which now took their place were quite subtle conceptually, and it required some deep thinking to see how to apply them correctly. Past difficulties which had led to rejection of Laplace's work, were seen finally as only misapplications, arising usually from failure to define the problem unambiguously or to appreciate the cogency of seemingly trivial side information, and easy to correct once this is recognized. The various relations between our "extended logic" approach and the usual "random variable" one appear in almost every Chapter, in many different forms.

Eventually, the material grew to far more than could be presented in a short series of lectures, and the work evolved out of the pedagogical phase; with the clearing up of old difficulties accomplished, we found ourselves in possession of a powerful tool for dealing with new problems. Since about 1970 the accretion has continued at the same pace, but fed instead by the research activity of the writer and his colleagues. We hope that the final result has retained enough of its hybrid origins to be usable either as a textbook or as a reference work; indeed, several generations of students have carried away earlier versions of our notes, and in turn taught it to their students.

In view of the above, we repeat the sentence that Charles Darwin wrote in the Introduction to his *Origin of Species*: "I hope that I may be excused for entering on these personal details, as I give them to show that I have not been hasty in coming to a decision." But it might be thought that work done thirty years ago would be obsolete today. Fortunately, the work of Jeffreys, Pólya and Cox was of a fundamental, timeless character whose truth does not change and whose importance grows with time. Their perception about the nature of inference, which was merely curious thirty years ago, is very important in a half-dozen different areas of science today; and it will be crucially important in all areas 100 years hence.

Foundations: From many years of experience with its applications in hundreds of real problems, our views on the foundations of probability theory have evolved into something quite complex, which cannot be described in any such simplistic terms as "pro-this" or "anti-that." For example, our system of probability could hardly be more different from that of Kolmogorov, in style, philosophy, and purpose. What we consider to be fully half of probability theory as it is needed in current applications—the principles for assigning probabilities by logical analysis of incomplete information—is not present at all in the Kolmogorov system.

Yet when all is said and done we find ourselves, to our own surprise, in agreement with Kolmogorov and in disagreement with his critics, on nearly all technical issues. As noted in Appendix A, each of his axioms turns out to be, for all practical purposes, derivable from the Pólya-Cox

[‡] Some of the material in the early Chapters was issued in 1958 by the Socony-Mobil Oil Company as Number 4 in their series "Colloquium Lectures in Pure and Applied Science."

desiderata of rationality and consistency. In short, we regard our system of probability as not contradicting Kolmogorov's; but rather seeking a deeper logical foundation that permits its extension in the directions that are needed for modern applications. In this endeavor, many problems have been solved, and those still unsolved appear where we should naturally expect them: in breaking into new ground.

As another example, it appears at first glance to everyone that we are in very close agreement with the de Finetti system of probability. Indeed, the writer believed this for some time. Yet when all is said and done we find, to our own surprise, that little more than a loose philosophical agreement remains; on many technical issues we disagree strongly with de Finetti. It appears to us that his way of treating infinite sets has opened up a Pandora's box of useless and unnecessary paradoxes; nonconglomerability and finite additivity are examples discussed in Chapter 15.

Infinite set paradoxing has become a morbid infection that is today spreading in a way that threatens the very life of probability theory, and requires immediate surgical removal. In our system, after this surgery, such paradoxes are avoided automatically; they cannot arise from correct application of our basic rules, because those rules admit only finite sets and infinite sets that arise as well-defined and well-behaved limits of finite sets. The paradoxing was caused by (1) jumping directly into an infinite set without specifying any limiting process to define its properties; and then (2) asking questions whose answers depend on how the limit was approached.

For example, the question: "What is the probability that an integer is even?" can have any answer we please in (0, 1), depending on what limiting process is to define the "set of all integers" (just as a conditionally convergent series can be made to converge to any number we please, depending on the order in which we arrange the terms).

In our view, an infinite set cannot be said to possess any "existence" and mathematical properties at all—at least, in probability theory—until we have specified the limiting process that is to generate it from a finite set. In other words, we sail under the banner of Gauss, Kronecker, and Poincaré rather than Cantor, Hilbert, and Bourbaki. We hope that readers who are shocked by this will study the indictment of Bourbakism by the mathematician Morris Kline (1980), and then bear with us long enough to see the advantages of our approach. Examples appear in almost every Chapter.

Comparisons: For many years there has been controversy over "frequentist" versus "Bayesian" methods of inference, in which the writer has been an outspoken partisan on the Bayesian side. The record of this up to 1981 is given in an earlier book (Jaynes, 1983). In these old works there was a strong tendency, on both sides, to argue on the level of philosophy or ideology. We can now hold ourselves somewhat aloof from this because, thanks to recent work, there is no longer any need to appeal to such arguments. We are now in possession of proven theorems and masses of worked-out numerical examples. As a result, the superiority of Bayesian methods is now a thoroughly demonstrated fact in a hundred different areas. One can argue with a philosophy; it is not so easy to argue with a computer printout, which says to us: "Independently of all your philosophy, here are the facts of actual performance." We point this out in some detail whenever there is a substantial difference in the final results. Thus we continue to argue vigorously for the Bayesian methods; but we ask the reader to note that our arguments now proceed by citing facts rather than proclaiming a philosophical or ideological position.

However, neither the Bayesian nor the frequentist approach is universally applicable, so in the present more general work we take a broader view of things. Our theme is simply: *Probability Theory as Extended Logic*. The "new" perception amounts to the recognition that the mathematical rules of probability theory are not merely rules for calculating frequencies of "random variables";

they are also the unique consistent rules for conducting inference (*i.e.* plausible reasoning) of any kind, and we shall apply them in full generality to that end.

It is true that all "Bayesian" calculations are included automatically as particular cases of our rules; but so are all "frequentist" calculations. Nevertheless, our basic rules are broader than either of these, and in many applications our calculations do not fit into either category.

To explain the situation as we see it presently: The traditional "frequentist" methods which use only sampling distributions are usable and useful in many particularly simple, idealized problems; but they represent the most proscribed special cases of probability theory, because they presuppose conditions (independent repetitions of a "random experiment" but no relevant prior information) that are hardly ever met in real problems. This approach is quite inadequate for the current needs of science.

In addition, frequentist methods provide no technical means to eliminate nuisance parameters or to take prior information into account, no way even to use all the information in the data when sufficient or ancillary statistics do not exist. Lacking the necessary theoretical principles, they force one to "choose a statistic" from intuition rather than from probability theory, and then to invent *ad hoc* devices (such as unbiased estimators, confidence intervals, tail-area significance tests) not contained in the rules of probability theory. Each of these is usable within a small domain for which it was invented but, as Cox's theorems guarantee, such arbitrary devices always generate inconsistencies or absurd results when applied to extreme cases; we shall see dozens of examples.

All of these defects are corrected by use of Bayesian methods, which are adequate for what we might call "well-developed" problems of inference. As Harold Jeffreys demonstrated, they have a superb analytical apparatus, able to deal effortlessly with the technical problems on which frequentist methods fail. They determine the optimal estimators and algorithms automatically while taking into account prior information and making proper allowance for nuisance parameters and, being exact, they do not break down—but continue to yield reasonable results—in extreme cases. Therefore they enable us to solve problems of far greater complexity than can be discussed at all in frequentist terms. One of our main purposes is to show how all this capability was contained already in the simple product and sum rules of probability theory interpreted as extended logic, with no need for—indeed, no room for—any *ad hoc* devices.

But before Bayesian methods can be used, a problem must be developed beyond the "exploratory phase" to the point where it has enough structure to determine all the needed apparatus (a model, sample space, hypothesis space, prior probabilities, sampling distribution). Almost all scientific problems pass through an initial exploratory phase in which we have need for inference, but the frequentist assumptions are invalid and the Bayesian apparatus is not yet available. Indeed, some of them never evolve out of the exploratory phase. Problems at this level call for more primitive means of assigning probabilities directly out of our incomplete information.

For this purpose, the Principle of Maximum Entropy has at present the clearest theoretical justification and is the most highly developed computationally, with an analytical apparatus as powerful and versatile as the Bayesian one. To apply it we must define a sample space, but do not need any model or sampling distribution. In effect, entropy maximization creates a model for us out of our data, which proves to be optimal by so many different criteria* that it is hard to imagine

* These concern efficient information handling; for example, (1) The model created is the simplest one that captures all the information in the constraints (Chapter 11); (2) It is the unique model for which the constraints would have been sufficient statistics (Chapter 8); (3) If viewed as constructing a sampling distribution for subsequent Bayesian inference from new data D, the only property of the measurement

circumstances where one would not want to use it in a problem where we have a sample space but no model.

Bayesian and Maximum Entropy methods differ in another respect. Both procedures yield the optimal inferences from the information that went into them, but we may choose a model for Bayesian analysis; this amounts to expressing some prior knowledge—or some working hypothesis—about the phenomenon being observed. Usually such hypotheses extend beyond what is directly observable in the data, and in that sense we might say that Bayesian methods are—or at least may be—speculative. If the extra hypotheses are true, then we expect that the Bayesian results will improve on maximum entropy; if they are false, the Bayesian inferences will likely be worse.

On the other hand, Maximum Entropy is a nonspeculative procedure, in the sense that it invokes no hypotheses beyond the sample space and the evidence that is in the available data. Thus it predicts only observable facts (functions of future or past observations) rather than values of parameters which may exist only in our imagination. It is just for that reason that Maximum Entropy is the appropriate (safest) tool when we have very little knowledge beyond the raw data; it protects us against drawing conclusions not warranted by the data. But when the information is extremely vague it may be difficult to define any appropriate sample space, and one may wonder whether still more primitive principles than Maximum Entropy can be found. There is room for much new creative thought here.

For the present, there are many important and highly nontrivial applications where Maximum Entropy is the only tool we need. The planned second volume of this work is to consider them in detail; usually, they require more technical knowledge of the subject-matter area than do the more general applications studied in this volume. All of presently known statistical mechanics, for example, is included in this, as are the highly successful Maximum Entropy spectrum analysis and image reconstruction algorithms in current use. However, we think that in the future the latter two applications will evolve on into the Bayesian phase, as we become more aware of the appropriate models and hypothesis spaces, which enable us to incorporate more prior information.

We are conscious of having so many theoretical points to explain, that we fail to present as many practical worked-out numerical examples as we should. Fortunately, three recent books largely make up this deficiency, and so should be considered as adjuncts to the present work. "Bayesian Spectrum Analysis and Parameter Estimation" by G. L. Bretthorst [Springer Lecture Notes in Statistics #48 (1988)] and two works published in the Oxford University Science Publications series: ["Maximum Entropy in Action," ed B. Buck & V. A. Macaulay (1991), and "Data Analysis: A Bayesian Tutorial" by D. S. Sivia (1996)], are written from a viewpoint essentially identical with ours and present a wealth of real problems carried through to numerical solutions. Of course, these works do not contain nearly as much theoretical explanation as does the present one. Also, the Proceedings volumes of the various annual MAXENT workshops since 1981 consider a great variety of useful applications.

Mental Activity: As one would expect already from Pólya's examples, probability theory as extended logic reproduces many aspects of human mental activity, sometimes in surprising and even disturbing detail. In Chapter 5 we find our equations exhibiting the phenomenon of a person

errors in D that are used in that subsequent inference are the ones about which that sampling distribution contained some definite prior information (Chapter 7). Thus the formalism automatically takes into account all the information we have, but avoids assuming information that we do not have. This contrasts sharply with orthodox methods, where one does not think in terms of information at all, and in general violates both of these desiderata.

who tells the truth and is not believed, even though the disbelievers are reasoning consistently. The theory explains why and under what circumstances this will happen.

The equations also reproduce a more complicated phenomenon, divergence of opinions. One might expect that open discussion of public issues would tend to bring about a general consensus. On the contrary, we observe repeatedly that when some controversial issue has been discussed vigorously for a few years, society becomes polarized into two opposite extreme camps; it is almost impossible to find anyone who retains a moderate view. Probability theory as logic shows how two persons, given the same information, may have their opinions driven in opposite directions by it, and what must be done to avoid this.

In such respects, it is clear that probability theory is telling us something about the way our own minds operate when we form intuitive judgments, of which we may not have been consciously aware. Some may feel uncomfortable at these revelations; others may see in them useful tools for psychological, sociological, or legal research.

What is 'safe'? We are not concerned here only with abstract issues of mathematics and logic. One of the main practical messages of this work is the great effect of prior information on the conclusions that one should draw from a given data set. Currently much discussed issues such as environmental hazards or the toxicity of a food additive, cannot be judged rationally if one looks only at the current data and ignores the prior information that scientists have about the phenomenon. This can lead one to greatly overestimate or underestimate the danger.

A common error, when judging the effects of radioactivity or the toxicity of some substance, is to assume a linear response model without threshold (that is, without a dose rate below which there is no ill effect). Presumably there is no threshold effect for cumulative poisons like heavy metal ions (mercury, lead), which are eliminated only very slowly if at all. But for virtually every organic substance (such as saccharin or cyclamates), the existence of a finite metabolic rate means that there must exist a finite threshold dose rate, below which the substance is decomposed, eliminated, or chemically altered so rapidly that it has no ill effects. If this were not true, the human race could never have survived to the present time, in view of all the things we have been eating.

Indeed, every mouthful of food you and I have ever taken contained many billions of kinds of complex molecules whose structure and physiological effects have never been determined—and many millions of which would be toxic or fatal in large doses. We cannot doubt that we are daily ingesting thousands of substances that are far more dangerous than saccharin—but in amounts that are safe, because they are far below the various thresholds of toxicity. But at present there is hardly any substance except some common drugs, for which we actually know the threshold.

Therefore, the goal of inference in this field should be to estimate not only the slope of the response curve, but *far more importantly*, to decide whether there is evidence for a threshold; and if so, to estimate its magnitude (the "maximum safe dose"). For example, to tell us that a sugar substitute can produce a barely detectable incidence of cancer in doses a thousand times greater than would ever be encountered in practice, is hardly an argument against using the substitute; indeed, the fact that it is necessary to go to kilodoses in order to detect any ill effects at all, is rather conclusive evidence, not of the danger, but of the *safety*, of a tested substance. A similar overdose of sugar would be far more dangerous, leading not to barely detectable harmful effects, but to sure, immediate death by diabetic coma; yet nobody has proposed to ban the use of sugar in food.

Kilodose effects are irrelevant because we do not take kilodoses; in the case of a sugar substitute the important question is: *What are the threshold doses for toxicity of a sugar substitute and for* *sugar, compared to the normal doses*? If that of a sugar substitute is higher, then the rational conclusion would be that the substitute is actually safer than sugar, as a food ingredient. To analyze one's data in terms of a model which does not allow even the possibility of a threshold effect, is to prejudge the issue in a way that can lead to false conclusions however good the data. If we hope to detect any phenomenon, we must use a model that at least allows the *possibility* that it may exist.

We emphasize this in the Preface because false conclusions of just this kind are now not only causing major economic waste, but also creating unnecessary dangers to public health and safety. Society has only finite resources to deal with such problems, so any effort expended on imaginary dangers means that real dangers are going unattended. Even worse, the error is incorrectible by the currently most used data analysis procedures; a false premise built into a model which is never questioned, cannot be removed by any amount of new data. Use of models which correctly represent the prior information that scientists have about the mechanism at work can prevent such folly in the future.

But such considerations are not the only reasons why prior information is essential in inference; the progress of science itself is at stake. To see this, note a corollary to the last paragraph; that new data that we insist on analyzing in terms of old ideas (that is, old models which are not questioned) *cannot lead us out of the old ideas*. However many data we record and analyze, we may just keep repeating the same old errors, and missing the same crucially important things that the experiment was competent to find. That is what ignoring prior information can do to us; no amount of analyzing coin tossing data by a stochastic model could have led us to discovery of Newtonian mechanics, which alone determines those data.

But old data, when seen in the light of new ideas, can give us an entirely new insight into a phenomenon; we have an impressive recent example of this in the Bayesian spectrum analysis of nuclear magnetic resonance data, which enables us to make accurate quantitative determinations of phenomena which were not accessible to observation at all with the previously used data analysis by Fourier transforms. When a data set is mutilated (or, to use the common euphemism, 'filtered') by processing according to false assumptions, important information in it may be destroyed irreversibly. As some have recognized, this is happening constantly from orthodox methods of detrending or seasonal adjustment in Econometrics. But old data sets, if preserved unmutilated by old assumptions, may have a new lease on life when our prior information advances.

Style of Presentation: In Volume 1, expounding principles and elementary applications, most Chapters start with several pages of verbal discussion of the nature of the problem. Here we try to explain the constructive ways of looking at it, and the logical pitfalls responsible for past errors. Only then do we turn to the mathematics, solving a few of the problems of the genre to the point where the reader may carry it on by straightforward mathematical generalization. In Volume 2, expounding more advanced applications, we can concentrate from the start on the mathematics.

The writer has learned from much experience that this primary emphasis on the logic of the problem, rather than the mathematics, is necessary in the early stages. For modern students, the mathematics is the easy part; once a problem has been reduced to a definite mathematical exercise, most students can solve it effortlessly and extend it endlessly, without further help from any book or teacher. It is in the conceptual matters (how to make the initial connection between the real-world problem and the abstract mathematics) that they are perplexed and unsure how to proceed.

Recent history demonstrates that anyone foolhardy enough to describe his own work as "rigorous" is headed for a fall. Therefore, we shall claim only that we do not knowingly give erroneous arguments. We are conscious also of writing for a large and varied audience, for most of whom clarity of meaning is more important than "rigor" in the narrow mathematical sense.

There are two more, even stronger reasons for placing our primary emphasis on logic and clarity. Firstly, no argument is stronger than the premises that go into it, and as Harold Jeffreys noted, those who lay the greatest stress on mathematical rigor are just the ones who, lacking a sure sense of the real world, tie their arguments to unrealistic premises and thus destroy their relevance. Jeffreys likened this to trying to strengthen a building by anchoring steel beams into plaster. An argument which makes it clear intuitively *why* a result is correct, is actually more trustworthy and more likely of a permanent place in science, than is one that makes a great overt show of mathematical rigor unaccompanied by understanding.

Secondly, we have to recognize that there are no really trustworthy standards of rigor in a mathematics that has embraced the theory of infinite sets. Morris Kline (1980, p. 351) came close to the Jeffreys simile: "Should one design a bridge using theory involving infinite sets or the axiom of choice? Might not the bridge collapse?" The only real rigor we have today is in the operations of elementary arithmetic on finite sets of finite integers, and our own bridge will be safest from collapse if we keep this in mind.

Of course, it is essential that we follow this "finite sets" policy whenever it matters for our results; but we do not propose to become fanatics about it. In particular, the arts of computation and approximation are on a different level than that of basic principle; and so once a result is derived from strict application of the rules, we allow ourselves to use any convenient analytical methods for evaluation or approximation (such as replacing a sum by an integral) without feeling obliged to show how to generate an uncountable set as the limit of a finite one.

But we impose on ourselves a far stricter adherence to the mathematical rules of probability theory than was ever exhibited in the "orthodox" statistical literature, in which authors repeatedly invoke the aforementioned intuitive *ad hoc* devices to do, arbitrarily and imperfectly, what the rules of probability theory would have done for them uniquely and optimally. It is just this strict adherence that enables us to avoid the artificial paradoxes and contradictions of orthodox statistics, as described in Chapters 15 and 17.

Equally important, this policy often simplifies the computations in two ways: (A) The problem of determining the sampling distribution of a "statistic" is eliminated; the evidence of the data is displayed fully in the likelihood function, which can be written down immediately. (B) One can eliminate nuisance parameters at the beginning of a calculation, thus reducing the dimensionality of a search algorithm. If there are several parameters in a problem, this can mean orders of magnitude reduction in computation over what would be needed with a least squares or maximum likelihood algorithm. The Bayesian computer programs of Bretthorst (1988) demonstrate these advantages impressively, leading in some cases to major improvements in the ability to extract information from data, over previously used methods. But this has barely scratched the surface of what can be done with sophisticated Bayesian models. We expect a great proliferation of this field in the near future.

A scientist who has learned how to use probability theory directly as extended logic, has a great advantage in power and versatility over one who has learned only a collection of unrelated *ad hoc* devices. As the complexity of our problems increases, so does this relative advantage. Therefore we think that in the future, workers in all the quantitative sciences will be obliged, as a matter of practical necessity, to use probability theory in the manner expounded here. This trend is already well under way in several fields, ranging from econometrics to astronomy to magnetic resonance spectroscopy; but to make progress in a new area it is necessary to develop a healthy disrespect for tradition and authority, which have retarded progress throughout the 20'th century.

Finally, some readers should be warned not to look for hidden subtleties of meaning which are not present. We shall, of course, explain and use all the standard technical jargon of probability and statistics—because that is our topic. But although our concern with the nature of logical inference leads us to discuss many of the same issues, our language differs greatly from the stilted jargon of logicians and philosophers. There are no linguistic tricks and there is no "meta-language" gobbledygook; only plain English. We think that this will convey our message clearly enough to anyone who seriously wants to understand it. In any event, we feel sure that no further clarity would be achieved by taking the first few steps down that infinite regress that starts with: "What do you mean by 'exists'?"

Acknowledgments: In addition to the inspiration received from the writings of Jeffreys, Cox, Pólya, and Shannon, I have profited by interaction with some 300 former students, who have diligently caught my errors and forced me to think more carefully about many issues. Also, over the years my thinking has been influenced by discussions with many colleagues; to list a few (in the reverse alphabetical order preferred by some): Arnold Zellner, Eugene Wigner, George Uhlenbeck, John Tukey, William Sudderth, Stephen Stigler, Ray Smith, John Skilling, Jimmie Savage, Carlos Rodriguez, Lincoln Moses, Elliott Montroll, Paul Meier, Dennis Lindley, David Lane, Mark Kac, Harold Jeffreys, Bruce Hill, Mike Hardy, Stephen Gull, Tom Grandy, Jack Good, Seymour Geisser, Anthony Garrett, Fritz Fröhner, Willy Feller, Anthony Edwards, Morrie de Groot, Phil Dawid, Jerome Cornfield, John Parker Burg, David Blackwell, and George Barnard. While I have not agreed with all of the great variety of things they told me, it has all been taken into account in one way or another in the following pages. Even when we ended in disagreement on some issue, I believe that our frank private discussions have enabled me to avoid misrepresenting their positions, while clarifying my own thinking; I thank them for their patience.

E. T. Jaynes
July 1996

Chapter 1

PLAUSIBLE REASONING

"The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind." — *James Clerk Maxwell (1850)*

Suppose some dark night a policeman walks down a street, apparently deserted; but suddenly he hears a burglar alarm, looks across the street, and sees a jewelry store with a broken window. Then a gentleman wearing a mask comes crawling out through the broken window, carrying a bag which turns out to be full of expensive jewelry. The policeman doesn't hesitate at all in deciding that this gentleman is dishonest. But by what reasoning process does he arrive at this conclusion? Let us first take a leisurely look at the general nature of such problems.

Deductive and Plausible Reasoning

A moment's thought makes it clear that our policeman's conclusion was not a logical deduction from the evidence; for there may have been a perfectly innocent explanation for everything. It might be, for example, that this gentleman was the owner of the jewelry store and he was coming home from a masquerade party, and didn't have the key with him. But just as he walked by his store a passing truck threw a stone through the window; and he was only protecting his own property.

Now while the policeman's reasoning process was not logical deduction, we will grant that it had a certain degree of validity. The evidence did not make the gentleman's dishonesty *certain*, but it did make it extremely *plausible*. This is an example of a kind of reasoning in which we have all become more or less proficient, necessarily, long before studying mathematical theories. We are hardly able to get through one waking hour without facing some situation (*e.g.* will it rain or won't it?) where we do not have enough information to permit deductive reasoning; but still we must decide immediately what to do.

But in spite of its familiarity, the formation of plausible conclusions is a very subtle process. Although history records discussions of it extending over 24 centuries, probably nobody has ever produced an analysis of the process which anyone else finds completely satisfactory. But in this work we will be able to report some useful and encouraging new progress, in which conflicting intuitive judgments are replaced by definite theorems, and *ad hoc* procedures are replaced by rules that are determined uniquely by some very elementary—and nearly inescapable—criteria of rationality.

All discussions of these questions start by giving examples of the contrast between deductive reasoning and plausible reasoning. As is generally credited to the *Organon* of Aristotle (4'th century B. C.)† deductive reasoning (*apodeixis*) can be analyzed ultimately into the repeated application of two strong syllogisms:

---

† Today, several different views are held about the exact nature of Aristotle's contribution. Such issues are irrelevant to our present purpose, but the interested reader may find an extensive discussion of them in Lukasiewicz (1957).

$$\frac{\begin{array}{c}\text{If } A \text{ is true, then } B \text{ is true} \\ A \text{ is true}\end{array}}{\text{Therefore, } B \text{ is true}} \tag{1–1}$$

and its inverse:

$$\frac{\begin{array}{c}\text{If } A \text{ is true, then } B \text{ is true} \\ B \text{ is false}\end{array}}{\text{Therefore, } A \text{ is false}} \tag{1–2}$$

This is the kind of reasoning we would like to use all the time; but as noted, in almost all the situations confronting us we do not have the right kind of information to allow this kind of reasoning. We fall back on weaker syllogisms (*epagoge*):

$$\frac{\begin{array}{c}\text{If } A \text{ is true, then } B \text{ is true} \\ B \text{ is true}\end{array}}{\text{Therefore, } A \text{ becomes more plausible}} \tag{1–3}$$

The evidence does not prove that A is true, but verification of one of its consequences does give us more confidence in A. For example, let

$A \equiv$ "It will start to rain by 10 AM at the latest."
$B \equiv$ "The sky will become cloudy before 10 AM."

Observing clouds at 9:45 AM does not give us a logical certainty that the rain will follow; nevertheless our common sense, obeying the weak syllogism, may induce us to change our plans and behave *as if* we believed that it will, if those clouds are sufficiently dark.

This example shows also that the major premise, "If A then B" expresses B only as a *logical* consequence of A; and not necessarily a causal physical consequence, which could be effective only at a later time. The rain at 10 AM is not the physical cause of the clouds at 9:45 AM. Nevertheless, the proper logical connection is not in the uncertain causal direction (clouds $\Longrightarrow$ rain), but rather (rain $\Longrightarrow$ clouds) which is certain, although noncausal.

We emphasize at the outset that we are concerned here with *logical* connections, because some discussions and applications of inference have fallen into serious error through failure to see the distinction between logical implication and physical causation. The distinction is analyzed in some depth by H. A. Simon and N. Rescher (1966), who note that all attempts to interpret implication as expressing physical causation founder on the lack of contraposition expressed by the second syllogism (1–2). That is, if we tried to interpret the major premise as "A is the physical cause of B," then we would hardly be able to accept that "not-B is the physical cause of not-A." In Chapter 3 we shall see that attempts to interpret plausible inferences in terms of physical causation fare no better.

Another weak syllogism, still using the same major premise, is

$$\frac{\begin{array}{c}\text{If } A \text{ is true, then } B \text{ is true} \\ A \text{ is false}\end{array}}{\text{Therefore, } B \text{ becomes less plausible}} \tag{1–4}$$

In this case, the evidence does not prove that B is false; but one of the possible reasons for its being true has been eliminated, and so we feel less confident about B. The reasoning of a scientist, by which he accepts or rejects his theories, consists almost entirely of syllogisms of the second and third kind.

Now the reasoning of our policeman was not even of the above types. It is best described by a still weaker syllogism:

$$\frac{\begin{array}{c}\text{If } A \text{ is true, then } B \text{ becomes more plausible}\\ B \text{ is true}\end{array}}{\text{Therefore, } A \text{ becomes more plausible}} \tag{1–5}$$

But in spite of the apparent weakness of this argument, when stated abstractly in terms of A and B, we recognize that the policeman's conclusion has a very strong convincing power. There is something which makes us believe that in this particular case, his argument had almost the power of deductive reasoning.

These examples show that the brain, in doing plausible reasoning, not only decides whether something becomes more plausible or less plausible, but it evaluates the *degree* of plausibility in some way. The plausibility for rain by 10 AM depends very much on the darkness of those clouds. And the brain also makes use of old information as well as the specific new data of the problem; in deciding what to do we try to recall our past experience with clouds and rain, and what the weatherman predicted last night.

To illustrate that the policeman was also making use of the past experience of policemen in general, we have only to change that experience. Suppose that events like these happened several times every night to every policeman—and in every case the gentleman turned out to be completely innocent. Very soon, policemen would learn to ignore such trivial things.

Thus, in our reasoning we depend very much on *prior information* to help us in evaluating the degree of plausibility in a new problem. This reasoning process goes on unconsciously, almost instantaneously, and we conceal how complicated it really is by calling it *common sense*.

The mathematician George Pólya (1945, 1954) wrote three books about plausible reasoning, pointing out a wealth of interesting examples and showing that there are definite rules by which we do plausible reasoning (although in his work they remain in qualitative form). The above weak syllogisms appear in his third volume. The reader is strongly urged to consult Pólya's exposition, which was the original source of many of the ideas underlying the present work. We show below how Pólya's principles may be made quantitative, with resulting useful applications.

Evidently, the deductive reasoning described above has the property that we can go through long chains of reasoning of the type (1–1) and (1–2) and the conclusions have just as much certainty as the premises. With the other kinds of reasoning, (1–3)–(1–5), the reliability of the conclusion changes as we go through several stages. But in their quantitative form we shall find that in many cases our conclusions can still approach the certainty of deductive reasoning (as the example of the policeman leads us to expect). Pólya showed that even a pure mathematician actually uses these weaker forms of reasoning most of the time. Of course, when he publishes a new theorem, he will try very hard to invent an argument which uses only the first kind; but the reasoning process which led him to the theorem in the first place almost always involves one of the weaker forms (based, for example, on following up conjectures suggested by analogies). The same idea is expressed in a remark of S. Banach (quoted by S. Ulam, 1957): "*Good mathematicians see analogies between theorems; great mathematicians see analogies between analogies.*"

As a first orientation, then, let us note some very suggestive analogies to another field—which is itself based, in the last analysis, on plausible reasoning.

Analogies with Physical Theories

In physics, we learn quickly that the world is too complicated for us to analyze it all at once. We can make progress only if we dissect it into little pieces and study them separately. Sometimes, we can invent a mathematical model which reproduces several features of one of these pieces, and whenever this happens we feel that progress has been made. These models are called *physical theories*. As knowledge advances, we are able to invent better and better models, which reproduce more and more features of the real world, more and more accurately. Nobody knows whether there is some natural end to this process, or whether it will go on indefinitely.

In trying to understand common sense, we shall take a similar course. We won't try to understand it all at once, but we shall feel that progress has been made if we are able to construct idealized mathematical models which reproduce a few of its features. We expect that any model we are now able to construct will be replaced by more complete ones in the future, and we do not know whether there is any natural end to this process.

The analogy with physical theories is deeper than a mere analogy of method. Often, the things which are most familiar to us turn out to be the hardest to understand. Phenomena whose very existence is unknown to the vast majority of the human race (such as the difference in ultraviolet spectra of Iron and Nickel) can be explained in exhaustive mathematical detail—but all of modern science is practically helpless when faced with the complications of such a commonplace fact as growth of a blade of grass. Accordingly, we must not expect too much of our models; we must be prepared to find that some of the most familiar features of mental activity may be ones for which we have the greatest difficulty in constructing any adequate model.

There are many more analogies. In physics we are accustomed to finding that any advance in knowledge leads to consequences of great practical value, but of an unpredictable nature. Röntgen's discovery of X-rays led to important new possibilities of medical diagnosis; Maxwell's discovery of one more term in the equation for curl H led to practically instantaneous communication all over the earth.

Our mathematical models for common sense also exhibit this feature of practical usefulness. Any successful model, even though it may reproduce only a few features of common sense, will prove to be a powerful extension of common sense in some field of application. Within this field, it enables us to solve problems of inference which are so involved in complicated detail that we would never attempt to solve them without its help.

The Thinking Computer

Models have practical uses of a quite different type. Many people are fond of saying, "They will never make a machine to replace the human mind—it does many things which no machine could ever do." A beautiful answer to this was given by J. von Neumann in a talk on computers given in Princeton in 1948, which the writer was privileged to attend. In reply to the canonical question from the audience ["But of course, a mere machine can't really *think*, can it?"], he said: "*You insist that there is something a machine cannot do. If you will tell me precisely what it is that a machine cannot do, then I can always make a machine which will do just that*!"

In principle, the only operations which a machine cannot perform for us are those which we cannot describe in detail, or which could not be completed in a finite number of steps. Of course, some will conjure up images of Gödel incompleteness, undecidability, Turing machines which never stop, etc. But to answer all such doubts we need only point to the existence of the human brain, which *does* it. Just as von Neumann indicated, the only real limitations on making "machines which think" are our own limitations in not knowing exactly what "thinking" consists of.

But in our study of common sense we shall be led to some very explicit ideas about the mechanism of thinking. Every time we can construct a mathematical model which reproduces a part of common sense by prescribing a definite set of operations, this shows us how to "build a machine," (*i.e.* write a computer program) which operates on incomplete information and, by applying quantitative versions of the above weak syllogisms, does plausible reasoning instead of deductive reasoning.

Indeed, the development of such computer software for certain specialized problems of inference is one of the most active and useful current trends in this field. One kind of problem thus dealt with might be: given a mass of data, comprising 10,000 separate observations, determine in the light of these data and whatever prior information is at hand, the relative plausibilities of 100 different possible hypotheses about the causes at work.

Our unaided common sense might be adequate for deciding between two hypotheses whose consequences are very different; but for dealing with 100 hypotheses which are not very different, we would be helpless without a computer *and* a well-developed mathematical theory that shows us how to program it. That is, what determines, in the policeman's syllogism (1–5), whether the plausibility for A increases by a large amount, raising it almost to certainty; or only a negligibly small amount, making the data B almost irrelevant? The object of the present work is to develop the mathematical theory which answers such questions, in the greatest depth and generality now possible.

While we expect a mathematical theory to be useful in programming computers, the idea of a thinking computer is also helpful psychologically in developing the mathematical theory. The question of the reasoning process used by actual human brains is charged with emotion and grotesque misunderstandings. It is hardly possible to say anything about this without becoming involved in debates over issues that are not only undecidable in our present state of knowledge, but are irrelevant to our purpose here.

Obviously, the operation of real human brains is so complicated that we can make no pretense of explaining its mysteries; and in any event we are not trying to explain, much less reproduce, all the aberrations and inconsistencies of human brains. That is an interesting and important subject; but it is not the subject we are studying here. Our topic is the *normative principles of logic*; and not the principles of psychology or neurophysiology.

To emphasize this, instead of asking, "How can we build a mathematical model of human common sense?" let us ask, "How could we build a machine which would carry out useful plausible reasoning, following clearly defined principles expressing an idealized common sense?"

Introducing the Robot

In order to direct attention to constructive things and away from controversial irrelevancies, we shall invent an imaginary being. Its brain is to be designed *by us*, so that it reasons according to certain definite rules. These rules will be deduced from simple desiderata which, it appears to us, would be desirable in human brains; *i.e.* we think that a rational person, should he discover that he was violating one of these desiderata, would wish to revise his thinking.

In principle, we are free to adopt any rules we please; that is our way of *defining* which robot we shall study. Comparing its reasoning with yours, if you find no resemblance you are in turn free to reject our robot and design a different one more to your liking. But if you find a very strong resemblance, and decide that you want and trust this robot to help you in your own problems of inference, then that will be an accomplishment of the theory, not a premise.

Our robot is going to reason about propositions. As already indicated above, we shall denote various propositions by italicized capital letters, $\{A, B, C, \text{etc.}\}$, and for the time being we must require that any proposition used must have, to the robot, an unambiguous meaning and must be of the simple, definite logical type that must be either true or false. That is, until otherwise stated we shall be concerned only with two-valued logic, or Aristotelian logic. We do not require that the truth or falsity of such an "Aristotelian proposition" be ascertainable by any feasible investigation; indeed, our inability to do this is usually just the reason why we need the robot's help. For example, the writer personally considers both of the following propositions to be true:

$A \equiv$ "Beethoven and Berlioz never met."

$B \equiv$ "Beethoven's music has a better sustained quality than that of Berlioz, although Berlioz at his best is the equal of anybody."

But proposition B is not a permissible one for our robot to think about at present, while proposition A is, although it is unlikely that its truth or falsity could be definitely established today.‡ After our theory is developed, it will be of interest to see whether the present restriction to Aristotelian propositions such as A can be relaxed, so that the robot might help us also with more vague propositions like B (see Chapter 18 on the A_p-distribution).*

Boolean Algebra

To state these ideas more formally, we introduce some notation of the usual symbolic logic, or Boolean algebra, so called because George Boole (1854) introduced a *notation* similar to the following. Of course, the principles of deductive logic itself were well understood centuries before Boole, and as we shall see presently, all the results that follow from Boolean algebra were contained already as special cases in the rules of plausible inference given by Laplace (1812). The symbol

$$AB, \tag{1–6}$$

called the *logical product* or the *conjunction*, denotes the proposition "both A and B are true." Obviously, the order in which we state them does not matter; AB and BA say the same thing. The expression

$$A + B, \tag{1–7}$$

called the *logical sum* or *disjunction*, stands for "at least one of the propositions A, B is true" and has the same meaning as $B + A$. These symbols are only a shorthand way of writing propositions; and do not stand for numerical values.

‡ Their meeting is a chronological possibility, since their lives overlapped by 24 years; my reason for doubting it is the failure of Berlioz to mention any such meeting in his memoirs—on the other hand, neither does he come out and say definitely that they did *not* meet.

* The question how one is to make a machine in some sense "cognizant" of the conceptual meaning that a proposition like A has to humans, might seem very difficult, and much of Artificial Intelligence is devoted to inventing *ad hoc* devices to deal with this problem. However, we shall find in Chapter 4 that for us the problem is almost nonexistent; our rules for plausible reasoning automatically provide the means to do the mathematical equivalent of this.

Given two propositions A, B, it may happen that one is true if and only if the other is true; we then say that they have the same *truth value*. This may be only a simple tautology (*i.e.* A and B are verbal statements which obviously say the same thing), or it may be that only after immense mathematical labor is it finally proved that A is the necessary and sufficient condition for B. From the standpoint of logic it does not matter; once it is established, by any means, that A and B have the same truth value, then they are logically equivalent propositions, in the sense that any evidence concerning the truth of one pertains equally well to the truth of the other, and they have the same implications for any further reasoning.

Evidently, then, it must be the most primitive axiom of plausible reasoning that two propositions with the same truth-value are equally plausible. This might appear almost too trivial to mention, were it not for the fact that Boole himself (*loc. cit.*, p. 286) fell into error on this point, by mistakenly identifying two propositions which were in fact different—and then failing to see any contradiction in their different plausibilities. Three years later (Boole, 1857) he gave a revised theory which supersedes that in his book; for further comments on this incident, see Keynes (1921), pp. 167-168; Jaynes (1976), pp. 240-242.

In Boolean algebra, the equal sign is used to denote, not equal numerical value, but equal truth-value: $A = B$, and the "equations" of Boolean algebra thus consist of assertions that the proposition on the left-hand side has the same truth-value as the one on the right-hand side. The symbol "$\equiv$" means, as usual, "equals by definition."

In denoting complicated propositions we use parentheses in the same way as in ordinary algebra, to indicate the order in which propositions are to be combined (at times we shall use them also merely for clarity of expression although they are not strictly necessary). In their absence we observe the rules of algebraic hierarchy, familiar to those who use hand calculators: thus $AB + C$ denotes $(AB) + C$; and not $A(B + C)$.

The *denial* of a proposition is indicated by a bar:

$$\overline{A} \equiv \text{“}A \textit{ is false.”} \tag{1–8}$$

The relation between A, $\overline{A}$ is a reciprocal one:

$$A = \text{“}\overline{A} \textit{ is false,”} \tag{1–9}$$

and it does not matter which proposition we denote by the barred, which by the unbarred, letter. Note that some care is needed in the unambiguous use of the bar. For example, according to the above conventions,

$$\overline{AB} = \text{“}AB \textit{ is false.”} \tag{1–10}$$

$$\overline{A}\,\overline{B} = \text{“}\textit{Both } A \textit{ and } B \textit{ are false.”} \tag{1–11}$$

These are quite different propositions; in fact, $\overline{AB}$ is not the logical product $\overline{A}\,\overline{B}$, but the logical sum: $\overline{AB} = \overline{A} + \overline{B}$.

With these understandings, Boolean algebra is characterized by some rather trivial and obvious basic identities, which express the properties of:

$$\text{Idempotence:} \quad \begin{cases} AA = A \\ A + A = A \end{cases}$$

$$\text{Commutativity:} \quad \begin{cases} AB = BA \\ A + B = B + A \end{cases}$$

$$\text{Associativity:} \quad \begin{cases} A(BC) = (AB)C = ABC \\ A + (B + C) = (A + B) + C = A + B + C \end{cases} \tag{1–12}$$

$$\text{Distributivity:} \quad \begin{cases} A(B + C) = AB + AC \\ A + (BC) = (A + B)(A + C) \end{cases}$$

$$\text{Duality:} \quad \begin{cases} \text{If } C = AB, \text{ then } \overline{C} = \overline{A} + \overline{B} \\ \text{If } D = A + B, \text{ then } \overline{D} = \overline{A}\,\overline{B} \end{cases}$$

but by their application one can prove any number of further relations, some highly nontrivial. For example, we shall presently have use for the rather elementary theorem:

$$\text{If } \overline{B} = AD \text{ then } A\overline{B} = \overline{B} \text{ and } B\,\overline{A} = \overline{A}. \tag{1–13}$$

Implication. The proposition

$$A \Rightarrow B \tag{1–14}$$

to be read: "A implies B," does not assert that either A or B is true; it means only that $A\,\overline{B}$ is false, or what is the same thing, $(\overline{A} + B)$ is true. This can be written also as the logical equation $A = AB$. That is, given (1–14), if A is true then B must be true; or, if B is false then A must be false. This is just what is stated in the strong syllogisms (1–1) and (1–2).

On the other hand, if A is false, (1–14) says nothing about B: and if B is true, (1–14) says nothing about A. But these are just the cases in which our weak syllogisms (1–3), (1–4) do say something. In one respect, then, the term "weak syllogism" is misleading. The theory of plausible reasoning based on them is not a "weakened" form of logic; it is an *extension* of logic with new content not present at all in conventional deductive logic. It will become clear in the next Chapter [Eqs. (2–69), (2–70)] that our rules include deductive logic as a special case.

A Tricky Point: Note carefully that in ordinary language one would take "A implies B" to mean that B is logically deducible from A. But in formal logic, "A implies B" means only that the propositions A and AB have the same truth value. In general, whether B is logically deducible from A does not depend only on the propositions A and B; it depends on the totality of propositions $(A, A', A'', \cdots)$ that we accept as true and which are therefore available to use in the deduction. Devinatz (1968, p. 3) and Hamilton (1988, p. 5) give the truth table for the implication as a binary operation, illustrating that $A \Rightarrow B$ is false only if A is true and B is false; in all other cases $A \Rightarrow B$ is true!

This may seem startling at first glance; but note that indeed, if A and B are both true, then $A = AB$ and so $A \Rightarrow B$ is true; in formal logic every true statement implies every other true statement. On the other hand, if A is false, then AQ is also false for all Q, thus $A = AB$ and $A = A\overline{B}$ are both true, so $A \Rightarrow B$ and $A \Rightarrow \overline{B}$ are both true; a false proposition implies all propositions. If we tried to interpret this as logical deducibility, (*i.e.* both B and $\overline{B}$ are deducible from A), it would follow that every false proposition is logically contradictory. Yet the proposition: "Beethoven outlived Berlioz" is false but hardly logically contradictory (for Beethoven did outlive many people who were the same age as Berlioz).

Obviously, merely knowing that propositions A and B are both true does not provide enough information to decide whether either is logically deducible from the other, plus some unspecified "toolbox" of other propositions. The question of logical deducibility of one proposition from a set of others arises in a crucial way in the Gödel theorem discussed at the end of Chapter 2. This great difference in the meaning of the word "implies" in ordinary language and in formal logic is a tricky point that can lead to serious error if it is not properly understood; it appears to us that "implication" is an unfortunate choice of word and this is not sufficiently emphasized in conventional expositions of logic.

Adequate Sets of Operations

We note some features of deductive logic which will be needed in the design of our robot. We have defined four operations, or "connectives," by which, starting from two propositions A, B, other propositions may be defined: the logical product, or conjunction AB, the logical sum or disjunction $A + B$, the implication $A \Rightarrow B$, and the negation $\overline{A}$. By combining these operations repeatedly in every possible way, one can generate any number of new propositions, such as

$$C \equiv (A + \overline{B})(\overline{A} + A\,\overline{B}) + \overline{A}\,B(A + B). \qquad (1\text{–}15)$$

Many questions then occur to us: How large is the class of new propositions thus generated? Is it infinite, or is there a finite set that is closed under these operations? Can every proposition defined from A, B, be thus represented, or does this require further connectives beyond the above four? Or are these four already overcomplete so that some might be dispensed with? What is the smallest set of operations that is adequate to generate all such "logic functions" of A and B? If instead of two starting propositions A, B we have an arbitrary number $\{A_1, \ldots, A_n\}$, is this set of operations still adequate to generate all possible logic functions of $\{A_1, \ldots, A_n\}$?

All these questions are answered easily, with results useful for logic, probability theory, and computer design. Broadly speaking, we are asking whether, starting from our present vantage point, we can (1) increase the number of functions, (2) decrease the number of operations. The first query is simplified by noting that two propositions, although they may appear entirely different when written out in the manner (1–15), are not different propositions from the standpoint of logic if they have the same truth value. For example, it is left for the reader to verify that C in (1–15) is logically the same statement as the implication $C = (B \Rightarrow \overline{A})$.

Since we are, at this stage, restricting our attention to Aristotelian propositions, any logic function $C = f(A, B)$ such as (1–15) has only two possible "values," true and false; and likewise the "independent variables" A and B can take on only those two values.

At this point a logician might object to our notation, saying that the symbol A has been defined as standing for some fixed proposition, whose truth cannot change; so if we wish to consider logic functions, then instead of writing $C = f(A, B)$ we should introduce new symbols and write

$z = f(x, y)$ where x, y, z are "statement variables" for which various specific statements A, B, C may be substituted. But if A stands for some fixed but unspecified proposition, then it can still be either true or false. We achieve the same flexibility merely by the understanding that equations like (1–15) which define logic functions are to be true for all ways of defining A, B; *i.e.* instead of a statement variable we use a variable statement.

In relations of the form $C = f(A, B)$, we are concerned with logic functions defined on a discrete "space" S consisting of only $2^2 = 4$ points; namely those at which A and B take on the "values" $\{\mathrm{TT}, \mathrm{TF}, \mathrm{FT}, \mathrm{FF}\}$ respectively; and at each point the function $f(A, B)$ can take on independently either of two values $\{\mathrm{T}, \mathrm{F}\}$. There are, therefore, exactly $2^4 = 16$ different logic functions $f(A, B)$; and no more. An expression $B = f(A_1, \ldots, A_n)$ involving n propositions is a logic function on a space S of $M = 2^n$ points; and there are exactly 2^M such functions.

In the case $n = 1$, there are four logic functions $\{f_1(A), \ldots, f_4(A)\}$, which we can define by enumeration: listing all their possible values in a truth-table:

A	T	F
$f_1(A)$	T	T
$f_2(A)$	T	F
$f_3(A)$	F	T
$f_4(A)$	F	F

But it is obvious by inspection that these are just:

$$\begin{aligned} f_1(A) &= A + \overline{A} \\ f_2(A) &= A \\ f_3(A) &= \overline{A} \\ f_4(A) &= A\,\overline{A} \end{aligned} \tag{1–16}$$

so we prove by enumeration that the three operations: conjunction, disjunction, and negation are adequate to generate all logic functions of a single proposition.

For the case of general n, consider first the special functions each of which is true at one and only one point of S. For $n = 2$ there are $2^n = 4$ such functions:

$A,\ B$	TT	TF	FT	FF
$f_1(A, B)$	T	F	F	F
$f_2(A, B)$	F	T	F	F
$f_3(A, B)$	F	F	T	F
$f_4(A, B)$	F	F	F	T

It is clear by inspection that these are just the four basic conjunctions:

$$\begin{aligned} f_1(A, B) &= A\,B \\ f_2(A, B) &= A\,\overline{B} \\ f_3(A, B) &= \overline{A}\,B \\ f_4(A, B) &= \overline{A}\,\overline{B}. \end{aligned} \tag{1–17}$$

Consider now any logic function which is true on certain specified points of S; for example, $f_5(A,B)$ and $f_6(A,B)$ defined by

$A,\ B$	TT	TF	FT	FF
$f_5(A,B)$	F	T	F	T
$f_6(A,B)$	T	F	T	T

We assert that each of these functions is the logical sum of the conjunctions (1–17) that are true on the same points (this is not trivial; the reader should verify it in detail); thus

$$\begin{aligned} f_5(A,B) &= f_2(A,B) + f_4(A,B) \\ &= A\,\overline{B} + \overline{A}\,\overline{B} \\ &= (A+\overline{A})\,\overline{B} \\ &= \overline{B} \end{aligned} \tag{1–18}$$

and likewise,

$$\begin{aligned} f_6(A,B) &= f_1(A,B) + f_3(A,B) + f_4(A,B) \\ &= AB + \overline{A}\,B + \overline{A}\,\overline{B} \\ &= B + \overline{A}\,\overline{B} \\ &= \overline{A} + B. \end{aligned} \tag{1–19}$$

That is, $f_6(A,B)$ is the implication $f_6(A,B) = (A \Rightarrow B)$, with the truth table discussed above. Any logic function $f(A,B)$ that is true on at least one point of S can be constructed in this way as a logical sum of the basic conjunctions (1–17). There are $2^4 - 1 = 15$ such functions. For the remaining function, which is always false, it suffices to take the contradiction, $f_{16}(A,B) \equiv A\,\overline{A}$.

This method (called "reduction to *disjunctive normal form*" in logic textbooks) will work for any n. For example, in the case $n = 5$ there are $2^5 = 32$ basic conjunctions

$$\{ABCDE,\ ABCD\overline{E},\ ABC\overline{D}E, \ldots, \overline{A}\,\overline{B}\,\overline{C}\,\overline{D}\,\overline{E}\} \tag{1–20}$$

and $2^{32} = 4,294,967,296$ different logic functions $f_i(A,B,C,D,E)$; of which $4,294,967,295$ can be written as logical sums of the basic conjunctions, leaving only the contradiction

$$f_{4294967296}(A,B,C,D,E) = A\,\overline{A}. \tag{1–21}$$

Thus one can verify by "construction in thought" that the three operations

$$\{\text{ conjunction, disjunction, negation }\}; \qquad i.e. \qquad \{\text{ AND, OR, NOT }\} \tag{1–22}$$

suffice to generate all possible logic functions; or more concisely, they form an *adequate set*.

But the duality property (1–12) shows that a smaller set will suffice; for disjunction of $A,\ B$ is the same as denying that they are both false:

$$A + B = \overline{(\overline{A}\,\overline{B})}. \tag{1–23}$$

Therefore, the two operations (AND, NOT) already constitute an adequate set for deductive logic.‡ This fact will be essential in determining when we have an adequate set of rules for plausible reasoning, in the next Chapter.

It is clear that we cannot now strike out either of these operations, leaving only the other; *i.e.* the operation "AND" cannot be reduced to negations; and negation cannot be accomplished by any number of "AND" operations. But this still leaves open the possibility that both conjunction and negation might be reducible to some third operation, not yet introduced; so that a single logic operation would constitute an adequate set.

It comes as a pleasant surprise to find that there is not only one, but two such operations. The operation "NAND" is defined as the negation of "AND":

$$A \uparrow B \equiv \overline{AB} = \overline{A} + \overline{B} \tag{1–24}$$

which we can read as "A NAND B." But then we have at once:

$$\begin{aligned} \overline{A} &= A \uparrow A \\ AB &= (A \uparrow B) \uparrow (A \uparrow B) \\ A + B &= (A \uparrow A) \uparrow (B \uparrow B). \end{aligned} \tag{1–25}$$

Therefore, every logic function can be constructed with NAND alone. Likewise, the operation NOR defined by

$$A \downarrow B \equiv \overline{A + B} = \overline{A}\ \overline{B} \tag{1–26}$$

is also powerful enough to generate all logic functions:

$$\begin{aligned} \overline{A} &= A \downarrow A \\ A + B &= (A \downarrow B) \downarrow (A \downarrow B) \\ AB &= (A \downarrow A) \downarrow (B \downarrow B). \end{aligned} \tag{1–27}$$

One can take advantage of this in designing computer and logic circuits. A "logic gate" is a circuit having, besides a common ground, two input terminals and one output. The voltage relative to ground at any of these terminals can take on only two values; say +3 volts, or "up" representing "true"; and zero volts or "down," representing "false." A NAND gate is thus one whose output is up if and only if at least one of the inputs is down; or what is the same thing, down if and only if both inputs are up; while for a NOR gate the output is up if and only if both inputs are down.

One of the standard components of logic circuits is the "quad NAND gate," an integrated circuit containing four independent NAND gates on one semiconductor chip. Given a sufficient number of these and no other circuit components, it is possible to generate any required logic function by interconnecting them in various ways.

This short excursion into deductive logic is as far as we need go for our purposes. Further developments are given in many textbooks; for example, a modern treatment of Aristotelian logic is given by I. M. Copi (1994). For non-Aristotelian forms with special emphasis on Gödel incompleteness, computability, decidability, Turing machines, etc., see A. G. Hamilton (1988).

‡ For you to ponder: does it follow that these two commands are the only ones needed to write any computer program?

We turn now to our extension of logic, which is to follow from the conditions discussed next. We call them "desiderata" rather than "axioms" because they do not assert that anything is "true" but only state what appear to be desirable goals. Whether these goals are attainable without contradictions and whether they determine any unique extension of logic, are matters of mathematical analysis, given in Chapter 2.

The Basic Desiderata

To each proposition about which it reasons, our robot must assign some degree of plausibility, based on the evidence we have given it; and whenever it receives new evidence it must revise these assignments to take that new evidence into account. In order that these plausibility assignments can be stored and modified in the circuits of its brain, they must be associated with some definite physical quantity, such as voltage or pulse duration or a binary coded number, etc.—however our engineers want to design the details. For present purposes this means that there will have to be some kind of association between degrees of plausibility and real numbers:

(I) *Degrees of Plausibility are represented by real numbers.* (1–28)

Desideratum (I) is practically forced on us by the requirement that the robot's brain must operate by the carrying out of some definite physical process. However, it will appear (Appendix A) that it is also required theoretically; we do not see the possibility of any consistent theory without a property that is equivalent functionally to Desideratum (I).

We adopt a natural but nonessential convention; that a greater plausibility shall correspond to a greater number. It will be convenient to assume also a continuity property, which is hard to state precisely at this stage; but to say it intuitively: an infinitesimally greater plausibility ought to correspond only to an infinitesimally greater number.

The plausibility that the robot assigns to some proposition A will, in general, depend on whether we told it that some other proposition B is true. Following the notation of Keynes (1921) and Cox (1961), we indicate this by the symbol

$$A|B \tag{1–29}$$

which we may call "the conditional plausibility that A is true, given that B is true" or just, "A given B." It stands for some real number. Thus, for example,

$$A|BC \tag{1–30}$$

(which we may read: "A given BC") represents the plausibility that A is true, given that both B and C are true. Or,

$$A + B|CD \tag{1–31}$$

represents the plausibility that at least one of the propositions A and B is true, given that both C and D are true; and so on. We have decided to represent a greater plausibility by a greater number, so

$$(A|B) > (C|B) \tag{1–32}$$

says that, given B, A is more plausible than C. In this notation, while the symbol for plausibility is just of the form $A|B$ without parentheses, we often add parentheses for clarity of expression. Thus (1–32) says the same thing as

$$A|B \ > \ C|B \tag{1–33}$$

but its meaning is clearer to the eye.

In the interest of avoiding impossible problems, we are not going to ask our robot to undergo the agony of reasoning from impossible or mutually contradictory premises; there could be no "correct" answer. Thus, we make no attempt to define $A|BC$ when B and C are mutually contradictory. Whenever such a symbol appears, it is understood that B and C are compatible propositions.

Also, we do not want this robot to think in a way that is directly opposed to the way you and I think. So we shall design it to reason in a way that is at least *qualitatively* like the way humans try to reason, as described by the above weak syllogisms and a number of other similar ones.

Thus, if it has old information C which gets updated to C' in such a way that the plausibility for A is increased:

$$(A|C') \; > \; (A|C) \tag{1–34}$$

but the plausibility for B given A is not changed:

$$(B|AC') \; = \; (B|AC). \tag{1–35}$$

This can, of course, produce only an increase, never a decrease, in the plausibility that both A and B are true:

$$(AB|C') \; \geq \; (AB|C) \tag{1–36}$$

and it must produce a decrease in the plausibility that A is false:

$$(\overline{A}|C') \; < \; (\overline{A}|C). \tag{1–37}$$

This qualitative requirement simply gives the "sense of direction" in which the robot's reasoning is to go; it says nothing about *how much* the plausibilities change, except that our continuity assumption (which is also a condition for qualitative correspondence with common sense) now requires that if $A|C$ changes only infinitesimally, it can induce only an infinitesimal change in $AB|C$ and $\overline{A}|C$. The specific ways in which we use these qualitative requirements will be given in the next Chapter, at the point where it is seen why we need them. For the present we summarize them simply as:

(II) *Qualitative Correspondence with common sense.* (1–38)

Finally, we want to give our robot another desirable property for which honest people strive without always attaining; that it always reasons *consistently*. By this we mean just the three common colloquial meanings of the word "consistent":

(IIIa) $\left\{ \text{\textit{If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.}} \right\}$ (1–39a)

(IIIb) $\left\{ \text{\textit{The robot always takes into account all of the evidence it has relevant to a question. It does not arbitrarily ignore some of the information, basing its conclusions only on what remains. In other words, the robot is completely non-ideological.}} \right\}$ (1–39b)

(IIIc) $\left\{ \text{\textit{The robot always represents equivalent states of knowledge by equivalent plausibility assignments. That is, if in two problems the robot's state of knowledge is the same (except perhaps for the labeling of the propositions), then it must assign the same plausibilities in both.}} \right\}$ (1–39c)

Desiderata (I), (II), (IIIa) are the basic "structural" requirements on the inner workings of our robot's brain, while (IIIb), (IIIc) are "interface" conditions which show how the robot's behavior should relate to the outer world.

At this point, most students are surprised to learn that our search for desiderata is at an end. The above conditions, it turns out, uniquely determine the rules by which our robot must reason; *i.e.* there is only one set of mathematical operations for manipulating plausibilities which has all these properties. These rules are deduced in the next Chapter.

[At the end of most Chapters, we insert a Section of informal Comments in which are collected various side remarks, background material, etc. The reader may skip them without losing the main thread of the argument.]

COMMENTS

As politicians, advertisers, salesmen, and propagandists for various political, economic, moral, religious, psychic, environmental, dietary, and artistic doctrinaire positions know only too well, fallible human minds are easily tricked, by clever verbiage, into committing violations of the above desiderata. We shall try to ensure that they do not succeed with our robot.

We emphasize another contrast between the robot and a human brain. By Desideratum I, the robot's mental state about any proposition is to be represented by a real number. Now it is clear that our attitude toward any given proposition may have more than one "coordinate." You and I form simultaneous judgments not only as to whether it is plausible, but also whether it is desirable, whether it is important, whether it is useful, whether it is interesting, whether it is amusing, whether it is morally right, etc. If we assume that each of these judgments might be represented by a number, then a fully adequate description of a human state of mind would be represented by a vector in a space of a rather large number of dimensions.

Not all propositions require this. For example, the proposition "The refractive index of water is less than 1.3" generates no emotions; consequently the state of mind which it produces has very few coordinates. On the other hand, the proposition, "Your mother-in-law just wrecked your new car" generates a state of mind with many coordinates. Quite generally, the situations of everyday life are those involving many coordinates. It is just for this reason, we suggest, that the most familiar examples of mental activity are often the most difficult to reproduce by a model. Perhaps we have here the reason why science and mathematics are the most successful of human activities; they deal with propositions which produce the simplest of all mental states. Such states would be the ones least perturbed by a given amount of imperfection in the human mind.

Of course, for many purposes we would not want our robot to adopt any of these more "human" features arising from the other coordinates. It is just the fact that computers do *not* get confused by emotional factors, do *not* get bored with a lengthy problem, do *not* pursue hidden motives opposed to ours, that makes them safer agents than men for carrying out certain tasks.

These remarks are interjected to point out that there is a large unexplored area of possible generalizations and extensions of the theory to be developed here; perhaps this may inspire others to try their hand at developing "multidimensional theories" of mental activity, which would more and more resemble the behavior of actual human brains—not all of which is undesirable. Such a theory, if successful, might have an importance beyond our present ability to imagine.*

---

* Indeed, some psychologists think that as few as five dimensions might suffice to characterize a human personality; that is, that we all differ only in having different mixes of five basic personality traits which may

For the present, however, we shall have to be content with a much more modest undertaking. Is it possible to develop a consistent "one-dimensional" model of plausible reasoning? Evidently, our problem will be simplest if we can manage to represent a degree of plausibility uniquely by a single real number, and ignore the other "coordinates" just mentioned.

We stress that we are in no way asserting that degrees of plausibility in actual human minds have a unique numerical measure. Our job is not to postulate—or indeed to conjecture about—any such thing; it is to *investigate* whether it is possible, in our robot, to set up such a correspondence without contradictions.

But to some it may appear that we have already assumed more than is necessary, thereby putting gratuitous restrictions on the generality of our theory. Why must we represent degrees of plausibility by real numbers? Would not a "comparative" theory based on a system of qualitative ordering relations like $(A|C) > (B|C)$ suffice? This point is discussed further in Appendix A, where we describe other approaches to probability theory and note that some attempts have been made to develop comparative theories which it was thought would be logically simpler, or more general. But this turned out not to be the case; so although it is quite possible to develop the foundations in other ways than ours, the final results will not be different.

Common Language vs. Formal Logic

We should note the distinction between the statements of formal logic and those of ordinary language. It might be thought that the latter is only a less precise form of expression; but on examination of details the relation appears different. It appears to us that ordinary language, carefully used, need not be less precise than formal logic; but ordinary language is more complicated in its rules and has consequently richer possibilities of expression than we allow ourselves in formal logic.

In particular, common language, being in constant use for other purposes than logic, has developed subtle nuances—means of implying something without actually stating it—that are lost on formal logic. Mr. A, to affirm his objectivity, says, "I believe what I see." Mr. B retorts: "He doesn't see what he doesn't believe." From the standpoint of formal logic, it appears that they have said the same thing; yet from the standpoint of common language, those statements had the intent and effect of conveying opposite meanings.

Here is a less trivial example, taken from a mathematics textbook. Let L be a straight line in a plane, and S an infinite set of points in that plane, each of which is projected onto L. Now consider the statements:

(I) The projection of the limit is the limit of the projections.

(II) The limit of the projections is the projection of the limit.

These have the grammatical structures: "A is B" and "B is A," and so they might appear logically equivalent. Yet in that textbook, (I) was held to be true, and (II) not true in general, on the grounds that the limit of the projections may exist when the limit of the set does not.

As we see from this, in common language—even in mathematics textbooks—we have learned to read subtle nuances of meaning into the exact phrasing, probably without realizing it until an example like this is pointed out. We interpret "A is B" as asserting first of all, as a kind of

be genetically determined. But it seems to us that this must be grossly oversimplified; identifiable chemical factors continuously varying in both space and time (such as the distribution of glucose metabolism in the brain) affect mental activity but cannot be represented faithfully in a space of only five dimensions. Yet it may be that five numbers can capture enough of the truth to be useful for many purposes.

major premise, that A exists; and the rest of the statement is understood to be conditional on that premise. Put differently, in common grammar the verb "is" implies a distinction between subject and object, which the symbol "=" does not have in formal logic or in conventional mathematics. [But in computer languages we encounter such statements as "J = J + 1" which everybody seems to understand, but in which the "=" sign has now acquired that implied distinction after all.]

Another amusing example is the old adage: "Knowledge is Power," which is a very cogent truth, both in human relations and in thermodynamics. An ad writer for a chemical trade journal† fouled this up into: "Power is Knowledge," an absurd—indeed, obscene—falsity.

These examples remind us that the verb "is" has, like any other verb, a subject and a predicate; but it is seldom noted that this verb has two entirely different meanings. A person whose native language is English may require some effort to see the different meanings in the statements: "The room is noisy" and "There is noise in the room." But in Turkish these meanings are rendered by different words, which makes the distinction so clear that a visitor who uses the wrong word will not be understood. The latter statement is ontological, asserting the physical existence of something, while the former is epistemological, expressing only the speaker's personal perception.

Common language—or at least, the English language—has an almost universal tendency to disguise epistemological statements by putting them into a grammatical form which suggests to the unwary an ontological statement. A major source of error in current probability theory arises from an unthinking failure to perceive this. To interpret the first kind of statement in the ontological sense is to assert that one's own private thoughts and sensations are realities existing externally in Nature. We call this the "Mind Projection Fallacy," and note the trouble it causes many times in what follows. But this trouble is hardly confined to probability theory; as soon as it is pointed out, it becomes evident that much of the discourse of philosophers and Gestalt psychologists, and the attempts of physicists to explain quantum theory, are reduced to nonsense by the author falling repeatedly into the Mind Projection Fallacy.

These examples illustrate the care that is needed when we try to translate the complex statements of common language into the simpler statements of formal logic. Of course, common language is often less precise than we should want in formal logic. But everybody expects this and is on the lookout for it, so it is less dangerous.

It is too much to expect that our robot will grasp all the subtle nuances of common language, which a human spends perhaps twenty years acquiring. In this respect, our robot will remain like a small child—it interprets all statements literally and blurts out the truth without thought of whom this may offend.

It is unclear to the writer how difficult—and even less clear how desirable—it would be to design a newer model robot with the ability to recognize these finer shades of meaning. Of course, the question of principle is disposed of at once by the existence of the human brain which does this. But in practice von Neumann's principle applies; a robot designed by us cannot do it until someone develops a theory of "nuance recognition" which reduces the process to a definitely prescribed set of operations. This we gladly leave to others.

In any event, our present model robot is quite literally real, because today it is almost universally true that any nontrivial probability evaluation is performed by a computer. The person who programmed that computer was necessarily, whether or not he thought of it that way, designing part of the brain of a robot according to some preconceived notion of how the robot should behave.

† LC-CG magazine, March 1988, p. 211

But very few of the computer programs now in use satisfy all our desiderata; indeed, most are intuitive *ad hoc* procedures that were not chosen with any well-defined desiderata at all in mind.

Any such adhockery is presumably usable within some special area of application—that was the criterion for choosing it—but as the proofs of Chapter 2 will show, any adhockery which conflicts with the rules of probability theory must generate demonstrable inconsistencies when we try to apply it beyond some restricted area. Our aim is to avoid this by developing the general principles of inference once and for all, directly from the requirement of consistency, and in a form applicable to any problem of plausible inference that is formulated in a sufficiently unambiguous way.

Nitpicking

As is apparent from the above, in the present work we use the term "Boolean algebra" in its long-established meaning as referring to two-valued logic in which symbols like "A" stand for propositions. A compulsive nit-picker has complained to us that some mathematicians have used the term in a slightly different meaning, in which "A" could refer to a class of propositions. But the two usages are not in conflict; we recognize the broader meaning, but just find no reason to avail ourselves of it.

The set of rules and symbols that we have called "Boolean Algebra" is sometimes called "The Propositional Calculus." The term seems to be used only for the purpose of adding that we need also another set of rules and symbols called "The Predicate Calculus." However, these new symbols prove to be only abbreviations for short and familiar phrases. The "Universal Quantifier" is only an abbreviation for "for all"; the "existential quantifier" is an abbreviation for "there is a." If we merely write our statements in plain English, we are using automatically all of the predicate calculus that we need for our purposes, and doing it more intelligibly.

The validity of the second strong syllogism (in two-valued logic) is sometimes questioned. However, it appears that in current mathematics it is still considered valid reasoning to say that a supposed theorem is disproved by exhibiting a counter-example, that a set of statements is considered inconsistent if we can derive a contradiction from them, and that a proposition can be established by *Reductio ad Absurdum*, deriving a contradiction from its denial. This is enough for us; we are quite content to follow this long tradition. Our feeling of security in this stance comes from the conviction that, while logic may move forward in the future, it can hardly move backward. A new logic might lead to new results about which Aristotelian logic has nothing to say; indeed, that is just what we are trying to create here. But surely, if a new logic was found to conflict with Aristotelian logic in an area where Aristotelian logic is applicable, we would consider that a fatal objection to the new logic.

Therefore, to those who feel confined by two-valued deductive logic we can say only: "By all means, investigate other possibilities if you wish to; and please let us know about it as soon as you have found a new result that was not contained in two-valued logic or our extension of it, *and* is useful in scientific inference." Actually, there are many different and mutually inconsistent multiple-valued logics already in the literature. But in Appendix A we adduce arguments which suggest that they can have no useful content that is not already in two-valued logic; that is, that an n-valued logic applied to one set of propositions is either equivalent to a two-valued logic applied to an enlarged set, or else it contains internal inconsistencies.

Our experience is consistent with this conjecture; in practice, multiple-valued logics seem to be used, not to find new useful results, but rather in attempts to remove supposed difficulties with two-valued logic, particularly in quantum theory, fuzzy sets, and Artificial Intelligence. But on closer study, all such difficulties known to us have proved to be only examples of the Mind Projection Fallacy, calling for direct revision of the concepts rather than a new logic.

Chapter 2

THE QUANTITATIVE RULES

"Probability theory is nothing but common sense reduced to calculation."

— *Laplace, 1819*

We have now formulated our problem, and it is a matter of straightforward mathematics to work out the consequences of our desiderata: stated broadly,

I. Representation of degrees of plausibility by real numbers

II. Qualitative Correspondence with common sense

III. Consistency.

The present Chapter is devoted entirely to deduction of the quantitative rules for inference which follow from these desiderata. The resulting rules have a long, complicated, and astonishing history, full of lessons for scientific methodology in general (see Comments at the end of several Chapters).

The Product Rule

We first seek a consistent rule relating the plausibility of the logical product AB to the plausibilities of A and B separately. In particular, let us find $AB|C$. Since the reasoning is somewhat subtle, we examine this from several different viewpoints.

As a first orientation, note that the process of deciding that AB is true can be broken down into elementary decisions about A and B separately. The robot can

(1) Decide that B is true. $(B|C)$

(2) Having accepted B as true, decide that A is true. $(A|BC)$

Or, equally well,

(1′) Decide that A is true. $(A|C)$

(2′) Having accepted A as true, decide that B is true. $(B|AC)$

In each case we indicate above the plausibility corresponding to that step.

Now let us describe the first procedure in words. In order for AB to be a true proposition, it is necessary that B is true. Thus the plausibility $B|C$ should be involved. In addition, if B is true, it is further necessary that A should be true; so the plausibility $A|BC$ is also needed. But if B is false, then of course AB is false independently of whatever one knows about A, as expressed by $A|\overline{B}\,C$; if the robot reasons first about B, then the plausibility of A will be relevant only if B is true. Thus, if the robot has $B|C$ and $A|BC$ it will not need $A|C$. That would tell it nothing about AB that it did not have already.

Similarly, $A|B$ and $B|A$ are not needed; whatever plausibility A or B might have in the absence of information C could not be relevant to judgments of a case in which the robot knows that C is true. For example, if the robot learns that the earth is round, then in judging questions about cosmology today, it does not need to take into account the opinions it might have (*i.e.* the extra possibilities that it would need to take into account) if it did not know that the earth is round.

Of course, since the logical product is commutative, $AB = BA$, we could interchange A and B in the above statements; *i.e.* knowledge of $A|C$ and $B|AC$ would serve equally well to determine

$AB|C = BA|C$. That the robot must obtain the same value for $AB|C$ from either procedure, is one of our conditions of consistency, Desideratum (IIIa).

We can state this in a more definite form. $(AB|C)$ will be some function of $B|C$ and $A|BC$:

$$(AB|C) = F[(B|C), (A|BC)]. \tag{2–1}$$

Now if the reasoning we went through here is not completely obvious, let us examine some alternatives. We might suppose, for example, that

$$(AB|C) = F[(A|C), (B|C)] \tag{2–2}$$

might be a permissible form. But we can show easily that no relation of this form could satisfy our qualitative conditions of Desideratum (II). Proposition A might be very plausible given C, and B might be very plausible given C; but AB could still be very plausible or very implausible.

For example, it is quite plausible that the next person you meet has blue eyes and also quite plausible that this person's hair is black; and it is reasonably plausible that both are true. On the other hand it is quite plausible that the left eye is blue, and quite plausible that the right eye is brown; but extremely implausible that both of those are true. We would have no way of taking such influences into account if we tried to use a formula of this kind. Our robot could not reason the way humans do, even qualitatively, with that kind of functional relation.

But other possibilities occur to us. The method of trying out all possibilities—a kind of "proof by exhaustion"—can be organized as follows. Introduce the real numbers

$$u = (AB|C), \qquad v = (A|C), \qquad w = (B|AC), \qquad x = (B|C), \qquad y = (A|BC). \tag{2–3}$$

If u is to be expressed as a function of two or more of v, w, x, y, there are eleven possibilities. You can write out each of them, and subject each one to various extreme conditions, as in the brown and blue eyes (which was the abstract statement: A implies that B is false). Other extreme conditions are $A = B$, $A = C$, $C \Rightarrow \overline{A}$, etc. Carrying out this somewhat tedious analysis, Tribus (1969) finds that all but two of the possibilities can exhibit qualitative violations of common sense in some extreme case. The two which survive are $u = F(x, y)$ and $u = F(w, v)$, just the two functional forms already suggested by our previous reasoning.

We now apply the qualitative requirement discussed in Chapter 1; given any change in the prior information $C \to C'$ such that B becomes more plausible but A does not change:

$$B|C' > B|C, \tag{2–4}$$

$$A|BC' = A|BC, \tag{2–5}$$

common sense demands that AB could only become more plausible, not less:

$$AB|C' \geq AB|C \tag{2–6}$$

with equality if and only if $A|BC$ corresponds to impossibility. Likewise, given prior information C'' such that

$$B|C'' = B|C \tag{2–7}$$

$$A|BC'' > A|BC \tag{2–8}$$

we require that
$$AB|C'' \geq AB|C \tag{2–9}$$
in which the equality can hold only if B is impossible, given C (for then AB might still be impossible given C'', although $A|BC$ is not defined). Furthermore, the function $F(x, y)$ must be continuous; for otherwise an arbitrarily small increase in one of the plausibilities on the right-hand side of (2–1) could result in a large increase in $AB|C$.

In summary, $F(x, y)$ must be a continuous monotonic increasing function of both x and y. If we assume it is differentiable [this is not necessary; see the discussion following (2–13)], then we have
$$F_1(x, y) \equiv \frac{\partial F}{\partial x} \geq 0 \tag{2–10a}$$
with equality if and only if y represents impossibility; and also
$$F_2(x, y) \equiv \frac{\partial F}{\partial y} \geq 0 \tag{2–10b}$$
with equality permitted only if x represents impossibility. Note for later purposes that in this notation, F_i denotes differentiation with respect to the i'th argument of F, whatever it may be.

Next we impose the Desideratum (IIIa) of "structural" consistency. Suppose we try to find the plausibility $(ABC|D)$ that three propositions would be true simultaneously. Because of the fact that Boolean algebra is associative: $ABC = (AB)C = A(BC)$, we can do this in two different ways. If the rule is to be consistent, we must get the same result for either order of carrying out the operations. We can say first that BC will be considered a single proposition, and then apply (2–1):
$$(ABC|D) = F[(BC|D), (A|BCD)] \tag{2–11}$$
and then in the plausibility $(BC|D)$ we can again apply (2–1) to give
$$(ABC|D) = F\{F[(C|D), (B|CD)], (A|BCD)\}. \tag{2–12a}$$
But we could equally well have said that AB shall be considered a single proposition at first. From this we can reason out in the other order to obtain a different expression:
$$(ABC|D) = F[(C|D), (AB|CD)] = F\{(C|D), F[(B|CD), (A|BCD)]\}. \tag{2–12b}$$
If this rule is to represent a consistent way of reasoning, the two expressions (2–12a), (2–12b) must always be the same. A necessary condition that our robot will reason consistently in this case therefore takes the form of a functional equation,
$$F[F(x, y), z] = F[x, F(y, z)]. \tag{2–13}$$
This equation has a long history in mathematics, starting from a work of N. H. Abel in 1826. Aczél (1966), in his monumental work on functional equations, calls it, very appropriately, "The Associativity Equation," and lists a total of 98 references to works that discuss it or use it. Aczél derives the general solution, Eq. (2–27) below, without assuming differentiability; unfortunately, the proof fills eleven pages (256-267) of his book. We give here the shorter proof by R. T. Cox (1961), which assumes differentiability; see also the discussion in Appendix B.

It is evident that (2–13) has a trivial solution, $F(x, y) = \text{const}$. But that violates our monotonicity requirement (2–10) and is in any event useless for our purposes. Unless (2–13) has a nontrivial solution, this approach will fail; so we seek the most general nontrivial solution. Using the abbreviations

$$u \equiv F(x, y), \qquad v \equiv F(y, z), \tag{2–14}$$

but still considering (x, y, z) the independent variables, the functional equation to be solved is

$$F(x, v) = F(u, z). \tag{2–15}$$

Differentiating with respect to x and y we obtain, in the notation of (2–10),

$$\begin{aligned} F_1(x, v) &= F_1(u, z) F_1(x, y) \\ F_2(x, v) F_1(y, z) &= F_1(u, z) F_2(x, y). \end{aligned} \tag{2–16}$$

Elimination of $F_1(u, z)$ from these equations yields

$$G(x, v) F_1(y, z) = G(x, y) \tag{2–17}$$

where we use the notation $G(x, y) \equiv F_2(x, y)/F_1(x, y)$. Evidently, the left-hand side of (2–17) must be independent of z. Now (2–17) can be written equally well as

$$G(x, v) F_2(y, z) = G(x, y) G(y, z), \tag{2–18}$$

and denoting the left-hand sides of (2–17), (2–18) by U, V respectively we verify that $\partial V/\partial y = \partial U/\partial z$. Thus, $G(x, y)G(y, z)$ must be independent of y. The most general function $G(x, y)$ with this property is

$$G(x, y) = r \frac{H(x)}{H(y)} \tag{2–19}$$

where r is a constant, and the function $H(x)$ is arbitrary. In the present case, $G > 0$ by monotonicity of F, and so we require that $r > 0$, and $H(x)$ may not change sign in the region of interest. Using (2–19), Eqs. (2–17) and (2–18) become

$$F_1(y, z) = \frac{H(v)}{H(y)} \tag{2–20}$$

$$F_2(y, z) = r \frac{H(v)}{H(z)} \tag{2–21}$$

and the relation $dv = dF(y, z) = F_1 dy + F_2 dz$ takes the form

$$\frac{dv}{H(v)} = \frac{dy}{H(y)} + r \frac{dz}{H(z)} \tag{2–22}$$

or, on integration,

$$w[F(y, z)] = w(v) = w(y) w^r(z) \tag{2–23}$$

where

$$w(x) \equiv \exp \left\{ \int^x \frac{dx}{H(x)} \right\}. \tag{2–24}$$

The absence of a lower limit on the integral signifies an arbitrary multiplicative factor in w. But taking the function $w(\cdot)$ of (2–15) and applying (2–23), we obtain $w(x)w^r(v) = w(u)w^r(z)$; applying (2–23) again, our functional equation now reduces to

$$w(x)w^r(y)[w(z)]^{r^2} = w(x)w^r(y)w^r(z). \tag{2–25}$$

Thus we obtain a nontrivial solution only if $r = 1$, and our final result can be expressed in either of the two forms:

$$w[F(x,y)] = w(x)w(y) \tag{2–26}$$

$$F(x,y) = w^{-1}[w(x)w(y)]. \tag{2–27}$$

Associativity and commutativity of the logical product thus require that the relation sought must take the functional form

$$w(AB|C) = w(A|BC)w(B|C) = w(B|AC)w(A|C) \tag{2–28}$$

which we shall call henceforth the *product rule*. By its construction (2–24), $w(x)$ must be a positive continuous monotonic function, increasing or decreasing according to the sign of $H(x)$; at this stage it is otherwise arbitrary.

The result (2–28) has been derived as a necessary condition for consistency in the sense of Desideratum (IIIa). Conversely, it is evident that (2–28) is also sufficient to ensure this consistency for any number of joint propositions. For example, there are an enormous number of different ways in which $(ABCDEFG|H)$ could be expanded by successive partitions in the manner of (2–12); but if (2–28) is satisfied, they will all yield the same result.

The requirements of qualitative correspondence with common sense impose further conditions on the function $w(x)$. For example, in the first given form of (2–28) suppose that A is certain, given C. Then in the "logical environment" produced by knowledge of C, the propositions AB and B are the same, in the sense that one is true if and only if the other is true. By our most primitive axiom of all, discussed in Chapter 1, propositions with the same truth value must have equal plausibility:

$$AB|C = B|C \tag{2–29}$$

and also we will have

$$A|BC = A|C \tag{2–30}$$

because if A is already certain given C (*i.e.* C implies A), then given any other information B which does not contradict C, it is still certain. In this case, (2–28) reduces to

$$w(B|C) = w(A|C)w(B|C) \tag{2–31}$$

and this must hold no matter how plausible or implausible B is to the robot. So our function $w(x)$ must have the property that

$$\text{Certainty is represented by } w(A|C) = 1. \tag{2–32}$$

Now suppose that A is impossible, given C. Then the proposition AB is also impossible given C:

$$AB|C = A|C \tag{2–33}$$

and if A is already impossible given C (*i.e.* C implies $\overline{A}$), then given any further information B which does not contradict C, A would still be impossible:

$$A|BC = A|C. \tag{2–34}$$

In this case, Eq. (2–28) reduces to

$$w(A|C) = w(A|C)w(B|C) \tag{2–35}$$

and again this equation must hold no matter what plausibility B might have. There are only two possible values of $w(A|C)$ that could satisfy this condition; it could be 0 or $+\infty$ (the choice $-\infty$ is ruled out because then by continuity $w(B|C)$ would have to be capable of negative values; (2–35) would then be a contradiction).

In summary, qualitative correspondence with common sense requires that $w(x)$ be a positive continuous monotonic function. It may be either increasing or decreasing. If it is increasing, it must range from zero for impossibility up to one for certainty. If it is decreasing, it must range from ∞ for impossibility down to one for certainty. Thus far, our conditions say nothing at all about how it varies between these limits.

However, these two possibilities of representation are not different in content. Given any function $w_1(x)$ which is acceptable by the above criteria and represents impossibility by ∞, we can define a new function $w_2(x) \equiv 1/w_1(x)$, which will be equally acceptable and represents impossibility by zero. Therefore, there will be no loss of generality if we now adopt the choice $0 \le w(x) \le 1$ as a *convention*; that is, as far as content is concerned, all possibilities consistent with our desiderata are included in this form. [As the reader may check, we could just as well have chosen the opposite convention; and the entire development of the theory from this point on, including all its applications, would go through equally well, with equations of a less familiar form but exactly the same content.]

The Sum Rule

Since the propositions now being considered are of the Aristotelian logical type which must be either true or false, the logical product $A\overline{A}$ is always false, the logical sum $A + \overline{A}$ always true. The plausibility that A is false must depend in some way on the plausibility that it is true. If we define $u \equiv w(A|B)$, $v \equiv w(\overline{A}|B)$, there must exist some functional relation

$$v = S(u). \tag{2–36}$$

Evidently, qualitative correspondence with common sense requires that $S(u)$ be a continuous monotonic decreasing function in $0 \le u \le 1$, with extreme values $S(0) = 1$, $S(1) = 0$. But it cannot be just any function with these properties, for it must be consistent with the fact that the product rule can be written for either AB or $A\overline{B}$:

$$w(AB|C) = w(A|C)w(B|AC) \tag{2–37}$$

$$w(A\overline{B}|C) = w(A|C)w(\overline{B}|AC). \tag{2–38}$$

Thus, using (2–36) and (2–38), Eq. (2–37) becomes

$$w(AB|C) = w(A|C)S[w(\overline{B}|AC)] = w(A|C)S\left[\frac{w(A\overline{B}|C)}{w(A|C)}\right]. \tag{2–39}$$

Again, we invoke commutativity: $w(AB|C)$ is symmetric in A, B, and so consistency requires that

$$w(A|C)S\left[\frac{w(A\overline{B}|C)}{w(A|C)}\right] = w(B|C)S\left[\frac{w(B\overline{A}|C)}{w(B|C)}\right]. \tag{2–40}$$

This must hold for all propositions A, B, C; in particular, (2–40) must hold when

$$\overline{B} = AD \tag{2–41}$$

where D is any new proposition. But then we have the truth-values noted before in (1–13):

$$A\overline{B} = \overline{B}, \qquad\qquad B\overline{A} = \overline{A}, \tag{2–42}$$

and in (2–40) we may write

$$\begin{aligned} w(A\overline{B}|C) &= w(\overline{B}|C) = S[w(B|C)] \\ w(B\overline{A}|C) &= w(\overline{A}|C) = S[w(A|C)]. \end{aligned} \tag{2–43}$$

Therefore, using now the abbreviations

$$x \equiv w(A|C), \qquad\qquad y \equiv w(B|C) \tag{2–44}$$

Eq. (2-25) becomes a functional equation

$$xS\left[\frac{S(y)}{x}\right] = yS\left[\frac{S(x)}{y}\right], \qquad \begin{array}{l} 0 \le S(y) \le x \\ 0 \le x \le 1 \end{array} \tag{2–45}$$

which expresses a scaling property that $S(x)$ must have in order to be consistent with the product rule. In the special case $y = 1$, this reduces to

$$S[S(x)] = x \tag{2–46}$$

which states that $S(x)$ is a self-reciprocal function; $S(x) = S^{-1}(x)$. Thus, from (2–36) it follows also that $u = S(v)$. But this expresses only the evident fact that the relation between A, $\overline{A}$ is a reciprocal one; it does not matter which proposition we denote by the simple letter, which by the barred letter. We noted this before in (1–8); if it had not been obvious before, we should be obliged to recognize it at this point.

The domain of validity given in (2–45) is found as follows. The proposition D is arbitrary, and so by various choices of D we can achieve all values of $w(D|AC)$ in

$$0 \le w(D|AC) \le 1. \tag{2–47}$$

But $S(y) = w(AD|C) = w(A|C)w(D|AC)$, and so (2–47) is just $(0 \le S(y) \le x)$, as stated in (2–45). This domain is symmetric in x, y; it can be written equally well with them interchanged. Geometrically, it consists of all points in the xy plane lying in the unit square $(0 \le x, y \le 1)$ and on or above the curve $y = S(x)$.

Indeed, the shape of that curve is determined already by what (2–45) says for points lying infinitesimally above it. For if we set $y = S(x) + \epsilon$, then as $\epsilon \to 0^+$ two terms in (2–45) tend to $S(1) = 0$, but at different rates. Therefore everything depends on the exact way in which $S(1-\delta)$ tends to zero as $\delta \to 0$. To investigate this, we define a new variable $q(x, y)$ by

$$\frac{S(x)}{y} = 1 - \exp\{-q\}. \tag{2–48}$$

Then we may choose $\delta = \exp\{-q\}$, define the function $J(q)$ by

$$S(1-\delta) = S(1 - \exp\{-q\} = \exp\{-J(q)\}, \tag{2–49}$$

and find the asymptotic form of $J(q)$ as $q \to \infty$.

Considering now x, q as the independent variables, we have from (2–48)

$$S(y) = S[S(x)] + \exp\{-q\}S(x)S'[S(x)] + O(\exp\{-2q\}). \tag{2–50}$$

Using (2–46) and its derivative $S'[S(x)]S'(x) = 1$, this reduces to

$$\frac{S(y)}{x} = 1 - \exp\{-(\alpha + q)\} + O(\exp\{-2q\}) \tag{2–51}$$

where

$$\alpha(x) \equiv \log\left[\frac{-xS'(x)}{S(x)}\right] > 0. \tag{2–52}$$

With these substitutions our functional equation (2–45) becomes

$$J(q+\alpha) - J(q) = \log\left[\frac{x}{S(x)}\right] + \log(1 - \exp\{-q\}) + O(\exp\{-2q\}), \qquad \begin{matrix} 0 < q < \infty \\ 0 < x \le 1 \end{matrix}. \tag{2–53}$$

As $q \to \infty$ the last two terms go to zero exponentially fast, so $J(q)$ must be asymptotically linear

$$J(q) \sim a + bq + O(\exp\{-q\}), \tag{2–54}$$

with positive slope

$$b = \alpha^{-1}\log\left[\frac{x}{S(x)}\right]. \tag{2–55}$$

In (2–54) there is no periodic term with period α, because (2–53) must hold for a continuum of different values of x, and therefore for a continuum of values of $\alpha(x)$. But by definition, J is a function of q only, so the right-hand side of (2–55) must be independent of x. This gives, using (2–52),

$$\frac{x}{S(x)} = \left[\frac{-xS'(x)}{S(x)}\right]^b, \qquad 0 < b < \infty \tag{2–56}$$

or rearranging, $S(x)$ must satisfy the differential equation

$$S^{m-1}dS + x^{m-1}dx = 0, \tag{2–57}$$

where $m \equiv 1/b$ is some positive constant. The only solution of this satisfying $S(0) = 1$ is

$$S(x) = (1 - x^m)^{1/m}, \qquad \begin{array}{c} 0 \le x \le 1 \\ 0 < m < \infty \end{array} \tag{2–58}$$

and conversely, we verify at once that (2–58) is a solution of (2–45).

The result (2–58) was first derived by R. T. Cox (1946) by a different argument which assumed $S(x)$ twice differentiable. Again, Aczél (1966) derives the same result without assuming differentiability. [But to assume differentiability in the present application seems to us a very innocuous step, for if the functional equations had led us to nondifferentiable functions, we would have rejected this whole theory as a qualitative violation of common sense]. In any event, (2–58) is the most general function satisfying the functional equation (2–45) and the left boundary condition $S(0) = 1$; whereupon we are encouraged to find that it automatically satisfies the right boundary condition $S(1) = 0$.

Since our derivation of the functional equation (2–45) used the special choice (2–41) for B, we have shown thus far only that (2–58) is a necessary condition to satisfy the general consistency requirement (2–40). To check its sufficiency, substitute (2–58) into (2–40). We obtain

$$w^m(A|C) - w^m(A\overline{B}|C) = w^m(B|C) - w^m(B\overline{A}|C), \tag{2–59}$$

a trivial identity by virtue of (2–28) and (2–38). Therefore, (2–58) is the necessary and sufficient condition on $S(x)$ for consistency in the sense (2–40).

Our results up to this point can be summarized as follows. Associativity of the logical product requires that some monotonic function $w(x)$ of the plausibility $x = A|B$ must obey the product rule (2–28). Our result (2–58) states that this same function must also obey a sum rule:

$$w^m(A|B) + w^m(\overline{A}|B) = 1 \tag{2–60}$$

for some positive m. Of course, the product rule itself can be written equally well as

$$w^m(AB|C) = w^m(A|C)w^m(B|AC) = w^m(B|C)w^m(A|BC) \tag{2–61}$$

but then we see that the value of m is actually irrelevant; for whatever value is chosen, we can define a new function

$$p(x) \equiv w^m(x) \tag{2–62}$$

and our rules take the form

$$p(AB|C) = p(A|C)p(B|AC) = p(B|C)p(A|BC) \tag{2–63}$$

$$p(A|B) + p(\overline{A}|B) = 1. \tag{2–64}$$

In fact, this entails no loss of generality, for the only requirement we have imposed on the function $w(x)$ is that it is a continuous monotonic increasing function ranging from $w = 0$ for impossibility to $w = 1$ for certainty. But if $w(x)$ satisfies this, then so also does $w^m(x)$, $0 < m < \infty$. Therefore, to say that we could use different values of m does not give us any freedom that we did not have already in the arbitrariness of $w(x)$. All possibilities allowed by our desiderata are contained in

(2–63) and (2–64) in which $p(x)$ is any continuous monotonic increasing function with the range $0 \le p(x) \le 1$.

Are further relations needed to yield a complete set of rules for plausible inference, adequate to determine the plausibility of any logic function $f(A_1, \ldots, A_n)$ from those of $\{A_1, \ldots, A_n\}$? We have, in the product rule (2–63) and sum rule (2–64), formulas for the plausibility of the conjunction AB and the negation $\overline{A}$. But we noted, in the discussion following Eq. (1–23), that conjunction and negation are an adequate set of operations, from which all logic functions can be constructed.

Therefore, one would conjecture that our search for basic rules should be finished; it ought to be possible, by repeated applications of the product rule and sum rule, to arrive at the plausibility of any proposition in the Boolean algebra generated by $\{A_1, \ldots, A_n\}$.

To verify this, we seek first a formula for the logical sum $A + B$. Applying the product rule and sum rule repeatedly, we have

$$\begin{aligned} p(A+B|C) &= 1 - p(\overline{A}\,\overline{B}|C) = 1 - p(\overline{A}|C)p(\overline{B}|\overline{A}C) \\ &= 1 - p(\overline{A}|C)[1 - p(B|\overline{A}C)] = p(A|C) + p(\overline{A}B|C) \\ &= p(A|C) + p(B|C)p(\overline{A}|BC) = p(A|C) + p(B|C)[1 - p(A|BC)] \end{aligned} \tag{2–65}$$

and finally,

$$p(A+B|C) = p(A|C) + p(B|C) - p(AB|C). \tag{2–66}$$

This generalized sum rule is one of the most useful in applications. Evidently, the primitive sum rule (2–64) is a special case of (2–66), with the choice $B = \overline{A}$.

Exercise 2.1. Is it possible to find a general formula for $p(C|A+B)$, analogous to (2–66), from the product and sum rules? If so, derive it; if not, explain why this cannot be done.

Exercise 2.2. Now suppose we have a set of propositions $\{A_1, \cdots, A_n\}$ which on information X are mutually exclusive: $p(A_iA_j|X) = p(A_i|X)\,\delta_{ij}$. Show that $p(C|(A_1 + A_2 + \cdots + A_nX)$ is a weighted average of the separate plausibilities $p(C|A_iX)$:

$$p(C|(A_1 + \cdots + A_nX) = p(C|A_1X + A_2X + \cdots + A_nX) = \frac{\sum_i p(A_i|X)\,p(C|A_iX)}{\sum_i p(A_i|X)}. \tag{2–67}$$

To extend the result (2–66), we noted following (1–17) that any logic function other than the trivial contradiction can be expressed in disjunctive normal form, as a logical sum of the basic conjunctions such as (1–17). Now the plausibility of any one of the basic conjunctions $\{Q_i,\ 1 \le i \le 2^n\}$ is determined by repeated applications of the product rule; and then repeated application of (2–66) will yield the plausibility of any logical sum of the Q_i. In fact, these conjunctions are mutually exclusive, so we shall find [Eq. (2–85) below] that this reduces to a simple sum $\sum_i p(Q_i|C)$ of at most $(2^n - 1)$ terms.

So, just as conjunction and negation are an adequate set of operations for deductive logic, the above product and sum rules are an adequate set for plausible inference, in the following sense. Whenever the background information is enough to determine the plausibilities of the basic conjunctions, our rules are adequate to determine the plausibility of every proposition in the Boolean algebra generated by $\{A_1, \cdots, A_n\}$. Thus, in the case $n = 4$ we need the plausibilities of $2^4 = 16$

basic conjunctions, whereupon our rules will determine the plausibility of each of the $2^{16} = 65,536$ propositions in the Boolean algebra.

But this is almost always more than we need in a real application; if the background information is enough to determine the plausibility of a few of the basic conjunctions, this may be adequate for the small part of the Boolean algebra that is of concern to us.

Qualitative Properties

Now let us check to see how the theory based on (2–63) and (2–64) is related to the theory of deductive logic and the various qualitative syllogisms from which we started in Chapter 1. In the first place it is obvious that in the limit as $p(A|B) \to 0$ or $p(A|B) \to 1$, the sum rule (2–64) expresses the primitive postulate of Aristotelian logic: if A is true, then $\overline{A}$ must be false, etc.

Indeed, all of that logic consists of the two strong syllogisms (1–1), (1–2) and all that follows from them; using now the implication sign (1–14) to state the major premise:

$$\begin{array}{c} A \Rightarrow B \\ A \text{ is true} \\ \hline B \text{ is true} \end{array} \qquad\qquad \begin{array}{c} A \Rightarrow B \\ B \text{ is false} \\ \hline A \text{ is false} \end{array} \tag{2–68}$$

and the endless stream of their consequences. If we let C stand for their major premise:

$$C \equiv \text{“}A \Rightarrow B\text{”} \tag{2–69}$$

then these syllogisms correspond to our product rule (2–63) in the forms

$$p(B|AC) = \frac{p(AB|C)}{p(A|C)}, \qquad\qquad p(A|\overline{B}C) = \frac{p(A\overline{B}|C)}{p(\overline{B}|C)} \tag{2–70}$$

respectively. But from (2–68) we have $p(AB|C) = p(A|C)$ and $p(A\overline{B}|C) = 0$, and so (2–70) reduces to

$$p(B|AC) = 1, \qquad\qquad p(A|\overline{B}C) = 0 \tag{2–71}$$

as stated in the syllogisms (2–68). Thus the relation is simply: *Aristotelian deductive logic is the limiting form of our rules for plausible reasoning, as the robot becomes more and more certain of its conclusions.*

But our rules have also what is not contained in deductive logic: a quantitative form of the weak syllogisms (1–3) and (1–4). To show that those original qualitative statements always follow from the present rules, note that the first weak syllogism

$$\begin{array}{c} A \Rightarrow B \\ B \text{ is true} \\ \hline \text{Therefore, } A \text{ becomes more plausible} \end{array} \tag{2–72}$$

corresponds to the product rule (2–63) in the form

$$p(A|BC) = p(A|C)\frac{p(B|AC)}{p(B|C)}. \tag{2–73}$$

But from (2–68), $p(B|AC) = 1$, and since $p(B|C) \le 1$, (2–73) gives

$$p(A|BC) \ge p(A|C) \tag{2–74}$$

as stated in the syllogism. Likewise, the syllogism (1–4)

$$\begin{array}{c} A \Rightarrow B \\ A \text{ is false} \\ \hline \text{Therefore, } B \text{ becomes less plausible} \end{array} \tag{2–75}$$

corresponds to the product rule in the form

$$p(B|\overline{A}C) = p(B|C)\frac{p(\overline{A}|BC)}{p(\overline{A}|C)}. \tag{2–76}$$

But from (2–74) it follows that $p(\overline{A}|BC) \leq p(\overline{A}|C)$; and so (2–76) gives

$$p(B|\overline{A}C) \leq p(B|C) \tag{2–77}$$

as stated in the syllogism.

Finally, the policeman's syllogism (1–5), which seemed very weak when stated abstractly, is also contained in our product rule, stated in the form (2–73). Letting C now stand for the background information [not noted explicitly in (1–5) because the need for it was not yet apparent], the major premise, "If A is true, then B becomes more plausible," now takes the form

$$p(B|AC) > p(B|C) \tag{2–78}$$

and (2–73) gives at once

$$p(A|BC) > p(A|C) \tag{2–79}$$

as stated in the syllogism.

But now we have more than the mere qualitative statement (2–79). In Chapter 1 we wondered, without answering: What determines whether the evidence B elevates A almost to certainty, or has a negligible effect on its plausibility? The answer from (2–73) is that, since $p(B|AC)$ cannot be greater than unity, a large increase in the plausibility of A can occur only when $p(B|C)$ is very small. Observing the gentleman's behavior (B) makes his guilt (A) seem virtually certain, because that behavior is otherwise so very unlikely on the background information; no policeman has ever seen an innocent person behaving that way. On the other hand, if knowing that A is true can make only a negligible increase in the plausibility of B, then observing B can in turn make only a negligible increase in the plausibility of A.

We could give many more comparisons of this type; indeed, the complete qualitative correspondence of these rules with common sense has been noted and demonstrated by many writers, including Keynes (1921), Jeffreys (1939), Pólya (1945, 1954), Cox R. T. (1961), Tribus (1969), de Finetti (1974), and Rosenkrantz (1977). The treatment of Pólya was described briefly in our Preface and Chapter 1, and we have just recounted that of Cox more fully. However, our aim now is to push ahead to quantitative applications; so we return to the basic development of the theory.

Numerical Values

We have found so far the most general consistent rules by which our robot can manipulate plausibilities, granted that it must associate them with real numbers, so that its brain can operate by the carrying out of some definite physical process. While we are encouraged by the familiar formal appearance of these rules and their qualitative properties just noted, two evident circumstances show that our job of designing the robot's brain is not yet finished.

In the first place, while the rules (2–63), (2–64) place some limitations on how plausibilities of different propositions must be related to each other, it would appear that we have not yet found any *unique* rules, but rather an infinite number of possible rules by which our robot can do plausible reasoning. Corresponding to every different choice of a monotonic function $p(x)$, there seems to be a different set of rules, with different content.

Secondly, nothing given so far tells us what actual numerical values of plausibility should be assigned at the beginning of a problem, so that the robot can get started on its calculations. How is the robot to make its initial encoding of the background information into definite numerical values of plausibilities? For this we must invoke the "interface" desiderata (IIIb), (IIIc) of (1–39), not yet used.

The following analysis answers both of these questions, in a way both interesting and unexpected. Let us ask for the plausibility $(A_1 + A_2 + A_3|B)$ that at least one of three propositions $\{A_1, A_2, A_3\}$ is true. We can find this by two applications of the extended sum rule (2–66), as follows. The first application gives

$$p(A_1 + A_2 + A_3|B) = p(A_1 + A_2|B) + p(A_3|B) - p(A_1A_3 + A_2A_3|B) \tag{2–80}$$

where we first considered $(A_1 + A_2)$ as a single proposition, and used the logical relation

$$(A_1 + A_2)A_3 = A_1A_3 + A_2A_3. \tag{2–81}$$

Applying (2–66) again, we obtain seven terms which can be grouped as follows:

$$\begin{aligned} p(A_1 + A_2 + A_3|B) = {} & p(A_1|B) + p(A_2|B) + p(A_3|B) \\ & - p(A_1A_2|B) - p(A_2A_3|B) - p(A_3A_1|B) \\ & + p(A_1A_2A_3|B). \end{aligned} \tag{2–82}$$

Now suppose these propositions are mutually exclusive; *i.e.* the evidence B implies that no two of them can be true simultaneously:

$$p(A_iA_j|B) = p(A_i|B)\delta_{ij}. \tag{2–83}$$

Then the last four terms of (2–82) vanish, and we have

$$p(A_1 + A_2 + A_3|B) = p(A_1|B) + P(A_2|B) + P(A_3|B). \tag{2–84}$$

Adding more propositions A_4, A_5, etc., it is easy to show by induction that if we have n mutually exclusive propositions $\{A_1 \cdots A_n\}$, (2–84) generalizes to

$$p(A_1 + \cdots + A_m|B) = \sum_{i=1}^{m} p(A_i|B), \qquad 1 \le m \le n \tag{2–85}$$

a rule which we will be using constantly from now on.

In conventional expositions, Eq. (2–85) is usually introduced first as the basic but, as far as one can see, arbitrary axiom of the theory. The present approach shows that this rule is deducible from simple qualitative conditions of consistency. The viewpoint which sees (2–85) as the primitive, fundamental relation is one which we are particularly anxious to avoid (see Comments at the end of this Chapter).

Now suppose that the propositions $\{A_1 \dots A_n\}$ are not only mutually exclusive but also exhaustive; *i.e.* the background information B stipulates that one and only one of them must be true. In that case the sum (2–85) for $m = n$ must be unity:

$$\sum_{i=1}^{n} p(A_i|B) = 1. \tag{2–86}$$

This alone is not enough to determine the individual numerical values $p(A_i|B)$. Depending on further details of the information B, many different choices might be appropriate, and in general finding the $p(A_i|B)$ by logical analysis of B can be a difficult problem. It is, in fact, an open-ended problem, since there is no end to the variety of complicated information that might be contained in B; and therefore no end to the complicated mathematical problems of translating that information into numerical values of $p(A_i|B)$. As we shall see, this is one of the most important current research problems; every new principle we can discover for translating information B into numerical values of $p(A_i|B)$ will open up a new class of useful applications of this theory.

There is, however, one case in which the answer is particularly simple, requiring only direct application of principles already given. But we are entering now into a very delicate area, a cause of confusion and controversy for over a century. In the early stages of this theory, as in elementary geometry, our intuition runs so far ahead of logical analysis that the point of the logical analysis is often missed. The trouble is that intuition leads us to the same final conclusions far more quickly; but without any correct appreciation of their range of validity. The result has been that the development of this theory has been retarded for some 150 years because various workers have insisted on debating these issues on the basis, not of demonstrative arguments, but of their conflicting intuitions.

At this point, therefore, we must ask the reader to suppress all intuitive feelings you may have, and allow yourself to be guided solely by the following logical analysis. The point we are about to make cannot be developed too carefully; and unless it is clearly understood, we will be faced with tremendous conceptual difficulties from here on.

Consider two different problems. Problem I is the one just formulated; we have a given set of mutually exclusive and exhaustive propositions $\{A_1 \dots A_n\}$ and we seek to evaluate $p(A_i|B)_I$. Problem II differs in that the labels A_1, A_2 of the first two propositions have been interchanged. These labels are, of course, entirely arbitrary; it makes no difference which proposition we choose to call A_1 and which A_2. In Problem II, therefore, we also have a set of mutually exclusive and exhaustive propositions $\{A'_1 \dots A'_n\}$, given by

$$\begin{aligned} A'_1 &\equiv A_2 \\ A'_2 &\equiv A_1 \\ A'_k &\equiv A_k, \quad 3 \le k \le n \end{aligned} \tag{2–87}$$

and we seek to evaluate the quantities $p(A'_i|B)_{II}, \ i = 1, 2, \dots, n$.

In interchanging the labels we have generated a different but closely related problem. It is clear that, whatever state of knowledge the robot had about A_1 in Problem I, it must have the same state of knowledge about A'_2 in Problem II, for they are the same proposition, the given information B

is the same in both problems, and it is contemplating the same totality of propositions $\{A_1 \dots A_n\}$ in both problems. Therefore we must have

$$p(A_1|B)_I = p(A'_2|B)_{II} \tag{2–88}$$

and similarly

$$p(A_2|B)_I = p(A'_1|B)_{II}. \tag{2–89}$$

We will call these the *transformation equations*. They describe only how the two problems are related to each other, and therefore they must hold whatever the information B might be; in particular, however plausible or implausible the propositions A_1, A_2 might seem to the robot in Problem I.

But now suppose that information B is indifferent between propositions A_1 and A_2; *i.e.* if it says something about one, it says the same thing about the other, and so it contains nothing that would give the robot any reason to prefer either one over the other. In this case, Problems I and II are not merely related, but entirely equivalent; *i.e.* the robot is in exactly the same state of knowledge about the set of propositions $\{A'_1 \dots A'_n\}$ in Problem II, *including their labeling*, as it is about the set $\{A_1 \dots A_n\}$ in Problem I.

Now we invoke our Desideratum of Consistency in the sense (IIIc) in (1–39). This stated that equivalent states of knowledge must be represented by equivalent plausibility assignments. In equations, this statement is

$$p(A_i|B)_I = p(A'_i|B)_{II}, \qquad i = 1, 2, \dots, n \tag{2–90}$$

which we shall call the *symmetry equations*. But now, combining equations (2–88), (2–89), (2–90) we obtain

$$p(A_1|B)_I = p(A_2|B)_I. \tag{2–91}$$

In other words, propositions A_1 and A_2 must be assigned equal plausibilities in Problem I (and, of course, also in Problem II).

At this point, depending on your personality and background in this subject, you will be either greatly impressed or greatly disappointed by the result (2–91). The argument we have just given is the first "baby" version of the group invariance principle for assigning plausibilities; it will be extended greatly in Chapter 6, when we consider the general problem of assigning "noninformative priors."

More generally, let $\{A''_1 \dots A''_n\}$ be any permutation of $\{A_1 \dots A_n\}$ and let Problem III be that of determining the $p(A''_i|B)$. If the permutation is such that $A''_k \equiv A_i$, there will be n transformation equations of the form

$$p(A_i|B)_I = p(A''_k|B)_{III} \tag{2–92}$$

which show how Problems I and III are related to each other; and these relations will hold whatever the given information B.

But if information B is now indifferent between all the propositions A_i, then the robot is in exactly the same state of knowledge about the set of propositions $\{A''_1 \dots A''_n\}$ in Problem III as it was about the set $\{A_1 \dots A_n\}$ in Problem I; and again our desideratum of consistency demands that it assign equivalent plausibilities in equivalent states of knowledge, leading to the n symmetry conditions

$$p(A_k|B)_I = p(A''_k|B)_{III}, \qquad k = 1, 2, \dots, n. \tag{2–93}$$

From (2–92) and (2–93) we obtain n equations of the form

$$p(A_i|B)_I = p(A_k|B)_I. \tag{2–94}$$

Now these relations must hold whatever the particular permutation we used to define Problem III. There are $n!$ such permutations, and so there are actually $n!$ equivalent problems among which, for given i, the index k will range over all of the $(n-1)$ others in (2–94). Therefore, the only possibility is that all of the $p(A_i|B)_I$ be equal (indeed, this is required already by consideration of a single permutation if it is cyclic of order n). Since the $\{A_1 \dots A_n\}$ are exhaustive, Eq. (2–86) will hold, and the only possibility is therefore

$$p(A_i|B)_I = \frac{1}{n}, \qquad (1 \le i \le n) \tag{2–95}$$

and we have finally arrived at a set of definite numerical values! Following Keynes (1921), we shall call this result the *Principle of Indifference*.

Perhaps, in spite of our admonitions, the reader's intuition had already led to just this conclusion, without any need for the rather tortuous reasoning we have just been through. If so, then at least that intuition is consistent with our desiderata. But merely writing down (2–95) intuitively gives one no appreciation of the importance and uniqueness of this result. To see the uniqueness, note that if the robot were to assign any values different from (2–95), then by a mere permutation of labels we could exhibit a second problem in which the robot's state of knowledge is the same, but in which it is assigning different plausibilities.

To see the importance, note that (2–95) actually answers both of the questions posed at the beginning of this Section. It shows—in one particular case which can be greatly generalized—how the information given the robot can lead to definite numerical values, so that a calculation can get started. But it also shows something even more important because it is not at all obvious intuitively; the information given the robot determines the numerical values of the quantities $p(x) = p(A_i|B)$, and not the numerical values of the plausibilities $x = A_i|B$ from which we started. This, also, will be found to be true in general.

Recognizing this gives us a beautiful answer to the first question posed at the beginning of this Section; after having found the product and sum rules, it still appeared that we had not found any unique rules of reasoning, because every different choice of a monotonic function $p(x)$ would lead to a different set of rules (*i.e.* a set with different content). But now we see that no matter what function $p(x)$ we choose, we shall be led to the same result (2–95), and the same numerical value of p. Furthermore, the robot's reasoning processes can be carried out entirely by manipulation of the quantities p, as the product and sum rules show; and the robot's final conclusions can be stated equally well in terms of the p's instead of the x's.

So, we now see that different choices of the function $p(x)$ correspond only to different ways we could design the robot's internal memory circuits. For each proposition A_i about which it is to reason, it will need a memory address in which it stores some number representing the degree of plausibility of A_i, on the basis of all the data it has been given. Of course, instead of storing the number p_i it could equally well store any strict monotonic function of p_i. But no matter what function it used internally, the externally observable behavior of the robot would be just the same.

As soon as we recognize this it is clear that, instead of saying that $p(x)$ is an arbitrary monotonic function of x, it is much more to the point to turn this around and say that:

The plausibility $x \equiv A|B$ is an arbitrary monotonic function of p, defined in $(0 \le p \le 1)$.

It is p that is rigidly fixed by the data, not x.

The question of uniqueness is therefore disposed of automatically by the result (2–95); in spite of first appearances, there is actually only one consistent set of rules by which our robot can do plausible reasoning, and for all practical purposes, the plausibilities $x \equiv A|B$ from which we started have faded entirely out of the picture! We will just have no further use for them.

Having seen that our theory of plausible reasoning can be carried out entirely in terms of the quantities p, we finally introduce their technical names; from now on, we will call these quantities *probabilities*. The word "probability" has been studiously avoided up to this point, because while the word does have a colloquial meaning to the proverbial "man on the street," it is for us a technical term, which ought to have a precise meaning. But until it had been demonstrated that these quantities are uniquely determined by the data of a problem, we had no grounds for supposing that the quantities p were possessed of any precise meaning.

We now see that they define a particular scale on which degrees of plausibility can be measured. Out of all possible monotonic functions which could in principle serve this purpose equally well, we choose this particular one, not because it is more "correct," but because it is more convenient; *i.e.* it is the quantities p that obey the simplest rules of combination, the product and sum rules. Because of this, numerical values of p are directly determined by our information.

This situation is analogous to that in thermodynamics, where out of all possible empirical temperature scales t, which are monotonic functions of each other, we finally decide to use the Kelvin scale T; not because it is more "correct" than others but because it is more convenient; *i.e.* the laws of thermodynamics take their simplest form [$dU = TdS - PdV$, $dG = -SdT + VdP$, etc.] in terms of this particular scale. Because of this, numerical values of Kelvin temperatures are "rigidly fixed" in the sense of being directly measurable in experiments, independently of the properties of any particular substance like water or mercury.

Another rule, equally appealing to our intuition, follows at once from (2–95). Consider the traditional "Bernoulli Urn" of probability theory; ours is known to contain ten balls of identical size and weight, labeled $\{1, 2, \ldots, 10\}$. Three balls (numbers 4, 6, 7) are black, the other seven are white. We are to shake the Urn and draw one ball blindfolded. The background information B in (2–95) consists of the statements in the last two sentences. What is the probability that we draw a black one?

Define the propositions: $A_i \equiv$ "The i'th ball is drawn," $(1 \le i \le 10)$. Since the background information is indifferent to these ten possibilities, (2–95) applies and the robot assigns

$$p(A_i|B) = \frac{1}{10}, \quad 1 \le i \le 10. \tag{2–96}$$

The statement that we draw a black ball is that we draw number 4, 6, or 7;

$$p(\text{Black}|B) = p(A_4 + A_6 + A_7|B). \tag{2–97}$$

But these are mutually exclusive propositions (*i.e.* they assert mutually exclusive events) so (2–85) applies and the robot's conclusion is

$$p(\text{Black}|B) = \frac{3}{10} \tag{2–98}$$

as intuition had told us already. More generally, if there are N such balls, and the proposition A is defined to be true on any specified subset of M of them, $(0 \le M \le N)$, false on the rest, we have

$$p(A|B) = \frac{M}{N}. \tag{2–99}$$

This was the original mathematical *definition* of probability, as given by James Bernoulli (1713) and used by most writers for the next 150 years. For example, Laplace's great *Théorie analytique des probabilités* (1812) opens with this sentence: "The Probability for an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible."

Exercise 2.3. As soon as we have the numerical values $a = P(A|C)$ and $b = P(B|C)$, the product and sum rules place some limits on the possible numerical values for their conjunction and disjunction. Supposing that $a \leq b$, show that the probability for the conjunction cannot exceed that of the least probable proposition: $0 \leq P(AB|C) \leq a$, and the probability for the disjunction cannot be less than that of the most probable proposition: $b \leq P(A + B|C) \leq 1$. Then show that, if $a + b > 1$, there is a stronger inequality for the conjunction; and if $a + b < 1$ there is a stronger one for the disjunction. These necessary general inequalities are helpful in detecting errors in calculations.

Notation and Finite Sets Policy

Now we can introduce the notation to be used in the remainder of this work (discussed more fully in Appendix B). Henceforth, our formal probability symbols will use the capital P:

$$P(A|B) \tag{2–100}$$

which signifies that the arguments are *propositions*. Probabilities whose arguments are numerical values are generally denoted by other functional symbols such as

$$f(r|np) \tag{2–101}$$

which denote ordinary mathematical functions. The reason for making this distinction is to avoid ambiguity in the meaning of our symbols, which has been a recent problem in this field.

However, in agreement with the customary loose notation in the existing literature, we sometimes relax our standards enough to allow the probability symbols with small p: $p(x|y)$ or $p(A|B)$ or $p(x|B)$ to have arguments which can be either propositions or numerical values, in any mix. Thus the meaning of expressions with small p can be judged only from the surrounding context.

It is very important to note that our consistency theorems have been established only for probabilities assigned on *finite sets* of propositions. In principle, every problem must start with such finite set probabilities; extension to infinite sets is permitted only when this is the result of a well-defined and well-behaved limiting process from a finite set. More generally, in any mathematical operations involving infinite sets the safe procedure is the finite sets policy:

> *Apply the ordinary processes of arithmetic and analysis only to expressions with a finite number of terms. Then after the calculation is done, observe how the resulting finite expressions behave as the number of terms increases indefinitely.*

In laying down this rule of conduct, we are only following the policy that mathematicians from Archimedes to Gauss have considered clearly necessary for nonsense avoidance in all of mathematics. But more recently, the popularity of infinite set theory and measure theory have led some to disregard it and seek shortcuts which purport to use measure theory directly. Note, however, that this rule of conduct is consistent with the original Lebesgue definition of measure, and *when a well-behaved limit exists* it leads us automatically to correct "measure theoretic" results. Indeed, this is how Lebesgue found his first results.

The danger is that the present measure theory notation presupposes the infinite limit already accomplished, but contains no symbol indicating which limiting process was used. Yet as noted in our Preface, different limiting processes—equally well-behaved—lead in general to different results. When there is no well-behaved limit, any attempt to go directly to the limit can result in nonsense, *the cause of which cannot be seen as long as one looks only at the limit, and not at the limiting process.*

This little Sermon is an introduction to Chapter 15 on Infinite Set Paradoxes, where we shall see some of the results that have been produced by those who ignored this rule of conduct, and tried to calculate probabilities directly on an infinite set without considering any limit from a finite set. The results are at best ambiguous, at worst nonsensical.

COMMENTS

It has taken us two Chapters of close reasoning to get back to the point (2–99) from which Laplace started some 180 years ago. We shall try to understand the intervening period, as a weird episode of history, throughout the rest of the present work. The story is so complicated that we can unfold it only gradually, over the next ten Chapters. To make a start on this, let us consider some of the questions often raised about the use of probability theory as an extension of logic.

"Subjective" vs. "Objective"

These words are abused so much in probability theory that we try to clarify our use of them. In the theory we are developing, any probability assignment is necessarily "subjective" in the sense that it describes only a state of knowledge, and not anything that could be measured in a physical experiment. Inevitably, someone will demand to know: "*Whose* state of knowledge?" The answer is always: "The robot—or anyone else who is given the same information and reasons according to the desiderata used in our derivations in this Chapter."

Anyone who has the same information but comes to a different conclusion than our robot, is necessarily violating one of those desiderata. While nobody has the authority to forbid such violations, it appears to us that a rational person, should he discover that he was violating one of them, would wish to revise his thinking (in any event, he would surely have difficulty in persuading anyone else, who was aware of that violation, to accept his conclusions).

Now it was just the function of our interface desiderata (IIIb), (IIIc) to make these probability assignments completely "objective" in the sense that they are independent of the personality of the user. They are a means of describing (or what is the same thing, of encoding) the *information* given in the statement of a problem, independently of whatever personal feelings (hopes, fears, value judgments, etc.) you or I might have about the propositions involved. It is "objectivity" in this sense that is needed for a scientifically respectable theory of inference.

Gödel's Theorem

To answer another inevitable question, we recapitulate just what has and what has not been proved in this Chapter. The main constructive requirement which determined our product and sum rules was the desideratum (IIIa) of "structural consistency." Of course, this does not mean that our rules have been proved consistent; it means only that any other rules which represent degrees of plausibility by real numbers, but which differ in content from ours, will lead necessarily either to inconsistencies or violations of our other desiderata.

A famous theorem of Kurt Gödel (1931) states that no mathematical system can provide a proof of its own consistency. Does this prevent us from ever proving the consistency of probability theory as logic? We are not prepared to answer this fully, but perhaps we can clarify the situation a little.

First, let us be sure that "inconsistency" means the same thing to us and to a logician. What we had in mind was that if our rules were inconsistent, then it would be possible to derive contradictory results from valid application of them; for example, by applying the rules in two equally valid ways, one might be able to derive both $P(A|BC) = 1/3$ and $P(A|BC) = 2/3$. Cox's functional equations sought to guard against this. Now when a logician says that a system of axioms $\{A_1, A_2, \ldots, A_n\}$ is inconsistent, he means that a contradiction can be deduced from them; *i.e.* some proposition Q and its denial $\overline{Q}$ are both deducible. Indeed, this is not really different from our meaning.

To understand the above Gödel result, the essential point is the principle of elementary logic that a contradiction $\overline{A}\,A$ implies all propositions, true and false. [For, given any two propositions A and B, we have $A \Rightarrow (A + B)$, therefore $\overline{A}\,A \Rightarrow \overline{A}(A + B) = \overline{A}\,A + \overline{A}\,B \Rightarrow B$.] Then let $A = A_1\,A_2 \cdots A_n$ be the system of axioms underlying a mathematical theory and T any proposition, or theorem, deducible from them:†

$$A \Rightarrow T. \tag{2–102}$$

Now whatever T may assert, the fact that T can be deduced from the axioms cannot prove that there is no contradiction in them, since if there were a contradiction, T could certainly be deduced from them!

This is the essence of the Gödel theorem, as it pertains to our problems. As noted by R. A. Fisher (1956), it shows us the intuitive reason why Gödel's result is true. We do not suppose that any logician would accept Fisher's simple argument as a proof of the full Gödel theorem; yet for most of us it is more convincing than Gödel's long and complicated proof.‡

Now suppose that the axioms contain an inconsistency. Then the opposite of T and therefore the contradiction $\overline{T}\,T$ can also be deduced from them:

$$A \Rightarrow \overline{T}. \tag{2–103}$$

So if there is an inconsistency, its existence can be proved by exhibiting any proposition T and its opposite $\overline{T}$ that are both deducible from the axioms. However, in practice it may not be easy to find a T for which one sees how to prove both T and $\overline{T}$.

† In Chapter 1 we noted the tricky distinction between the weak property of formal implication and the strong one of logical deducibility; by "implications of a proposition C" we really mean "propositions logically deducible from C *and the totality of other background information.*" Conventional expositions of Aristotelian logic are, in our view, flawed by their failure to make explicit mention of background information, which is usually essential to our reasoning, whether inductive or deductive. But in the present argument, we can understand A as including all the propositions that constitute that background information; then "implication" and "logical deducibility" are the same thing.

‡ The 1957 Edition of Harold Jeffreys' *Scientific Inference* has a short summary of Gödel's original reasoning which is far clearer and easier to read than any other "explanation" we have seen. The full theorem refers to other matters of concern in 1931, but of no interest to us right now; the above discussion has abstracted the part of it that we need to understand for our present purposes.

Evidently, we could prove the consistency of a set of axioms if we could find a feasible procedure which is guaranteed to locate an inconsistency if one exists; so Gödel's theorem seems to imply that no such procedure exists. Actually, it says only that no such procedure *derivable from the axioms of the system being tested* exists.

Yet we shall find that probability theory comes close to this; it is a powerful analytical tool which can search out a set of propositions and detect a contradiction in them if one exists. The principle is that probabilities conditional on contradictory premises do not exist (the hypothesis space is reduced to the empty set). Therefore, put our robot to work; *i.e.* write a computer program to calculate probabilities $p(B|E)$ conditional on a set of propositions $E = (E_1\, E_2\, \ldots\, E_n)$. Even though no contradiction is apparent from inspection, if there is a contradiction hidden in E, the computer program will crash.

We discovered this "empirically," and after some thought realized that it is not a reason for dismay, but rather a valuable diagnostic tool that warns us of unforeseen special cases in which our formulation of a problem can break down.

If the computer program does not crash, but prints out valid numbers, then we know that the conditioning propositions E_i are mutually consistent, and we have accomplished what one might have thought to be impossible in view of Gödel's theorem. But of course our use of probability theory appeals to principles not derivable from the propositions being tested, so there is no difficulty; it is important to understand what Gödel's theorem does and does not prove.

When Gödel's theorem first appeared, with its more general conclusion that a mathematical system may contain certain propositions that are undecidable within that system, it seems to have been a great psychological blow to logicians, who saw it at first as a devastating obstacle to what they were trying to achieve.

Yet a moment's thought shows us that many quite simple questions are undecidable by deductive logic. There are situations in which one can prove that a certain property must exist in a finite set, even though it is impossible to exhibit any member of the set that has that property. For example, two persons are the sole witnesses to an event; they give opposite testimony about it and then both die. Then we know that one of them was lying, but it is impossible to determine which one.

In this example, the undecidability is not an inherent property of the proposition or the event; it signifies only the incompleteness of our own information. But this is equally true of abstract mathematical systems; when a proposition is undecidable in such a system, that means only that its axioms do not provide enough *information* to decide it. But new axioms, external to the original set, might supply the missing information and make the proposition decidable after all.

In the future, as science becomes more and more oriented to thinking in terms of information content, Gödel's result will be seen as more of a platitude than a paradox. Indeed, from our viewpoint "undecidability" merely signifies that a problem is one that calls for *inference* rather than deduction. Probability theory as extended logic is designed specifically for such problems.

These considerations seem to open up the possibility that, by going into a wider field by invoking principles external to probability theory, one might be able to prove the consistency of our rules. At the moment, this appears to us to be an open question.

Needless to say, no inconsistency has ever been found from correct application of our rules, although some of our calculations will put them to a severe test. Apparent inconsistencies have always proved, on closer examination, to be misapplications of the rules. On the other hand, guided by Cox's theorems which tell us where to look, we have never had the slightest difficulty in exhibiting the inconsistencies in the *ad hoc* rules which abound in the literature, which differ in content from ours and whose sole basis is the intuitive judgment of their inventors. Examples are found throughout the sequel, but particularly in Chapters 5, 15, 17.

Venn Diagrams

Doubtless, some readers will ask, "After the rather long and seemingly unmotivated derivation of the extended sum rule (2–66), which in our new notation now takes the form:

$$P(A+B|C) = P(A|C) + P(B|C) - P(AB|C) \tag{2–104}$$

why did we not illustrate it by the Venn diagram? That makes its meaning so much clearer." [Here we draw two circles labeled A and B, with intersection labeled AB, all within a circle C.]

The Venn diagram is indeed a useful device, illustrating—in one special case—why the negative term appears in (2–104). But it can also mislead, because it suggests to our intuition more than the actual content of (2–104). Looking at the Venn diagram, we are encouraged to ask, "What do the points in the diagram mean?" If the diagram is intended to illustrate (2–104), then the probability for A is, presumably, represented by the area of circle A; for then the total area covered by circles A, B is the sum of their separate areas, minus the area of overlap, corresponding exactly to (2–104).

Now the circle A can be broken down into non-overlapping subregions in many different ways; what do these subregions mean? Since their areas are additive, if the Venn diagram is to remain applicable they must represent a refinement of A into the disjunction of some mutually exclusive sub-propositions. We can—if we have no mathematical scruples about approaching infinite limits—imagine this subdivision carried down to the individual points in the diagram. Therefore these points must represent some ultimate "elementary" propositions ω_i into which A can be resolved.† Of course, consistency then requires us to suppose that B and C can also be resolved into these same propositions ω_i.

Already, we have jumped to the conclusion that the propositions to which we assign probabilities correspond to sets of points in some space, that the logical disjunction $A+B$ stands for the union of the sets, the conjunction AB for their intersection, that the probabilities are an additive measure over those sets. But the general theory we are developing has no such structure; all these things are properties only of the Venn diagram.

In developing our theory of inference we have taken special pains to avoid restrictive assumptions which would limit its scope; it is to apply, in principle, to any propositions with unambiguous meaning. In the special case where those propositions happen to be statements about sets, the Venn diagram is an appropriate illustration of (2–104). But most of the propositions about which we reason, for example,

$$A \equiv \text{“It will rain today,”} \tag{2–105}$$

$$B \equiv \text{“The roof will leak”} \tag{2–106}$$

are simply declarative statements of fact, which may or may not be resolvable into a disjunction of more elementary propositions within the context of our problem.

† A physicist refuses to call them "atomic" propositions, for obvious reasons.

Of course, one can always force such a resolution by introducing irrelevancies; for example, even though the above-defined B has nothing to do with penguins, we could still resolve it into the disjunction:

$$B = BC_1 + BC_2 + BC_3 + \cdots + BC_N \tag{2–107}$$

where $C_k \equiv$ "The number of penguins in Antarctica is k." By choosing N sufficiently large, we will surely be making a valid statement of Boolean algebra; but this is idle and it cannot help us to reason about a leaky roof.

Even if a meaningful resolution exists in our problem, it may not be of any use to us. For example, the proposition "Rain Today" could be resolved into an enumeration of every conceivable trajectory of each individual raindrop; but we do not see how this could help a meteorologist trying to forecast rain. In real problems, there is a natural end to this resolving, beyond which it serves no purpose and degenerates into an empty formal exercise. We shall give an explicit demonstration of this later (Chapter 8), in the scenario of Sam's Broken Thermometer: does the exact way in which it broke matter for the conclusions that Sam should draw from his corrupted data?

But in some cases there is a resolution so relevant to the context of the problem that it becomes a useful calculational device; Eq. (2–98) was a trivial example. We shall be glad to take advantage of this whenever we can, but we cannot expect it in general.

Even when both A and B can be resolved in a way meaningful and useful in our problem, it would seldom be the case that they are resolvable into the *same* set of elementary propositions ω_i. And we always reserve the right to enlarge our context by introducing more propositions $D, E, F, \ldots$ into the discussion; and we could hardly ever expect that all of them would continue to be expressible as disjunctions of the *same* original set of elementary propositions ω_i. To assume this would be to place a quite unnecessary restriction on the generality of our theory.

Therefore, the conjunction AB should be regarded simply as the statement that both A and B are true; it is a mistake to try to read any more detailed meaning, such as an intersection of sets, into it in every problem. Then $p(AB|C)$ should also be regarded as an elementary quantity in its own right, not necessarily resolvable into a sum of still more elementary ones (although if it is so resolvable this may be a good way of calculating it). We have adhered to the original notation $A + B$, AB of Boole, instead of the more common $A \vee B$, $A \wedge B$, or $A \cup B$, $A \cap B$ which everyone associates with a set-theory context, in order to head off this confusion as much as possible.

So, rather than saying that the Venn diagram justifies or explains (2–104), we prefer to say that (2–104) explains and justifies the Venn diagram, in one special case. But the Venn diagram has played a major role in the history of probability theory, as we note next.

The "Kolmogorov Axioms"

In 1933, A. N. Kolmogorov presented an approach to probability theory phrased in the language of set theory and measure theory. This language was just then becoming so fashionable that today many mathematical results are named, not for the discoverer, but for the one who first restated them in that language. For example, in the theory of continuous groups the term "Hurwitz invariant integral" disappeared, to be replaced by "Haar measure." Because of this custom, some modern works—particularly by mathematicians—can give one the impression that probability theory started with Kolmogorov.

Kolmogorov formalized and axiomatized the picture suggested by the Venn diagram, which we have just described. At first glance, this system appears so totally different from ours that some discussion is needed to see the close relation between them. In Appendix A we describe the Kolmogorov system and show that, for all practical purposes the four axioms concerning his probability measure, first stated arbitrarily (for which Kolmogorov has been criticized) have all been derived in this Chapter as necessary to meet our consistency requirements. As a result, we shall find ourselves defending Kolmogorov against his critics on many technical points. The reader who first learned probability theory on the Kolmogorov basis is urged to read Appendix A at this point.

However, our system of probability differs conceptually from that of Kolmogorov in that we do not interpret propositions in terms of sets, but we do interpret probability distributions as carriers of incomplete information. Partly as a result, our system has analytical resources not present at all in the Kolmogorov system. This enables us to formulate and solve many problems—particularly the so-called "ill posed" problems and "generalized inverse" problems—that would be considered outside the scope of probability theory according to the Kolmogorov system. These problems are just the ones of greatest interest in current applications.

Chapter 3

ELEMENTARY SAMPLING THEORY

At this point, the mathematical material we have available consists of the basic product and sum rules

$$P(AB|C) = P(A|BC)P(B|C) = P(B|AC)P(A|C) \tag{3–1}$$

$$P(A|B) + P(\overline{A}|B) = 1 \tag{3–2}$$

from which we derived the extended sum rule

$$P(A + B|C) = P(A|C) + P(B|C) - P(AB|C) \tag{3–3}$$

and with the Desideratum (IIIc) of consistency, the principle of indifference: if on background information B the hypotheses $(H_1, H_2 \cdots H_N)$ are mutually exclusive and exhaustive, and B does not favor any one of them over any other, then

$$P(H_i|B) = \frac{1}{N}, \qquad 1 \le i \le N. \tag{3–4}$$

From (3–3) and (3–4) we then derived the Bernoulli urn rule; if B specifies that A is true on some subset of M of the H_i, false on the remaining $(N - M)$, then

$$P(A|B) = \frac{M}{N}. \tag{3–5}$$

It is important to realize how much of probability theory can be derived from no more than this.

In fact, essentially all of conventional probability theory as currently taught, plus many important results that are often thought to lie beyond the domain of probability theory, can be derived from the above foundation. We devote the next several Chapters to demonstrating this in some detail, and then in Chapter 11 resume the basic development of our robot's brain, with a better understanding of what additional principles are needed for advanced applications.

The first applications of the theory given in this Chapter are, to be sure, rather simple and naïve compared to the serious scientific inference that we hope to achieve later. Nevertheless, our reason for considering them in close detail is not mere pedagogical form. Failure to understand the logic of these simplest applications has been one of the major factors retarding the progress of scientific inference—and therefore of science itself—for many decades. Therefore we urge the reader, even one who considers himself already familiar with elementary sampling theory, to digest the contents of this Chapter carefully before proceeding to more complicated problems.

Sampling Without Replacement

Let us make the Bernoulli Urn scenario a little more specific by defining the propositions:

$B \equiv$ "An urn contains N balls, identical in every respect except that they carry numbers $(1, 2 \cdots N)$ and M of them are colored red, the remaining $(N-M)$ white, $0 \le M \le N$. We draw a ball from the urn blindfolded, observe and record its color, lay it aside, and repeat the process until n balls have been drawn, $0 \le n \le N$."

$R_i \equiv$ "Red ball on the i'th draw."

$W_i \equiv$ "White ball on the i'th draw," Since according to B only red or white can be drawn, we have

$$P(R_i|B) + P(W_i|B) = 1, \qquad 1 \le i \le N \tag{3–6}$$

which amounts to saying that, in the "logical environment" created by knowledge of B, the propositions are related by negation:

$$\overline{R}_i = W_i, \qquad \overline{W}_i = R_i \tag{3–7}$$

and for the first draw, (3–5) becomes

$$P(R_1|B) = \frac{M}{N}, \tag{3–8}$$

$$P(W_1|B) = 1 - \frac{M}{N}. \tag{3–9}$$

Let us understand clearly what this means. The probability assignments (3–8), (3–9) are not assertions of any physical property of the urn or its contents; they are a description of the *state of knowledge* of the robot prior to the drawing. Indeed, were the robot's state of knowledge different from B as just defined (for example, if it knew the actual positions of the red and white balls in the urn, or if it did not know the true values of N and M), then its probability assignments for R_1 and W_1 would be different; but the real properties of the urn would be just the same.

It is therefore illogical to speak of "verifying" (3–8) by performing experiments with the urn; that would be like trying to verify a boy's love for his dog by performing experiments on the dog. At this stage, we are concerned with the logic of consistent reasoning from incomplete information; not with assertions of physical fact about what will be drawn from the urn (which are in any event impossible just because of the incompleteness of the information B).

Eventually, our robot will be able to make some very confident physical predictions which can approach, but (except in degenerate cases) not actually reach, the certainty of logical deduction; but the theory needs to be developed further before we are in a position to say what quantities can be well predicted, and what kind of information is needed for this. Put differently, relations between probabilities assigned by the robot in various states of knowledge, and observable facts in experiments, may not be assumed arbitrarily; we are justified in using only those relations that can be deduced from the rules of probability theory, as we now seek to do.

Changes in the robot's state of knowledge appear already when we ask for probabilities referring to the second draw. For example, what is the robot's probability for red on the first two draws? From the product rule, this is

$$P(R_1R_2|B) = P(R_1|B)P(R_2|R_1B). \tag{3–10}$$

In the last factor, the robot must take into account that one red ball has been removed at the first draw, so there remain $(N-1)$ balls of which $(M-1)$ are red. Therefore

$$P(R_1R_2|B) = \frac{M}{N}\frac{M-1}{N-1}. \tag{3–11}$$

Continuing in this way, the probability for red on the first r consecutive draws is

$$\begin{aligned} P(R_1R_2\cdots R_r|B) &= \frac{M(M-1)\cdots(M-r+1)}{N(N-1)\cdots(N-r+1)} \\ &= \frac{M!(N-r)!}{(M-r)!N!}, \qquad r \le M. \end{aligned} \tag{3–12}$$

The restriction $r \le M$ is not necessary if we understand that we define factorials by the gamma function relation $n! = \Gamma(n+1)$, for then the factorial of a negative integer is infinite, and (3–12) is zero automatically when $r > M$.

The probability for white on the first w draws is similar but for the interchange of M and $(N-M)$:

$$P(W_1 W_2 \cdots W_w|B) = \frac{(N-M)!(N-w)!}{(N-M-w)!N!}. \tag{3–13}$$

Then, the probability for white on draws $(r+1, r+2 \cdots r+w)$ given that we got red on the first r draws, is given by (3–13) taking into account that N and M have been reduced to $(N-r)$ and $(M-r)$:

$$P(W_{r+1} \cdots W_{r+w}|R_1 \cdots R_r B) = \frac{(N-M)!(N-r-w)!}{(N-M-w)!(N-r)!} \tag{3–14}$$

and so, by the product rule, the probability for obtaining r red followed by $w = n - r$ white in n draws is from (3–12), (3–14),

$$P(R_1 \cdots R_r W_{r+1} \cdots W_n|B) = \frac{M!(N-M)!(N-n)!}{(M-r)!(N-M-w)!N!}, \tag{3–15}$$

a term $(N-r)!$ having cancelled out.

Although this result was derived for a particular order of drawing red and white balls, the probability for drawing exactly r red balls in any specified order in n draws is the same. To see this, write out the expression (3–15) more fully, in the manner

$$\frac{M!}{(M-r)!} = M(M-1)\cdots(M-r+1) \tag{3–16}$$

and similarly for the other ratios of factorials in (3–15). The right-hand side becomes

$$\frac{M(M-1)\cdots(M-r+1)(N-M)(N-M-1)\cdots(N-M-w+1)}{N(N-1)\cdots(N-n+1)}. \tag{3–17}$$

Now suppose that r red and $(n-r) = w$ white are drawn, in any other order. The probability for this is the product of n factors; every time red is drawn there is a factor (number of red balls in urn)/(total number of balls), and similarly for drawing a white one. The number of balls in the urn decreases by one at each draw; therefore for the k'th draw a factor $(N-k+1)$ appears in the denominator, whatever the colors of the previous draws.

Just before the k'th red ball is drawn, whether this occurs at the k'th draw or any later one, there are $(M-k+1)$ red balls in the urn; so drawing the k'th one places a factor $(M-k+1)$ in the numerator. Just before the k'th white ball is drawn, there are $(N-M-k+1)$ white balls in the urn, and so drawing the k'th white one places a factor $(N-M-k+1)$ in the numerator, regardless of whether this occurs at the k'th draw or any later one. Therefore, by the time all n balls have been drawn, of which r were red, we have accumulated exactly the same factors in numerator and denominator as in (3–17); different orders of drawing them only permute the order of the factors in the numerator. The probability for drawing exactly r balls in any specified order in n draws, is therefore given by (3–15).

Note carefully that in this result the product rule was expanded in a particular way that showed us how to organize the calculation into a product of factors, each of which is a probability at one specified draw, *given the results of all the previous draws*. But the product rule could have been expanded in many other ways, which would give factors conditional on other information than the previous draws; the fact that all these calculations must lead to the same final result is a nontrivial consistency property, which the derivations of Chapter 2 sought to ensure.

Next, we ask: what is the robot's probability for drawing exactly r red balls in n draws, regardless of order? Different orders of appearance of red and white balls are mutually exclusive possibilities, so we must sum over all of them; but since each term is equal to (3–15), we merely multiply it by the binomial coefficient

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{3–18}$$

which represents the number of possible orders of drawing r red balls in n draws or as we shall call it, the *multiplicity* of the event r. For example, to get 3 red in 3 draws can happen in only

$$\binom{3}{3} = 1 \tag{3–19}$$

way, namely $R_1R_2R_3$; the event $r = 3$ has a multiplicity of 1. But to get 2 red in 3 draws can happen in

$$\binom{3}{2} = 3 \tag{3–20}$$

ways, namely $R_1R_2W_3$, $R_1W_2R_3$, $W_1R_2R_3$, so the event $r = 2$ has a multiplicity of 3.

Exercise 3.1. Why isn't the multiplicity factor (3–18) just $n!$? After all, we started this discussion by stipulating that the balls, in addition to having colors, also carry labels $(1, 2 \cdots N)$, so that different permutations of the red balls among themselves, which give the $r!$ in the denominator of (3–18), are distinguishable arrangements. *Hint:* in (3–15) we are not specifying which red balls and which white ones are to be drawn.

Taking the product of (3–15) and (3–18), the many factorials can be reorganized into three binomial coefficients. Defining $A \equiv$ "Exactly r red balls in n draws, in any order" and the function

$$h(r|N, M, n) \equiv P(A|B) \tag{3–21}$$

we have

$$h(r|N, M, n) = \frac{\binom{M}{r}\binom{N-M}{n-r}}{\binom{N}{n}} \tag{3–22}$$

which we shall usually abbreviate to $h(r)$. By the convention $x! = \Gamma(x+1)$ it vanishes automatically when $r > M$, or $r > n$, or $(n-r) > (N-M)$, as it should.

We are here doing a little notational acrobatics for reasons explained in Appendix B. The point is that in our formal probability symbols $P(A|B)$ with the capital P, the arguments A, B

always stand for propositions, which can be quite complicated verbal statements. If we wish to use ordinary numbers for arguments, then for consistency we should define new functional symbols such as $h(r|N, M, n)$. To try to use a notation like $P(r|NMn)$, thereby losing sight of the qualitative stipulations contained in A and B, has led to serious errors from misinterpretation of the equations (such as the marginalization paradox discussed later). However, as already indicated in Chapter 2, we follow the custom of most contemporary works by using probability symbols of the form $p(A|B)$, or $p(r|n)$ with small p, in which we permit the arguments to be either propositions or algebraic variables; in this case, the meaning must be judged from the context.

The fundamental result (3–22) is called the *hypergeometric distribution* because it is related to the coefficients in the power series representation of the Gauss hypergeometric function

$$F(a, b, c; t) = \sum_{r=0}^{\infty} \frac{\Gamma(a+r)\Gamma(b+r)\Gamma(c)}{\Gamma(a)\Gamma(b)\Gamma(c+r)} \frac{t^r}{r!}. \tag{3–23}$$

If either a or b is a negative integer, the series terminates and this is a polynomial. It is easily verified that the *generating function*

$$G(t) \equiv \sum_{r=0}^{n} h(r|N, M, n) t^r \tag{3–24}$$

is equal to

$$G(t) = \frac{F(-M, -n, c; t)}{F(-M, -n, c; 1)} \tag{3–25}$$

with $c = N - M - n + 1$. The evident relation $G(1) = 1$ is from (3–24) just the statement that the hypergeometric distribution is correctly normalized. In consequence of (3–25), $G(t)$ satisfies the second-order hypergeometric differential equation and has many other properties useful in calculations.

Although the hypergeometric distribution $h(r)$ appears complicated, it has some surprisingly simple properties. The most probable value of r is found to within one unit by setting $h(r') = h(r' - 1)$ and solving for r'. We find

$$r' = \frac{(n+1)(M+1)}{N+2}. \tag{3–26}$$

If r' is an integer, then r' and $r' - 1$ are jointly the most probable values. If r' is not an integer, then there is a unique most probable value

$$\hat{r} = \mathrm{INT}(r') \tag{3–27}$$

that is, the next integer below r'. Thus the most probable fraction $f = r/n$ of red balls in the sample drawn is nearly equal to the fraction $F = M/N$ originally in the urn, as one would expect intuitively. This is our first crude example of a physical prediction: a relation between a quantity F specified in our information, and a quantity f measurable in a physical experiment, derived from the theory.

r	$h(r)$	$H(r)$
0	0.000593	0.000593
1	0.007237	0.007830
2	0.037993	0.045824
3	0.113096	0.158920
4	0.211413	0.370333
5	0.259334	0.629667
6	0.211413	0.841080
7	0.113096	0.954177
8	0.037993	0.992170
9	0.007237	0.999407
10	0.000593	1.000000

Table 3.1: $N, M, n = 100, 10, 50$

r	$h(r)$	$H(r)$
0	0.000593	0.000593
1	0.007237	0.007830
2	0.037993	0.045824
3	0.113096	0.158920
4	0.211413	0.370333
5	0.259334	0.629667
6	0.211413	0.841080
7	0.113096	0.954177
8	0.037993	0.992170
9	0.007237	0.999407
10	0.000593	1.000000

Table 3.2: $N, M, n = 100, 50, 10$

The width of the distribution $h(r)$ gives an indication of the accuracy with which the robot can predict r. Many such questions are answered by calculating the *cumulative probability distribution*, which is the probability for finding R or fewer red balls. If R is an integer, that is

$$H(R) \equiv \sum_{r=0}^{R} h(r), \tag{3–28}$$

but for later formal reasons we define $H(x)$ to be a staircase function for all non-negative real x; thus $H(x) \equiv H(R)$, where $R = \mathrm{INT}(x)$ is the greatest integer $\leq x$.

The *median* of a probability distribution such as $h(r)$ is defined to be a number m such that equal probabilities are assigned to the propositions $(r < m)$ and $(r > m)$. Strictly speaking, according to this definition a discrete distribution has in general no median. If there is an integer R for which $H(R-1) = 1 - H(R)$ and $H(R) > H(R-1)$, then R is the unique median. If there is an integer R for which $H(R) = 1/2$, then any r in $(R \leq r < R')$ is a median, where R' is the next higher jump point of $H(x)$; otherwise there is none.

But for most purposes we may take a more relaxed attitude and approximate the strict definition. If n is reasonably large, then it makes reasonably good sense to call that value of R for which $H(R)$ is closest to $1/2$, the "median." In the same relaxed spirit, the values of R for which $H(R)$ is closest to $1/4$, $3/4$ may be called the "lower quartile" and "upper quartile," and if $n \gg 10$ we may call the value of R for which $H(R)$ is closest to $k/10$ the "k'th decile," and so on. As $n \to \infty$ these loose definitions come into conformity with the strict one.

Usually, the fine details of $H(R)$ are unimportant and for our purposes it is sufficient to know the median and the quartiles. Then the (median) $\pm$ (interquartile distance) will provide a good enough idea of the robot's prediction and its probable accuracy. That is, on the information given to the robot, the true value of r is about as likely to lie in this interval as outside it. Likewise, the robot assigns a probability of $(5/6) - (1/6) = 2/3$ (in other words, odds of $2:1$) that r lies between the first and fifth hexile, odds of $8:2 = 4:1$ that it is bracketed by the first and ninth decile; and so on.

Although one can develop rather messy approximate formulas for these distributions which were much used in the past, it is easier today to calculate the exact distribution by computer. For

r	$h(r)$	$H(r)$
0	0.000527	0.000527
1	0.006594	0.007121
2	0.035460	0.042581
3	0.108070	0.150651
4	0.206715	0.357367
5	0.259334	0.616700
6	0.216111	0.832812
7	0.118123	0.950934
8	0.040526	0.991461
9	0.007880	0.999341
10	0.000659	1.000000

Table 3.3: Hypergeometric Distribution, $N, M, n = 99, 50, 10$.

example Press, W. H., *et al*, (1986) list two routines that will calculate the generalized complex hypergeometric distribution for any values of a, b and c. Tables 3.1 and 3.2 give the hypergeometric distribution for $N = 100$, $M = 50$, $n = 10$ and $N = 100$, $M = 10$, $n = 50$. In the latter case, it is not possible to draw more than 10 red balls, so the entries for $r > 10$ are all $h(r) = 0$, $H(r) = 1$ and are not tabulated. One is struck immediately by the fact that the entries for positive $h(r)$ are identical; the hypergeometric distribution has the symmetry property

$$h(r|N, M, n) = h(r|N, n, M) \tag{3–29}$$

under interchange of M and n. Whether we draw 10 balls from an urn containing 50 red ones, or 50 from an urn containing 10 red ones, the probability for finding r red ones in the sample drawn is the same. This is readily verified by closer inspection of (3–22), and it is evident from the symmetry in a, b of the hypergeometric function (3–23).

Another symmetry evident from the table is the symmetry of the distribution about its peak: $h(r|100, 50, 10) = h(10 - r|100, 50, 10)$. However, this is not so in general; changing N to 99 results in a slightly unsymmetrical peak as we see from Table 3.3. The symmetric peak in Table 3.1 arises as follows: if we interchange M and $(N - M)$ and at the same time interchange r and $(n - r)$ we have in effect only interchanged the words "red" and "white," so the distribution is unchanged:

$$h(n - r|N, N - M, n) = h(r|N, M, n). \tag{3–30}$$

But when $M = N/2$, this reduces to the symmetry

$$h(n - r|N, M, n) = h(r|N, M, n) \tag{3–31}$$

observed in Table 3.1. By (3–29) the peak must be symmetric also when $n = N/2$.

The hypergeometric distribution has two more symmetries not at all obvious intuitively or even visible in (3–22). Let us ask the robot for its probability $P(R_2|B)$ of red on the second draw. This is not the same calculation as (3–8), because the robot knows that, just prior to the second draw, there are only $(N - 1)$ balls in the urn, not N. But it does not know what color of ball was removed on the first draw, so it does not know whether the number of red balls now in the urn is M or $(M - 1)$. Then the basis for the Bernoulli urn result (3–5) is lost, and it might appear that the problem is indeterminate.

Yet it is quite determinate after all; the following is our first example of one of the useful techniques in probability calculations, which derives from the resolution of a proposition into disjunctions of simpler ones, as discussed in Chapters 1 and 2. The robot does know that either R_1 or W_1 is true, therefore a relation of Boolean algebra is

$$R_2 = (R_1 + W_1)R_2 = R_1R_2 + W_1R_2. \tag{3–32}$$

So we apply the sum rule and the product rule to get

$$\begin{aligned} P(R_2|B) &= P(R_1R_2|B) + P(W_1R_2|B) \\ &= P(R_2|R_1B)P(R_1|B) + P(R_2|W_1B)P(W_1|B). \end{aligned} \tag{3–33}$$

But

$$P(R_2|R_1B) = \frac{M-1}{N-1}, \qquad P(R_2|W_1B) = \frac{M}{N-1} \tag{3–34}$$

and so

$$P(R_2|B) = \frac{M-1}{N-1}\frac{M}{N} + \frac{M}{N-1}\frac{N-M}{N} = \frac{M}{N}. \tag{3–35}$$

The complications cancel out, and we have the same probability for red on the first and second draws. Let us see whether this continues. For the third draw we have

$$R_3 = (R_1 + W_1)(R_2 + W_2)R_3 = R_1R_2R_3 + R_1W_2R_3 + W_1R_2R_3 + W_1W_2R_3 \tag{3–36}$$

and so

$$\begin{aligned} P(R_3|B) &= \frac{M}{N}\frac{M-1}{N-1}\frac{M-2}{N-2} + \frac{M}{N}\frac{N-M}{N-1}\frac{M-1}{N-2} \\ &\quad + \frac{N-M}{N}\frac{M}{N-1}\frac{M-1}{N-2} + \frac{N-M}{N}\frac{N-M-1}{N-1}\frac{M}{N-2} \\ &= \frac{M}{N}. \end{aligned} \tag{3–37}$$

Again all the complications cancel out. The robot's probability for red at any draw, *if it does not know the result of any other draw*, is always the same as the Bernoulli urn result (3–5). This is the first non-obvious symmetry. We shall not prove this in generality here, because it is contained as a special case of a still more general result, Eq. (3–118) below.

The method of calculation illustrated by (3–32) and (3–36) is: resolve the quantity whose probability is wanted into mutually exclusive sub-propositions, then apply the sum rule and the product rule. If the sub-propositions are well chosen (*i.e.* if they have some simple meaning in the context of the problem), their probabilities are often calculable. If they are not well chosen (as in the example of the penguins at the end of Chapter 2), then of course this procedure cannot help us.

Logic Versus Propensity

This suggests a new question. In finding the probability for red at the k'th draw, knowledge of what color was found at some earlier draw is clearly relevant because an earlier draw affects the number M_k of red balls in the urn for the k'th draw. Would knowledge of the color for a later draw be relevant? At first glance it seems that it could not be, because the result of a later draw cannot influence the value of M_k. For example, a well-known exposition of statistical mechanics (Penrose, 1979) takes it as a fundamental axiom that probabilities referring to the present time can depend only on what happened earlier, not on what happens later. The author considers this to be a necessary physical condition of "causality."

Therefore we stress again, as we did in Chapter 1, that inference is concerned with *logical* connections, which may or may not correspond to causal physical influences. To show why knowledge of later events is relevant to the probabilities of earlier ones, consider an urn which is known (background information B) to contain only one red and one white ball: $N = 2$, $M = 1$. Given only this information, the probability for red on the first draw is $P(R_1|B) = 1/2$. But then if the robot learns that red occurs on the second draw, it becomes certain that it did not occur on the first:

$$P(R_1|R_2B) = 0. \tag{3–38}$$

More generally, the product rule gives us

$$P(R_jR_k|B) = P(R_j|R_kB)P(R_k|B) = P(R_k|R_jB)P(R_j|B). \tag{3–39}$$

But we have just seen that $P(R_j|B) = P(R_k|B) = M/N$ for all j, k, so

$$P(R_j|R_kB) = P(R_k|R_jB), \qquad \text{all } j, k. \tag{3–40}$$

Probability theory tells us that the results of later draws have precisely the same relevance as do the results of earlier ones! Even though performing the later draw does not physically affect the number M_k of red balls in the urn at the k'th draw, *information* about the result of a later draw has the same effect on our *state of knowledge* about what could have been taken on the k'th draw, as does information about an earlier one. This is our second non-obvious symmetry.

This result will be quite disconcerting to some schools of thought about the "meaning of probability." Although it is generally recognized that logical implication is not the same as physical causation, nevertheless there is a strong inclination to cling to the idea anyway, by trying to interpret a probability $P(A|B)$ as expressing some kind of partial causal influence of B on A. This is evident not only in the aforementioned work of Penrose, but more strikingly in the "propensity" theory of probability expounded by the philosopher Karl Popper.†

---

† In his presentation at the Ninth Colston Symposium, Popper (1957) describes his propensity interpretation as "purely objective" but avoids the expression "physical influence." Instead he would say that the probability for a particular face in tossing a die is not a physical property of the die [as Cramér (1946) insisted] but rather is an objective property of the whole experimental arrangement, the die plus the method of tossing. Of course, that the *result of the experiment* depends on the entire arrangement and procedure is only a truism. It was stressed repeatedly by Niels Bohr in connection with quantum theory, but presumably no scientist from Galileo on has ever doubted it. However, unless Popper really meant "physical influence," his interpretation would seem to be supernatural rather than objective. In a later article (Popper, 1959) he defines the propensity interpretation more completely; now a propensity is held to be "objective" and

It appears to us that such a relation as (3–40) would be quite inexplicable from a propensity viewpoint, although the simple example (3–38) makes its logical necessity obvious. In any event, the theory of logical inference that we are developing here differs fundamentally, in outlook and in results, from the theory of physical causation envisaged by Penrose and Popper. It is evident that logical inference can be applied in many problems where assumptions of physical causation would not make sense.

This does not mean that we are forbidden to introduce the notion of "propensity" or physical causation; the point is rather that logical inference is applicable and useful whether or not a propensity exists. If such a notion (*i.e.* that some such propensity exists) is formulated as a well-defined hypothesis, then our form of probability theory can analyze its implications. We shall do this in "Correction for Correlations" below. Also, we can test that hypothesis against alternatives in the light of the evidence, just as we can test any well-defined hypothesis. Indeed, one of the most common and important applications of probability theory is to decide whether there is evidence for a causal influence: is a new medicine more effective, or a new engineering design more reliable? Does a new anti-crime law reduce the incidence of crime? Our study of hypothesis testing starts in Chapter 4.

In all the sciences, logical inference is more generally applicable. We agree that physical influences can propagate only forward in time; but logical inferences propagate equally well in either direction. An archaeologist uncovers an artifact that changes his knowledge of events thousands of years ago; were it otherwise, archaeology, geology, and paleontology would be impossible. The reasoning of Sherlock Holmes is also directed to inferring, from presently existing evidence, what events must have transpired in the past. The sounds reaching your ears from a marching band 600 meters distant change your state of knowledge about what the band was playing two seconds earlier. Listening to a Toscanini recording of a Beethoven symphony changes your state of knowledge about the sounds Toscanini elicited from his orchestra many years ago.

As this suggests, and as we shall verify later, a fully adequate theory of nonequilibrium phenomena such as sound propagation, also requires that backward logical inferences be recognized and used, although they do not express physical causes. The point is that the best inferences we can make about any phenomenon—whether in physics, biology, economics, or any other field—must take into account all the relevant information we have, regardless of whether that information refers to times earlier or later than the phenomenon itself; this ought to be considered a platitude, not a paradox. At the end of this Chapter [Exercise 3.6] the reader will have an opportunity to demonstrate this directly, by calculating a backward inference that takes into account a forward causal influence.

More generally, consider a probability distribution $p(x_1 \cdots x_n|B)$, where x_i denotes the result of the i'th trial, and could take on, not just two values (red or white) but, say, the values $x_i = (1, 2 \cdots k)$ labeling k different colors. If the probability is invariant under any permutation of the x_i, then it depends only on the sample numbers $(n_1 \cdots n_k)$ denoting how many times the result

---

"physically real" even when applied to the individual trial. In the following we see by mathematical demonstration some of the logical difficulties that result from a propensity interpretation. Popper complains that in quantum theory one oscillates between "$\cdots$ an *objective* purely statistical interpretation and a *subjective* interpretation in terms of our incomplete knowledge" and thinks that the latter is reprehensible and the propensity interpretation avoids any need for it. He could not possibly be more mistaken. In Chapter 9 we answer this in detail at the conceptual level; obviously, *incomplete knowledge is the only working material a scientist has*! In Chapter 10 we consider the detailed physics of coin tossing and see just how the method of tossing affects the results by direct physical influence.

$x_i = 1$ occurs, how many times $x_i = 2$ occurs, etc. Such a distribution is called *exchangeable*; as we shall find later, exchangeable distributions have many interesting mathematical properties and important applications.

Returning to our Urn problem, it is clear already from the fact that the hypergeometric distribution is exchangeable, that every draw must have just the same relevance to every other draw regardless of their time order and regardless of whether they are near or far apart in the sequence. But this is not limited to the hypergeometric distribution; it is true of any exchangeable distribution (*i.e.* whenever the probability for a sequence of events is independent of their order). So with a little more thought these symmetries, so inexplicable from the standpoint of physical causation, become obvious after all as propositions of logic.

Let us calculate this effect quantitatively. Supposing $j < k$, the proposition $R_j R_k$ (red at both draws j and k) is in Boolean algebra the same as

$$R_j R_k = (R_1 + W_1) \cdots (R_{j-1} + W_{j-1})\, R_j (R_{j+1} + W_{j+1}) \cdots (R_{k-1} + W_{k-1}) R_k \tag{3–41}$$

which we could expand in the manner of (3–36) into a logical sum of

$$2^{j-1} \times 2^{k-j-1} = 2^{k-2} \tag{3–42}$$

propositions, each specifying a full sequence, such as

$$W_1 R_2 W_3 \cdots R_j \cdots R_k \tag{3–43}$$

of k results. The probability $P(R_j R_k|B)$ is the sum of all their probabilities. But we know that, given B, the probability for any one sequence is independent of the order in which red and white appear. Therefore we can permute each sequence, moving R_j to the first position, and R_k to the second. That is, replace the sequence $(W_1 \cdots R_j \cdots)$ by $(R_1 \cdots W_j \cdots)$, etc. Recombining them, we have $(R_1 R_2)$ followed by every possible result for draws $(3, 4 \cdots k)$. In other words, the probability for $R_j R_k$ is the same as that of

$$R_1 R_2 (R_3 + W_3) \cdots (R_k + W_k) = R_1 R_2 \tag{3–44}$$

and we have

$$P(R_j R_k|B) = P(R_1 R_2|B) = \frac{M(M-1)}{N(N-1)} \tag{3–45}$$

and likewise

$$P(W_j R_k|B) = P(W_1 R_2|B) = \frac{(N-M)M}{N(N-1)}. \tag{3–46}$$

Therefore by the product rule

$$P(R_k|R_j B) = \frac{P(R_j R_k|B)}{P(R_j|B)} = \frac{M-1}{N-1} \tag{3–47}$$

and

$$P(R_k|W_j B) = \frac{P(W_j R_k|B)}{P(W_j|B)} = \frac{M}{N-1} \tag{3–48}$$

for all $j < k$. By (3–40), the results (3–47), (3–48) are true for all $j \neq k$.

Since as noted this conclusion appears astonishing to many people, we shall belabor the point by explaining it still another time in different words. The robot knows that the urn contained originally M red balls and $(N-M)$ white ones. Then learning that an earlier draw gave red, it knows that one less red ball is available for the later draws. The problem becomes the same as if we had started with an urn of $(N-1)$ balls, of which $(M-1)$ are red; (3–47) corresponds just to the solution (3–37) adapted to this different problem.

But why is knowing the result of a later draw equally cogent? Because if the robot knows that red will be drawn at any later time, then in effect one of the red balls in the urn must be "set aside" to make this possible. The number of red balls which could have been taken in earlier draws is reduced by one, as a result of having this information. The above example (3–38) is an extreme special case of this, where the conclusion is particularly obvious.

Reasoning from Less Precise Information

Now let us try to apply this understanding to a more complicated problem. Suppose the robot learns that red will be found at least once in later draws, but not at which draw or draws this will occur. That is, the new information is, as a proposition of Boolean algebra,

$$R_{later} \equiv R_{k+1} + R_{k+2} + \cdots + R_n. \tag{3–49}$$

This information reduces the number of red available for the k'th draw by at least one, but it is not obvious whether R_{later} has exactly the same implications as does R_n. To investigate this we appeal again to the symmetry of the product rule:

$$P(R_k R_{later}|B) = P(R_k|R_{later}B)P(R_{later}|B) = P(R_{later}|R_k B)P(R_k|B) \tag{3–50}$$

which gives us

$$P(R_k|R_{later}B) = P(R_k|B)\frac{P(R_{later}|R_k B)}{P(R_{later}|B)} \tag{3–51}$$

and all quantities on the right-hand side are easily calculated.

Seeing (3–49) one might be tempted to reason as follows:

$$P(R_{later}|B) = \sum_{j=k+1}^{n} P(R_j|B) \tag{3–52}$$

but this is not correct because, unless $M = 1$, the events R_j are not mutually exclusive, and as we see from (2–82), many more terms would be needed. This method of calculation would be very tedious.

To organize the calculation better, note that the denial of R_{later} is the statement that white occurs at all the later draws:

$$\overline{R}_{later} = W_{k+1}\, W_{k+2} \cdots W_n. \tag{3–53}$$

So $P(\overline{R}_{later}|B)$ is the probability for white at all the later draws, regardless of what happens at the earlier ones (*i.e.* when the robot does not know what happens at the earlier ones). By exchangeability this is the same as the probability for white at the first $(n-k)$ draws, regardless of what happens at the later ones; from (3–13),

$$P(\overline{R}_{later}|B) = \frac{(N-M)!(N-n+k)!}{N!(N-M-n+k)!} = \binom{N-M}{n-k}\binom{N}{n-k}^{-1}. \tag{3–54}$$

Likewise $P(\overline{R}_{later}|R_k B)$ is the same result for the case of $(N-1)$ balls, $(M-1)$ of which are red:

$$P(\overline{R}_{later}|R_k B) = \frac{(N-M)!}{(N-1)!}\frac{(N-n+k-1)!}{(N-M-n+k)!} = \binom{N-M}{n-k}\binom{N-1}{n-k}^{-1}. \tag{3–55}$$

Now (3–51) becomes

$$P(R_k|R_{later}B) = \frac{M}{N-n+k} \times \frac{\binom{N-1}{n-k} - \binom{N-M}{n-k}}{\binom{N}{n-k} - \binom{N-M}{n-k}}. \tag{3–56}$$

As a check, note that if $n = k+1$, this reduces to $(M-1)/(N-1)$, as it should.

At the moment, however, our interest in (3–56) is not so much in the numerical values, but in understanding the logic of the result. So let us specialize it to the simplest case that is not entirely trivial. Suppose we draw $n = 3$ times from an urn containing $N = 4$ balls, $M = 2$ of which are white, and ask how knowledge that red occurs at least once on the second and third draws, affects the probability for red at the first draw. This is given by (3–56) with $N = 4,\ M = 2,\ n = 3,\ k = 1$:

$$P(R_1|R_2 + R_3, B) = \frac{6-2}{12-2} = \frac{2}{5} = \left(\frac{1}{2}\right)\frac{1-\frac{1}{3}}{1-\frac{1}{6}}, \tag{3–57}$$

the last form corresponding to (3–51). Compare this to the previously calculated probabilities:

$$P(R_1|B) = \frac{1}{2}, \qquad P(R_1|R_2B) = P(R_2|R_1B) = \frac{1}{3}. \tag{3–58}$$

What seems surprising is that

$$P(R_1|R_{later}B) > P(R_1|R_2B). \tag{3–59}$$

Most people guess at first that the inequality should go the other way; *i.e.* knowing that red occurs at least once on the later draws ought to decrease the chances of red at the first draw more than does the information R_2. But in this case the numbers are so small that we can check the calculation (3–51) directly. To find $P(R_{later}|B)$ by the extended sum rule (2–82) now requires only one extra term:

$$\begin{aligned} P(R_{later}|B) &= P(R_2|B) + P(R_3|B) - P(R_2R_3|B) \\ &= \frac{1}{2} + \frac{1}{2} - \frac{1}{2}\times\frac{1}{3} = \frac{5}{6}. \end{aligned} \tag{3–60}$$

We could equally well resolve R_{later} into mutually exclusive propositions and calculate

$$\begin{aligned} P(R_{later}|B) &= P(R_2W_3|B) + P(W_2R_3|B) + P(R_2R_3|B) \\ &= \frac{1}{2}\times\frac{2}{3} + \frac{1}{2}\times\frac{2}{3} + \frac{1}{2}\times\frac{1}{3} = \frac{5}{6}. \end{aligned} \tag{3–61}$$

The denominator $(1-1/6)$ in (3–57) has now been calculated in three different ways, with the same result. If the three results were not the same, we would have found an inconsistency in our rules, of the kind we sought to prevent by Cox's functional equation arguments in Chapter 2. This is a good example of what "consistency" means in practice, and it shows the trouble we would be in if our rules did not have it.

Likewise, we can check the numerator of (3–51) by an independent calculation:

$$\begin{aligned} P(R_{later}|R_1B) &= P(R_2|R_1B) + P(R_3|R_1B) - P(R_2R_3|R_1B) \\ &= \frac{1}{3} + \frac{1}{3} - \frac{1}{3} \times 0 = \frac{2}{3} \end{aligned} \tag{3–62}$$

and the result (3–57) is confirmed. So we have no choice but to accept the inequality (3–59) and try to understand it intuitively. Let us reason as follows: The information R_2 reduces the number of red balls available for the first draw by one, and it reduces the number of balls in the urn available for the first draw by one, giving $P(R_1|R_2B) = (M-1)/(N-1) = \frac{1}{3}$. The information R_{later} reduces the "effective number of red balls" available for the first draw by more than one, but it reduces the number of balls in the urn available for the first draw by 2 (because it assures the robot that there are two later draws in which two balls are removed). So let us try tentatively to interpret the result (3–57) as

$$P(R_1|R_{later}B) = \frac{(M)_{\mathit{eff}}}{N-2} \tag{3–63}$$

although we are not quite sure what this means. Given R_{later}, it is certain that at least one red ball is removed, and the probability that two are removed is by the product rule:

$$\begin{aligned} P(R_2R_3|R_{later}B) &= \frac{P(R_2R_3R_{later}|B)}{P(R_{later}|B)} = \frac{P(R_2R_3|B)}{P(R_{later}|B)} \\ &= \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{5}{6}} = \frac{1}{5} \end{aligned} \tag{3–64}$$

because R_2R_3 implies R_{later}; *i.e.* a relation of Boolean algebra is $(R_2R_3R_{later} = R_2R_3)$. Intuitively, given R_{later} there is probability $1/5$ that two red balls are removed, so the effective number removed is $1+(1/5) = 6/5$. The "effective" number remaining for draw 1 is $4/5$. Indeed, (3–63) then becomes

$$P(R_1|R_{later}B) = \frac{4/5}{2} = \frac{2}{5} \tag{3–65}$$

in agreement with our better motivated but less intuitive calculation (3–57).

Expectations

Another way of looking at this result appeals more strongly to our intuition and generalizes far beyond the present problem. We can hardly suppose that the reader is not already familiar with the idea of expectation, but this is the first time it has appeared in the present work, so we pause to define it. If a variable quantity X can take on the particular values $(x_1 \cdots x_n)$ in n mutually exclusive and exhaustive situations and the robot assigns corresponding probabilities $(p_1, p_2 \cdots p_n)$ to them, then the quantity

$$\langle X \rangle = E(X) = \sum_{i=1}^{n} p_i x_i \tag{3–66}$$

is called the *expectation* (in the older literature, *mathematical expectation* or *expectation value*) of X. It is a weighted average of the possible values, weighted according to their probabilities. Statisticians and mathematicians generally use the notation $E(X)$; but physicists, having already pre-empted E to stand for energy and electric field, use the bracket notation $\langle X \rangle$. We shall use both notations here; they have the same meaning but sometimes one is easier to read than the other.

Like most of the standard terms that arose out of the distant past, the term "expectation" seems singularly inappropriate to us; for it is almost never a value that anyone "expects" to find. Indeed, it is often known to be an impossible value. But we adhere to it because of centuries of precedent.

Given R_{later}, what is the expectation of the number of red balls in the urn for draw number one? There are three mutually exclusive possibilities compatible with R_{later}:

$$R_2 W_3, W_2 R_3, R_2 R_3 \tag{3–67}$$

for which M is $(1, 1, 0)$ respectively, and for which the probabilities are as in (3–64), (3–65):

$$P(R_2 W_3 | R_{later} B) = \frac{P(R_2 W_3|B)}{P(R_{later}|B)} = \frac{(1/2) \times (2/3)}{(5/6)} = \frac{2}{5}, \tag{3–68}$$

$$P(W_2 R_3 | R_{later} B) = \frac{2}{5}, \tag{3–69}$$

$$P(R_2 R_3 | R_{later} B) = \frac{1}{5}. \tag{3–70}$$

So

$$\langle M \rangle = 1 \times \frac{2}{5} + 1 \times \frac{2}{5} + 0 \times \frac{1}{5} = \frac{4}{5}. \tag{3–71}$$

Thus what we called intuitively the "effective" value of M in (3–63) is really the expectation of M.

We can now state (3–63) in a more cogent way: when the fraction $F = M/N$ of red balls is known, then the Bernoulli urn rule applies and $P(R_1|B) = F$. When F is unknown, the probability for red is the expectation of F:

$$P(R_1|B) = \langle F \rangle \equiv E(F). \tag{3–72}$$

If M and N are both unknown, the expectation is over the joint probability distribution for M and N.

That a probability is numerically equal to the expectation of a fraction will prove to be a general rule that holds as well in thousands of far more complicated situations, providing one of the most useful and common rules for physical prediction. We leave it as an exercise for the reader to show that the more general result (3–56) can also be calculated in the way suggested by (3–72).

Other Forms and Extensions

The hypergeometric distribution (3–22) can be written in various ways. The nine factorials can be organized into binomial coefficients also as follows:

$$h(r|N, M, n) = \frac{\binom{n}{r}\binom{N-n}{M-r}}{\binom{N}{M}}. \tag{3–73}$$

But the symmetry under exchange of M and n is still not evident; to see it one must write out (3–22) or (3–73) in full, displaying all the individual factorials.

We may also rewrite (3–22), as an aid to memory, in a more symmetric form: the probability for drawing exactly r red balls and w white ones in $n = r + w$ draws from an urn containing R red and W white, is

$$h(r) = \frac{\binom{R}{r}\binom{W}{w}}{\binom{R+W}{r+w}} \tag{3–74}$$

and in this form it is easily generalized. Suppose that instead of only two colors, there are k different colors of balls in the urn, N_1 of color 1, N_2 of color 2, $\cdots$ N_k of color k. The probability for drawing r_1 balls of color 1, r_2 of color 2, $\cdots$ r_k of color k in $n = \sum r_i$ draws is, as the reader may verify, the generalized hypergeometric distribution:

$$h(r_1 \cdots r_k | N_1 \cdots N_k) = \frac{\binom{N_1}{r_1} \cdots \binom{N_k}{r_k}}{\binom{\sum N_i}{\sum r_i}}. \tag{3–75}$$

Probability as a Mathematical Tool

From the result (3–75) one may obtain a number of identities obeyed by the binomial coefficients. For example, we may decide not to distinguish between colors 1 and 2; *i.e.* a ball of either color is declared to have color "a." Then from (3–75) we must have on the one hand,

$$h(r_a, r_3 \cdots r_k | N_a, N_3 \cdots N_k) = \frac{\binom{N_a}{r_a}\binom{N_3}{r_3} \cdots \binom{N_k}{r_k}}{\binom{\sum N_i}{\sum r_i}} \tag{3–76}$$

with

$$N_a = N_1 + N_2, \qquad\qquad r_a = r_1 + r_2. \tag{3–77}$$

But the event r_a can occur for any values of r_1, r_2 satisfying (3–77), and so we must have also, on the other hand,

$$h(r_a, r_3 \cdots r_k | N_a, N_3 \cdots N_k) = \sum_{r_1=0}^{r_a} h(r_1, r_a - r_1, r_3 \cdots r_k | N_1 \cdots N_k). \tag{3–78}$$

Then, comparing (3–76) and (3–78) we have the identity

$$\binom{N_a}{r_a} = \sum_{r_1=0}^{r_a} \binom{N_1}{r_1}\binom{N_2}{r_a - r_1}. \tag{3–79}$$

Continuing in this way, we can derive a multitude of more complicated identities obeyed by the binomial coefficients. For example,

$$\binom{N_1+N_2+N_3}{r_a} = \sum_{r_1=0}^{r_a}\sum_{r_2=0}^{r_1}\binom{N_1}{r_1}\binom{N_2}{r_2}\binom{N_3}{r_a-r_1-r_2}. \tag{3–80}$$

In many cases, probabilistic reasoning is a powerful tool for deriving purely mathematical results; more examples of this are given by Feller (1951, Chapters 2, 3) and in later Chapters of the present work.

The Binomial Distribution

Although somewhat complicated mathematically, the hypergeometric distribution arises from a problem that is very clear and simple conceptually; there are only a finite number of possibilities and all the above results are exact for the problems as stated. As an introduction to a mathematically simpler, but conceptually far more difficult problem, we examine a limiting form of the hypergeometric distribution.

The complication of the hypergeometric distribution arises because it is taking into account the changing contents of the urn; knowing the result of any draw changes the probability for red for any other draw. But if the number N of balls in the urn is very large compared to the number drawn ($N \gg n$), then this probability changes very little, and in the limit $N \to \infty$ we should have a simpler result, free of such dependencies. To verify this, we write the hypergeometric distribution (3–22) as

$$h(r|N,M,n) = \frac{\left[\frac{1}{N^r}\binom{M}{r}\right]\left[\frac{1}{N^{n-r}}\binom{N-M}{n-r}\right]}{\left[\frac{1}{N^n}\binom{N}{n}\right]}. \tag{3–81}$$

The first factor is

$$\frac{1}{N^r}\binom{M}{r} = \frac{1}{r!}\frac{M}{N}\left(\frac{M}{N}-\frac{1}{N}\right)\left(\frac{M}{N}-\frac{2}{N}\right)\cdots\left(\frac{M}{N}-\frac{r-1}{N}\right) \tag{3–82}$$

and in the limit $N \to \infty,\ M \to \infty,\ M/N \to f$ we have

$$\frac{1}{N^r}\binom{M}{r} \to \frac{f^r}{r!}. \tag{3–83}$$

Likewise

$$\frac{1}{N^{n-r}}\binom{M-1}{n-r} \to \frac{(1-f)^{n-r}}{(n-r)!} \tag{3–84}$$

$$\frac{1}{N^n}\binom{N}{n} \to \frac{1}{n!}. \tag{3–85}$$

In principle we should, of course, take the limit of the product in (3–81), not the product of the limits. But in (3–81) we have defined the factors so that each has its own independent limit, so the result is the same; the hypergeometric distribution goes into

$$h(r|N,M,n) \to b(r|n,f) \equiv \binom{n}{r} f^r (1-f)^{n-r} \tag{3–86}$$

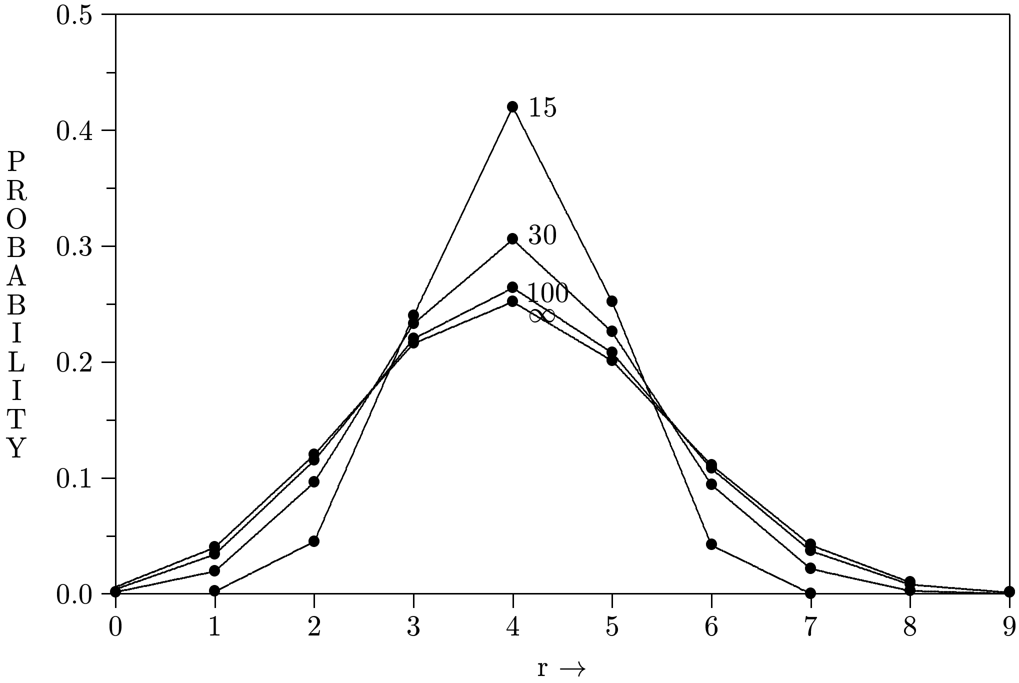

Fig. 3.1. The Hypergeometric Distribution for $N = 15$, 30, 100, ∞.

called the *binomial* distribution, because evaluation of the generating function (3–24) now reduces to

$$G(t) \equiv \sum_{r=0}^{n} b(r|n, f)t^r = (1 - f + ft)^n, \tag{3–87}$$

an example of Newton's binomial theorem.

Figure 3.1 compares three hypergeometric distributions with $N = 15, 30, 100$ and $M/N = 0.4, n = 10$ to the binomial distribution with $n = 10, f = 0.4$. All have their peak at $r = 4$, and all distributions have the same first moment $\langle r\rangle = E(r) = 4$, but the binomial distribution is broader.

The $N = 15$ hypergeometric distribution is zero for $r = 0$ and $r > 6$, since on drawing 10 balls from an urn containing only 6 red and 9 white, it is not possible to get fewer than one or more than 6 red balls. When $N > 100$ the hypergeometric distribution agrees so closely with the binomial that for most purposes it would not matter which one we used. Analytical properties of the binomial distribution are collected in Chapter 7. In Chapter 9 we find, in connection with significance tests, situations where the binomial distribution is exact for purely combinatorial reasons in a finite sample space, Eq. (9–46).

We can carry out a similar limiting process on the generalized hypergeometric distribution (3–75). It is left as an exercise to show that in the limit where all $N_i \to \infty$ in such a way that the fractions

$$f_i \equiv \frac{N_i}{\sum N_j} \tag{3–88}$$

tend to constants, (3–75) goes into the *multinomial distribution*

$$m(r_1 \cdots r_k|f_1 \cdots f_k) = \frac{r!}{r_1! \cdots r_k!} f_1^{r_1} \cdots f_k^{r_k}, \tag{3–89}$$

where $r \equiv \sum r_i$. And, as in (3–87) we can define a generating function of $(k-1)$ variables, from which we can prove that (3–89) is correctly normalized, and derive many other useful results.

Exercise 3.2. Suppose an urn contains $N = \sum N_i$ balls, N_1 of color 1, N_2 of color 2, $\cdots$ N_k of color k. We draw m balls without replacement; what is the probability that we have at least one of each color? Supposing $k = 5$, all $N_i = 10$, how many do we need to draw in order to have at least a 90% probability for getting a full set?

Exercise 3.3. Suppose that in the previous exercise k is initially unknown, but we know that the urn contains exactly 50 balls. Drawing out 20 of them, we find 3 different colors; now what do we know about k? We know from deductive reasoning (*i.e.* with certainty) that $3 \leq k \leq 33$; but can you set narrower limits $k_1 \leq k \leq k_2$ within which it is highly likely to be? [*Hint: this question goes beyond the sampling theory of this Chapter because, like most real scientific problems, the answer depends to some degree on our common sense judgments; nevertheless our rules of probability theory are quite capable of dealing with it, and persons with reasonable common sense cannot differ appreciably in their conclusions*].

Exercise 3.4. The M urns are now numbered 1 to M, and M balls, also numbered 1 to M, are thrown into them, one in each urn. If the numbers of a ball and its urn are the same, we have a match. Show that the probability for at least one match is

$$h = \sum_{k=1}^{M} (-1)^{k+1}/k! \tag{3–90}$$

As $M \to \infty$, this converges to $1 - 1/e = 0.632$. The result is surprising to many, because however large M is, there remains an appreciable probability for no match at all.

Exercise 3.5. N balls are tossed into M urns; there are evidently M^N ways this can be done. If the robot considers them all equally likely, what is its probability that each urn receives at least one ball?

Sampling With Replacement

Up to now, we have considered only the case where we sample without replacement; and that is evidently appropriate for many real situations. For example, in a quality control application, what we have called simply "drawing a ball" might consist really of taking a manufactured item such as an electric light bulb from a carton of them and testing it to destruction. In a chemistry experiment it might consist of weighing out a sample of an unknown protein, then dissolving it in hot sulfuric acid to measure its nitrogen content. In either case, there can be no thought of "drawing that same ball" again.

But suppose now that, being less destructive, we sample balls from the urn and, after recording the "color" (*i.e.* the relevant property) of each, we replace it in the urn before drawing the next ball. This case, of sampling with replacement, is enormously more complicated conceptually, but with some assumptions usually made, ends up being simpler mathematically, than sampling without replacement. For, let us go back to the probability for drawing two red balls in succession. Denoting by B' the same background information as before except for the added stipulation that the balls are to be replaced, we still have an equation like (3–9):

$$P(R_1 R_2|B') = P(R_1|B')P(R_2|R_1 B') \tag{3–91}$$

and the first factor is still, evidently, (M/N); but what is the second one?

Answering this would be, in general, a very difficult problem requiring much additional analysis if the background information B' includes some simple but highly relevant common-sense information that we all have. What happens to that red ball that we put back in the urn? If we merely dropped it into the urn, and immediately drew another ball, then it was left lying on the top of the other balls (or in the top layer of balls); and so it is more likely to be drawn again than any other specified ball, whose location in the urn is unknown. But this upsets the whole basis of our calculation, because the probability for drawing any particular (i'th) ball is no longer given by the Bernoulli Urn Rule which led to (3–11).

Digression: A Sermon on Reality vs. Models

The difficulty we face here is that many things which were irrelevant from symmetry as long as the robot's state of knowledge was invariant under any permutation of the balls, suddenly become relevant, and by one of our desiderata of rationality, the robot must take into account all the relevant information it has. But the probability for drawing any particular ball now depends on such details as the exact size and shape of the urn, the size of the balls, the exact way in which the first one was tossed back in, the elastic properties of balls and urn, the coefficients of friction between balls and between ball and urn, the exact way you reach in to draw the second ball, etc. In a symmetric situation, all of these details are irrelevant.

But even if all these relevant data were at hand, we do not think that a team of the world's best scientists and mathematicians, backed up by all the world's computing facilities, would be able to solve the problem; or would even know how to get started on it. Still, it would not be quite right to say that the problem is unsolvable *in principle*; only so complicated that it is not worth anybody's time to think about it. So what do we do?

In probability theory there is a very clever trick for handling a problem that becomes too difficult. We just solve it anyway by:

(1) Making it still harder;
(2) Redefining what we mean by "solving" it, so that it becomes something we *can* do;
(3) Inventing a dignified and technical-sounding word to describe this procedure, which has the psychological effect of concealing the real nature of what we have done, and making it appear respectable.

In the case of sampling with replacement, we apply this strategy by

(1) Supposing that after tossing the ball in, we shake up the urn. However complicated the problem was initially, it now becomes many orders of magnitude more complicated, because the solution now depends on every detail of the precise way we shake it, in addition to all the factors mentioned above;
(2) Asserting that the shaking has somehow made all these details irrelevant, so that the problem reverts back to the simple one where the Bernoulli Urn Rule applies;
(3) Inventing the dignified-sounding word *randomization* to describe what we have done. This term is, evidently, a euphemism whose real meaning is: *deliberately throwing away relevant information when it becomes too complicated for us to handle*.

We have described this procedure in laconic terms, because an antidote is needed for the impression created by some writers on probability theory, who attach a kind of mystical significance to it. For some, declaring a problem to be "randomized" is an incantation with the same purpose and effect as those uttered by an exorcist to drive out evil spirits; *i.e.* it cleanses their subsequent calculations and renders them immune to criticism. We agnostics often envy the True Believer, who thus acquires so easily that sense of security which is forever denied to us.

However, in defense of this procedure, we have to admit that it often leads to a useful approximation to the correct solution; *i.e.* the complicated details, while undeniably relevant in principle, might nevertheless have little numerical effect on the answers to certain particularly simple questions, such as the probability for drawing r red balls in n trials when n is sufficiently small. But from the standpoint of principle, an element of vagueness necessarily enters at this point; for while we may feel intuitively that this leads to a good approximation, we have no proof of this, much less a reliable estimate of the accuracy of the approximation, which presumably improves with more shaking.

The vagueness is evident particularly in the fact that different people have widely divergent views about how much shaking is required to justify step (2). Witness the minor furor surrounding a Government-sponsored and nationally televised game of chance some years ago, when someone objected that the procedure for drawing numbers from a fish bowl to determine the order of call-up of young men for Military Service was "unfair" because the bowl hadn't been shaken enough to make the drawing "truly random," whatever that means. Yet if anyone had asked the objector: "To *whom* is it unfair?" he could not have given any answer except, "To those whose numbers are on top; I don't know who they are." But after any amount of further shaking, this will still be true! So what does the shaking accomplish?

Shaking does not make the result "random," because that term is basically meaningless as an attribute of the real world; it has no clear definition applicable in the real world. The belief that "randomness" is some kind of real property existing in Nature is a form of the Mind Projection Fallacy which says, in effect, "I don't know the detailed causes—*therefore*—Nature does not know them." What shaking accomplishes is very different. It does not affect *Nature's* workings in any way; it only ensures that no *human* is able to exert any willful influence on the result. Therefore nobody can be charged with "fixing" the outcome.

At this point, you may accuse us of nit-picking, because you know that after all this sermonizing, we are just going to go ahead and use the randomized solution like everybody else does. Note, however, that our objection is not to the procedure itself, provided that we acknowledge honestly what we are doing; *i.e.* instead of solving the real problem, we are making a practical compromise and being, of necessity, content with an approximate solution. That is something we have to do in all areas of applied mathematics, and there is no reason to expect probability theory to be any different.

Our objection is to this belief that by randomization we somehow make our subsequent equations exact; so exact that we can then subject our solution to all kinds of extreme conditions and believe the results, applied to the real world. The most serious and most common error resulting from this belief is in the derivation of limit theorems (*i.e.* when sampling with replacement, nothing prevents us from passing to the limit $n \to \infty$ and obtaining the usual "laws of large numbers"). If we do not recognize the approximate nature of our starting equations, we delude ourselves into believing that we have proved things (such as the identity of probability and limiting frequency) that are just not true in real repetitive experiments.

The danger here is particularly great because mathematicians generally regard these limit theorems as the most important and sophisticated fruits of probability theory, and have a tendency to use language which implies that they are proving properties of the real world. Our point is that these theorems are valid properties *of the abstract mathematical model that was defined and analyzed*. The issue is: to what extent does that model resemble the real world? It is probably safe to say that no limit theorem is directly applicable in the real world, simply because no mathematical model captures every circumstance that is relevant in the real world. The person who believes that he is proving things about the real world, is a victim of the Mind Projection Fallacy.

Back to the Problem. Returning to the equations, what answer can we now give to the question posed after Eq. (3–91)? The probability $P(R_2|R_1B')$ of drawing a red ball on the second draw, clearly depends not only on N and M, but also on the fact that a red one has already been drawn and replaced. But this latter dependence is so complicated that we can't, in real life, take it into account; so we shake the urn to "randomize" the problem, and then declare R_1 to be irrelevant: $P(R_2|R_1B') = P(R_2|B') = M/N$. After drawing and replacing the second ball, we again shake the urn, declare it "randomized," and set $P(R_3|R_2R_1B') = P(R_3|B') = M/N$, etc. In this approximation, the probability for drawing a red one at *any* trial, is (M/N).

But this is not just a repetition of what we learned in (3–37); what is new here is that the result now holds *whatever information the robot may have about what happened in the other trials.* This leads us to write the probability for drawing exactly r red balls in n trials regardless of order, as

$$\binom{n}{r}\left(\frac{M}{N}\right)^r\left(\frac{N-M}{N}\right)^{n-r} \tag{3–92}$$

which is just the binomial distribution (3–86). Randomized sampling with replacement from an urn with finite N has approximately the same effect as passage to the limit $N \to \infty$ without replacement.

Evidently, for small n, this approximation will be quite good; but for large n these small errors can accumulate (depending on exactly how we shake the urn, etc.) to the point where (3–92) is misleading. Let us demonstrate this by a simple, but realistic, extension of the problem.

Correction for Correlations

Suppose that, from an intricate logical analysis, drawing and replacing a red ball increases the probability for a red one at the next draw by some small amount $\epsilon > 0$, while drawing and replacing a white one decreases the probability for a red one at the next draw by a (possibly equal) small quantity $\delta > 0$; and that the influence of earlier draws than the last one is negligible compared to ϵ or δ. You may call this effect a small "propensity" if you like; at least it expresses a physical causation that operates only forward in time. Then, letting C stand for all the above background information including the statements just made about correlations, and the information that we draw n balls, we have

$$\begin{aligned} P(R_k|R_{k-1}C) &= p+\epsilon, \qquad & P(R_k|W_{k-1}C) &= p-\delta \\ P(W_k|R_{k-1}C) &= 1-p-\epsilon, \qquad & P(W_k|W_{k-1}C) &= 1-p+\delta \end{aligned} \tag{3–93}$$

where $p \equiv M/N$. From this, the probability for drawing r red, $(n-r)$ white balls in any specified order, is easily seen to be:

$$p(p+\epsilon)^c(p-\delta)^{c'}(1-p+\delta)^w(1-p-\epsilon)^{w'} \tag{3–94}$$

if the first draw is red, while if the first is white, the first factor in (3–94) should be $(1-p)$. Here c is the number of red draws preceded by red ones, c' the number of red preceded by white, w the number of white draws preceded by white, and w' the number of white preceded by red. Evidently,

$$c + c' = \begin{bmatrix} r-1 \\ r \end{bmatrix}, \qquad w + w' = \begin{bmatrix} n-r \\ n-r-1 \end{bmatrix} \tag{3–95}$$

the upper and lower cases holding when the first draw is red or white, respectively.

When r and $(n-r)$ are small, the presence of ϵ and δ in (3–94) makes little difference, and it reduces for all practical purposes to

$$p^r(1-p)^{n-r} \tag{3–96}$$

as in the binomial distribution (3–92). But as these numbers increase, we can use relations of the form

$$\left(1 + \frac{\epsilon}{p}\right)^c \simeq \exp\left\{\frac{\epsilon c}{p}\right\} \tag{3–97}$$

and (3–94) goes into

$$p^r(1-p)^{n-r} \exp\left\{\frac{\epsilon c - \delta c'}{p} + \frac{\delta w - \epsilon w'}{1-p}\right\}. \tag{3–98}$$

The probability for drawing r red, $(n-r)$ white balls now depends on the order in which red and white appear, and for a given ϵ, when the numbers c, c', w, w' become sufficiently large, the probability can become arbitrarily large (or small) compared to (3–92).

We see this effect most clearly if we suppose that $N = 2M$, $p = 1/2$, in which case we will surely have $\epsilon = \delta$. The exponential factor in (3–98) then reduces to:

$$\exp\{2\epsilon[(c-c') + (w-w')]\}. \tag{3–99}$$

This shows that (1) as the number n of draws tends to infinity, the probability for results containing "long runs"; *i.e.* long strings of red (or white) balls in succession, becomes arbitrarily large compared to the value given by the "randomized" approximation; (2) this effect becomes appreciable when the numbers (ϵc), etc., become of order unity. Thus, if $\epsilon = 10^{-2}$, the randomized approximation can be trusted reasonably well as long as $n < 100$; beyond that, we might delude ourselves by using it. Indeed, it is notorious that in real repetitive experiments where conditions appear to be the same at each trial, such runs—although extremely improbable on the randomized approximation—are nevertheless observed to happen.

Now let us note how the correlations expressed by (3–93) affect some of our previous calculations. The probabilities for the first draw are of course the same as (3–8); now use the notation

$$p = P(R_1|C) = \frac{M}{N}, \qquad q = 1 - p = P(W_1|C) = \frac{N-M}{N}. \tag{3–100}$$

But for the second trial we have instead of (3–35)

$$\begin{aligned} P(R_2|C) &= P(R_2R_1|C) + P(R_2W_1|C) \\ &= P(R_2|R_1C)\,P(R_1|C) + P(R_2|W_1C)\,P(W_1|C) \\ &= (p+\epsilon)p + (p-\delta)q \\ &= p + (p\epsilon - q\delta) \end{aligned} \tag{3–101}$$

and continuing for the third trial,

$$\begin{aligned} P(R_3|C) &= P(R_3|R_2C)P(R_2|C) + P(R_3|W_2C)P(W_2|C) \\ &= (p+\epsilon)(p+p\epsilon-q\delta) + (p-\delta)(q-p\epsilon+q\delta) \\ &= p + (1+\epsilon+\delta)(p\epsilon-q\delta). \end{aligned} \tag{3–102}$$

We see that $P(R_k|C)$ is no longer independent of k; the correlated probability distribution is no longer exchangeable. But does $P(R_k|C)$ approach some limit as $k \to \infty$?

It would be almost impossible to guess the general $P(R_k|C)$ by induction, following the method (3–101), (3–102) a few steps further. For this calculation we need a more powerful method. If we write the probabilities for the k'th trial as a vector

$$V_k \equiv \begin{bmatrix} P(R_k|C) \\ P(W_k|C) \end{bmatrix} \tag{3–103}$$

then Equation (3–93) can be expressed in matrix form:

$$V_k = MV_{k-1}, \tag{3–104}$$

with

$$M = \begin{pmatrix} [p+\epsilon] & [p-\delta] \\ [q-\epsilon] & [q+\delta] \end{pmatrix}. \tag{3–105}$$

This defines a *Markov chain* of probabilities, and M is called the *transition matrix*. Now the slow induction of (3–101), (3–102) proceeds instantly to any distance we please:

$$V_k = M^{k-1}V_1. \tag{3–106}$$

So to have the general solution, we need only to find the eigenvectors and eigenvalues of M. The characteristic polynomial is

$$C(\lambda) \equiv \det(M_{ij} - \lambda\delta_{ij}) = \lambda^2 - \lambda(1+\epsilon+\delta) + (\epsilon+\delta) \tag{3–107}$$

so the roots of $C(\lambda) = 0$ are the eigenvalues

$$\begin{aligned} \lambda_1 &= 1 \\ \lambda_2 &= \epsilon + \delta. \end{aligned} \tag{3–108}$$

Now for any 2×2 matrix

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{3–109}$$

with an eigenvalue λ, the corresponding (non-normalized) right eigenvector is

$$x = (\, b\lambda - a \,) \tag{3–110}$$

for which we have at once $Mx = \lambda x$. Therefore, our eigenvectors are

$$x_1 = \begin{pmatrix} p-\delta \\ q-\epsilon \end{pmatrix}, \qquad x_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{3–111}$$

These are not orthogonal, since M is not a symmetric matrix. Nevertheless, if we use (3–111) to define the transformation matrix

$$S = \begin{pmatrix} [p-\delta] & 1 \\ [q-\epsilon] & -1 \end{pmatrix} \tag{3–112}$$

we find its inverse to be

$$S^{-1} = \frac{1}{1-\epsilon-\delta} \begin{pmatrix} 1 & 1 \\ [q-\epsilon] & -[p-\delta] \end{pmatrix} \tag{3–113}$$

and we can verify by direct matrix multiplication that

$$S^{-1} M S = \Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \tag{3–114}$$

where Λ is the diagonalized matrix. Then we have for any r, positive, negative, or even complex:

$$M^r = S \Lambda^r S^{-1} \tag{3–115}$$

or,

$$M^r = \frac{1}{1-\epsilon-\delta} \begin{pmatrix} p-\delta+[\epsilon+\delta]^r[q-\epsilon] & [p-\delta][1-(\epsilon+\delta)^r] \\ [q-\epsilon][1-(\epsilon+\delta)^r] & q-\epsilon+[\epsilon+\delta]^r[p-\delta] \end{pmatrix} \tag{3–116}$$

and since

$$V_1 = \begin{pmatrix} p \\ q \end{pmatrix} \tag{3–117}$$

the general solution (3–106) sought is

$$P(R_k|C) = \frac{(p-\delta)-(\epsilon+\delta)^{k-1}(p\epsilon-q\delta)}{1-\epsilon-\delta}. \tag{3–118}$$

We can check that this agrees with (3–100), (3–101), (3–102). From examining (3–118) it is clear why it would have been almost impossible to guess the general formula by induction. When $\epsilon = \delta = 0$, this reduces to $P(R_k|C) = p$, supplying the proof promised after Eq. (3–37).

Although we started this discussion by supposing that ϵ and δ were small and positive, we have not actually used that assumption and so, whatever their values, the solution (3–118) is exact for the abstract model that we have defined. This enables us to include two interesting extreme cases. If not small, ϵ and δ must be at least bounded, because all quantities in (3–93) must be probabilities (that is, in $[0,1]$). This requires that

$$-p \le \epsilon \le q, \qquad -q \le \delta \le p \tag{3–119}$$

or

$$-1 \le \epsilon + \delta \le 1. \tag{3–120}$$

But from (3–119), $\epsilon + \delta = 1$ if and only if $\epsilon = q$, $\delta = p$, in which case the transition matrix reduces to the unit matrix

$$M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{3–121}$$

and there are no "transitions." This is a degenerate case in which the positive correlations are so strong that whatever color happens to be drawn on the first trial, is certain to be drawn also on all succeeding ones:

$$P(R_k|C) = p, \qquad \text{all } k. \tag{3–122}$$

Likewise, if $\epsilon + \delta = -1$, then the transition matrix must be

$$M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{3–123}$$

and we have nothing but transitions; *i.e.* the negative correlations are so strong that the colors are certain to alternate after the first draw:

$$P(R_k|C) = \left\{ \begin{matrix} p, & \quad k \text{ odd} \\ q, & \quad k \text{ even} \end{matrix} \right\}. \tag{3–124}$$

This case is unrealistic because intuition tells us rather strongly that ϵ and δ should be positive quantities; surely, whatever the logical analysis used to assign the numerical value of ϵ, leaving a red ball in the top layer must *increase*, not decrease, the probability of red on the next draw. But if ϵ and δ must not be negative, then the lower bound in (3–120) is really zero, which is achieved only when $\epsilon = \delta = 0$. Then M in (3–105) becomes singular, and we revert to the binomial distribution case already discussed.

In the intermediate and realistic cases where $0 < |\epsilon + \delta| < 1$, the last term of (3–118) attenuates exponentially with k, and in the limit

$$P(R_k|C) \to \frac{p - \delta}{1 - \epsilon - \delta}. \tag{3–125}$$

But although these single-trial probabilities settle down to steady values as in an exchangeable distribution, the underlying correlations are still at work and the limiting distribution is not exchangeable. To see this, let us consider the conditional probabilities $P(R_k|R_jC)$. These are found by noting that the Markov chain relation (3–104) holds whatever the vector V_{k-1}; *i.e.* whether or not it is the vector generated from V_1 as in (3–106). Therefore, if we are given that red occurred on the j'th trial, then

$$V_j = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{3–126}$$

and we have from (3–104)

$$V_k = M^{k-j} V_j, \qquad j \le k \tag{3–127}$$

from which, using (3–115),

$$P(R_k|R_jC) = \frac{(p - \delta) + (\epsilon + \delta)^{k-j}\,(q - \epsilon)}{1 - \epsilon - \delta}, \qquad j < k \tag{3–128}$$

which approaches the same limit (3–125). The forward inferences are about what we might expect; the steady value (3–125) plus a term that decays exponentially with distance. But the backward inferences are different; note that the general product rule holds, as always:

$$P(R_k R_j|C) = P(R_k|R_j C)\, P(R_j|C) = P(R_j|R_k C)\, P(R_k|C). \tag{3–129}$$

Therefore, since we have seen that $P(R_k|C) \neq P(R_j|C)$, it follows that

$$P(R_j|R_k C) \neq P(R_k|R_j C). \tag{3–130}$$

The backward inference is still possible, but it is no longer the same formula as the forward inference as it would be in an exchangeable sequence.

As we shall see later, this example is the simplest possible "baby" version of a very common and important physical problem; an irreversible process in the "Markovian approximation." Another common technical language would call it an *autoregressive model* of first order. It can be generalized greatly to the case of matrices of arbitrary dimension and many-step or continuous, rather than single-step, memory influences. But for reasons noted earlier (confusion of inference and causality in the literature of statistical mechanics) the backward inference part of the solution is almost always missed. Some try to do backward inference by extrapolating the forward solution backward in time, with quite bizarre and unphysical results. Therefore the reader is, in effect, conducting new research in doing the following exercise.

Exercise 3.6. Find the explicit formula $P(R_j|R_k C)$ for the backward inference corresponding to the result (3–128) by using (3–118) and (3–129). Then (a) Explain the reason for the difference between forward and backward inferences in simple intuitive terms. (b) In what way does the backward inference differ from the forward inference extrapolated backward? Which is more reasonable intuitively? (c) Do backward inferences also decay to steady values? If so, is a property somewhat like exchangeability restored for events sufficiently separated? For example, if we consider only every tenth draw or every hundredth draw, do we approach an exchangeable distribution on this subset?

Simplification

The above formulas (3–100)–(3–130) hold for any ϵ, δ satisfying the inequalities (3–119). But surveying them, we note that a remarkable simplification occurs if they satisfy

$$p\epsilon = q\delta. \tag{3–131}$$

For then we have

$$\frac{p-\delta}{1-\epsilon-\delta} = p, \qquad \frac{q-\epsilon}{1-\epsilon-\delta} = q, \qquad \epsilon+\delta = \frac{\epsilon}{q} \tag{3–132}$$

and our main results (3–118), (3–128) collapse to

$$P(R_k|C) = p, \qquad \text{all } k \tag{3–133}$$

$$P(R_k|R_j C) = P(R_j|R_k C) = p + q\left(\frac{\epsilon}{q}\right)^{|k-j|}, \qquad \text{all } k, j. \tag{3–134}$$

The distribution is still not exchangeable, since the conditional probabilities (3–134) still depend on the separation $|k-j|$ of the trials; but the symmetry of forward and backward inferences is restored even though the causal influences ϵ, δ operate only forward. Indeed, we see from our derivation of (3–40) that this forward—backward symmetry is a necessary consequence of (3–133) whether or not the distribution is exchangeable.

What is the meaning of this magic condition (3–131)? It does not make the matrix M assume any particularly simple form, and it does not turn off the effect of the correlations. What it does is to make the solution (3–133) invariant; that is, the initial vector (3–117) is then equal but for normalization to the eigenvector x_1 in (3–111), so the initial vector remains unchanged by the matrix (3–105).

In general, of course, there is no reason why this simplifying condition should hold. Yet in the case of our urn, we can see a kind of rationale for it. Suppose that when the urn has initially N balls, they are in L layers. Then after withdrawing one ball, there are about $n = (N-1)/L$ of them in the top layer, of which we expect about np to be red, $nq = n(1-p)$ white. Now we toss the drawn ball back in. If it was red, the probability of getting red at the next draw if we do not shake the urn, is about

$$\frac{np+1}{n+1} = p + \frac{1-p}{n} + O\left(\frac{1}{n^2}\right) \tag{3–135}$$

and if it is white, the probability for getting white at the next draw is about

$$\frac{n(1-p)+1}{n+1} = 1 - p + \frac{p}{n} + O\left(\frac{1}{n^2}\right). \tag{3–136}$$

Comparing with (3–93) we see that we could estimate ϵ and δ by

$$\epsilon \simeq q/n\ , \qquad\qquad \delta \simeq p/n \tag{3–137}$$

whereupon our magic condition (3–131) is satisfied. Of course, the argument just given is too crude to be called a derivation, but at least it indicates that there is nothing inherently unreasonable about (3–131). We leave it for the reader to speculate about what significance and use this curious fact might have, and whether it generalizes beyond the Markovian approximation.

We have now had a first glimpse of some of the principles and pitfalls of standard sampling theory. All the results we have found will generalize greatly, and will be useful parts of our "toolbox" for the applications to follow.

COMMENTS

In most real physical experiments we are not, literally, drawing from any "urn." Nevertheless, the idea has turned out to be a useful conceptual device, and in the 250 years since Bernoulli's *Ars Conjectandi* it has appeared to scientists that many physical measurements are very much like "drawing from Nature's urn." But to some the word "urn" has gruesome connotations and in much of the literature one finds such expressions as "drawing from a population."

In a few cases, such as recording counts from a radioactive source, survey sampling, and industrial quality control testing, one is quite literally drawing from a real, finite population, and the urn analogy is particular apt. Then the probability distributions just found, and their limiting forms and generalizations noted in Chapter 7, will be appropriate and useful. In some cases, such as agricultural experiments or testing the effectiveness of a new medical procedure, our credulity can be strained to the point where we see a vague resemblance to the urn problem.

But in other cases, such as flipping a coin, making repeated measurements of the temperature and wind velocity, the position of a planet, the weight of a baby, or the price of a commodity, the urn analogy seems so farfetched as to be dangerously misleading. Yet in much of the literature one still uses urn distributions to represent the data probabilities, and tries to justify that choice by visualizing the experiment as drawing from some "hypothetical infinite population" which is entirely a figment of our imagination. Functionally, the main consequence of this is strict independence of successive draws, regardless of all other circumstances. Obviously, this is not sound reasoning, and a price must be paid eventually in erroneous conclusions.

This kind of conceptualizing often leads one to suppose that these distributions represent not just our prior state of knowledge about the data, but the *actual* long-run variability of the data in such experiments. Clearly, such a belief cannot be justified; anyone who claims to know in advance the long-run results in an experiment that has not been performed, is drawing on a vivid imagination, not on any fund of actual knowledge of the phenomenon. Indeed, if that infinite population is only imagined, then it seems that we are free to imagine any population we please.

But from a mere act of the imagination we cannot learn anything about the real world. To suppose that the resulting probability assignments have any real physical meaning is just another form of the Mind Projection Fallacy. In practice this diverts our attention to irrelevancies and away from the things that really matter (such as information about the real world that is not expressible in terms of any sampling distribution, or does not fit into the urn picture; but which is nevertheless highly cogent for the inferences we want to make). Usually, the price paid for this folly is missed opportunities; had we recognized that information, more accurate and/or more reliable inferences could have been made.

Urn-type conceptualizing is capable of dealing with only the most primitive kind of information, and really sophisticated applications require us to develop principles that go far beyond the idea of urns. But the situation is quite subtle, because as we stressed before in connection with Gödel'stheorem, an erroneous argument does not necessarily lead to a wrong conclusion. In fact, as we shall find in Chapter 9, highly sophisticated calculations sometimes lead us back to urn-type distributions, for purely mathematical reasons that have nothing to do conceptually with urns or populations. The hypergeometric and binomial distributions found in this Chapter will continue to reappear, because they have a fundamental mathematical status quite independent of arguments that we used to find them here.[†]

On the other hand, we could imagine a different problem in which we would have full confidence in urn-type reasoning leading to the binomial distribution, although it probably never arises in the real world. If we had a large supply $\{U_1, U_2 \cdots U_n\}$ of urns known to have identical contents and those contents known with certainty in advance—and then we used a fresh new urn for each draw—then we would assign $P(A) = M/N$ for every draw, strictly independently of what we know about any other draw. Such prior information would take precedence over any amount of data. If we did not know the contents (M, N) of the urns—but we knew they all had identical contents—this strict independence would be lost, because then every draw from one urn would tell us something about the contents of the other urns, although it does not physically influence them.

[†] In a similar way, exponential functions appear in all parts of analysis because of their fundamental mathematical properties, although their conceptual basis varies widely.

From this we see once again that logical dependence is in general very different from causal physical dependence. We belabor this point so much because it is not recognized at all in most expositions of probability theory, and this has led to errors, as is suggested by Exercise 3.6. In Chapter 4 we shall see a more serious error of this kind [discussion following (4–29)]. But even when one manages to avoid actual error, to restrict probability theory to problems of physical causation is to lose its most important applications. The extent of this restriction—and the magnitude of the missed opportunity—does not seem to be realized by those who are victims of this fallacy.

Indeed, most of the problems we have solved in this Chapter are not considered to be within the scope of probability theory—and do not appear at all—in those expositions which regard probability as a physical phenomenon. Such a view restricts one to a small subclass of the problems which can be dealt with usefully by probability theory as logic. For example, in the "physical probability" theory it is not even considered legitimate to speak of the probability for an outcome at a specified trial; yet that is exactly the kind of thing about which it is necessary to reason in conducting scientific inference. The calculations of this Chapter have illustrated this many times.

In summary: in each of the applications to follow, one must consider whether the experiment is really "like" drawing from an urn; if it is not, then we must go back to first principles and apply the basic product and sum rules in the new context. This may or may not yield the urn distributions.

A Look Ahead

The probability distributions found in this Chapter are called *sampling distributions*, or *direct probabilities*, which names indicate that they are of the form: given some hypothesis H about the phenomenon being observed (in the case just studied, the contents (M, N) of the urn), what is the probability that we shall obtain some specified data D (in this case, some sequence of red and white balls)? Historically, the term "direct probability" has long had the additional connotation of reasoning from a supposed physical cause to an observable effect. But we have seen that not all sampling distributions can be so interpreted. In the present work we shall not use this term, but use "sampling distribution" in the general sense of *reasoning from some specified hypothesis to potentially observable data*, whether the link between hypothesis and data is logical or causal.

Sampling distributions make predictions, such as the hypergeometric distribution (3–22), about potential observations (for example, the possible values and relative probabilities of different values of r). If the correct hypothesis is indeed known, then we expect the predictions to agree closely with the observations. If our hypothesis is not correct, they may be very different; then the nature of the discrepancy gives us a clue toward finding a better hypothesis. This is, very broadly stated, the basis for scientific inference. Just how wide the disagreement between prediction and observation must be in order to justify our rejecting the present hypothesis and seeking a new one, is the subject of *significance tests*. It was the need for such tests in astronomy that led Laplace and Gauss to study probability theory in the 18'th and 19'th centuries.

Although sampling theory plays a dominant role in conventional pedagogy, in the real world such problems are an almost negligible minority. In virtually all real problems of scientific inference we are in just the opposite situation; the data D are known but the correct hypothesis H is not. Then the problem facing the scientist is of the inverse type: given the data D, what is the probability that some specified hypothesis H is true? Exercise 3.3 above was a simple introduction to this kind of problem. Indeed, the scientist's motivation for collecting data is usually to enable him to learn something about the phenomenon, in this way.

Therefore, in the present work our attention will be directed almost exclusively to the methods for solving the inverse problem. This does not mean that we do not calculate sampling distributions;

we need to do this constantly and it may be a major part of our computational job. But it does mean that for us the finding of a sampling distribution is almost never an end in itself.

Although the basic rules of probability theory solve such inverse problems just as readily as sampling problems, they have appeared quite different conceptually to many writers. A new feature seems present, because it is obvious that the question: "What do you know about the hypothesis H after seeing the data D?" cannot have any defensible answer unless we take into account: "What did you know about H before seeing D?" But this matter of previous knowledge did not figure in any of our sampling theory calculations. When we asked: "What do you know about the data given the contents (M, N) of the urn?" we did not seem to consider: "What did you know about the data before you knew (M, N)?"

This apparent dissymmetry, it will turn out, is more apparent than real; it arises mostly from some habits of notation that we have slipped into, which obscure the basic unity of all inference. But we shall need to understand this very well before we can use probability theory effectively for hypothesis tests and their special cases, significance tests. In the next Chapter we turn to this problem.